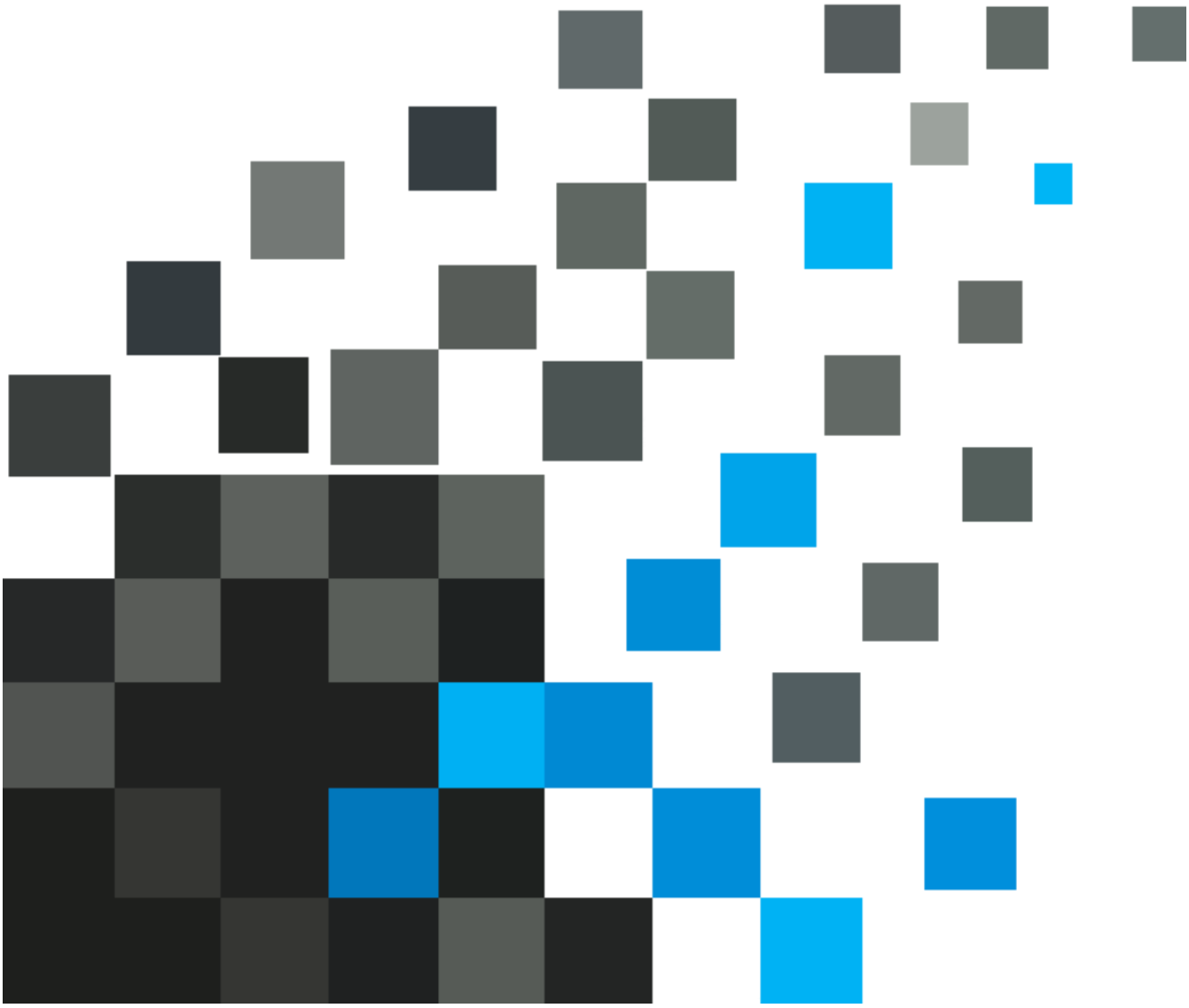
unesp

UNIVERSIDADE ESTADUAL PAULISTA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



AGENTE DE MODELO DE LINGUAGEM DE GRANDE ESCALA PARA DETECÇÃO E EXPLICAÇÃO DE DESINFORMAÇÃO: UMA ABORDAGEM BASEADA NO MÉTODO QUADRIPOLAR

GABRIEL NUNES HENKE

INSTITUTO DE GEOCIÊNCIAS E CIÊNCIAS EXATAS

RIO CLARO

UNIVERSIDADE ESTADUAL PAULISTA
“Júlio de Mesquita Filho”
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

GABRIEL NUNES HENKE

Agente de Modelo de Linguagem de Grande Escala Para Detecção e Explicação de Desinformação: Uma Abordagem Baseada no Método Quadripolar

Dissertação de mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Campus de Rio Claro, da Universidade Estadual Paulista “Júlio de Mesquita Filho”, como parte dos requisitos para a obtenção do título de mestre em Ciência da Computação.

Orientador: Prof. Dr. Denis Henrique Pinheiro Salvadeo

Rio Claro - SP

2026

H512a

Henke, Gabriel Nunes
Agente de modelo de linguagem de grande escala para detecção e explicação de desinformação: uma abordagem baseada no Método Quadripolar / Gabriel Nunes Henke. -- Rio Claro, 2026
203 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Denis Henrique Pinheiro Salvadeo

1. Método Quadripolar. 2. Large Language Models. 3. Transformers. 4. Desinformação. 5. Multiagente. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Impacto da Pesquisa:

Alinhada aos ODS 9 e 16, a pesquisa beneficia a sociedade ao criar uma infraestrutura de IA explicável para detecção de desinformação. Essa inovação tecnológica garante transparência algorítmica, protegendo o debate democrático e assegurando o acesso seguro e confiável a informações no ambiente digital.

Impact of the Research (English):

Aligned with SDGs 9 and 16, this research benefits society by creating an explainable AI infrastructure for misinformation detection. This technological innovation ensures algorithmic transparency, protecting democratic debate and ensuring secure and reliable access to information in the digital environment.

UNIVERSIDADE ESTADUAL PAULISTA
“Júlio de Mesquita Filho”
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

GABRIEL NUNES HENKE

Agente de Modelo de Linguagem de Grande Escala Para Detecção e Explicação de Desinformação: Uma Abordagem Baseada no Método Quadripolar

Dissertação de mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Campus de Rio Claro, da Universidade Estadual Paulista “Júlio de Mesquita Filho”, como parte dos requisitos para a obtenção do título de mestre em Ciência da Computação.

Comissão Examinadora

Prof(a). Dr(a). DENIS HENRIQUE PINHEIRO SALVADEO
IGCE / UNESP / Rio Claro (SP)

Prof(a). Dr(a). VERONICA OLIVEIRA DE CARVALHO
IGCE / UNESP / Rio Claro (SP)

Prof(a). Dr(a). CLÁUDIO ELÍZIO CALAZANS CAMPELO
CEEI / UFCG / Campina Grande (PB)

Conceito: Aprovado

Rio Claro (SP), 26 de fevereiro de 2026.

A Deus, minha gratidão eterna por ser o alicerce, a força e a sabedoria em cada passo desta caminhada.

Dedico este trabalho a todos que caminharam comigo nesta jornada.

Aos meus pais, pelo amor incondicional, pelo exemplo de perseverança e pelos inúmeros sacrifícios feitos ao longo da minha formação.

Aos meus amigos e familiares, pelo apoio nos momentos mais difíceis e pelas palavras de encorajamento quando a exaustão parecia vencer.

Aos meus avôs, José e Carlos. A coragem e a disciplina de vocês vive em mim, e sei que hoje celebram esta vitória junto às estrelas.

Dedico este trabalho à minha versão do passado, que enfrentou as batalhas mais duras com persistência. Que escolheu os caminhos difíceis, rompeu padrões e mudou de rota, apostando em uma história diferente.

Hoje, essa vitória é a prova de que a coragem de recomeçar tem o poder de transformar destinos.

“Importante, em verdade, é o homem que está na arena, com a face coberta de poeira, suor e sangue; que luta com bravura, erra e, seguidamente, tenta atingir o alvo. É aquele que, no sucesso, melhor conhece o triunfo final dos grandes feitos e que, se fracassa, pelo menos falha com ousadia, de modo que o seu lugar jamais será entre as almas tímidas, que não conhecem nem a vitória, nem a derrota.”

Theodore Roosevelt

RESUMO

A proliferação de desinformação no ecossistema digital, exacerbada pelo advento da inteligência artificial generativa, impõe desafios críticos à integridade dos processos democráticos e sociais. As abordagens tradicionais de aprendizado de máquina, embora escaláveis, limitam-se à classificação binária sem prover a fundamentação necessária para a compreensão do fenômeno. Esta dissertação propõe uma arquitetura inovadora que integra Modelos de Linguagem de Grande Escala (LLMs) ao Método Quadripolar, estruturando a verificação de fatos em quatro eixos analíticos: Epistemológico, Teórico, Morfológico e Técnico. A solução emprega um sistema baseado em Agentes de IA e técnicas de *Retrieval-Augmented Generation* (RAG) para decompor a investigação em etapas de coleta de evidências externas e raciocínio contextual. Os experimentos, conduzidos sobre o conjunto de dados *FakeRecogna* e um conjunto de dados de curadoria manual de notícias e golpes cibernéticos, revelaram *trade-offs* significativos entre eficiência e profundidade analítica. Os resultados demonstraram que a configuração baseada exclusivamente no Polo Técnico (C4) atingiu o melhor desempenho classificatório (F1-Score de 0,873) com maior viabilidade econômica para ambientes produtivos. Em contrapartida, a aplicação completa do Método Quadripolar (C6) (F1-Score de 0,853) estabeleceu-se como o estado da arte em explicabilidade, demonstrando robustez superior na distinção de casos ambíguos, como fraudes de engenharia social, conteúdos opinativos complexos e notícias considerando seus fatores temporais em que são apresentadas. O estudo conclui que a estruturação metodológica do raciocínio dos LLMs amplia a acurácia e explicabilidade da detecção, e também opera de forma para garantir a transparência e a auditabilidade das decisões em sistemas de combate à desinformação.

Palavras-chave: Desinformação; Grandes Modelos de Linguagem; Método Quadripolar; *Transformers*; *Embeddings*.

ABSTRACT

The proliferation of misinformation in the digital ecosystem, exacerbated by the advent of generative artificial intelligence, poses critical challenges to the integrity of democratic and social processes. Traditional machine learning approaches, while scalable, are limited to binary classification without providing the necessary foundation for understanding the phenomenon. This dissertation proposes an innovative architecture that integrates Large Language Models (LLMs) with the Quadripolar Method, structuring fact-checking into four analytical poles: Epistemological, Theoretical, Morphological, and Technical. The solution employs a system based on AI Agents and Retrieval-Augmented Generation (RAG) techniques to break down the investigation into stages of external evidence collection and contextual reasoning. The experiments, conducted on the FakeRecogna dataset and a manually curated dataset of news and cyber scams, revealed significant trade-offs between efficiency and analytical depth. The results showed that the configuration based exclusively on the Technical Pole (C4) achieved the best classification performance (F1-Score of 0.873) with greater economic viability for productive environments. In contrast, the complete application of the Quadripolar Method (C6) (F1-Score of 0.853) established itself as the state of the art in explainability, demonstrating superior robustness in distinguishing ambiguous cases, such as social engineering fraud, complex opinion content, and news considering the temporal factors in which they are presented. The study concludes that the methodological structuring of LLM reasoning increases the accuracy and explainability of detection, and also operates in a way that ensures the transparency and auditability of decisions in systems designed to combat disinformation.

Keywords: Misinformation; Large Language Models; Quadripolar Method; Transformers; Embeddings.

Title in english: Large-Scale Language Model Agent for the Detection and Explanation of Disinformation: An Approach Based on the Quadripolar Method.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

NLP	*Natural Language Processing*
RAG	*Retrieval Augmented Generation*
IA	Inteligência Artificial
LLMs	*Large Language Models*
MLM	*Masked Language Model*
NSP	*Next Sentence Prediction*
SFT	*Supervised Fine-Tuning*
RLHF	*Reinforcement Learning with Human Feedback*
GPU	*Graphics Processing Units*
GNNs	*Graph Neural Networks*
ANNs	*Artificial Neural Networks*
LSTM	*Long Short-Term Memory*
RNN	*Recurrent Neural Network*
TF-IDF	*Term Frequency–Inverse Document Frequency*
BoW	*Bag Of Words*
FFN	*Feed-Forward Network*

SUMÁRIO

1 INTRODUÇÃO

A disseminação de notícias falsas, conhecidas popularmente como *fake news*, fazem parte do contexto de desinformação, constitui um fenômeno crescente e complexo no ecossistema digital contemporâneo (ALGHAMDI; LUO; LIN, 2024). Definidas como informações fabricadas ou distorcidas com o intuito de enganar, manipular ou influenciar percepções sociais, políticas ou econômicas, as notícias falsas diferem de boatos de disseminação de desinformação mais restrita e lenta, tipicamente propagados de forma interpessoal, pela sua ampla capacidade de alcance e velocidade de propagação, potencializada pelas plataformas digitais. A natureza dinâmica da Internet, somada à centralidade das redes sociais na formação da opinião pública, criou um ambiente propício para a circulação massiva de conteúdos desinformativos, frequentemente indistinguíveis de notícias legítimas aos olhos do público comum.

Nos últimos anos, o avanço tecnológico tem impulsionado significativamente a sofisticação das notícias falsas. O surgimento de tecnologias como inteligência artificial generativa, *deepfakes* e ferramentas automatizadas de produção e disseminação de conteúdo falso ampliou exponencialmente a escala e o impacto da desinformação (ALGHAMDI; LUO; LIN, 2024). Essas tecnologias tendem a facilitar a criação de materiais falsos de alta qualidade e difícil detecção, e possibilitando sua circulação automatizada em grande volume, por meio de robôs (*bots*) e redes coordenadas. Como resultado, a capacidade de influenciar debates públicos, interferir em processos democráticos e moldar comportamentos sociais foi intensificada, elevando o desafio para as entidades reguladoras, acadêmicas e de verificação de fatos (MANZOOR; SINGLA; NIKITA, 2019).

A problemática associada à proliferação de notícias falsas é multifacetada e severa. Além dos danos imediatos à reputação de indivíduos e instituições, a desinformação pode comprometer a confiança pública em fontes legítimas de informação, deteriorar o debate democrático, fomentar polarizações e gerar consequências sociais, políticas e econômicas de longo prazo (ALGHAMDI; LUO; LIN, 2024). A dificuldade em identificar e mitigar a disseminação de conteúdos falsos evidencia a urgência da implementação de estratégias técnicas, regulatórias e educativas para combater a desinformação. Nesse contexto, torna-se imperativo compreender as dinâmicas de produção, difusão e impacto das notícias falsas, bem como desenvolver mecanismos eficazes para sua detecção e contenção no ambiente digital.

A revisão conduzida por Alghamdi, Luo e Lin (2024) documenta padrões alarmantes na disseminação de desinformação política. O estudo destaca que, durante a eleição presidencial estadunidense de 2016, *fake news* políticas acumularam mais de 37 milhões de compartilhamentos no *Facebook* em apenas três meses (ALGHAMDI; LUO; LIN, 2024). As

vinte fake news eleitorais mais difundidas nas redes sociais registraram 8.711.000 comentários, reações e compartilhamentos, superando o engajamento combinado das vinte notícias mais discutidas nas 19 principais redações jornalísticas do período (ALGHAMDI; LUO; LIN, 2024). Adicionalmente, Alghamdi, Luo e Lin (2024) aponta que fake news se propagam até seis vezes mais rápido do que notícias verdadeiras no Twitter, fenômeno potencializado pelo fato de que, em 2016, 62% dos americanos já obtinham notícias por meio de redes sociais (ALGHAMDI; LUO; LIN, 2024). Um fator estrutural relevante identificado pela revisão é a atuação de agentes automatizados. Na semana anterior ao pleito de 2016, cerca de 19 milhões de contas falsas publicaram *tweets* de apoio a um dos candidatos, amplificando artificialmente a circulação de conteúdos desinformativos (ALGHAMDI; LUO; LIN, 2024).

Com o avanço acelerado da proliferação de notícias falsas, tornou-se imperativo o desenvolvimento de métodos cada vez mais robustos para sua detecção e mitigação (AHMED; TRAORE; SAAD, 2017). A necessidade de respostas tecnológicas à altura da complexidade e da velocidade da desinformação impulsionou a aplicação extensiva de técnicas de *Machine Learning* (aprendizado de máquina) realizadas de formas tradicionais como classificação baseada em árvores de decisão (BREIMAN, 2001), especialmente para a classificação automática de conteúdos textuais (ALGHAMDI; LUO; LIN, 2024). Essas abordagens têm-se mostrado fundamentais para a construção de sistemas capazes de diferenciar informações verídicas de conteúdos fabricados ou manipulados, contribuindo para a preservação da integridade informacional em ambientes digitais (ALGHAMDI; LUO; LIN, 2024). Além disso, a evolução das ameaças associadas à desinformação demandou não apenas o refinamento dos algoritmos de detecção, mas também a ampliação das bases de dados utilizadas para treinamento e validação, assegurando maior representatividade e eficácia na identificação de padrões de falsidade textual (FAYAZ *et al.*, 2022).

Dentro do contexto do uso de *machine learning* tradicional, destaca-se a importância da engenharia de *features* como etapa crítica no processo de detecção de notícias falsas (FAYAZ *et al.*, 2022). Essa prática envolve a extração e seleção criteriosa de atributos que capturam características linguísticas, semânticas, sintáticas e contextuais relevantes à identificação de notícias falsas (ALGHAMDI; LUO; LIN, 2024). Elementos como a análise de polaridade emocional, complexidade textual, presença de informações sensacionalistas, frequência de termos específicos e padrões de citação são amplamente explorados em conjuntos de dados específicos (ALGHAMDI; LUO; LIN, 2024). A correta definição e modelagem dessas *features* determinam o desempenho dos algoritmos de classificação, influenciando diretamente métricas como acurácia, precisão e *recall* (TING, 2010) dos sistemas de detecção.

O avanço dos algoritmos de aprendizado de máquina determinísticos na detecção de notícias falsas demonstra eficácia em cenários com padrões explícitos e regras pré-definidas são suficientes para identificar desinformação. No entanto, a ascensão de

modelos generativos de Inteligência Artificial (IA), como GPT *Generative Pre-trained Transformers* (YENDURI; RAMALINGAM; SELVI, 2024) e Gemini (Gemini Team *et al.*, 2024), introduz uma camada de complexidade mais robusta, uma vez que esses sistemas são capazes de produzir conteúdos sintéticos altamente coerentes e semanticamente consistentes, dificultando a distinção entre informações verídicas e falsas (ALGHAMDI; LUO; LIN, 2024). Essa evolução demanda a adoção de abordagens mais sofisticadas, que incorporem a análise superficial de palavras isoladas ou estruturas linguísticas, e também uma compreensão profunda do contexto discursivo, da intencionalidade e da consistência factual. Nesse sentido, técnicas baseadas em *deep learning* (aprendizado profundo) (REUTHER *et al.*, 2019), particularmente modelos *transformer* (VASWANI *et al.*, 2017), emergem como soluções promissoras devido à sua capacidade de processamento de linguagem natural (PLN) em nível semântico, permitindo a identificação de nuances e inconsistências sutis que métodos tradicionais não conseguem capturar (ALGHAMDI; LUO; LIN, 2024).

A aplicação de arquiteturas *transformer*, como Gemini e GPT, tem se destacado na tarefa de detecção de desinformação devido à sua habilidade de modelar relações contextuais de longa distância em textos, capturando dependências semânticas e pragmáticas que são críticas para distinguir entre informações autênticas e manipuladas (OAD *et al.*, 2024). Esses modelos, treinados baseados em grandes conjuntos de dados textuais, são capazes de extrair representações vetoriais ricas em significado, possibilitando a identificação de padrões enganosos, como contradições internas, tom persuasivo tendencioso ou incoerências factuais (OAD *et al.*, 2024). Técnicas de *fine-tuning* (ANISUZZAMAN *et al.*, 2025) podem adaptar esses modelos para domínios específicos, aumentando sua precisão em cenários complexos, como notícias políticas ou científicas.

A detecção de notícias falsas em larga escala configura-se como um desafio computacionalmente complexo, especialmente quando considerados os requisitos de escalabilidade (ALGHAMDI; LUO; LIN, 2024), multimodalidade e processamento de grandes volumes de dados (*big data*) (ALGHAMDI; LUO; LIN, 2024). A natureza dinâmica da desinformação, que circula em múltiplas plataformas (redes sociais, portais de notícias e aplicativos de mensagens) (ALGHAMDI; LUO; LIN, 2024), exige arquiteturas de processamento em larga escala (LI; ZHANG; MALTHOUSE, 2024), integrando diferentes modalidades, como texto, imagem, vídeo e metadados. Além disso, o processamento de Inteligência Artificial (IA) de modo geral exige demandas técnicas avançadas de paralelização, distribuição de carga e otimização de recursos computacionais (LI; ZHANG; MALTHOUSE, 2024).

A Inteligência Artificial avança significativamente a cada ano, com modelos mais robustos sendo introduzidos em uma curta janela de tempo (REUTHER *et al.*, 2019). Esses avanços são impulsionados por melhorias constantes em técnicas de aprendizado de máquina, aumento na capacidade de processamento de dados e a disponibilidade de grandes volumes de informações para treinar esses modelos.

A partir de 2018, o desenvolvimento na área de IA ganhou um impulso significativo com o surgimento e popularização da IA generativa (Gemini Team *et al.*, 2024). Essa categoria de modelos é capaz de criar conteúdo original, incluindo textos, imagens, músicas e até vídeos, com uma qualidade que muitas vezes se aproxima ou até supera a produção humana.

Tecnologias que são denominadas como *Large Language Models* (LLM) (LI; ZHANG; MALTHOUSE, 2024) são caracterizadas pela capacidade dos modelos de entender e gerar linguagem natural. Esses modelos são treinados em grandes volumes de dados textuais e possuem bilhões de parâmetros, o que lhes permite captar nuances linguísticas, contextos complexos e padrões de linguagem. Como resultado, LLMs podem realizar uma ampla gama de tarefas, como tradução de idiomas, criação de conteúdo, resposta a perguntas e resumo de textos, com uma precisão que continua a melhorar à medida que esses modelos evoluem (WÖLFLEIN *et al.*, 2025). A capacidade desses modelos de compreender, prever e gerar texto de maneira coerente e relevante, é o que os torna uma ferramenta poderosa e versátil em diversas aplicações. Os modelos Gemini (Gemini Team *et al.*, 2024), e GPT (OpenAI, 2023) demonstram o potencial vasto dessa nova era, possibilitando aplicações inovadoras em diversas áreas, incluindo comunicação, entretenimento, educação e negócios. Esta evolução reconfigura os paradigmas de interação humano-computador e também suscita questões fundamentais concernentes à ética, à privacidade e às perspectivas futuras das relações de trabalho.

Acoplado aos avanços da IA, observa-se a crescente utilização de agentes inteligentes, especialmente aqueles baseados em LLMs, no contexto da detecção e combate à desinformação (LI; ZHANG; MALTHOUSE, 2024). De modo geral, um agente é definido como uma entidade computacional autônoma capaz de perceber o ambiente em que está inserido, processar informações e executar ações orientadas a objetivos específicos, podendo adaptar seu comportamento de acordo com as mudanças do ambiente (WÖLFLEIN *et al.*, 2025). Quando impulsionados por LLMs, esses agentes adquirem capacidades avançadas de compreensão semântica, geração de linguagem natural e tomada de decisão baseada em grandes volumes de dados textuais. No domínio da detecção de notícias falsas, agentes baseados em LLMs podem realizar tarefas complexas, como análise automática de credibilidade de informações, rastreamento de fontes, verificação de fatos em tempo real e geração de relatórios explicativos sobre a veracidade de conteúdos (LI; ZHANG; MALTHOUSE, 2024). Esses agentes tendem a aumentar a escalabilidade das estratégias de mitigação de desinformação e também introduzem um novo patamar de sofisticação, ao serem capazes de interagir dinamicamente com usuários e sistemas, aprimorando continuamente seus modelos internos a partir de *feedbacks* e novos dados (LI; ZHANG; MALTHOUSE, 2024).

Esse avanço se deve ao progresso significativo nos estudos de modelos de

linguagem, combinado com o aumento exponencial do poder computacional necessário para treinar esses modelos, com o uso de GPUs (unidades de processamento gráfico) e TPUs (unidades de processamento de tensor) de alta performance, permitindo que redes neurais cada vez maiores e mais complexas sejam treinadas em prazos razoáveis (REUTHER *et al.*, 2019). Nos últimos anos, o desenvolvimento de arquiteturas avançadas de redes neurais, como *Transformers*, possibilitou a criação de modelos de linguagem cada vez mais sofisticados e precisos, capazes de entender e gerar conteúdo com alta qualidade e coerência.

A arquitetura dos *transformers* foi introduzida por (VASWANI *et al.*, 2017), que traz o inovador mecanismo de *attention*, que é a base da arquitetura *transformer*, permitindo que o modelo analise a importância relativa de cada palavra em uma sequência de texto (VASWANI *et al.*, 2017). Durante o processamento, o mecanismo calcula pesos de atenção (*attention scores*) que indicam quanto o modelo deve "prestar atenção" em diferentes partes da entrada para realizar sua tarefa (VASWANI *et al.*, 2017). Em vez de tratar todas as palavras igualmente, como em modelos tradicionais, o *transformer* usa essas pontuações para focar nas palavras mais relevantes em cada contexto (VASWANI *et al.*, 2017). Essa abordagem torna o processamento mais eficiente, permitindo paralelização e como resultado, os *transformers* se tornaram extremamente eficazes em tarefas como tradução automática, geração de texto e compreensão de linguagem natural (VASWANI *et al.*, 2017).

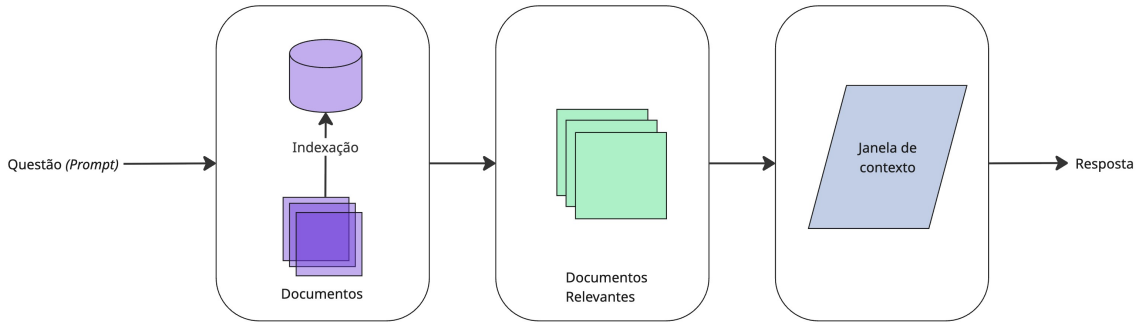
No contexto de *transformers*, existem técnicas e metodologias para trabalhar junto ao contexto dos LLMs, o *Retrieval Augmented Generation* (RAG) é uma das técnicas que possibilita a geração de texto para gerar uma resposta contextualizada. O RAG (GAO *et al.*, 2024) é um processo avançado de geração de texto que capacita modelos a produzir conteúdo a partir de uma combinação de geração baseada em aprendizado profundo e recuperação de informações relevantes de bases de dados extensas e externas às utilizadas para o treinamento do LLM. Esse método funciona integrando mecanismos de busca que recuperam dados pertinentes ao contexto ou ao tema solicitado, que são então utilizados pelo modelo de geração para criar textos mais informativos, precisos e contextualizados.

O RAG é especialmente eficaz em tarefas que exigem conhecimento atualizado ou especializado, pois permite ao modelo acessar e incorporar informações recentes e específicas diretamente de fontes confiáveis. Isso melhora significativamente a qualidade e a relevância do texto gerado, superando as limitações dos modelos puramente generativos que dependem exclusivamente do conhecimento armazenado durante o treinamento. Além disso, essa abordagem pode ser adaptada para uma ampla variedade de aplicações, desde assistentes virtuais que fornecem respostas detalhadas e precisas até sistemas de suporte ao cliente que oferecem soluções informadas e contextualmente apropriadas.

A implementação de um *pipeline* de um RAG é essencial para otimizar a gestão de documentos e a recuperação de informações em diversos contextos empresariais e

organizacionais. Esse fluxo começa com a recuperação de informações, na qual técnicas avançadas de busca e indexação localizam dados relevantes a partir de grandes repositórios de informações estruturadas e não estruturadas. Em seguida, esses dados são processados por modelos de geração de texto que sintetizam a informação recuperada, produzindo respostas ou relatórios detalhados e contextualmente apropriados. Por exemplo, no atendimento ao cliente, o RAG pode fornecer respostas imediatas e precisas a consultas complexas, melhorando significativamente a experiência do usuário e reduzindo o tempo de espera. Em contextos corporativos, como financeiro ou jurídico, o *pipeline* pode automatizar a extração de informações críticas e a geração de relatórios customizados, adaptando-se às necessidades específicas de cada situação. A personalização é um aspecto vital, permitindo que o RAG atenda a diferentes departamentos dentro de uma organização, enquanto sua escalabilidade possibilita a aplicação tanto em pequenas empresas quanto em grandes corporações. Com sua capacidade de inovar e melhorar a eficiência operacional, o RAG posiciona a organização de maneira competitiva no mercado, demonstrando um compromisso com a modernização e a excelência na gestão de informações. A Figura 1 representa o fluxo essencial de RAG.

Figura 1 – Fluxo e etapas de *Retrieval-Augmented-Generation*



Fonte: Do Autor

Um dos métodos para realizar a recuperação da informação em um conjunto de documentos, é utilizar o modelo BERT para geração de *embeddings* (REIMERS; GUREVYCH, 2019). Os *embeddings* são conjuntos numericos que representam a similaridade entre uma *query* ou *prompt* de entrada e o conjunto de documentos que serão associados para a resposta, a partir dessa modelagem é gerado um score de similaridade sintatica e semântica para o contexto de pergunta e resposta. Um outro exemplo de método para a recuperação de documentos dentro desse contexto é o BM25, que se trata de um algoritmo tradicional de recuperação da informação (TROTMAN; PUURULA; BURGESS, 2014), é uma variante do modelo de espaço vetorial que utiliza a frequência dos termos para avaliar a relevância dos documentos em relação a uma consulta. O BM25, ou *Best Matching* 25,

aplica uma fórmula probabilística que pondera a frequência dos termos e a normalização do comprimento do documento, permitindo um balanceamento eficiente entre a quantidade de termos correspondentes e a penalização de documentos excessivamente longos.

No contexto de RAG, a etapa de *Augmentation*, por sua vez, possibilita a geração de saídas mais estilizadas e personalizadas, adaptando-se a diferentes formatos de dados de entrada e demandando conteúdos específicos por meio de conjuntos de dados direcionados. O RAG, introduzido por Lewis (LEWIS *et al.*, 2020), representa um marco na área de processamento de linguagem natural. Essa inovadora técnica combina a força de modelos de linguagem pré-treinados (LLMs) com a capacidade de busca de informações relevantes, transcendendo as limitações dos modelos tradicionais.

Um outro método para trabalhar no contexto de LLM, é realizar o *fine-tuning* (ajuste de pesos do modelo para um objetivo específico (ANISUZZAMAN *et al.*, 2025)) do modelo LLM em si, com o conteúdo contextual para a geração de resposta, é ilustrado pela Figura 2. Existem diferenças significativas entre o RAG e o *fine-tuning* para este fim (LEWIS *et al.*, 2020). O que torna imprescindível a análise crítica sob qual o melhor caminho para ser seguido em determinado cenário de trabalho.

Figura 2 – *Fine-tuning* em um modelo de LLM

Fonte: Do Autor

A sinergia resultante tanto do *fine-tuning* quanto de um processo de *Retrieval* à partir de um conjunto de documentos previamente preparados, na etapa de *Generation*, aumenta a capacidade de generalização do modelo, embora também aumente o consumo de recursos.

O RAG se diferencia pela utilização de memórias paramétrica e não paramétrica. A memória paramétrica, geralmente um modelo LLM (*Large Language Model*) pré-treinado, fornece a base para a geração de texto. Já a memória não paramétrica, como um índice vetorial denso ou esparso de dados de um determinado contexto, atua como um repositório de informações externas. A chave reside na integração dessas duas memórias através de um modelo de recuperação pré-treinado.

Em resumo, o RAG funciona como um sistema híbrido que combina a busca por informações com a geração de texto. Através de um mecanismo de recuperação, o modelo extrai as partes mais relevantes de documentos relevantes de um *corpus*. Na sequência, essas informações extraídas são utilizadas por um LLM para gerar respostas precisas e informativas.

A investigação de fenômenos complexos como a desinformação demanda alicerces metodológicos que transcendam a dimensão puramente tecnológica, exigindo uma estruturação científica capaz de articular teoria e prática. Neste contexto, o método quadripolar, proposto originalmente por (BRUYNE; HERMAN; SCHOUTHEETE, 1977), estabelece-se como uma abordagem fundamental no âmbito da Ciência da Informação, organizando o processo investigativo em quatro eixos dialéticos e complementares: epistemológico, teórico, morfológico e técnico (GOUVEIA, 2022). A aplicação deste método permite uma análise integral do objeto de estudo: o polo epistemológico garante a vigilância crítica sobre os conceitos e as premissas do conhecimento produzido; o polo teórico sistematiza os modelos explicativos e hipóteses que fundamentam a interpretação do fenômeno; o polo morfológico examina a configuração estrutural e as formas de manifestação do objeto; e, finalmente, o polo técnico instrumentaliza a coleta e o tratamento de dados para a construção de evidências (GOUVEIA, 2022). Essa participação quadripolar equilibra a abstração conceitual com a verificação prática, assegurando que a solução proposta opere funcionalmente e também esteja ancorada em um rigor científico que fortalece a interdisciplinaridade e a robustez dos resultados obtidos (GOUVEIA, 2022).

A partir do contexto estabelecido pelo RAG, introduz-se o conceito de agentes. Segundo (SUN; HUANG; POMPILI, 2025), um agente constitui um elemento de inteligência artificial capaz de realizar decisões, interpretações, ações e cálculos. Nesse contexto, emerge a estrutura de multiagentes descrita por (SUN; HUANG; POMPILI, 2025), base do presente projeto. Trata-se de uma arquitetura que combina múltiplos agentes, cada qual com função e objetivo específicos, operando de forma independente e autônoma, sendo cada agente responsável por gerar insumos para a produção de um resultado final.

1.1 O MÉTODO QUADRIPOLAR

O Método Quadripolar (MQ), proposto por (BRUYNE; HERMAN; SCHOUTHEETE, 1977), constitui uma estrutura metodológica para análise de objetos de estudo mediante integração de abordagens qualitativas e quantitativas em perspectiva interdisciplinar. O método opera sobre quatro dimensões fundamentais (polos): epistemológica, teórica, morfológica e técnica.

Gouveia (2022) examina aplicações do MQ em investigações de caráter aplicado-metodológico, demonstrando sua adequação ao estudo de objetos complexos. O método funciona como *framework* integrador, orientando desde a formulação do problema de pes-

quisa até a coleta e análise de dados, incluindo a avaliação conceitual e a verificação do estado do objeto investigado. Embora predominantemente qualitativo, o MQ incorpora técnicas quantitativas subordinadas à coerência epistemológica, na qual a validação científica resulta da articulação entre teoria e prática. Esta característica torna o método apropriado para pesquisas em Ciência da Informação e Infocomunicação, nas quais fenômenos digitais complexos demandam tanto reflexão crítica quanto aplicabilidade técnica (GOUVEIA, 2022).

O MQ distingue-se de métodos lineares por sua natureza iterativa (GOUVEIA, 2022), permitindo influência mútua entre os quatro polos durante a investigação. Esta propriedade se mostra necessária em estudos que requerem flexibilidade metodológica, como aqueles envolvendo comportamentos informacionais que necessitam de lidar com um determinado grau de incerteza, nos quais modelos rígidos limitariam a compreensão de realidades em transformação. O MQ configura-se, assim, como instrumento para produção de conhecimento em contextos interdisciplinares e sociotecnológicos (GOUVEIA, 2022).

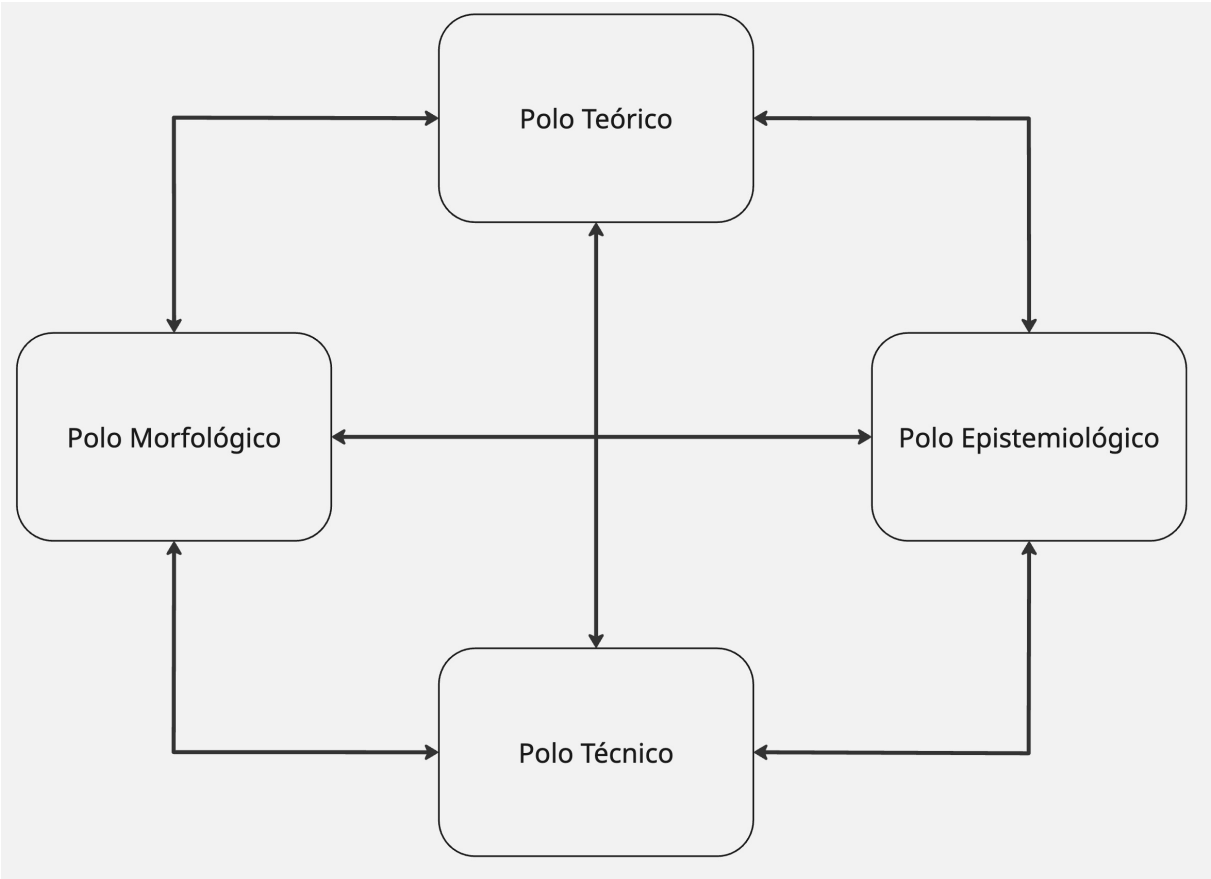
A Figura 3 representa o fluxo iterativo entre os polos, nos quais fundamentação teórica, modelagem conceitual, reflexão epistemológica e operacionalização técnica mantêm fluxo contínuo de influências recíprocas, sem sequência linear rígida. A coerência metodológica deriva da articulação simultânea entre os elementos, permitindo ajustes durante o processo investigativo.

A disposição não sequencial dos polos reflete a natureza dialética do MQ, na qual teoria e prática se informam mutuamente (BRUYNE; HERMAN; SCHOUTHEETE, 1977), enquanto a epistemologia valida escolhas morfológicas e técnicas. Questões emergentes no polo técnico podem demandar revisões no polo teórico, exigindo realinhamentos no polo morfológico e justificativas no polo epistemológico.

1.2 OBJETIVO E RELEVÂNCIA

Diante disso, o presente projeto consiste em uma arquitetura inovadora que combina IA com o método quadripolar, do qual cada polo é operacionalizado por um Agente baseado em *Large Language Model* (LLM) especializado em uma função distinta, integrando técnicas de *Retrieval-Augmented Generation* (RAG) e acesso dinâmico a bases de dados externas via APIs (*Application Programming Interface*) de busca. O escopo consiste em uma arquitetura baseada em multiagentes, que será detalhada nos parágrafos a seguir. Ela é projetada para decompor o processo de detecção de notícias falsas em etapas lógicas, sendo cada etapa correspondente a um dos quatro polos do método quadripolar: coleta de evidências, contextualização de conceitos, análise do formato e viés e, finalmente, veredito explicativo. Cada agente da estrutura, representa uma ação específica dentro desse fluxo, permitindo uma análise multifacetada e interpretável do conteúdo suspeito. A etapa de veredito classifica a veracidade da informação e gera explicações baseadas nas evidências recuperadas e nos padrões identificados durante o processamento. Esta abordagem híbrida

Figura 3 – Os quatro polos do método quadripolar



Fonte: Adaptado de (GOUVEIA, 2022)

integra a capacidade generativa de LLMs com a precisão de sistemas de recuperação de informação, fundamentada em princípios da Ciência da Informação. O objetivo é superar limitações de modelos com baixa explicabilidade e de arquiteturas de inferência baseadas em etapa única de instrução via *prompt*, proporcionando uma solução escalável e adaptável a diferentes domínios de desinformação.

A relevância deste projeto emerge da crescente complexidade e impacto da disseminação de notícias falsas em múltiplos domínios sociais, políticos e econômicos. Pesquisas recentes, como a de (ALGHAMDI; LUO; LIN, 2024), revelam que conteúdos fraudulentos, particularmente em contextos eleitorais, atingem índices alarmantes de engajamento, amplificando a desinformação de forma sistêmica nas plataformas digitais. Tais fenômenos distorcem percepções públicas que geram instabilidade social e desinformam comunidades inteiras. Em face dessa conjuntura, torna-se imperativo o desenvolvimento de soluções que detectem a falsidade textual, e que também consigam justificar suas conclusões de maneira compreensível e auditável. Nesse sentido, a proposta de um agente estruturado com base no método quadripolar oferece uma abordagem inovadora, pois permite a decomposição do processo analítico em etapas distintas e verificáveis. Dessa maneira, o projeto pode contribuir de forma significativa para a construção de um ecossistema

digital mais seguro, fundamentado em práticas de checagem de fatos automatizadas, escaláveis e fundamentadas, reforçando a integridade das informações circulantes e promovendo uma cultura de responsabilização informacional.

A detecção precisa e escalável de notícias falsas é um desafio multidimensional, conforme destacado por (AHMED; TRAORE; SAAD, 2017), pois envolve a análise linguística de textos, e também a interpretação e investigação de contextos sociopolíticos em constante mutação. Sistemas tradicionais de aprendizado de máquina, embora aplicáveis à classificação inicial, apresentam limitações na captura de nuances semânticas e discursivas, na coleta de evidências em tempo real, no cruzamento e análise de dados não estruturados e na adaptação a táticas emergentes de desinformação, tais como *deepfakes* textuais e narrativas híbridas parcialmente verídicas. O estudo de (ALGHAMDI; LUO; LIN, 2024) reforça essa lacuna ao demonstrar que conteúdos enganosos exploram vieses cognitivos e mecanismos de plataformas para maximizar engajamento, exigindo modelos capazes de processar grandes volumes de dados em tempo quase real. A arquitetura proposta aborda esse requisito através da modularidade multiagentes, que pode permitir distribuição computacional e paralelização de tarefas, um fator essencial para aplicações em ambientes de alta demanda, como monitoramento eleitoral ou crises de saúde pública. Adicionalmente, a precisão é incrementada pela combinação de LLMs contextualizados, que seguem uma linha de raciocínio especializada, reduzindo falsos positivos através de checagens cruzadas entre polos.

A Figura 4 ilustra a inserção do Método Quadripolar no ecossistema multiagente proposto para a detecção de notícias falsas. Cada polo do método (morfológico, teórico, epistemológico e técnico) é operacionalizado como um agente autônomo dotado de um LLM e de uma cadeia de raciocínio própria. O Agente Técnico, em particular, dispõe ainda de módulos de reranqueamento, expansão de consulta, RAG e ferramentas externas. Esses quatro agentes são coordenados por um Agente de Veredito, que integra as análises parciais por meio de sub-agentes, compondo assim a camada de arquitetura multiagente do sistema.

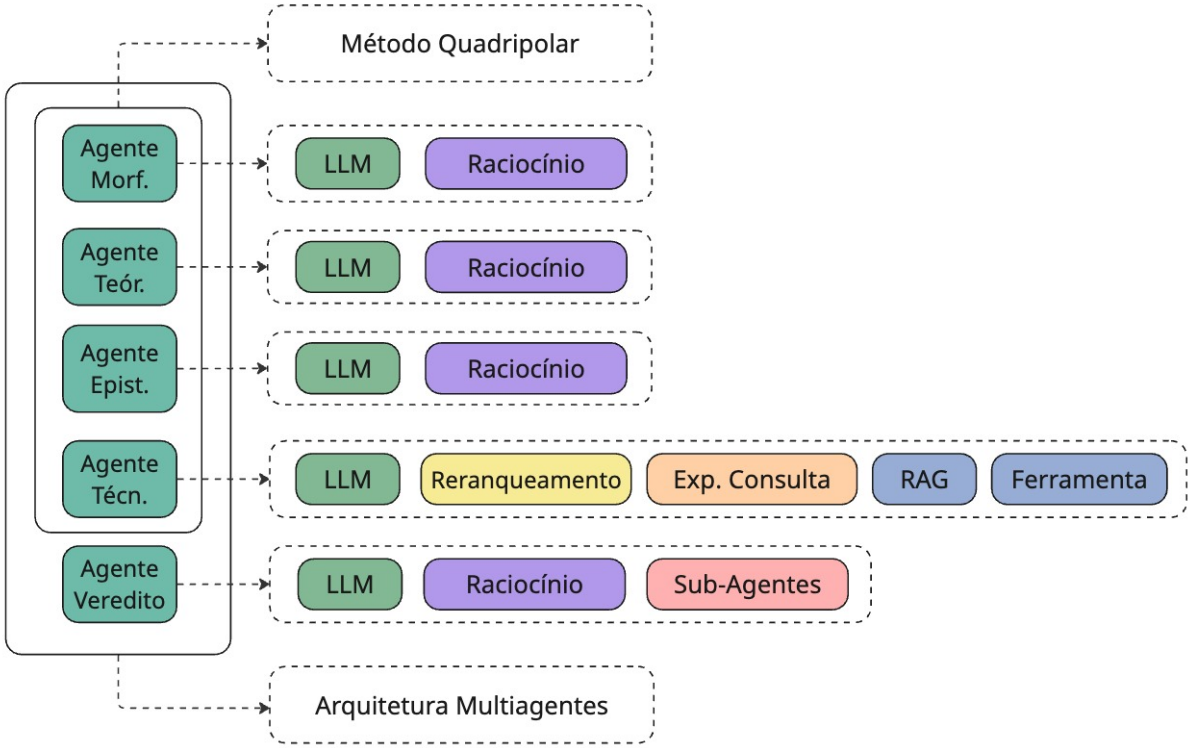
Os componentes estruturais ilustrados na Figura 4 delineiam a arquitetura do sistema proposto, organizando-se em camadas hierárquicas e módulos funcionais que interagem sinergicamente:

- **Método Quadripolar**: Estrutura analítica baseada no *framework* de De Bruyne (BRUYNE; HERMAN; SCHOUTHEETE, 1977), que organiza a investigação em quatro polos: Epistemológico, Teórico, Morfológico e Técnico. Esta abordagem permite examinar a informação considerando sua origem, contexto, forma e materialidade.
- **Arquitetura Multiagentes**: Sistema distribuído composto por agentes autônomos especializados que colaboram na execução de tarefas complementares de verificação,

conforme o paradigma de sistemas multiagentes aplicados a processamento de linguagem natural (SUN; HUANG; POMPILI, 2025).

- **Raciocínio**: Processo de inferência estruturado que orienta a tomada de decisão dos agentes, permitindo a decomposição de problemas complexos em etapas lógicas sequenciais e a síntese de conclusões fundamentadas (SUN; HUANG; POMPILI, 2025).
- ***LLMs (Large Language Models)***: Modelos de linguagem de grande escala que realizam a interpretação semântica de texto, inferência lógica e geração de explicações em linguagem natural, constituindo a camada de processamento cognitivo do sistema (VASWANI *et al.*, 2017).
- **Agente**: Unidade funcional autônoma que executa tarefas específicas (decomposição de conteúdo, validação de fontes) através da orquestração de ferramentas e da interface com o LLM subjacente (DONG *et al.*, 2024).
- **Expansão de Consulta (*Query Expansion*)**: Técnica de enriquecimento semântico que gera termos relacionados e contextos associados à consulta original, ampliando a cobertura da recuperação de informações (NAGPAL, 2022).
- **Reranqueamento (*Re-ranking*)**: Etapa de refinamento que reordena documentos recuperados utilizando modelos de *cross-encoder* (BAŞ *et al.*, 2022), atribuindo escores de relevância baseados na correspondência semântica entre consulta e evidência.
- ***RAG (Retrieval-Augmented Generation)***: Arquitetura que integra geração de linguagem natural com recuperação de informações externas, ancorando as respostas do modelo em dados factuais recuperados (GAO *et al.*, 2024).
- **Ferramenta (*Tools*)**: Conjunto de recursos computacionais disponíveis aos agentes, incluindo extratores de metadados, validadores de URL e interfaces com bases de dados externas (WÖLFLEIN *et al.*, 2025).

Figura 4 – Inserção do Método Quadripolar no Ecossistema de Agentes Baseados em LLMs



Fonte: Do Autor

A Figura 4 reflete a proposta de um sistema híbrido, aplicando técnicas avançadas de IA que são potencializadas pelo método quadripolar e uma sequência de aplicação de conceitos relevantes para o presente projeto.

A integração desses componentes resulta em um uma arquitetura que articula explicabilidade, auditabilidade, precisão e transparência (Arquitetura Quadripolar). A estrutura permite que a contribuição de cada polo seja auditada e verificada, possibilitando seu aperfeiçoamento. A precisão é obtida mediante validação cruzada entre evidências recuperadas dinamicamente (via RAG) e a análise contextual estruturada pelo método quadripolar, garantindo que as decisões sejam fundamentadas em múltiplas camadas de verificação. A transparência é assegurada pela rastreabilidade explícita em cada etapa do processamento, da recuperação de dados à geração do veredito, facilitando a auditoria e a interpretabilidade das saídas do sistema. Essas características visam superar as limitações de abordagens monolíticas, oferecendo uma solução adaptável a cenários complexos e em evolução de desinformação.

1.3 ORGANIZAÇÃO DO TEXTO

Este trabalho está estruturado em sete capítulos, abordando desde conceitos fundamentais até a proposta metodológica:

- **Capítulo 2 - Transformers e LLMs: Arquitetura e Estrutura**
  Apresenta a definição dos modelos de *transformers*, descreve sua arquitetura básica e explora em detalhes a estrutura e o funcionamento do modelo BERT e GPT que são referências fundamentais para o desenvolvimento de sistemas baseados em LLMs.

- **Capítulo 3 - *Retrieval-Augmented Generation* (RAG)**
  Introduz o conceito de RAG, detalhando seu funcionamento, principais componentes, técnicas avançadas e a relevância dessa abordagem para o aprimoramento da geração de respostas em modelos de linguagem.

- **Capítulo 4 - Detecção de Desinformação usando Aprendizado de Máquina**
  Aborda trabalhos relacionados à aplicação de Aprendizado de máquina em contextos de desinformação e notícias falsas, e discute os fundamentos de *Machine Learning* e Inteligência Artificial.

- **Capítulo 5 - Proposta Metodológica**
  Propõe uma metodologia baseada na construção de um agente inteligente, fundamentado no Método Quadripolar, destinado à detecção de notícias falsas por meio da integração de técnicas de IA e RAG.

- **Capítulo 6 - Resultados**
  Elucida todos os resultados de todas as análises pertinentes definidas na Proposta Metodológica, resultados Qualitativos, Quantitativos, e Estudos de Casos definidos para a aplicação da arquitetura quadripolar.

- **Capítulo 7 - Considerações Finais**
  Sintetiza os principais pontos conclusivos diante da abordagem de aplicação do método quadripolar para detecção e explicação de notícias falsas e apresenta as reflexões finais sobre o desenvolvimento da pesquisa, adentrando em oportunidades futuras.

2 *TRANSFORMERS* E LLMS: ARQUITETURA E ESTRUTURA

A capacidade de geração de texto está relacionada à habilidade de produzir linguagem natural de forma semelhante à comunicação humana. Essa capacidade é viabilizada pelos modelos de linguagem, que interpretam e reproduzem padrões linguísticos a partir de grandes volumes de dados. A arquitetura subjacente a esses modelos baseia-se nos *transformers*, estruturas que possibilitam o aprendizado de representações contextuais por meio do processamento de milhões de *tokens* e do ajuste de parâmetros (MOHAMMADABADI *et al.*, 2025). É esse processo de aprendizado que permite a sincronização entre a produção textual da máquina e o entendimento humano.

No âmbito do processamento de linguagem natural, os *tokens* constituem as unidades fundamentais de informação manipuladas pelos modelos computacionais, não se limitando a palavras inteiras, mas abrangendo também caracteres, fragmentos de palavras ou até mesmo *bytes*. Em arquiteturas modernas baseadas em *Transformers*, a tokenização emprega estratégias avançadas de segmentação em subunidades (*subword tokenization*) (VASWANI *et al.*, 2017) para lidar eficientemente com a complexidade morfológica e o problema de palavras fora do vocabulário. Um exemplo elucidativo desse processo é a decomposição de termos compostos ou flexões raras: uma palavra como “inconstitucionalmente” pode ser fragmentada pelo modelo na sequência de *subtokens* “in”, “constitucional” e “mente”. Essa abordagem permite que o sistema processe e infira o significado do termo completo a partir da semântica de seus morfemas constituintes, garantindo uma capacidade de generalização superior e uma representação mais robusta da linguagem.

2.1 O QUE SÃO LLMS

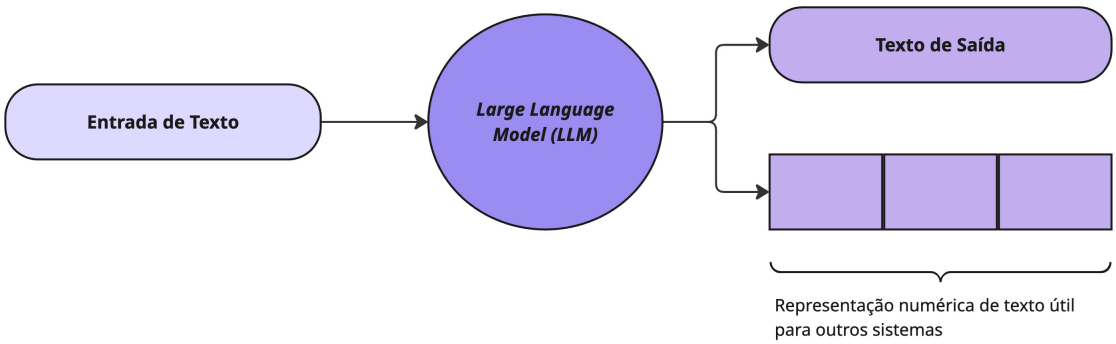
Os LLMs são modelos de aprendizado de máquina fundamentados em aprendizado profundo e estruturados sobre a arquitetura *Transformer*. Seu treinamento envolve bilhões de parâmetros e utiliza grandes conjuntos de dados textuais, compostos por artigos, livros, páginas da *web* e conteúdos de domínios específicos (VASWANI *et al.*, 2017). Essa combinação de arquitetura e volume de dados confere aos LLMs a capacidade de gerar texto e compreender linguagem natural.

O objetivo central de um modelo de linguagem é desenvolver a habilidade de prever a próxima palavra em uma sequência de texto, com base no contexto fornecido pelas palavras anteriores (MOHAMMADABADI *et al.*, 2025). Isso é possível graças ao treinamento em vastos conjuntos de dados textuais, onde o modelo aprende padrões e relações entre palavras, frases e ideias. Ao aprimorar essa capacidade preditiva, o modelo não apenas facilita a geração de texto coerente e fluido, mas também pode capturar nuances

de significado semântico, sintaxe e até mesmo contexto cultural (LI; ZHANG; MALTHOUSE, 2024). Essa capacidade torna esses modelos úteis para uma variedade de tarefas, como tradução automática, resumo de textos, geração de respostas em conversas, simulando de forma cada vez mais próxima a linguagem humana natural.

Um dos marcos disruptivos no desenvolvimento de Grandes Modelos de Linguagem (LLMs) foi a consolidação da arquitetura *Transformers* (VASWANI *et al.*, 2017). Neste cenário, o Google AI Language introduziu o modelo BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN *et al.*, 2018), estabelecendo um novo paradigma ao permitir o treinamento bidirecional profundo. Essa inovação trouxe avanços significativos no aprendizado de máquina, impulsionando o estado da arte em tarefas complexas como sistemas de Perguntas e Respostas (*Question Answering*), Classificação de Texto e mecanismos de busca semântica, além de auxiliar na identificação de entidades em contextos de engenharia de *software*. A Figura 5 ilustra o fluxo operacional deste modelo: a partir de um texto de entrada, o BERT processa as informações contextuais para gerar representações numéricas vetoriais (*embeddings*) e, subsequentemente, os resultados de saída pertinentes à tarefa específica.

Figura 5 – Fluxo de LLM para transformação de texto.



Fonte: Do Autor

2.2 VISÃO GERAL DO *TRANSFORMER*

Antes da arquitetura *Transformer*, as redes neurais recorrentes (RNNs) eram amplamente utilizadas em tarefas de processamento de linguagem natural. Essas redes, embora poderosas, apresentavam limitações, especialmente em lidar com dependências de longo prazo e no processamento sequencial. Em dezembro de 2017, o paper "*Attention is All You Need*" (VASWANI *et al.*, 2017) propôs uma nova abordagem para superar essas

limitações, eliminando a necessidade de recorrência e introduzindo uma nova forma de processamento de informação, sendo ele, o mecanismo de atenção (*attention*).

A principal inovação do *Transformer* foi a substituição da recorrência, que era uma característica central das RNNs, pelo mecanismo de atenção (VASWANI *et al.*, 2017). Esse mecanismo permitia que o modelo focasse diretamente em diferentes partes da entrada, independentemente de sua posição na sequência, permitindo assim uma compreensão mais flexível e paralela das dependências entre palavras. No entanto, ao remover a recorrência, o modelo se distanciava da sequencialidade inerente à linguagem. Para compensar isso, os *Transformers* necessitam de várias camadas de atenção para alcançar um bom desempenho, enquanto as RNNs normalmente precisavam de menos camadas. Sua estrutura principal é composta pelas etapas de *Encoding, Decoding, Positional Encoding*, que serão detalhadas nesta seção.

Positional Encoding* - Ordem dos *tokens

A ausência de mecanismos de recorrência na arquitetura original do *Transformer* implica que o modelo não possui, nativamente, o conhecimento sobre a ordem sequencial dos *tokens* (por exemplo, distinguir a posição de sujeito ou objeto em uma frase). Para solucionar essa limitação e permitir o processamento paralelo sem perda de contexto, a arquitetura introduz a etapa de Codificação Posicional (*Positional Encoding*) (VASWANI *et al.*, 2017).

Conforme ilustrado na Figura 6, o processo ocorre em paralelo, os *embeddings* de entrada (vetores densos que representam o significado semântico da palavra) são somados elemento a elemento a vetores de posição específicos (VASWANI *et al.*, 2017). Estes vetores posicionais são gerados deterministicamente através de funções matemáticas que determinam a posição do elemento em um espaço vetorial geométrico (VASWANI *et al.*, 2017).

Para compreender a mecânica da injeção de posicionamento, é necessário observar a operação em nível vetorial. Para elucidar o funcionamento da mecânica de Posição, elucidada no trabalho de Vaswani *et al.* (2017), a Figura 6, foi desenhada pelo autor deste presente texto para melhor exemplificar o funcionamento do processo para uma sentença de entrada $E = \{token1, token2, token3\}$ (ex: “O jogador joga“), sendo, seguindo as etapas de Mapeamento Semântico, Geração da Posição, e Fusão:

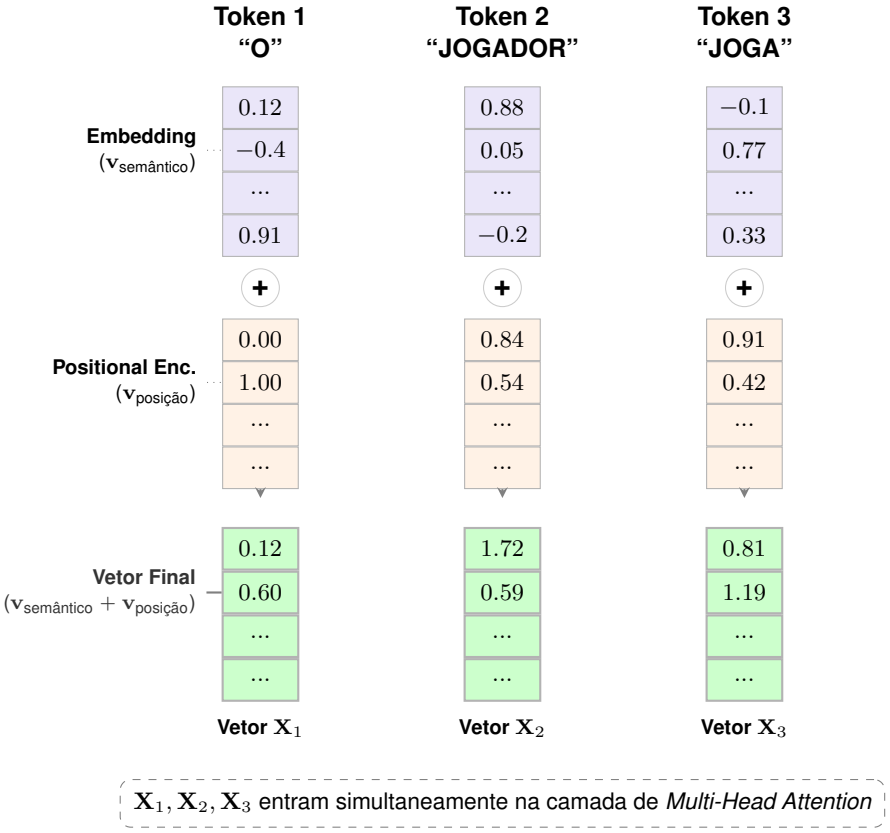
1. **Mapeamento Semântico:** Cada *token* é convertido em um vetor denso de dimensão Z de 512 dimensões.

2. **Geração da Posição:** Para cada índice de posição na frase (0, 1, 2...), gera-se um vetor único da mesma dimensão Z, utilizando frequências de seno e cosseno.

3. **Fusão:** Ocorre a soma vetorial $\mathbf{v}_{\text{final}} = \mathbf{v}_{\text{semântico}} + \mathbf{v}_{\text{posicão}}$.

O parâmetro Z define a dimensionalidade constante do espaço vetorial em que a informação é processada. O vetor $\mathbf{v}_{\text{semântico}}$ representa o *embedding* de conteúdo, enquanto o $\mathbf{v}_{\text{posição}}$ codifica a ordenação sequencial dos *tokens* dentro da sentença. A resultante dessa operação, $\mathbf{v}_{\text{final}}$, que constitui a representação rica do *token*, representando o seu significado semântico individual e também o significado semantico com relação à sentença toda.

Figura 6 – Detalhamento da Soma Vetorial: *Embeddings* + *Positional Encoding* = Vetor Final



Fonte: Do Autor.

A Figura 6 elucida que o *Positional Encoding* altera levemente os valores originais do *embedding* de forma padronizada.

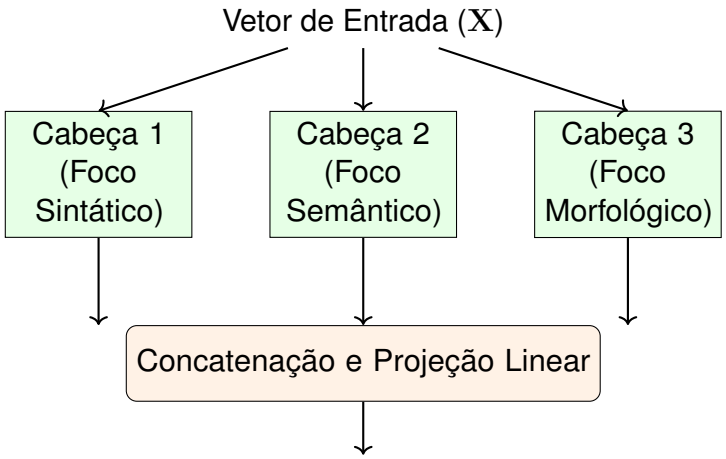
Os vetores $\mathbf{X}_1$, $\mathbf{X}_2$ e $\mathbf{X}_3$ representam os insumos finais de cada *token* após a união das informações de conteúdo e ordem. Esses vetores encapsulam simultaneamente o valor semântico ($\mathbf{v}_{\text{semântico}}$) e a posição do elemento dentro da sentença ($\mathbf{v}_{\text{posição}}$), permitindo que o modelo identifique, por exemplo, que o termo “O” ocupa a primeira posição da sentença. Essa representação unificada é o que possibilita o processamento paralelo pela camada de *Multi-Head Attention* (VASWANI *et al.*, 2017), servindo como a base numérica sobre a qual o mecanismo de atenção calculará as correlações entre todos os elementos da sequência de entrada.

O modelo aprende, durante o treinamento, a distinguir qual parte do sinal $\mathbf{X}_i$ vem da semântica e qual parte vem da posição, graças às frequências fixas utilizadas na codificação posicional.

Uma vez que os dados estão devidamente codificados, o *Transformer* aplica seu mecanismo central: o *Multi-Head Attention* (Atenção Multi-Cabeça), que é responsável por dividir o processamento de atenção em múltiplas "cabeças" independentes operando em paralelo.

A lógica deste processamento é detalhada na Figura 7. O diagrama é um exemplo ilustrativo para elucidar como o vetor de entrada é duplicado e projetado em diferentes subespaços de atenção. Cada "Cabeça" (*Head*) especializa-se em capturar um tipo distinto de relação linguística, por exemplo.

Figura 7 – Ilustração da Arquitetura de Processamento Paralelo (*Multi-Head*)



Fonte: Do Autor.

Conforme o fluxo apresentado na Figura 7, enquanto a *Cabeça 1* pode estar focada na concordância gramatical (sintaxe), a *Cabeça 2* pode simultaneamente analisar a relação sujeito-objeto (semântica). Os resultados dessas análises paralelas são, ao final, concatenados e projetados linearmente para formar uma única saída rica em contexto.

2.2.1 Transformação com modelo de linguagem

Antes de qualquer processamento no *Transformer*, o texto de entrada deve ser convertido em vetores numéricos (*embeddings*), enriquecidos com a *Positional Encoding*.

Considerando um modelo *Transformer* treinado para tradução automática do inglês para o português. Suponha que a frase de entrada em inglês seja:

Brazil has the best soccer players.

1. Entrada Codificada

A frase de entrada é processada pelo codificador (*encoder*) para gerar uma sequência de representações contínuas:

$$\mathbf{F} = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_5, \mathbf{p}_6, \mathbf{p}_7\},$$

onde cada $\mathbf{p}_i$ é uma representação contínua para cada palavra da frase "*Brazil has the best soccer players.*" Estes vetores **F** encapsulam o significado e contexto de toda a frase e servirão como referência para o decodificador.

Passo 1: Início da Decodificação

O processo de decodificação começa com um *token* especial de início de sequência (por exemplo, `<start>`):

$$\text{Entrada}_{\textit{(decoder)}} = \{\texttt{<inicio>}\}$$

Passo 2: Geração da Primeira Palavra

Usando a entrada `<inicio>`, o decodificador gera a primeira palavra da saída. Supondo que a palavra gerada seja "Brasil":

$$\text{Saída}_{\textit{(decoder)}} = \{\texttt{<inicio>}, \text{Brasil}\}$$

Passo 3: Geração de Palavras Subsequentemente (Natureza Autoregressiva)

O *token* "Brasil" é adicionado à entrada do decodificador e o processo é repetido. A sequência de entrada para o decodificador agora é:

$$\text{Entrada}_{\textit{(decoder)}} = \{\texttt{<inicio>}, \text{Brasil}\}$$

O decodificador gera a próxima palavra, por exemplo, "tem". Assim, a sequência de saída é atualizada para:

$$\text{Saída}_{\textit{(decoder)}} = \{\texttt{<inicio>}, \text{Brasil}, \text{tem}\}$$

Passo 4: Repetição do Processo

O *decoder* continua esse processo *token* a *token*, utilizando as representações codificadas (**Z**) como referência, até encontrar um *token* de fim de sequência (por exemplo, `<fim>`):

$$\text{Saída Final}_{\textit{(decoder)}} = \{\texttt{<inicio>}, \text{Brasil}, \text{tem}, \text{os}, \text{melhores}, \text{jogadores}, \text{de}, \text{futebol}, \texttt{<fim>}\}$$

Neste exemplo, o decodificador traduz a frase em inglês "*Brazil has the best soccer players.*" para o português "*Brasil tem os melhores jogadores de futebol*". O processo envolve iniciar com um *token* especial, gerar palavras sequencialmente e usar a atenção para considerar as representações codificadas da frase original. Cada palavra gerada é adicionada ao contexto, e a sequência é ajustada até que o *token* de fim de sequência seja produzido, esse comportamento é conhecido como natureza autoregressiva.

A natureza autorregressiva é um componente principal dos modelos de linguagem baseados na arquitetura *Transformer* (VASWANI *et al.*, 2017). Conforme elucidado no trabalho de (LEE, 2023), esse componente traz para comportamento de uma sequência sendo tratada como uma sucessão de decisões probabilísticas, em que a probabilidade de cada novo elemento (*token*) é derivada do histórico de elementos (*tokens*) precedente. Matematicamente, Lee (2023) define como o modelo estima a predição do próximo *token* por meio da Equação 2.1:

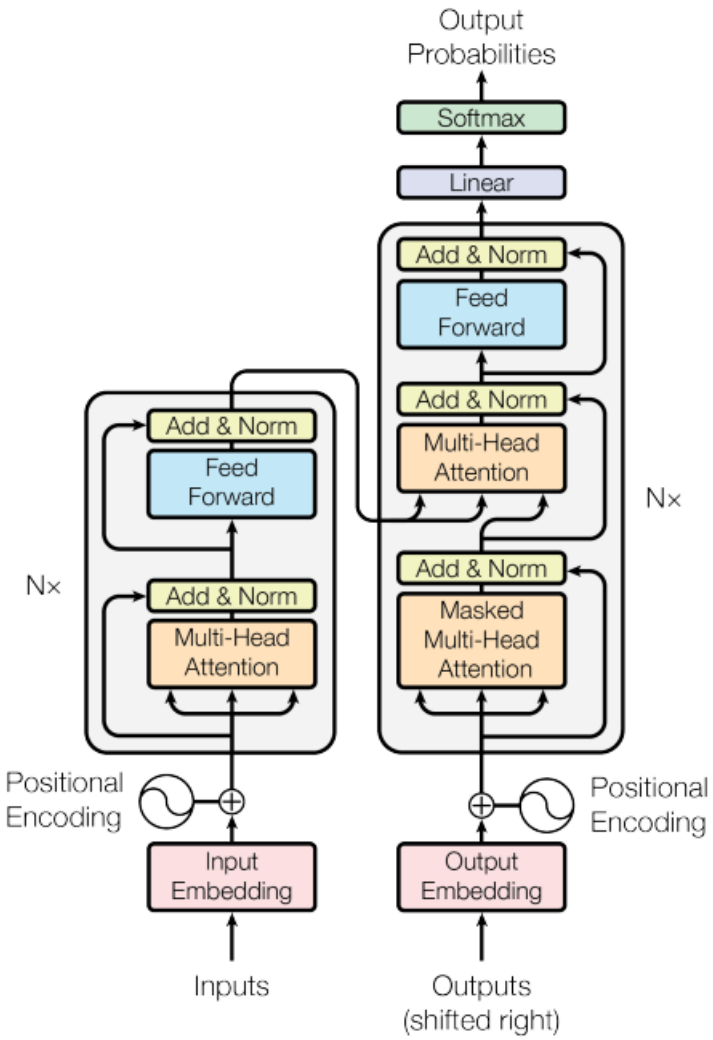
$$P(x_t \mid x_{<t}) = f_{\mathrm{O}}(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{2.1}$$

em que x_t define o *token* atual sob inferência e $x_{<t}$ representa o conjunto de *tokens* anteriores que compõem o contexto. A função f_{O}, parametrizada pelos pesos da rede neural (LEE, 2023), processa essa sequência prévia para projetar um conjunto de probabilidade sobre o vocabulário, garantindo que a saída seja construída de forma iterativa e semanticamente dependente do que foi processado anteriormente.

O *Transformer* de Vaswani *et al.* (2017) é composto por dois blocos principais: o *encoder* e o *decoder*, que trabalham em conjunto para transformar a entrada em saída. O bloco *encoder* é formado por múltiplos *encoders*, e o *decoder* por múltiplos *decoders*, todos com a mesma estrutura, mas com pesos diferentes (VASWANI *et al.*, 2017), cada *encoder* tem uma subcamada de *Self-Attention*, que permite ao modelo focar em diferentes partes da entrada simultaneamente, já os *decoders* possuem essas mesmas subcamadas, mas com a adição de uma camada extra de atenção, que auxilia o modelo a se concentrar nas partes mais relevantes da sequência de entrada, especialmente útil para tarefas de tradução (VASWANI *et al.*, 2017).

A Figura 8 representa a camada de *Encoder* e *Decoder*, juntamente com o contexto específico de blocos de *Attention*.

Figura 8 – Arquitetura do Modelo *Transformer*



Fonte: (VASWANI *et al.*, 2017)

Encoder

O codificador (*encoder*) do *Transformer* de Vaswani *et al.* (2017) consiste em camadas $N = 6$. Cada camada tem dois subníveis:

1. Subnível de *Self-Attention* e *Multi-Head Attention*

Este mecanismo é o componente que processa o contexto do *encoder*. A autoatenção (*Self-Attention*) permite que o modelo calcule a relevância de cada *token* em relação a todos os outros na sequência, independentemente da distância entre eles. O processo baseia-se na projeção da entrada em três vetores distintos: Consultas (*Queries* - Q), Chaves (*Keys* - K) e Valores (*Values* - V). O peso de atenção é calculado pelo produto escalar entre a consulta e a chave, escalado pela raiz quadrada da dimensão (d_k) e normalizado via função *softmax*. A extensão para *Multi-Head Attention* executa este cálculo em subespaços paralelos, permitindo que o modelo capture simultaneamente diferentes tipos de relações

semânticas e sintáticas em posições distintas da frase, cujos resultados são posteriormente concatenados (VASWANI *et al.*, 2017).

2. Subnível de *Feed-Forward Network* (FFN)

Vaswani *et al.* (2017) destaca que o elemento *Feed-Forward* é um componente de processamento individual para cada palavra da sentença, que busca a informação de cada posição e a submete a uma espécie de análise profunda daquele elemento. Para isso, o modelo expande o vetor para um espaço matemático maior, aplica um cálculo para captar nuances sutis da linguagem e, por fim, o comprime de volta ao tamanho original. Esse processo de "expandir e reduzir a dimensão do vetor" traz inteligência adicional para resolver problemas complexos, trazendo um nível de refinamento maior para o contexto.

3. Mecanismo de *Add & Norm*

Para garantir que o modelo consiga aprender padrões complexos sem se perder em cálculos instáveis, Vaswani *et al.* (2017) utilizou o bloco *Add & Norm*, ou conhecida como *LayerNorm*. Esse componente combina o conteúdo processado pelo modelo com o que ele acabou de processar, mantendo um padrão matemático (VASWANI *et al.*, 2017), seguindo a formulação:

$$\text{Saída} = \mathit{LayerNorm}(\mathbf{x} + \mathit{Sublayer}(\mathbf{x})). \tag{2.2}$$

Abaixo, o funcionamento desse mecanismo é detalhado em dois pontos:

- **Conexão Residual ("*Add*"):** Foca em aplicar um ajuste sobre a entrada original **x**. Ao somar **x** ao resultado *Sublayer(**x**)*, o modelo garante que a informação principal não seja esquecida, resolvendo o problema de redes profundas que "perdem o contexto" (ou *vanishing gradient*), que se trata quando uma rede neural obtém valores pequenos como sobra de um cálculo de ajuste em um processo de retropropagação, durante o treinamento (VASWANI *et al.*, 2017).
- **Normalização de Camada ("*Norm*"):** Após a soma, os números podem assumir escalas diferentes. A *LayerNorm* reajusta esses valores dentre -1 e 1, ou seja, valores de baixa magnitude (VASWANI *et al.*, 2017). Isso mantém os dados estáveis e evita que os cálculos da rede neural fiquem imprevisíveis.

Decoder

O decodificador (*decoder*) de Vaswani *et al.* (2017) é composto por 6 camadas. Além dos dois subníveis presentes no codificador, o decodificador introduz um terceiro subnível essencial, projetado para realizar a atenção *multi-head attention* sobre o resultado

do *Encoder*. Assim como no bloco anterior, empregam-se conexões residuais seguidas de normalização de camada em cada subnível (VASWANI *et al.*, 2017). O mecanismo de auto-atenção no decodificador é modificado para impedir que uma posição tenha acesso a posições subsequentes na sequência. Esse subnível articulado por Vaswani *et al.* (2017) (denominada *Masked Self-Attention*), é o componente que realiza o deslocamento dos *embeddings* de saída em uma posição, também tem a propriedade autoregressiva, assegurando que a predição para a posição seguinte da sequência dependa exclusivamente das saídas conhecidas nas posições anteriores a (VASWANI *et al.*, 2017).

1. Subcamada *Masked Self-Attention*

Esta subcamada implementa a lógica *autoregressiva* através de uma máscara que restringe o fluxo de informação. O objetivo é garantir que, durante o treinamento, o modelo não "veja o futuro". No *encoder* a atenção é bidirecional (todo *token* vê todo *token*), no *decoder* a atenção é causal, ou seja, é necessário que a técnica de modelagem realizada, tenha como necessidade ou não a visibilidade de *tokens* futuros, dependendo da mecânica de modelagem e do objetivo do modelo *Transformer* (VASWANI *et al.*, 2017).

Para exemplificar o funcionamento do componente *Masked Self-Attention* de Vaswani *et al.* (2017), considere por exemplo que o modelo está sendo treinado para gerar a frase "*O Brasil tem os melhores jogadores de futebol*". No momento em que o modelo está processando o *token* "Brasil" (posição 2), a máscara impede que ele acesse a informação do *token* "tem" (posição 3). Matematicamente, isso é realizado definindo os valores de atenção correspondentes às posições futuras como $-\infty$ (infinito) antes da aplicação da função *softmax*, resultando em uma probabilidade zero de atenção para esses elementos. Sem esse mecanismo, o modelo simplesmente copiaria a resposta alvo em vez de aprender a prever o próximo passo da sequência baseando-se no contexto anterior.

2. Subcamada *Encoder-Decoder*

Vaswani *et al.* (2017) traz a subcamada de *encoder-decoder* em sua arquitetura *Transformer*, como um componente que faz a integração entre dois blocos de processamento de conteúdo e a mecânica de atenção, permitindo que o *decoder* alinhe sua geração com o contexto extraído pelo *encoder*. Neste mecanismo, as Consultas (**Q**) provêm da camada anterior do próprio *decoder*, enquanto as Chaves (**K**) e os Valores (**V**) são projetados a partir da saída final do *encoder* (**Z**).

Isso permite que cada posição no decodificador atenda a todas as posições da sequência de entrada. Retomando o exemplo realizado anterior de tradução de uma frase do inglês para o portugûes (VASWANI *et al.*, 2017), se o *encoder* processou a frase em inglês "*Brazil has the best soccer players.*", e o *decoder* gerou até agora a palavra "Brazil", a camada de *Encoder-Decoder* permitirá que o modelo preste atenção nos vetores

correspondentes a “has” e “the” da entrada original para decidir que a próxima palavra correta a ser gerada é “best”. É através desta camada que o modelo aprende o alinhamento semântico entre a entrada (contexto) e a saída (geração).

Para exemplificar o funcionamento do componente *Encoder-Decoder* de (VASWANI *et al.*, 2017), considere a tradução de “*Brazil has the best soccer players.*” para “*Brasil tem os melhores jogadores de futebol*”. Suponha que o *encoder* já processou a frase em inglês e gerou os vetores de contexto $\mathbf{Z} = \{\mathbf{z}_{Brazil}, \mathbf{z}_{has}, \mathbf{z}_{the}, \mathbf{z}_{best}, \mathbf{z}_{soccer}, \mathbf{z}_{players}\}$. O *decoder* já gerou a palavra “the” e agora precisa predizer a próxima palavra (“best”):

1. **Geração da Consulta (Q):** A representação da palavra atual no *decoder* (“Brasil”) é transformada em um vetor de Consulta **Q**. Este vetor representa a pergunta: “*Dado que acabei de gerar um nome de um país, qual é a próxima informação relevante na frase original?*”.
2. **Comparação com Chaves (K):** O vetor **Q** é comparado (via produto escalar) com todas as Chaves do *encoder* ($\mathbf{k}_{Brazil}$, $\mathbf{k}_{has}, \mathbf{k}_{the}, \mathbf{k}_{best}, \mathbf{k}_{soccer}, \mathbf{k}_{players}$).
3. **Cálculo dos Pesos de Atenção:** O modelo calcula a relevância. A afinidade entre $\mathbf{k}_{Brazil}$ e $\mathbf{k}_{has}$ será alta, pois “has” é o verbo que deve seguir o artigo. A afinidade com “the” será menor neste momento.
4. **Agregação de Valores (V):** O mecanismo produz uma soma ponderada dos Valores do *encoder* ($\mathbf{v}_{Brazil}, \mathbf{v}_{has}, \mathbf{v}_{the}, \mathbf{v}_{best}, \mathbf{v}_{soccer}, \mathbf{v}_{players}$), dando mais peso ao vetor de “soccer”.
5. **Resultado:** O vetor resultante carrega a informação semântica de “*soccer*” trazida do inglês, mas adaptada ao contexto gramatical de “Brasil” do português, orientando a camada final a escolher a palavra “jogadores” como a próxima na sequência.

2.2.2 Arquitetura *Decoder-Only*

A arquitetura *decoder-only* representa um avanço em relação ao modelo *encoder-decoder* de Vaswani *et al.* (2017), sendo a estrutura amplamente utilizada por modelos de nova geração, como os da família GPT (ROBERTS, 2024). Seus componentes são similares aos da arquitetura *encoder-decoder*, possuem natureza autorregressiva e utiliza os mecanismos de *Masked Self-Attention*, *Feed-Forward* e normalização, conforme detalhado anteriormente nesta seção. A diferença é que esta arquitetura elimina a conexão com o *encoder*. Ao remover a necessidade de um codificador e de seus mecanismos de atenção, o modelo ganha em eficiência, pois não requer um número elevado de camadas dentro da rede neural (ROBERTS, 2024).

Embora a arquitetura original do *Transformer* proposta por (VASWANI *et al.*, 2017) tenha estabelecido uma estrutura híbrida de *Encoder-Decoder*, o estado da arte em

Grandes Modelos de Linguagem (LLMs) voltados para tarefas generativas e de raciocínio convergiu para a arquitetura *Decoder-only*. Esta abordagem utiliza o componente *Masked Self-Attention* para previsão autoregressiva de *tokens* e fundamenta os modelos centrais utilizados nesta pesquisa, como a família GPT da OpenAI (OpenAI, 2023), cujo potencial de aprendizado em escala foi demonstrado por (BROWN; MANN; RYDER, 2020) e aprimorado no GPT-4 (OpenAI, 2023), e a família Gemini do Google DeepMind, que também adota decodificadores aprimorados para processamento multimodal conforme o relatório técnico de Gemini Team *et al.* (2024), além dos modelos abertos LLaMA introduzidos por (TOUVRON *et al.*, 2023).

Em contraste com os modelos generativos, outras variações arquiteturais permanecem relevantes para tarefas específicas de Processamento de Linguagem Natural (PLN). A arquitetura *Encoder-only*, cujo mais popular é o modelo BERT proposto por Devlin *et al.* (2018), abdica da geração sequencial em favor de uma análise bidirecional profunda do contexto, e são utilizados em tarefas em que essa analise profunda e bidirecional sejam relevantes para o domínio em questão. São popularmente conhecidos como modelos de *embeddings* (GAO *et al.*, 2024). Paralelamente, a estrutura clássica *Encoder-Decoder* persiste em modelos como o T5 (*Text-to-Text Transfer Transformer*), apresentado por Raffel *et al.* (2020), que unifica diversos problemas de NLP em uma abordagem de entrada e saída textual, mantendo a fidelidade à proposta original de tradução e sumarização.

2.2.3 Equação de *Attention*

O mecanismo central do *Transformer* é a função de *Attention*, que permite ao modelo decidir e calcular em quais partes da sequência de entrada focar. A formulação matemática completa desta função é descrita pela Equação 2.3 (VASWANI *et al.*, 2017):

$$\textbf{\textit{Attention}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2.3}$$

O processamento desta equação ocorre em quatro etapas lógicas, operando sobre as matrizes de Consultas (**Q**), Chaves (**K**) e Valores (**V**):

1. **Cálculo de Similaridade ($\mathbf{QK}^T$):** Inicialmente, realiza-se o produto entre as consultas **Q** e a transposta das chaves $\mathbf{K}^T$. Esta operação calcula o produto escalar entre cada par de vetores, gerando uma matriz de valores que representam a correlação ou afinidade entre os *tokens* da sequência (VASWANI *et al.*, 2017).

2. **Escalonamento ($\frac{1}{\sqrt{d_k}}$):** As pontuações resultantes são divididas pela raiz quadrada da dimensão das chaves (d_k). Em dimensões elevadas, o produto escalar tende a produzir valores de grandes, com magnitude maior, que traria um problema para

o padrão de dimensão da sequência. O escalonamento previne isso, mantendo os valores em tamanhos padronizados durante o cálculo. (VASWANI *et al.*, 2017).

3. **Normalização Probabilística (*Softmax*):** É utilizada para converter os valores resultantes em uma distribuição de probabilidade. Isso garante que os pesos de atenção sejam positivos e que sua soma seja igual a 1, definindo a relevância relativa de cada *token* (VASWANI *et al.*, 2017).

4. **Agregação de Contexto:** Por fim, a matriz de probabilidades multiplica a matriz de Valores (**V**). O resultado é uma soma dos vetores de valor, onde a informação dos *tokens* considerados mais relevantes é preservada e enfatizada na saída final da camada (VASWANI *et al.*, 2017).

2.2.4 Equação *Multi-Head Attention*

O mecanismo de *Multi-Head Attention* é utilizado para lidar com a dificuldade de focar em vetores de diferentes posições e diferentes aspectos semânticos simultaneamente. Em uma única operação de atenção, a média ponderada dos vetores pode perder informações cruciais que ocorrem em diferentes partes da sentença. O *Multi-Head Attention* resolve isso permitindo que o modelo execute mecanismos de atenção (“cabeças“) em paralelo conforme ilustrado pela Figura 7. Cada um focado em um subespaço de representação distinto (VASWANI *et al.*, 2017).

Para exemplificar a necessidade deste componente elucidado por (VASWANI *et al.*, 2017), considere a frase: “*Brasil tem os melhores jogadores de futebol*”. Para que o modelo compreenda corretamente a palavra “melhores”, é necessário capturar diferentes tipos de relações:

- **Cabeça 1:** Precisa focar fortemente na relação entre “futebol” e “Brasil” (Qual a relação de país com o jogo?).
- **Cabeça 2 (Relação Sintática/Verbal):** Precisa associar “jogadores” à atribuição “melhores” ou o estado ser de “mais habilidosos daquele lugar”, “que lugar?”, ‘Brasil‘.

Se houvesse apenas uma cabeça de atenção, o modelo teria que dividir os pesos de probabilidade entre “Brasil”, “tem” e “futebol”, potencialmente resultando em uma representação ambígua. Com múltiplas cabeças, a Cabeça 1 pode dedicar sua atenção à palavra “melhores”, enquanto a Cabeça 2 foca em “jogadores”, permitindo uma compreensão mais ampla do entendimento da frase e contexto.

Matematicamente, isso é realizado projetando as entradas ($\mathbf{Q}, \mathbf{K}, \mathbf{V}$) através de matrizes de pesos aprendidas (W_i^Q, W_i^K, W_i^V) para dimensões menores. A atenção é calculada independentemente em cada uma dessas projeções. Ao final, os vetores de saída

de todas as cabeças são concatenados (*Concat*) e projetados novamente por uma matriz linear final W^O para restaurar a dimensão original do modelo. A Equação 2.4 formaliza este processo (VASWANI *et al.*, 2017):

$$\begin{aligned} \textbf{\textit{MultiHead}}(\textbf{Q, K, V}) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{onde} \quad \text{head}_i &= \textit{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{2.4}$$

2.3 TRANSFORMAÇÃO E PREDIÇÃO DE LINGUAGEM COM BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) representa um marco na evolução dos Modelos de Linguagem (LLMs), fundamentando-se na arquitetura de redes neurais profundas do tipo *Transformer*, especificamente em sua pilha de codificadores. O BERT inova ao processar o contexto de forma bidirecional, analisando simultaneamente os elementos que antecedem e sucedem cada *token* na sequência textual (DEVLIN *et al.*, 2018). O ciclo de desenvolvimento do modelo opera sob o paradigma de transferência de aprendizado, que inicialmente, ocorre o pré-treinamento em larga escala de maneira autossupervisionada (*self-supervised*), onde o modelo apreende estruturas linguísticas universais a partir de dados não rotulados, ganhando um forte contexto para todas nuances existente no seu conjunto de dados. Subsequentemente, executa-se o *fine-tuning*, onde os pesos pré-treinados são refinados com dados supervisionados para otimizar o desempenho em tarefas específicas (DEVLIN *et al.*, 2018).

Durante a fase de pré-treinamento, a capacidade cognitiva do BERT é construída através de duas funções de objetivo simultâneas, sendo a principal delas a *Masked Language Modeling* (MLM). Nesta tarefa uma porcentagem dos *tokens* de entrada (tipicamente 15%) é ocultada aleatoriamente e substituída pelo *token* especial `[MASK]`. O desafio imposto à rede neural é predizer o *token* original mascarado baseando-se exclusivamente no contexto fornecido pelas palavras visíveis adjacentes. Essa abordagem força o modelo a construir representações contextuais ricas e densas, fundindo o contexto à esquerda e à direita para resolver ambiguidades semânticas e sintáticas (DEVLIN *et al.*, 2018).

Complementarmente ao MLM, o modelo é treinado na tarefa de *Next Sentence Prediction* (NSP), uma classificação binária essencial para o entendimento de relacionamentos entre sentenças. Neste processo, o modelo recebe pares de segmentos textuais e deve inferir se o segundo segmento segue logicamente o primeiro no documento original, capturando assim a coerência discursiva e dependências de longo prazo diante do contexto. A combinação dessas duas estratégias (MLM e NSP) confere ao BERT uma robustez de representação linguística, que permite que, na etapa de *fine-tuning*, ele seja adaptado com alta eficiência para uma vasta gama de aplicações de compreensão de linguagem natural, como sistemas de perguntas e respostas (QA) e análise de sentimentos, bastando a adição de uma camada de saída específica para a tarefa alvo (DEVLIN *et al.*, 2018).

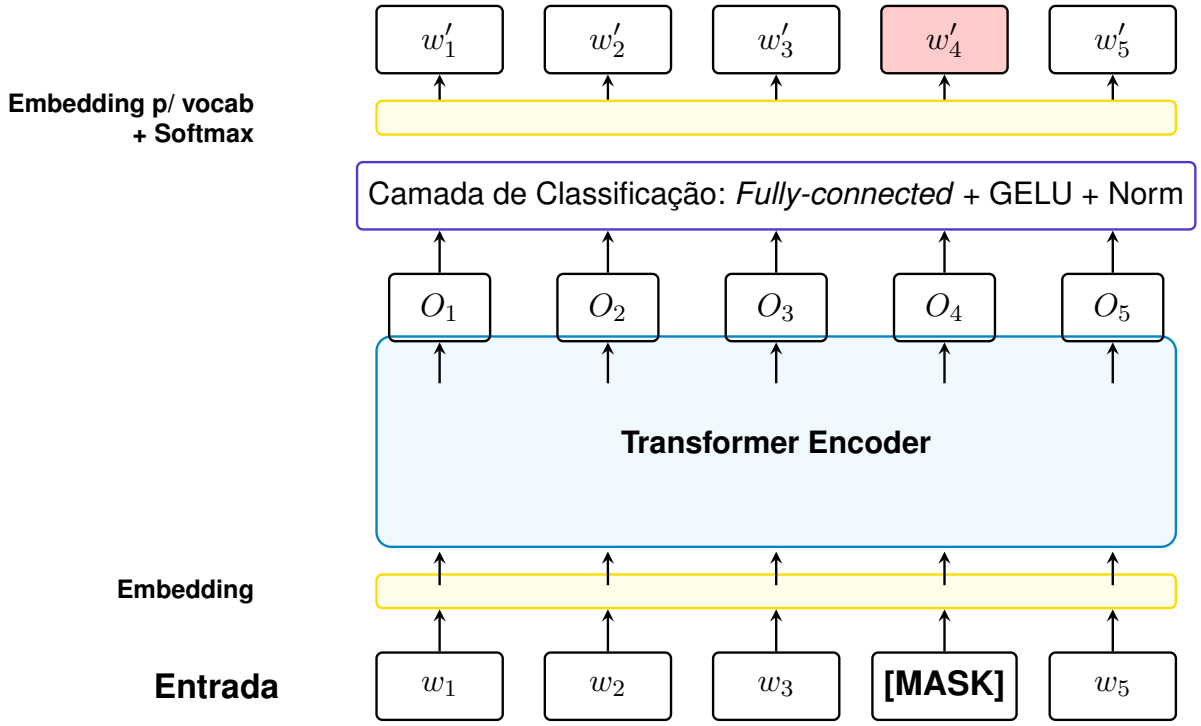
2.3.1 *Masked Language Modeling* (MLM)

A estratégia de *Masked Language Modeling* (MLM) é o componente do pré-treinamento bidirecional do BERT. O MLM permite que o modelo visualize o contexto completo da sentença simultaneamente, de forma bidirecional. Nesta abordagem, uma parte dos *tokens* é mascarada aleatoriamente e substituída pelo *token* especial `[MASK]`. O objetivo fundamental da rede neural é predizer qual era o *token* original, baseando-se exclusivamente no contexto fornecido pelas palavras não mascaradas ao seu redor (DEVLIN *et al.*, 2018).

O fluxo arquitetural deste processo é detalhado na Figura 9. Na camada inferior, observa-se a sequência de entrada onde o termo na quarta posição (w_4) foi substituído por `[MASK]`. Estes dados são convertidos em vetores (*embeddings*) e processados através das camadas do *Transformer Encoder*, que gera representações vetoriais contextuais de saída (O_1 a O_5) (DEVLIN *et al.*, 2018)

Conforme ilustrado e abordado por Devlin *et al.* (2018), para a posição mascarada o vetor resultante (O_4) é encaminhado para uma camada de classificação, composta por uma rede conectada, função de ativação GELU e normalização. Finalmente, o resultado é projetado sobre o vocabulário completo através de uma função *Softmax*, gerando a probabilidade de saída (w'_4) que determina a palavra predita.

Figura 9 – Arquitetura e fluxo de dados do BERT para a tarefa de MLM: entrada mascarada, processamento pelo Encoder e classificação do token oculto.



Fonte: (DEVLIN *et al.*, 2018).

Segundo Devlin *et al.* (2018), a função *Softmax* desempenha um papel central em redes neurais, atuando como uma generalização multidimensional da função logística. No contexto de modelos *Transformer* como o BERT, ela é aplicada tanto nas camadas de saída para classificação quanto nos mecanismos de atenção interna para ponderar a relevância entre *tokens*. A Equação 2.5 define a operação que transforma um vetor de pontuações brutas, denotados por T, em uma distribuição de probabilidade (DEVLIN *et al.*, 2018).

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \tag{2.5}$$

Nesta estrutura matemática, P_i representa a probabilidade atribuída à classe ou posição i. O numerador, $e^{S \cdot T_i}$, aplica a função exponencial ao valor do T_i ponderado pelo fator de escala S (DEVLIN *et al.*, 2018), esta operação é fundamental para projetar valores reais (que podem ser negativos) em um espaço estritamente positivo e para acentuar a distinção entre pontuações altas e baixas. O denominador, $\sum_j e^{S \cdot T_j}$, funciona como um normalizador global, assegurando que a soma de todas as probabilidades calculadas seja igual a 1, constituindo uma distribuição válida (DEVLIN *et al.*, 2018). O hiperparâmetro S atua como um regulador da “temperatura“ da distribuição, valores elevados de S aumentam a disparidade entre as classes (tornando a distribuição mais “aguda“ ou confiante), enquanto valores reduzidos suavizam as probabilidades, distribuindo a atenção de forma mais uniforme (DEVLIN *et al.*, 2018).

A probabilidade da palavra *Pi* estar classificada no intervalo de resposta, dado o *prompt* de texto de entrada, é calculada como um produto escalar entre $e^{S \cdot T_i}$ e S. O intervalo da probabilidade é calculado entre as posições *i* e *j* do vetor, que define o resultado da previsão.

Na Tabela 1 são exemplificados os formatos de textos de entrada e saída para a operação de MaskedLM.

Tabela 1 – Exemplos de entrada e saída para o contexto de *Masked Language* em português, sendo *P* a Probabilidade e [MASK] a palavra ocultada pelo modelo para que seja prevista.

Texto de entrada	**Texto de saída**	***P (palavra-1)***	***P (palavra-2)***	***P (palavra-3)***
*Os Estados Unidos são o país mais **[MASK]** do mundo*	*Os Estados Unidos são o país mais **rico** do mundo*	*rico*: 0.147	*poderoso*: 0.121	*visitado*: 0.097
*São Paulo é o melhor **[MASK]***	*São Paulo é o melhor **time***	*time*: 0.854	*estado*: 0.112	*santo*: 0.015
*O Brasil tem o **[MASK]** time de futebol do mundo*	*O Brasil tem o **melhor** time de futebol do mundo*	*melhor*: 0.505	*maior*: 0.122	*pior*: 0.068
*O **[MASK]** do mundo*	*O **fim** do mundo*	*fim*: 0.268	*topo*: 0.023	*começo*: 0.016

Fonte: Do Autor.

2.3.2 *Next Sentence Prediction* (NSP)

Paralelamente ao MLM, o BERT é submetido a uma tarefa de treinamento em nível de sentença denominada *Next Sentence Prediction* (NSP). Esta etapa é crucial para que o modelo desenvolva a capacidade de compreender relações discursivas, coerência textual e dependências lógicas de longo prazo, habilidades indispensáveis para aplicações como *Question Answering* (QA) e Inferência em Linguagem Natural (DEVLIN *et al.*, 2018).

Coforme detalhado por Devlin *et al.* (2018), a tarefa de NSP é modelada como uma classificação binária. O modelo recebe como entrada um par de sentenças $(1, 2)$ e deve predizer se 2 é a continuação lógica e imediata de 1 no texto original. Para garantir o aprendizado adequado, o conjunto de dados de treinamento é construído de forma balanceada:

- Em 50% dos casos, a sentença 2 é a que de fato sucede 1 no conjunto de dados. O modelo deve classificá-la com o rótulo ***IsNext***.
- Nos outros 50%, a sentença 1 é uma frase aleatória extraída do documento. O modelo deve classificá-la como ***NotNext***.

Exemplo Prático:

- *Input (IsNext):* [CLS] O homem foi à loja. [SEP] Ele comprou leite. [SEP]
- *Input (NotNext):* [CLS] O homem foi à loja. [SEP] Pinguins não voam. [SEP]

A representação vetorial de entrada do BERT de Devlin *et al.* (2018) para processar esses pares, ilustrada na Figura 10, é composta pela soma elementar de três camadas distintas de *embeddings*:

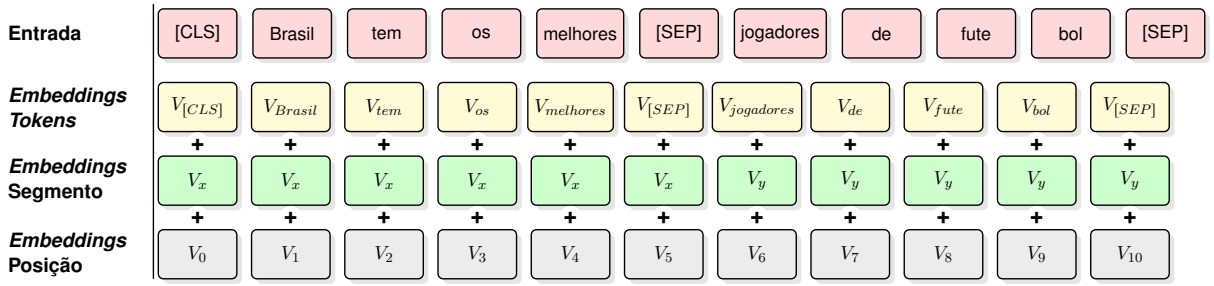
1. ***Token Embeddings***: Representam o conteúdo semântico das palavras (ou sub-palavras) da sequência (ex: “Brasil”, “tem”, “os”, “melhores”, “jogadores”, “de”, “futebol”).

2. ***Embeddings* de Segmento**: Vetores aprendidos (V_x e V_y) que funcionam como identificadores de sentença. Eles sinalizam ao modelo a qual oração cada *token* pertence, permitindo a distinção clara entre a Sentença 1 e a Sentença 2.

3. ***Embeddings* de Posição**: A codificação posicional ($V_0, V_1 \dots V_n$) que injeta a noção de ordem sequencial na arquitetura, uma vez que o mecanismo de atenção é invariante à posição.

O *token* especial `[CLS]`, segundo (DEVLIN *et al.*, 2018), é inserido no início da sequência para agregar a representação global do par, servindo como entrada para o classificador binário de NSP. O *token* `[SEP]` atua como um delimitador explícito entre as sentenças.

Figura 10 – Estrutura de representação de entrada do BERT para NSP. A entrada final é a soma vetorial dos *embeddings* de Token, Segmento e Posição.



Fonte: Adaptado de (DEVLIN *et al.*, 2018).

2.3.3 Versões do BERT

Atualmente, existem muitas versões do BERT pré-treinado disponíveis, cada uma adaptada para diferentes aplicações e necessidades. No entanto, no artigo original do BERT (DEVLIN *et al.*, 2018), foram apresentadas duas versões principais do BERT, sendo elas BERTbase e BERTlarge. Essas versões diferem significativamente em suas arquiteturas neurais e capacidade de processamento.

O *BERTbase* foi desenvolvido com uma arquitetura composta por 12 codificadores, cada uma com 12 cabeças de atenção. No total, esta configuração resulta em 110 milhões de parâmetros treináveis. Essa versão foi projetada para ser mais leve e eficiente,

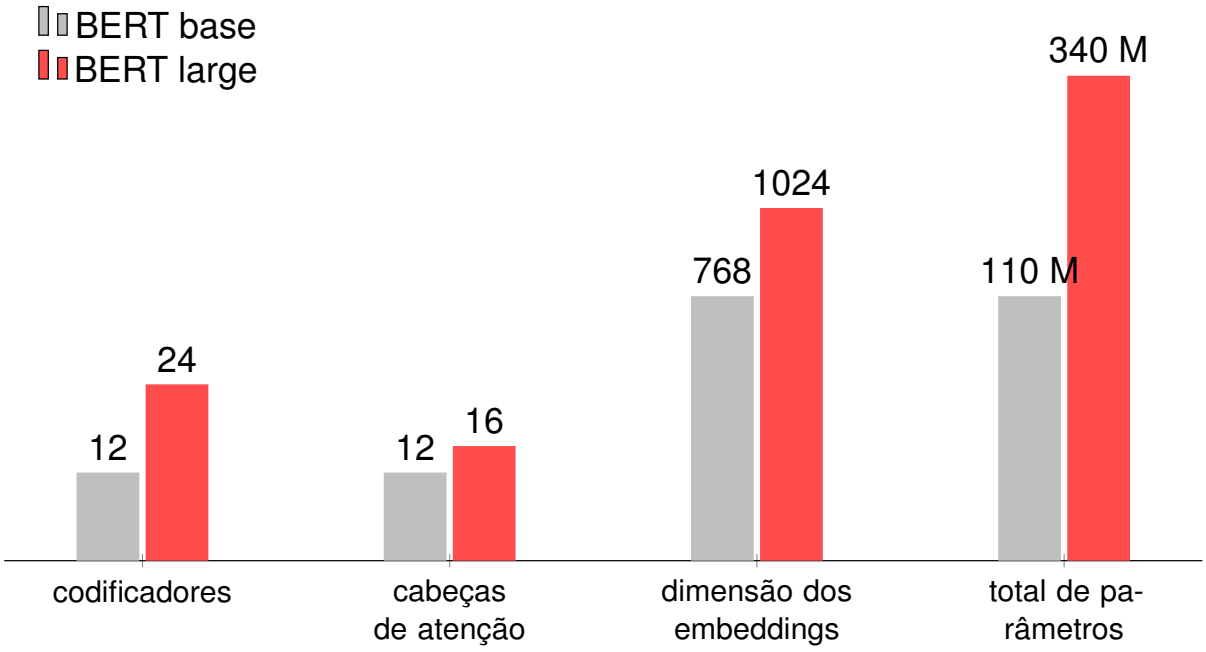
facilitando seu uso em uma variedade de tarefas sem demandar excessivamente recursos computacionais (DEVLIN *et al.*, 2018), a Figura 11 ilustra as configurações do *BERTbase*.

Devlin *et al.* (2018) demonstra que o *BERTlarge* apresenta uma arquitetura mais robusta, entregando 24 codificadores e 16 cabeças de atenção. Essa configuração expansiva leva o BERTlarge a um total de 340 milhões de parâmetros, trazendo aumento significativo na complexidade e no número de parâmetros permite ao BERTlarge capturar nuances mais finas e realizar inferências mais precisas, resultando em um desempenho superior em diversos *benchmarks* de compreensão de linguagem natural, a Figura 11 ilustra as configurações do *BERTlarge*.

Nos testes de precisão, o *BERTlarge* demonstrou um avanço notável em relação ao *BERTbase* (JOSHI *et al.*, 2019), evidenciando que a profundidade e a capacidade do modelo são fatores críticos para o sucesso em tarefas complexas de processamento de linguagem natural. Essa diferença de desempenho é particularmente notável em tarefas que requerem um entendimento profundo do contexto e da semântica, como resposta a perguntas, análise de sentimentos e tradução automática.

A criação dessas duas versões do BERT foi um marco significativo no campo da inteligência artificial e do processamento de linguagem natural, abrindo caminho para uma nova era de modelos de linguagem pré-treinados. A pesquisa e os desenvolvimentos subsequentes baseados no BERT continuaram a expandir e refinar essas ideias, resultando em uma proliferação de modelos derivados, cada um otimizado para diferentes aplicações e contextos específicos. A Figura 11 representa o gráfico comparativo entre BERTlarge e BERTbase, dado para cada um deles, a quantidade de *encoders*, a quantidade de *self-attention heads*, dimensionalidade dos *embeddings*, e total de parâmetros, conforme elucidado por Devlin *et al.* (2018).

Figura 11 – Comparativo do $BERT_{base}$ e $BERT_{large}$.



Fonte: Adaptado de (DEVLIN *et al.*, 2018)

2.3.4 *Fine-Tuning* no contexto do BERT

O processo de *fine-tuning* consiste no ajuste fino de um modelo de aprendizado de máquina previamente treinado em um vasto conjunto de dados genérico. Esta técnica adapta os pesos do modelo para uma tarefa ou domínio específico, modificando-os levemente para otimizar o desempenho em um cenário particular. No contexto de modelos de linguagem baseados em *Transformers*, o *fine-tuning* é a etapa crucial que transforma um conhecimento linguístico abrangente em especialização, permitindo que o modelo desempenhe tarefas complexas, como responder a perguntas de um nicho específico ou analisar sentimentos em determinado contexto, com alta precisão e eficiência de dados.

A arquitetura do BERT foi concebida para operar fundamentalmente em duas etapas distintas: o pré-treinamento não supervisionado e o *fine-tuning* supervisionado (DEVLIN *et al.*, 2018). Enquanto o pré-treinamento absorve representações de linguagem bidirecionais profundas de fontes massivas, é na etapa de *fine-tuning* que o modelo se torna funcional para aplicações práticas. Diferentemente de abordagens anteriores que exigiam arquiteturas complexas e específicas para cada tarefa, o BERT permite que a mesma estrutura base pré-treinada seja adaptada para uma variedade de objetivos finais com a adição de camadas de saída simples e custos computacionais relativamente baixos.

O mecanismo de *fine-tuning* do BERT envolve, usualmente, a introdução de uma camada densa (*fully connected*) específica para a tarefa sobre a representação final do modelo. Por exemplo, em tarefas de classificação, utiliza-se a representação vetorial do

token especial `[CLS]` como entrada para essa nova camada, seguida de uma função de ativação apropriada (como *softmax*) (DEVLIN *et al.*, 2018). O aspecto crucial é que, durante este treinamento supervisionado com dados rotulados, todos os parâmetros, tanto os da nova camada específica quanto os pesos profundos do BERT pré-treinado, são atualizados conjuntamente, permitindo que as representações genéricas se especializem para a sintaxe e semântica do domínio alvo (DEVLIN *et al.*, 2018).

A flexibilidade do *fine-tuning* permite que o BERT atinja o estado da arte em diversas aplicações, desde a classificação de texto e inferência em linguagem natural até sistemas complexos de Perguntas e Respostas. Além das tarefas de predição direta, o *fine-tuning* é essencial para adaptar o BERT como um gerador de *embeddings* semânticos de alta qualidade em domínios específicos (DEVLIN *et al.*, 2018). Esses vetores densos especializados tornam-se componentes vitais em arquiteturas modernas, como nos fluxos de Geração Aumentada por Recuperação (*Retrieval Augmented Generation* - RAG), onde a precisão na recuperação da informação depende diretamente da qualidade da representação vetorial ajustada ao conjunto de dados da organização.

Para a arquitetura e fluxo de dados no processo de *Fine-Tuning* do BERT, a Figura 12 ilustra a transição do modelo pré-treinado para a especialização, destacando a injeção do vetor de contexto ([CLS]) na camada específica e o ciclo de atualização conjunta dos parâmetros.

2.3.5 Utilização do SENTENCE-Bert (SBERT) para geração de *Embeddings*

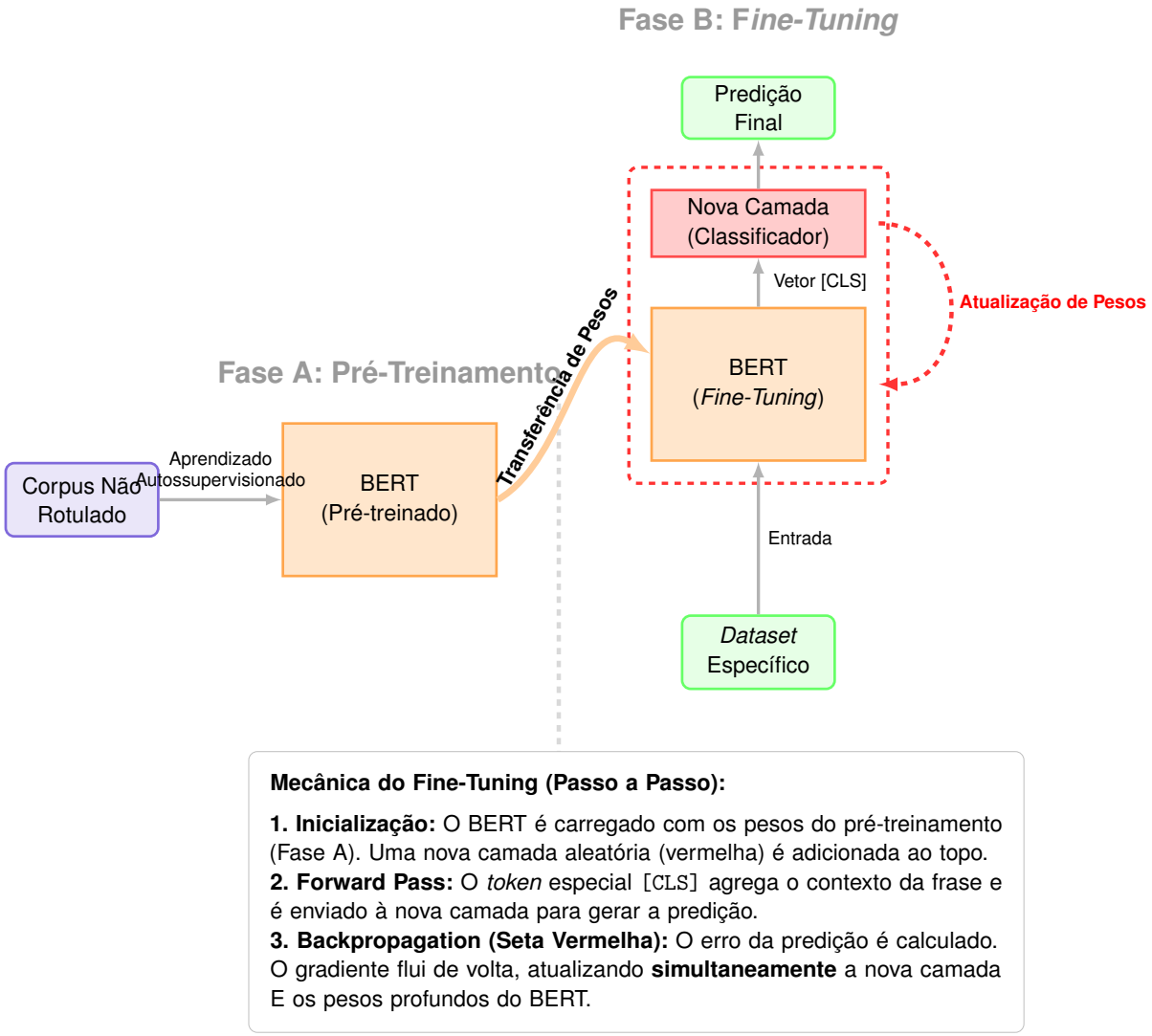
O BERT tornou-se uma ferramenta amplamente utilizada em uma variedade de tarefas de processamento de linguagem natural, como análise de sentimentos e resposta a perguntas. Além dessas aplicações, o BERT é cada vez mais valorizado para a criação de *embeddings* de palavras (vetores numéricos que capturam os significados semânticos das palavras).

Por exemplo, na análise de sentimentos, o BERT pode ser utilizado para determinar o tom emocional de um texto, como classificar uma avaliação de produto como positiva ou negativa. Em sistemas de resposta a perguntas, o BERT pode interpretar uma pergunta e encontrar a resposta mais relevante em um conjunto de dados de documentos.

Além disso, os *embeddings* gerados pelo BERT são usados em tarefas de similaridade semântica, onde a proximidade entre os vetores representa a similaridade no significado das palavras. Isso é útil em aplicações como a pesquisa de informações, onde o objetivo é encontrar documentos ou passagens que sejam semanticamente semelhantes à consulta do usuário, ou em sistemas de recomendação, onde os itens recomendados são aqueles que têm maior afinidade semântica com as preferências do usuário.

Dessa forma, o BERT não só aprimora a precisão em várias tarefas de NLP

Figura 12 – Fluxo de Pré-treinamento e *Fine-Tuning* do BERT para uma tarefa de classificação



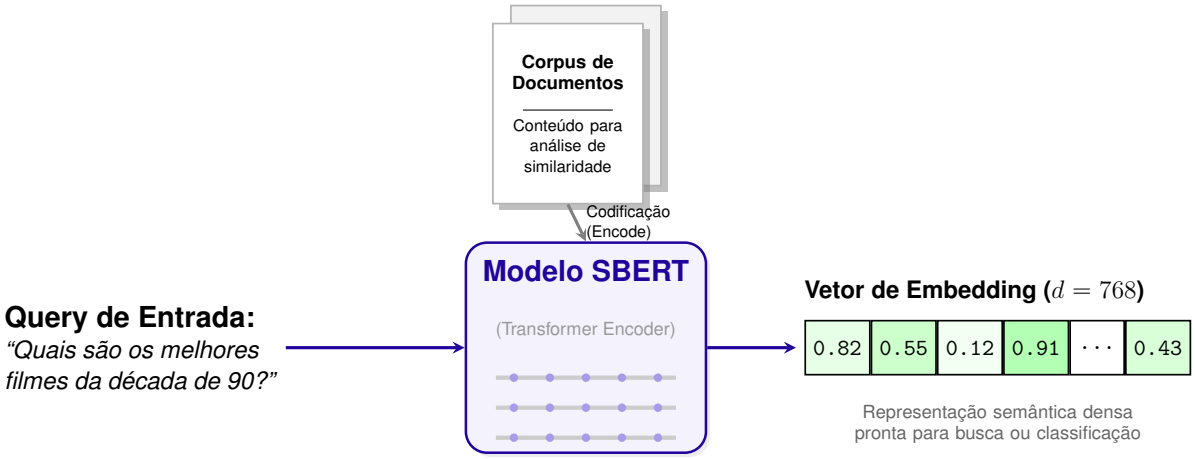
Fonte: Adaptado de (DEVLIN *et al.*, 2018)

(*Natural Language Processing*), mas também facilita a construção de representações semânticas ricas que podem ser aplicadas em diversos contextos.

De acordo com o estado da arte do SBERT (Sentence-BERT) (REIMERS; GUREVYCH, 2019), este modelo pode ser utilizado em uma ampla gama de tarefas, incluindo classificações, regressões, análise de sentimentos, entre outras. O SBERT é uma ramificação do modelo BERT, com uma etapa de *fine-tuning*, e tem seu principal objetivo a classificação de sentenças. Com isso, a *Query* de entrada podem ser relacionadas diretamente em formato semântico com documentos e conteúdo que estejam relacionados para um possivel gerador de resposta, utilizando o SBERT, esse significado semântico é obitido como resultado final no formato de *embeddings* (REIMERS; GUREVYCH, 2019). A Figura 13 representa a utilização do modelo LLM SBERT para geração de *embeddings*, com o objetivo de encontrar a relação semantica entre um texto de entrada, sendo realizado como

pergunta, e documentos ou conteúdos que estejam relacionado ao contexto da resposta.

Figura 13 – SBERT sendo utilizado para geração de *embeddings*, contemplando a relação semântica entre *Query* e Documento.



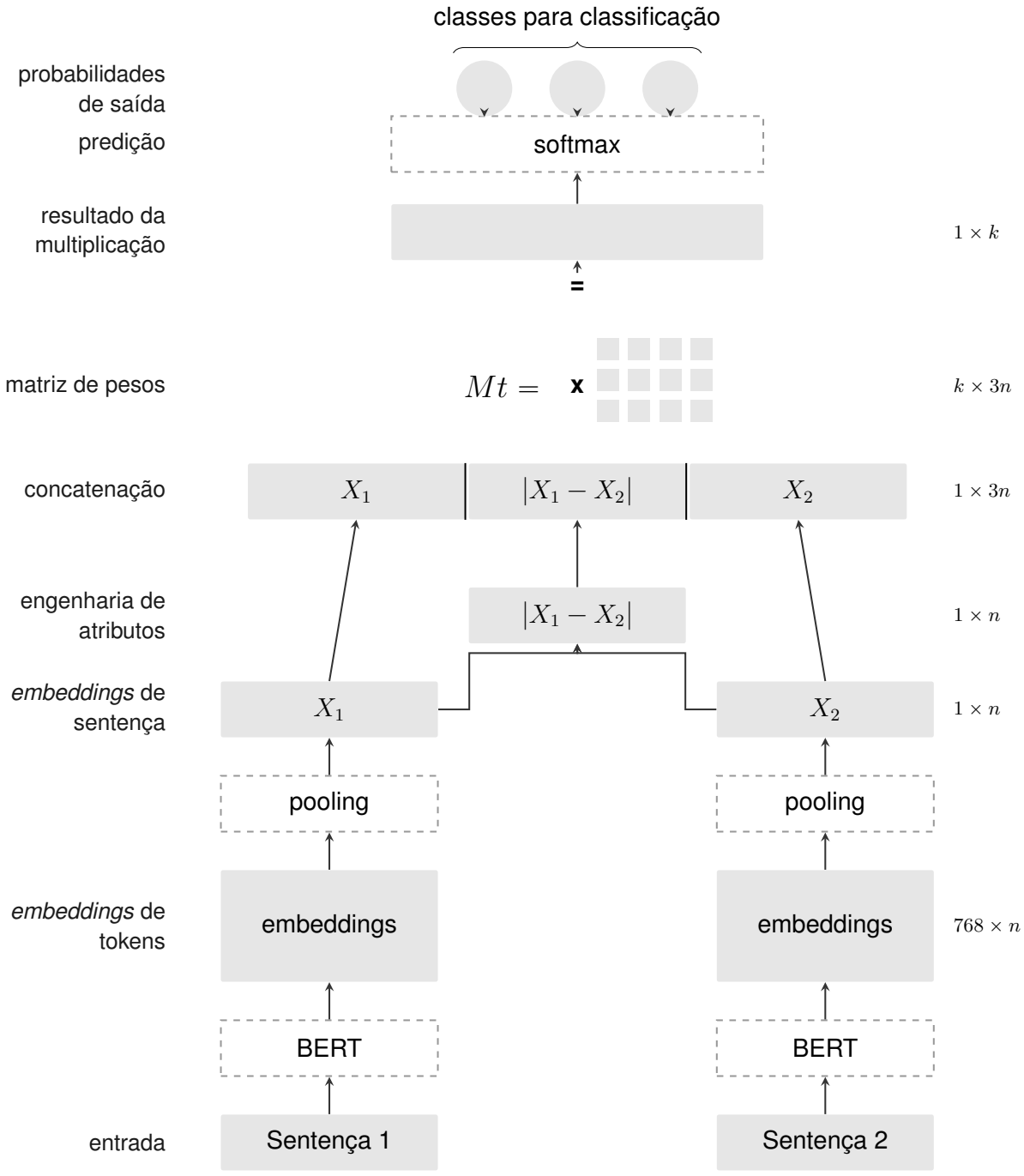
Fonte: Do Autor

Em tarefas de classificação, o SBERT pode categorizar textos em diversas classes, como identificar tópicos de discussão ou determinar a polaridade de opiniões. Em regressão, ele pode ser usado para prever valores contínuos, como a previsão de notas em resenhas de produtos. Na análise de sentimentos, o SBERT é eficaz em detectar e avaliar as emoções expressas em textos (REIMERS; GUREVYCH, 2019).

Todavia, *fine-tuning* supervisionado permite que o SBERT seja adaptado para tarefas específicas ao treinar o modelo com dados rotulados. Já o *fine-tuning* não supervisionado pode ser realizado através de métodos como o aprendizado contrastivo, onde o modelo aprende a partir das relações entre os pares de textos sem a necessidade de rótulos explícitos (REIMERS; GUREVYCH, 2019).

Para tarefas do escopo de RAG, a geração de *embeddings* pode ser baseada na arquitetura do SBERT como sendo contextualizado para uma tarefa de classificação, tendo em vista que nesse contexto o modelo trabalhará em uma situação probabilística a partir da proximidade de vetores. A Figura 14 ilustra o fluxo de tomada de decisão do modelo para realizar uma tarefa de classificação.

Figura 14 – Arquitetura do SBERT aplicado a tarefas de classificação.



Fonte: Adaptado de (REIMERS; GUREVYCH, 2019)

O processo de classificação no SBERT inicia-se com dois parâmetros de entrada, a **Sentença 1** e a **Sentença 2**. Ambas atravessam a arquitetura do BERT (ou outro modelo *transformer* base) para a geração de *embeddings* contextuais. A partir da saída do modelo, aplica-se uma camada de *pooling* (geralmente a média de todos os *tokens* de saída) para condensar a informação semântica de cada sentença em um vetor de tamanho fixo. Denotamos o vetor resultante da Sentença 1 como $X1$ e o da Sentença 2 como $X2$, onde ambos pertencem ao espaço $\mathbb{R}^n$ (n sendo a dimensão do *embedding*, ex: 768 para o

BERT-base (REIMERS; GUREVYCH, 2019)).

A inovação crucial da arquitetura de classificação do SBERT reside na forma como esses vetores são combinados antes da camada final (REIMERS; GUREVYCH, 2019). Conforme ilustrado na Figura 14, o modelo constrói um vetor de entrada para o classificador através da concatenação de três componentes distintos:

1. O vetor da própria Sentença 1 ($X1$);
2. O vetor da própria Sentença 2 ($X2$);
3. A diferença absoluta elemento a elemento entre os dois vetores ($|X1 - X2|$).

Esta operação de concatenação resulta em um vetor combinado de dimensão n. A inclusão da diferença absoluta é estrategicamente importante, pois ela captura a distância e as discrepâncias semânticas entre as duas sentenças em cada dimensão do espaço vetorial, o que é vital para tarefas de similaridade e inferência natural (REIMERS; GUREVYCH, 2019).

Este vetor concatenado é projetado para o mesmo nível vetorial em que as classes são representadas, através da multiplicação por uma matriz de pesos $M_t \in \mathbb{R}^{k \times 3n}$, na qual k representa o número de classes alvo (REIMERS; GUREVYCH, 2019).

O produto desta multiplicação gera um vetor de pontuações (*logits*) de tamanho k, que é submetido à função de ativação *Softmax* para produzir a probabilidade sobre as classes. A operação completa, que define a probabilidade da classe prevista, é descrita pela Equação 2.6 (REIMERS; GUREVYCH, 2019):

$$o = \text{softmax}(M_t[\mathbf{X1}; \mathbf{X2}; |\mathbf{X1} - \mathbf{X2}|]). \tag{2.6}$$

Na Equação 2.6, a notação $[\cdot; \cdot]$ representa a operação de concatenação vetorial. O modelo é treinado de forma supervisionada (*fine-tuning*) minimizando a função de perda de entropia cruzada (*cross-entropy loss*) (REIMERS; GUREVYCH, 2019), que penaliza a divergência entre a probabilidade predita o e o rótulo verdadeiro da classe.

2.4 *GENERATIVE PRE-TRAINED TRANSFORMER* (GPT)

O *Generative Pre-trained Transformer* (GPT) consolidou-se como um marco na Inteligência Artificial, evoluindo de modelos puramente textuais para sistemas multimodais (capacidade de processar imagem, texto e áudio) complexos. A versão mais recente documentada tecnicamente até a data deste presente trabalho, o GPT-4, é descrita oficialmente pela OpenAI como um modelo de larga escala que utiliza a arquitetura de *Transformers* (VASWANI *et al.*, 2017), tem a portabilidade de entender entradas de texto e imagem para

produzir saídas textuais (OpenAI, 2023). O relatório técnico confirma que o sistema mantém a fundação de modelos de linguagem autoregressivos, treinados para prever o próximo *token* em um documento (OpenAI, 2023).

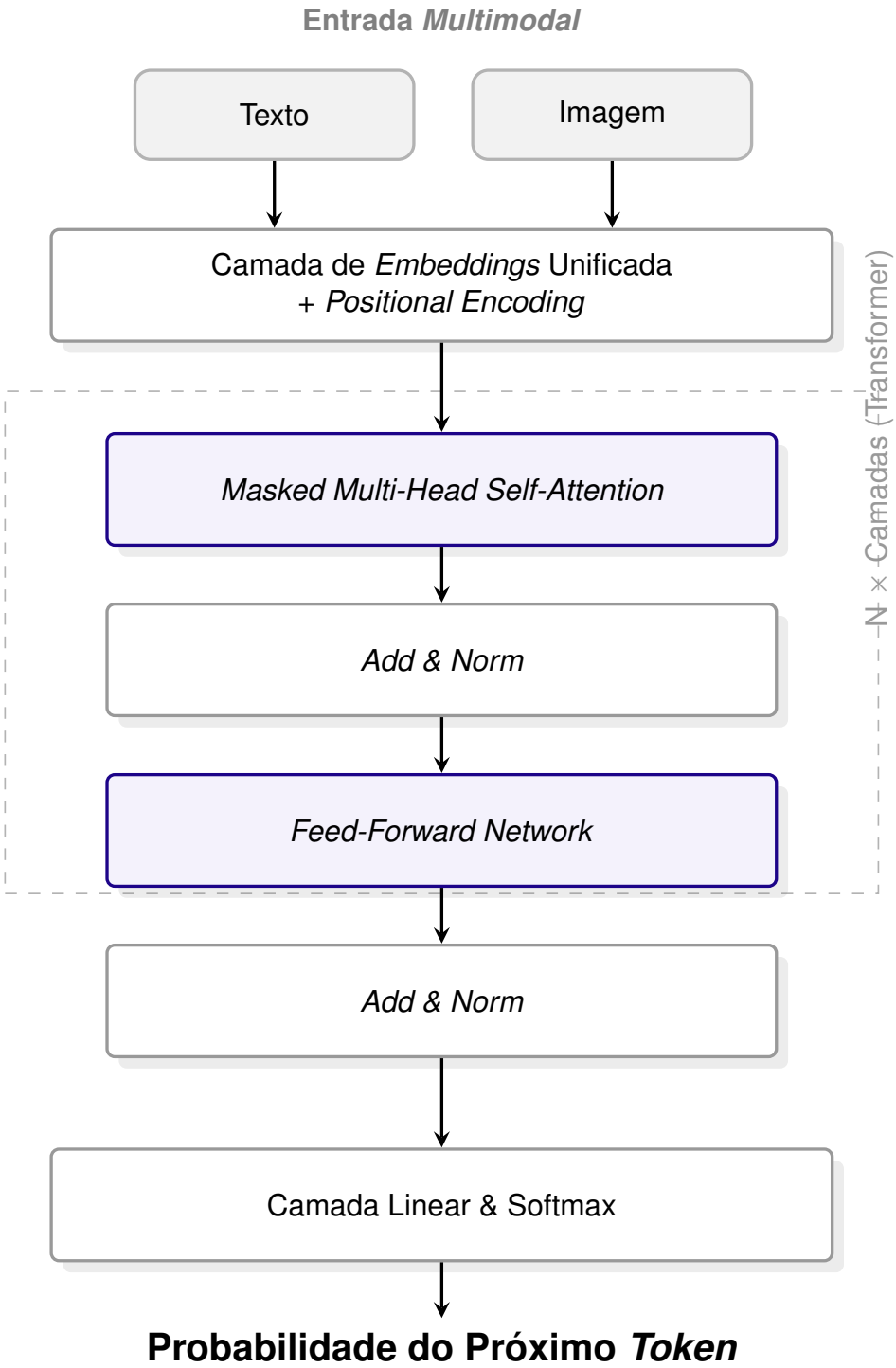
A Figura 15 ilustra a arquitetura do GPT-4 operando através de um fluxo sequencial de processamento multimodal. O sistema recebe entradas de diferentes naturezas (texto, imagem e áudio), que são convertidas em uma representação vetorial densa e unificada na camada de *embeddings*, enriquecida com codificação posicional para preservação da ordem sequencial (OpenAI, 2023). O núcleo do processamento ocorre em uma pilha profunda de decodificadores (*Decoder*), onde cada camada aplica mecanismos de autoatenção mascarada (*Masked Multi-Head Self-Attention*) para capturar dependências contextuais, seguidos por redes *Feed-Forward* para otimizar a capacidade computacional através de processamento especializado. O fluxo encerra-se com a projeção linear e a aplicação da função *Softmax*, resultando na distribuição de probabilidade para a predição autoregressiva do próximo *token*.

A arquitetura do GPT-4 segue o paradigma de pré-treinamento não supervisionado em grandes volumes de dados, seguido por um refinamento supervisionado (*fine-tuning*). Durante o pré-treinamento, o modelo utiliza dados disponíveis publicamente (como dados da internet) e dados licenciados de terceiros para aprender a prever o próximo *token* (OpenAI, 2023). Este processo dota o modelo de uma compreensão generalista da linguagem e do raciocínio, que é posteriormente refinada através do Aprendizado por Reforço com *Feedback* Humano (RLHF). O RLHF é citado como um componente crítico para alinhar as saídas do modelo com a intenção do usuário, melhorando a factualidade e a aderência a normas de segurança (OpenAI, 2023).

O artigo destaca que o diferencial do GPT-4 está na infraestrutura de treinamento que permite um cenário de aprendizado “previsível“ para o modelo. A OpenAI desenvolveu métodos de otimização que permitem prever o desempenho final do modelo utilizando uma fração da computação necessária para o treinamento completo (OpenAI, 2023). Em termos de capacidades, o modelo demonstra desempenho de nível humano em diversos exames profissionais e acadêmicos, superando significativamente versões anteriores como o GPT-3.5, especialmente em tarefas que exigem raciocínio complexo e manipulação de entradas visuais.

Apesar dos avanços, o relatório técnico enfatiza que o GPT-4 mantém limitações semelhantes às de modelos anteriores, como a tendência a alucinações (geração de fatos incorretos) e erros de raciocínio. A natureza de “caixa-preta” da arquitetura final, combinada com a falta de transparência sobre os dados de treinamento específicos, impõem desafios para a interpretabilidade completa de suas decisões. Para mitigar riscos, o desenvolvimento do GPT-4 incluiu extensivos testes (*red teaming*) com especialistas de domínio e a implementação de filtros de segurança assistidos pelo próprio modelo, visando reduzir a geração

Figura 15 – Representação da arquitetura do GPT-4 baseado em *Transformers* com capacidade multimodal



Fonte: Adaptado de (OpenAI, 2023; VASWANI *et al.*, 2017).

de conteúdo prejudicial ou enviesado (OpenAI, 2023).

3 *RETRIEVAL-AUGMENTED GENERATION* (RAG)

O *Retrieval-Augmented Generation* (RAG) é uma arquitetura híbrida que integra técnicas de recuperação de informações (*retrieval*) com modelos generativos de linguagem (*generation*), visando aprimorar a qualidade e a desenvoltura das respostas produzidas por sistemas que utilizam *Large Language Models* (LLMs) (LEWIS *et al.*, 2020) para geração de texto. Essa abordagem surge como uma solução para mitigar as limitações inerentes aos modelos puramente generativos, que frequentemente dependem exclusivamente de conhecimento internalizado durante o treinamento, realizado por seu próprio treinamento base, ou adaptado via *fine-tuning*, podendo gerar conteúdos desatualizados, imprecisos ou inconsistentes em contextos que exigem dados externos. Ao incorporar um módulo de recuperação, o RAG permite que o modelo acesse fontes de conhecimento atualizadas e específicas durante o processo de inferência, enriquecendo assim a base contextual utilizada para a geração de respostas.

A primeira etapa do RAG, denominada *retrieval*, consiste na consulta a um repositório de dados externo, como um banco de dados documental ou um corpus de informação indexado para recuperar conteúdos de texto semanticamente relevantes à *query* ou ao contexto de entrada (KARPUKHIN, 2020). Essa operação é realizada por meio de modelos de aprendizado de máquina que realizam a vetorização do conteúdo, ou também por algoritmos que mapeiam tanto a consulta quanto os documentos em um espaço vetorial de alta ou baixa dimensionalidade, onde a similaridade (distância vetorial) é calculada através de métricas como a similaridade do cosseno e distância euclidiana (KARPUKHIN, 2020). A eficácia dessa etapa é crítica, pois documentos mal recuperados podem introduzir ruído ou irrelevância no processo subsequente de geração, comprometendo a coerência e a precisão da saída final.

Uma vez recuperados os documentos mais pertinentes, o sistema procede à etapa de *augmentation*, na qual os textos selecionados são concatenados ao contexto original da consulta, formando um *prompt* ampliado. Esse *prompt* enriquecido serve como entrada para o modelo generativo, com uma arquitetura *Transformer* como GPT (OpenAI, 2023) ou Gemini (GEMINI; ANIL; AL., 2025). Ao processar tanto a *query* quanto as informações recuperadas, gera uma resposta contextualmente fundamentada (SHI, 2023). A integração explícita de dados externos pode reduzir a dependência do conhecimento pré-treinado do modelo, e também possibilitar a citação de fontes verificáveis, aumentando a transparência e a confiabilidade do sistema.

Do ponto de vista computacional, o RAG apresenta vantagens significativas em relação a abordagens puramente generativas ou baseadas em *fine-tuning* estático (GAO *et al.*, 2024). Enquanto modelos convencionais exigem retreinamento ou ajuste contínuo para

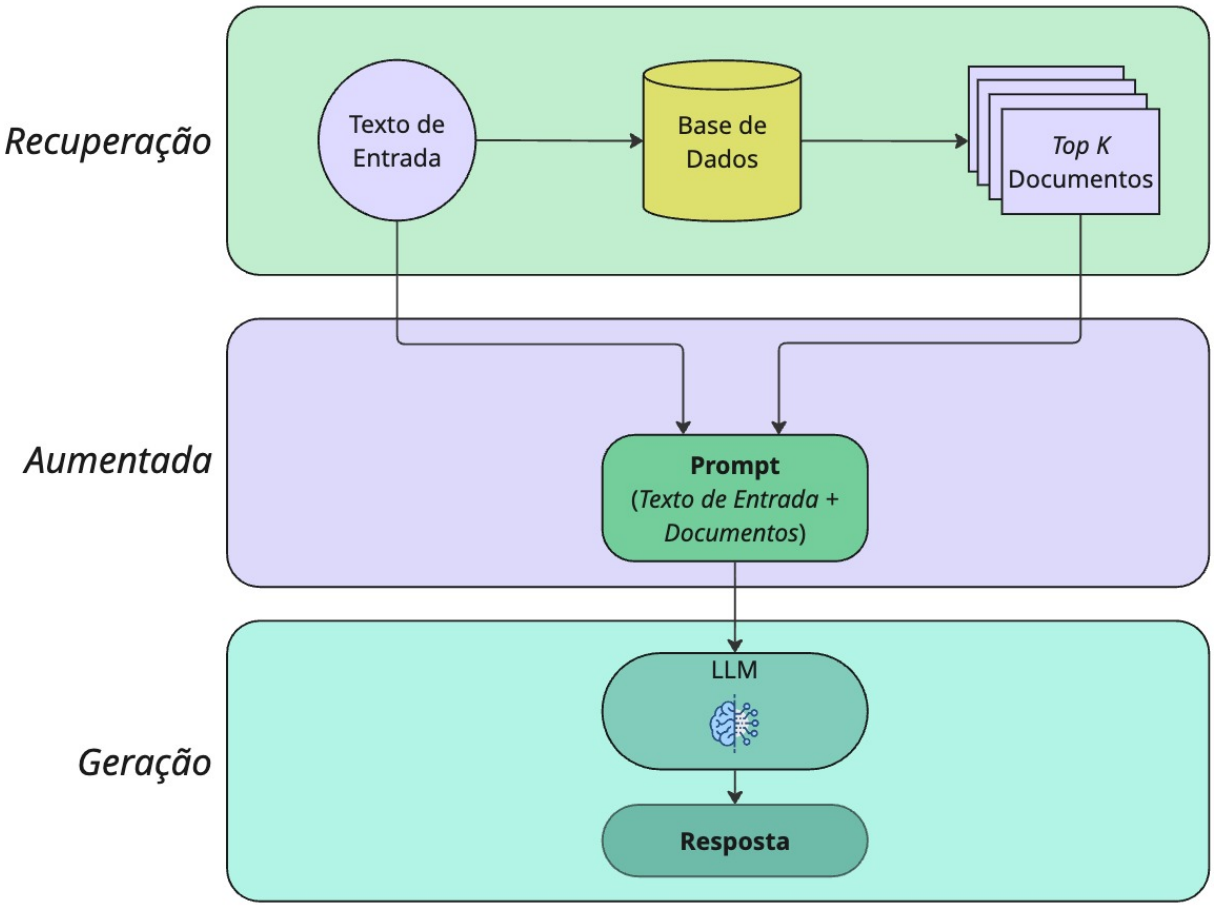
incorporar novas informações, o RAG dinamiza o acesso ao conhecimento por meio de sua componente de recuperação, que pode ser atualizada independentemente do modelo generativo (GAO *et al.*, 2024). Além disso, a separação entre recuperação e geração permite a otimização modular de cada etapa—como o uso de bancos vetoriais que trazem índices eficientes, como FAISS (DOUZE *et al.*, 2025) para *retrieval* e a seleção de LLMs com capacidades generativas distintas, adaptando-se a diferentes requisitos de latência e acurácia.

A aplicabilidade do RAG estende-se a domínios onde a precisão factual e a atualidade são primordiais, como agentes virtuais, sistemas de perguntas e respostas (QA) e sumarização documental (GAO *et al.*, 2024). Estudos empíricos demonstram que modelos baseados em RAG superam abordagens tradicionais em tarefas que demandam acesso a conhecimentos específicos ou em constante evolução, como respostas a perguntas sobre eventos recentes ou tópicos especializados (THULKE, 2021). Contudo, desafios persistem, incluindo a escalabilidade da recuperação em grandes escala, avaliação do conteúdo retornado na etapa de *retrieval*, a mitigação de viéses nos documentos recuperados e a otimização da interação entre as etapas de *retrieval* e *generation*.

A Figura 16 ilustra de maneira esquemática a arquitetura do RAG, detalhando seu fluxo em três etapas interconectadas: (1) Recuperação (*Retrieval*), onde uma consulta (*query*) é processada por um processo de busca para extrair documentos relevantes de uma base de dados externa, representada por vetores de *embeddings* e um *ranking* de trechos textuais; (2) Aumento de Contexto (*Augmentation*), que combina a *query* original com os documentos recuperados em um *prompt* ampliado, destacado visualmente pela fusão de caixas de texto; e (3) Geração (*Generation*), em que um LLM processa o contexto enriquecido para produzir uma resposta precisa e contextualizada.

A recuperação de informações (*information retrieval* - IR) pode ser abordada por meio de duas estratégias principais: métodos baseados em *embeddings* e métodos tradicionais baseados em correspondência lexical. A primeira abordagem emprega modelos de linguagem pré-treinados, como BERT (DEVLIN *et al.*, 2018) ou *SentenceTransformers* (REIMERS; GUREVYCH, 2019), para gerar representações vetoriais densas (*embeddings*) que capturam relações semânticas entre os termos e os documentos. Esses *embeddings* permitem a recuperação de informações mesmo na ausência de correspondência exata de palavras, graças à sua capacidade de inferir similaridades contextuais e sinônimos (KARPUKHIN, 2020). No entanto, essa abordagem demanda recursos computacionais significativos, tanto para o pré-processamento dos textos quanto para o cálculo de similaridade em espaços vetoriais de alta dimensionalidade, o que pode impactar a latência em sistemas quando operados em larga escala (LIN *et al.*, 2021b). Além disso, sua eficácia depende criticamente da qualidade do LLM adotado e do domínio de aplicação, requerendo *fine-tuning* em cenários específicos para maximizar a precisão.

Figura 16 – Visão Estrutural do RAG: *Retrieval*, *Augmentation* e *Generation*



Fonte: Do Autor

Em contrapartida, os métodos tradicionais de recuperação baseiam-se na eficiência de índices invertidos (estruturas de dados que mapeiam cada termo do vocabulário aos documentos que o contêm, agilizando a busca) e utilizam métricas estatísticas lexicais para o ranqueamento: o algoritmo TF-IDF (*Term Frequency-Inverse Document Frequency*) pondera a relevância equilibrando a frequência do termo no documento com sua raridade no *corpus*, enquanto o BM25 (*Best Matching 25*) representa uma evolução probabilística robusta que refina esse cálculo ao introduzir a saturação da frequência de termos (impedindo que repetições excessivas inflem desproporcionalmente o peso) e a normalização pelo comprimento do documento, favorecendo correspondências mais precisas em textos de tamanhos variados (ROBERTSON; ZARAGOZA, 2009).

Esses métodos são altamente eficientes em termos computacionais, pois operam sobre representações esparsas e permitem buscas rápidas em grandes volumes de dados (MANNING; RAGHAVAN; SCHÜTZE, 2008). Contudo, sua principal limitação reside na incapacidade de capturar relações semânticas mais profundas, como sinônimos, polissemias ou variações linguísticas, restringindo-se a correspondências superficiais de tokens (ZAMANI *et al.*, 2018). Embora soluções híbridas, como a combinação de BM25

com *embeddings* (LIN *et al.*, 2021a), tenham sido propostas para mitigar essas deficiências, a escolha entre as abordagens ainda depende do equilíbrio entre precisão semântica, escalabilidade e custo computacional inerente ao contexto de aplicação.

3.1 *RETRIEVAL:* BUSCA SEMÂNTICA BASEADA EM *EMBEDDINGS*

A busca baseada em *embeddings* constitui uma das técnicas utilizadas na etapa de recuperação de informações em sistemas RAG, fundamentado na transformação de unidades linguísticas como palavras, frases ou documentos que são representados por vetores de alta dimensionalidade (MIKOLOV *et al.*, 2013). Esses vetores, gerados por LLMs, capturam relações semânticas e sintáticas complexas, permitindo que termos com significados similares, mas expressões superficiais distintas, sejam mapeados para representações vetoriais próximas no espaço *embedding* (GOLDBERG, 2017). A eficácia dessa abordagem reside na capacidade dos modelos, como BERT (DEVLIN *et al.*, 2018) e GPT (BROWN; MANN; RYDER, 2020), de aprender representações contextuais.

A geração de *embeddings* contextuais, pilar fundamental para a eficácia da busca semântica, é viabilizada por arquiteturas baseadas em *Transformers* (VASWANI *et al.*, 2017) que utilizam mecanismos de atenção para ponderar dinamicamente a influência recíproca entre os termos de uma sentença. Modelos como o BERT consolidam essa competência linguística através de um pré-treinamento extensivo em conjuntos de dados diversificados, empregando tarefas como a Modelagem de Linguagem Mascarada (MLM) e a Previsão de Próxima Sentença (NSP) (DEVLIN *et al.*, 2018) para apreender estruturas sintáticas e semânticas complexas. Durante a inferência, esse aprendizado prévio permite que o texto de entrada seja processado em camadas profundas, resultando na produção de vetores numéricos dinâmicos que ajustam sua representação conforme o contexto, enriquecendo assim a precisão dos sistemas de recuperação de informação.

Na prática, a recuperação de informações via *embeddings* envolve a comparação de similaridade entre vetores, tipicamente utilizando métricas como a similaridade de cosseno ou distância euclidiana (MANNING; RAGHAVAN; SCHÜTZE, 2008). Isso acontece após a etapa de geração de *embeddings* ser executada para o conjunto de dados, sejam eles um conteúdo específico em uma base de dados, ou uma *query* que está sendo executada durante um processo de RAG. Sistemas de busca modernos, como os baseados no *Elasticsearch* com integração a modelos BERT, convertem tanto as *querys* quanto os documentos em *embeddings* e realizam uma busca por vizinhança no espaço vetorial (JOHNSON; DOUZE; JéGOU, 2021). Adicionalmente, métodos de *fine-tuning* adaptam os *embeddings* para domínios específicos, aumentando a precisão em tarefas especializadas (GURURANGAN *et al.*, 2020).

No processamento de uma consulta em sistemas de recuperação de informação jurídica, como a expressão "responsabilidade civil por danos ambientais", formulada para

fins ilustrativos neste presente texto, os *tokens* são definidos como as unidades lexicais resultantes da segmentação da consulta, sendo estes fundamentais para a interpretação semântica da intenção de busca do usuário e para a subsequente recuperação dos documentos pertinentes (REIMERS; GUREVYCH, 2019). O método léxico retornaria documentos contendo exatamente esses termos, ignorando variações como "responsabilização por prejuízos ecológicos". Já um sistema baseado em *embeddings* do BERT mapearia ambos os textos para vetores semanticamente próximos, recuperando jurisprudências relevantes mesmo que não compartilhem lexemas idênticos.

Para sistematizar o fluxo operacional de um sistema de recuperação baseado em similaridade semântica, a Tabela 2 descreve as etapas fundamentais do processamento de linguagem natural aplicadas neste contexto (MANNING; RAGHAVAN; SCHÜTZE, 2008). A estruturação abrange desde a decomposição granular da entrada, considerando a tokenização em subpalavras (*subtokens*), até a geração de representações vetoriais densas (*embeddings*) via arquitetura *Transformer*. Crucialmente, a tabela evidencia a estratégia de segmentação de documentos longos em fragmentos menores (*chunks*) para a composição da base vetorial, demonstrando como o cálculo de similaridade de cosseno é aplicado para ranquear estes segmentos em relação à consulta do usuário, permitindo uma recuperação de informação contextual e precisa.

Tabela 2. Etapas do Processamento de Linguagem Natural para Recuperação de Informação baseada em Similaridade Semântica.

Fase do Processo	**Descrição Detalhada**	**Exemplo/Resultado Chave**
Conteúdo Textual	Sentença ou documento original que servirá como base para a análise ou consulta.	Frase: *"Nesta quinta, jogadores do Cruzeiro tiveram seu alto desempenho notado pelas mídias..."*
Tokenização	Segmentação do texto em unidades menores. Em modelos modernos (*Transformers*), isso frequentemente envolve a quebra em subpalavras (*subtokens* ou *wordpieces*) para lidar com vocabulário complexo e variações morfológicas.	*Tokens*: "Nesta", "quin" "ta", "joga", "dores", "do", "Cruzeiro", "tiveram"
Processamento pelo *Transformer*	Geração de representações vetoriais contextuais (*embeddings*) para cada *token*. Utiliza-se arquiteturas *Transformer* com camadas de atenção *multi-head*. A auto-atenção (*self-attention*) permite ao modelo ponderar a influência de diferentes *tokens* na representação de um *token* específico (VASWANI *et al.*, 2017).	Saída: Vetores densos, representados por números.
Base de Dados de *Embeddings*	Repositório contendo os documentos processados. Devido ao limite de contexto dos modelos, os documentos longos são divididos em fragmentos menores (*chunks*). Cada *chunk* gera um vetor semântico independente, permitindo uma recuperação mais precisa de trechos específicos.	Doc1_Chunk1 = $[0, 3, \dots]$ Doc1_Chunk2 = $[0, 1, \dots]$ DocN_Chunk1 = $[EXEMPLO]$
Consulta (Q)	A sentença ou termo de busca é processado de forma análoga ao *input* textual, resultando em um vetor de *embedding* que representa semanticamente a consulta.	$Q = [0,21, -0,45, \dots, 1, 2]$
Cálculo da Similaridade de Cosseno	Mensuração da similaridade entre o vetor da consulta (Q) e os vetores dos *chunks* armazenados na base. A similaridade de cosseno avalia o cosseno do ângulo entre dois vetores, variando de -1 a 1.	Fórmula: $\cos(\theta) = \frac{Q \cdot C_i}{\|Q\| \cdot \|C_i\|}$ Ex: $\cos(Q, \text{Chunk1}) = 0.92$
Ranking e Retorno	Os trechos (*chunks*) são ordenados em função do valor de similaridade de cosseno em relação à consulta, em ordem decrescente. Os trechos mais relevantes (e seus documentos de origem) são apresentados.	Ordem: $\text{Chunk}_{\text{Doc1}} > \text{Chunk}_{\text{Doc3}} > \text{Chunk}_{\text{Doc2}}$

Fonte: Do Autor.

3.2 *RETRIEVAL:* BUSCA BASEADA EM ESTATÍSTICAS LEXICAIS

Os algoritmos de recuperação de informação baseados em representações léxicas, como o TF-IDF (*Term Frequency-Inverse Document Frequency*) e o BM25 (*Best Matching 25*), constituem abordagens clássicas amplamente utilizadas na construção de sistemas de busca e recuperação textual. Fundamentados em estatísticas de frequência de ocorrência de termos, esses métodos operam sobre estruturas de dados esparsas, sendo os índices invertidos (conforme mencionado no início deste Capítulo 3), permitindo

consultas extremamente rápidas e eficientes, mesmo em grandes volumes de documentos (MANNING; RAGHAVAN; SCHÜTZE, 2008). O TF-IDF calcula a relevância de um termo em um documento ponderando a frequência do termo no documento (TF) pela frequência inversa nos documentos da coleção (IDF), enquanto o BM25 introduz ajustes adicionais para considerar o comprimento do documento e a saturação da frequência dos termos (ROBERTSON; ZARAGOZA, 2009). Essas propriedades conferem aos métodos léxicos uma elevada escalabilidade, sendo frequentemente empregados em sistemas de produção devido à sua baixa complexidade computacional e à ausência de necessidade de treinamento supervisionado (ASKARI *et al.*, 2023).

As abordagens léxicas apresentam limitações intrínsecas relacionadas à superficialidade da representação textual adotada. Em particular, tanto o TF-IDF quanto o BM25 (ROBERTSON; ZARAGOZA, 2009) operam com base na correspondência exata de *tokens*, o que os torna incapazes de capturar relações semânticas mais profundas entre palavras ou expressões. Isso implica que termos com significados semelhantes, mas vocabulários distintos, como sinônimos ou expressões parafrásticas, não são associados durante o processo de recuperação, comprometendo a relevância dos resultados retornados (ZAMANI *et al.*, 2018). Adicionalmente, essas abordagens são sensíveis a variações morfológicas, ortográficas ou linguísticas, e tendem a apresentar desempenho inferior em domínios especializados ou em contextos de linguagem natural mais complexa.

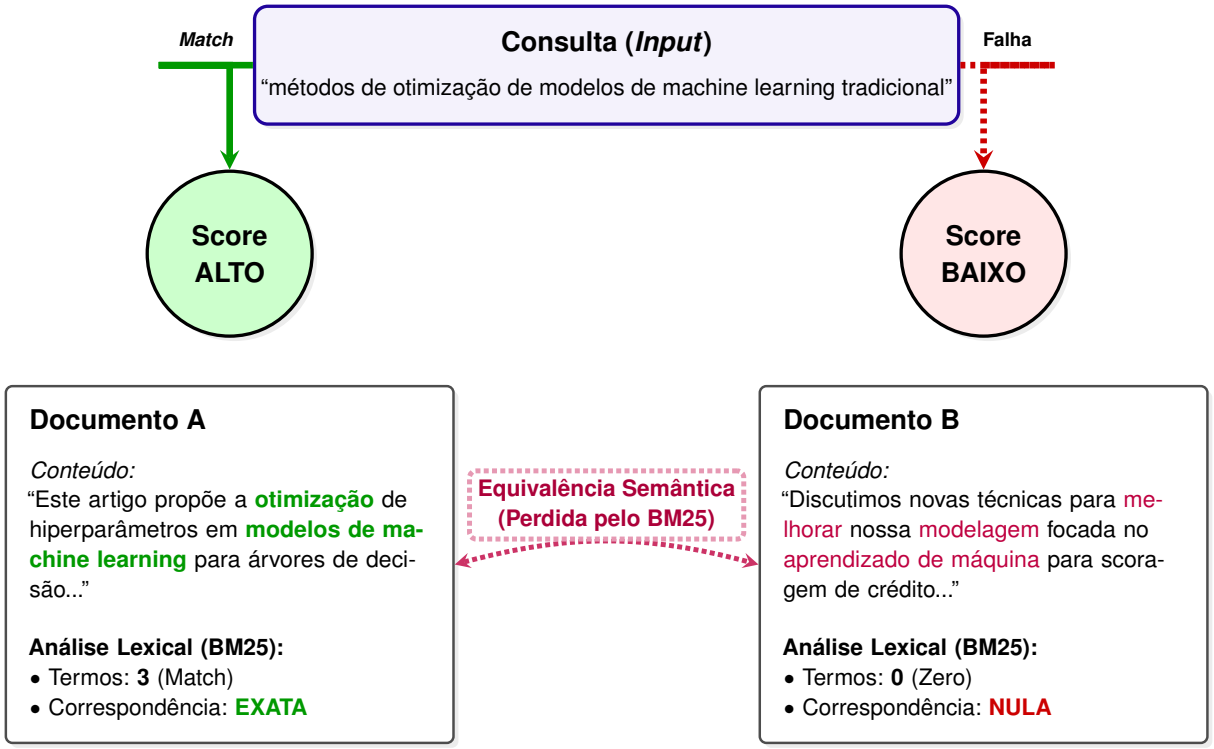
Em vez de percorrer cada documento integralmente a cada nova busca, o sistema mapeia antecipadamente cada termo único do vocabulário para uma lista de documentos onde ele ocorre (ROBERTSON; ZARAGOZA, 2009). Para ordenar a relevância desses documentos, historicamente utilizou-se o algoritmo TF-IDF. Esta métrica estatística pondera a importância de uma palavra equilibrando dois fatores: a frequência com que o termo aparece no documento específico (TF), recompensando a densidade, e a frequência com que aparece em todo o *corpus* (IDF), penalizando palavras comuns que oferecem pouco poder discriminativo (ROBERTSON; ZARAGOZA, 2009).

O estado da arte na recuperação lexical, contudo, é representado pelo BM25 (*Best Matching 25*) (ROBERTSON; ZARAGOZA, 2009). Este algoritmo é uma evolução probabilística do TF-IDF que introduz refinamentos cruciais para corrigir distorções de relevância: a saturação da frequência de termos, que impede que a repetição excessiva de uma mesma palavra aumente indefinidamente a pontuação (o décimo uso da palavra “jogador” não adiciona tanto valor quanto o primeiro), e a normalização pelo comprimento do documento, que penaliza textos excessivamente longos onde a ocorrência do termo pode ser apenas incidental, favorecendo documentos mais concisos e focados. Apesar dessa sofisticação estatística, o BM25 permanece restrito à correspondência exata de caracteres (ZAMANI *et al.*, 2018).

Para exemplificar essa limitação, um cenário onde um pesquisador busca pela

consulta: “métodos de otimização de modelos de machine learning tradicional”. O motor de busca lexical atomiza a frase e varre o índice invertido em busca de correspondências literais. Um Artigo A, que contenha a frase exata “otimização de modelos de machine learning”, receberá um *score* elevado. Em contrapartida, um Artigo B que discuta o mesmo tópico utilizando a terminologia “novas técnicas para melhorar nossa modelagem”, sem repetir as palavras da consulta, terá uma pontuação baixa ou nula, apesar de ter o mesmo sentido semântico. Para o algoritmo estatístico, os vetores de caracteres de “machine learning” e “aprendizado de máquina” são ortogonais e não possuem relação. Esse fenômeno, ilustrado na Figura 17, é conhecido como problema de descasamento de vocabulário e evidencia a falta de conexão semântica dos algoritmos tradicionais de busca textual(ZAMANI *et al.*, 2018).

Figura 17 – Impacto do Descasamento de Vocabulário na Recuperação de Informação em Buscas Léxicas.



Fonte: Do Autor.

A visualização proposta na Figura 17 sintetiza o desafio central da recuperação de informação tradicional. Enquanto o Documento A é corretamente recuperado devido à presença explícita das palavras-chave da consulta, o Documento B, que possui relevância semântica equivalente, mas utiliza terminologia distinta, é penalizado ou ignorado pelo algoritmo BM25.

3.3 *RETRIEVAL:* BUSCA HÍBRIDA

A busca híbrida é uma abordagem estratégica para superar as limitações tanto da recuperação léxica quanto da semântica, integrando seus respectivos pontos fortes em um pipeline unificado. Enquanto a busca léxica tradicional, baseada em técnicas como BM25 e TF-IDF, oferece alta eficiência computacional e precisão em correspondências literais, ela falha ao lidar com sinônimos, variações morfológicas e ambiguidade semântica. Por outro lado, a busca semântica, geralmente implementada com *embeddings* densos provenientes de modelos como BERT, captura relações de significado mesmo na ausência de sobreposição lexical, oferecendo maior cobertura conceitual e flexibilidade. A combinação dessas duas abordagens em um sistema híbrido permite capturar múltiplas dimensões da relevância e melhorar significativamente o desempenho em tarefas de recuperação de informação (VERBERNE, 2023).

A principal vantagem da busca híbrida está na sua capacidade de incorporar múltiplas fontes de sinal, sintático e semântico, para refinar o ranqueamento de documentos (ZHAN *et al.*, 2021). Uma estratégia amplamente adotada consiste em utilizar a recuperação inicial com BM25 para gerar um conjunto candidato de documentos, seguido por *re-ranking* com *embeddings* baseados em modelos de linguagem conhecidos como modelos *re-rankers*. Alternativamente, os *scores* gerados por modelos léxicos e semânticos podem ser combinados em um modelo de aprendizado para *ranking*, ponderando cada sinal com base em sua contribuição empírica para a relevância. Essa abordagem já demonstrou ganhos consistentes em avaliações de desempenho, onde os métodos híbridos superam isoladamente tanto a busca léxica, quanto a busca semântica (ZHAN *et al.*, 2021). Também é possível realizar esse fluxo de maneira estruturada via *pipeline*, sem a utilização de modelos *re-ranker*, essa situação pode ficar a critério de cada abordagem e necessidade de aplicação.

No entanto, a implementação de busca híbrida exige cuidados em ambientes de larga escala, onde latência e custo computacional são críticos. Estratégias como indexação eficiente de *embeddings* e pré-processamento seletivo (ex: filtragem por léxico antes da aplicação de modelos semânticos) ajudam a otimizar o *trade-off* entre desempenho e acurácia (BHAGDEV *et al.*, 2008). Assim, a busca híbrida consolida-se como um componente versátil, capaz de adaptar-se a domínios complexos sem comprometer a intenção original da consulta.

A Figura 18 ilustra um fluxo de busca híbrida, que pode ser aplicado a um sistema de RAG, composto por três estágios sequenciais: cálculo de similaridade vetorial, ranqueamento e fusão de resultados. No primeiro estágio (painel superior da Figura 18), observa-se a representação dual dos documentos e consultas por meio de vetores densos (*embeddings*), que codificam informações semânticas através de modelos de *embedding*, como SBERT (REIMERS; GUREVYCH, 2019), e vetores esparsos (como TF-IDF

ou BM25 (ROBERTSON; ZARAGOZA, 2009)), que quantificam a relevância léxica com base em estatísticas de frequência termo-documento. Essa abordagem bifacetada permite a complementaridade entre análise contextual e métricas tradicionais de recuperação de informação.

O estágio intermediário (painel central da Figura 18) demonstra o processo de ranqueamento, onde são selecionados separadamente os *Top K* resultados mais relevantes para cada modalidade de busca gerando duas listas ordenadas distintas: uma baseada em similaridade semântica (vetores densos) e outra em correspondência termo-estatística (vetores esparsos). Finalmente, o último estágio (painel inferior da Figura 18) apresenta a fusão estratégica desses resultados, que pode ocorrer através de métodos como: *Score fusion* (combinação linear ou não-linear das pontuações), *Re-ranking* semântico da lista léxica inicial, Interpolação adaptativa baseada em confiança do modelo (AZAD; DEEPAK, 2019).

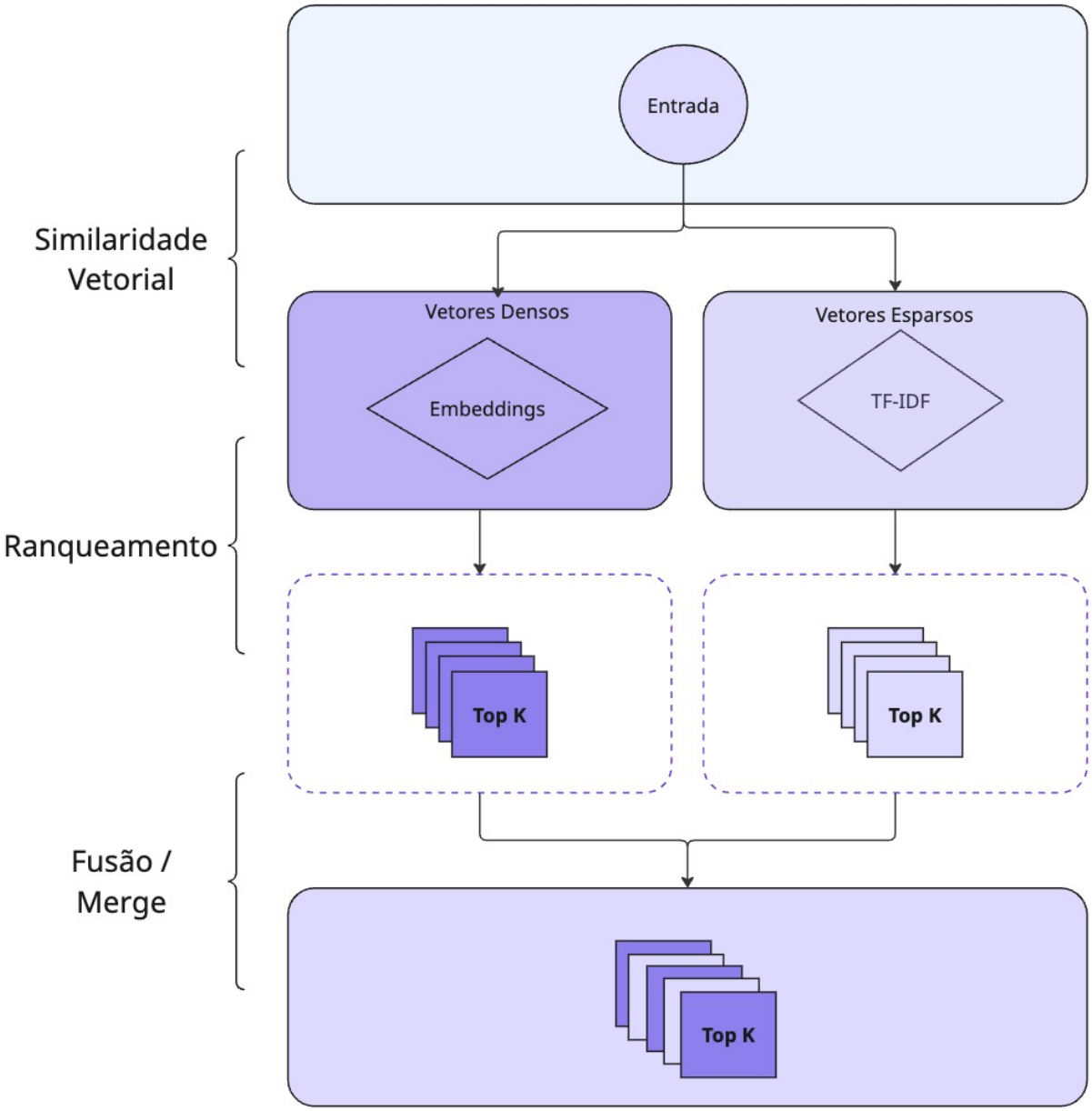
Como ilustrado na Figura 18, esta arquitetura possibilita a sinergia entre precisão léxica e compreensão contextual, que podem superar limitações inerentes a cada abordagem isolada. A disposição visual dos componentes na figura reforça o fluxo sequencial do processo, que podem ser feitos de forma paralela ou não, a depender do sistema e da necessidade de implementação para o caso.

3.4 *RETRIEVAL:* BUSCA POR *MULTI QUERY*

A técnica de busca baseada em *multiquery* representa um avanço significativo no contexto de RAG, especialmente quando aplicada a modelos de linguagem de grande escala (LLMs). Conforme destacado por (MA *et al.*, 2023), a reescrita de *queries* por meio de um LLM auxiliar permite reformular o *prompt* de entrada, adaptando-o semanticamente para melhorar a recuperação de documentos relevantes. Essa abordagem visa reduzir a lacuna entre a intenção do usuário e a necessidade de recuperação, um desafio crítico em sistemas de busca aumentada por recuperação (MA *et al.*, 2023). Ao reescrever a *query* original, é possível refiná-la lexical e semanticamente, potencializando a eficácia da busca em bases vetoriais e elevando o *score* de relevância dos documentos retornados.

A estratégia de *multiquery* estende esse princípio ao gerar múltiplas variações da *query* inicial, cada uma submetida a buscas independentes. Como observado por (MA *et al.*, 2023), essa técnica “clarifica a necessidade de recuperação a partir do texto de entrada“, permitindo que o sistema explore diferentes nuances semânticas. A diversificação das *queries* amplia a cobertura de documentos recuperados, mitigando limitações de ambiguidade ou escopo restrito. Como destacado por (MA *et al.*, 2023) esse método “alivia o ônus do aprimoramento de capacidade de recuperação“, uma vez que a pluralidade de *queries* compensa eventuais deficiências do *retriever* monolítico.

Figura 18 – Abordagem Híbrida para Busca: Integrando *Embeddings* Semânticos e Métricas Léxicas



Fonte: Do Autor

A implementação dessa técnica depende da integração entre um modelo de LLM reescritor (*rewriter*) treinável e o *pipeline* de RAG. No estudo de (MA *et al.*, 2023) é proposto o uso de “um modelo de linguagem pequeno e treinável“ para essa tarefa, otimizado via *reinforcement learning* com base no desempenho do LLM principal. Essa abordagem tende a reduzir o custo computacional (MA *et al.*, 2023), e também pode garantir que as *queries* reescritas estejam alinhadas com as necessidades do sentido que será realizada a busca. Os autores destacam que “um modelo menor pode ser competente para reescrita de queries“, evidenciando a viabilidade de soluções eficientes em cenários de recursos limitados.

Os benefícios da *multiquery* são particularmente evidentes em tarefas que demandam alto conhecimento factual, como domínios de QA. Conforme demonstrado por (MA *et al.*, 2023), a reescrita adaptativa “melhora consistentemente o desempenho do LLM aumentado por recuperação“, com ganhos mensuráveis em desempenho e qualidade das respostas. A agregação de resultados de múltiplas *queries* enriquece o contexto disponível para o LLM, e também minimiza erros de alucinação e desalinhamento temporal (MA *et al.*, 2023). Assim, a técnica consolida-se como uma solução escalável para sistemas que exigem precisão e atualização dinâmica de conhecimento.

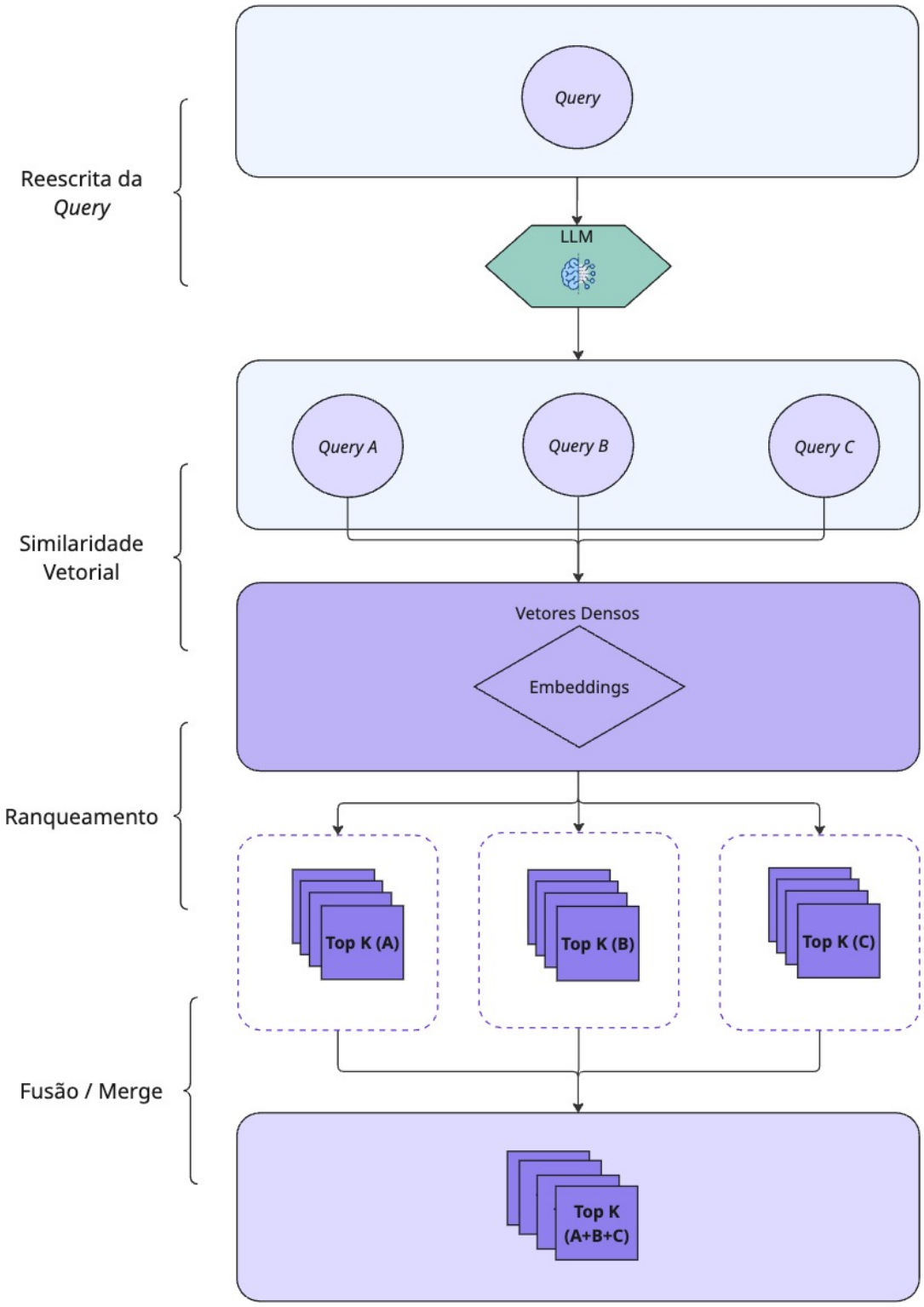
A Figura 19 ilustra o fluxo de processamento de uma estratégia de *multiquery* em um sistema de RAG. Inicialmente, a *query* de entrada (*Query Input*) é processada por um modelo de LLM, que a reescreve em múltiplas variações (*Query A, Query B, Query C*), cada uma capturando nuances semânticas distintas. Essas *queries* são então convertidas em *embeddings* vetoriais densos, permitindo a busca por similaridade (Similaridade Vetorial) em um banco de dados indexado. Na imagem é ilustrada uma única base vetorial, mas essa estrutura pode ser adaptada e a busca pode ser realizada para N bancos vetoriais distintos, sendo eles um para cada *query*, por exemplo. Para cada *query*, são recuperados os *Top K* documentos mais relevantes (*Top K (A), Top K (B), Top K (C)*), os quais são posteriormente consolidados em um único conjunto (*Top K (A+B+C)*) por meio de uma etapa de fusão (*Merge*). Esse processo otimiza o *ranking* (Ranqueamento) dos resultados, assegurando que a saída final agregue a diversidade e a precisão das consultas individuais, melhorando assim a qualidade do contexto fornecido ao LLM para geração de respostas.

A estratégia de *multiquery* demonstra-se como uma solução versátil e adaptável para sistemas de RAG, permitindo customizações específicas conforme as exigências do domínio ou aplicação. Ao decompor a consulta original em variações semanticamente enriquecidas, isso tende a ampliar a cobertura de recuperação de documentos. Conforme evidenciado na literatura (MA *et al.*, 2023), tal flexibilidade opera em sinergia com modelos de linguagem de menor porte, otimizando custos computacionais sem comprometer a precisão. Portanto, a implementação de *multiquery* em *pipelines* de RAG se posiciona como um elemento escalável, capaz de equilibrar desempenho, eficiência e personalização das soluções.

3.5 TÉCNICAS AVANÇADAS PARA RAG

Diversas técnicas avançadas podem ser aplicadas para aprimorar o fluxo de *Retrieval Augmented Generation* (RAG), incluindo expansão de consultas (*query expansion*) (NAGPAL, 2022), re-ranqueamento (*re-ranking*) (BAŞ *et al.*, 2022), atomização de documentos (RAINA; GALES, 2024), resumo de documentos (KINGER *et al.*, 2025), entre outras como a realização do *fine-tuning* no modelo de *embedding* (DEVLIN *et al.*, 2018). Essas abordagens visam melhorar tanto a precisão quanto a eficiência na recuperação

Figura 19 – Abordagem de *Multiquery* para Busca



Fonte: Do Autor

e geração de informações, permitindo que o modelo forneça respostas mais relevantes e contextualizadas, aproveitando melhor os dados externos disponíveis.

A expansão de consultas (*query expansion*) é uma técnica que visa melhorar

a precisão da recuperação de informações, expandindo a consulta original com termos adicionais ou sinônimos relevantes, ajudando o modelo a cobrir melhor diferentes formas de expressão de um mesmo conceito. Isso aumenta a chance de recuperar documentos mais pertinentes ao contexto da pergunta inicial, como descrito por (NAGPAL, 2022).

O re-ranqueamento (*re-rank*) envolve a ordenação de documentos recuperados com base na sua relevância, geralmente utilizando um modelo mais sofisticado para re-avaliar os resultados preliminares. (GLASS; ROSSIELLO, 2022) demonstraram que isso pode melhorar significativamente a qualidade das respostas geradas, garantindo que os documentos mais úteis e relacionados apareçam no topo da lista de recuperação.

O resumo ou encurtamento de documentos é uma técnica que visa condensar o conteúdo de um documento extenso em uma versão mais curta e concisa, preservando suas principais informações e ideias centrais (KINGER *et al.*, 2025). O modelo gera novas frases com base no entendimento do conteúdo, reescrevendo as informações de forma mais sintetizada e fluida. Essas técnicas são particularmente úteis para reduzir a sobrecarga de leitura e facilitar o consumo rápido de grandes volumes de texto, sendo amplamente aplicadas em áreas como análise de notícias, relatórios empresariais e artigos acadêmicos.

Já a atomização de documentos refere-se à prática de dividir documentos em fragmentos menores, ou “átomos“, como seções, parágrafos ou até frases. Isso permite uma recuperação mais precisa e direta, já que o modelo pode acessar partes específicas de um documento, em vez de processá-lo por completo, como apresentado por (RAINA; GALES, 2024). Esse nível de granularidade facilita que o modelo acesse informações mais relevantes e localizadas, gerando respostas mais coerentes e detalhadas.

3.5.1 *Fine-Tuning* no modelo *Embedding*

A integração do *fine-tuning* em modelos de *embedding* representa uma estratégia fundamental para a adaptação de LLMs a domínios específicos, permitindo a otimização da representação vetorial de textos com base em características contextuais únicas. Segundo (DEVLIN *et al.*, 2018), o *fine-tuning* possibilita o ajuste fino dos parâmetros do modelo de *embedding*, refinando sua capacidade de capturar nuances semânticas e sintáticas presentes em corpora especializados. Essa abordagem é particularmente relevante em cenários onde a heterogeneidade documental exige uma representação vetorial consistente, reduzindo discrepâncias entre documentos de formatos e conteúdos diversos. Além disso, como destacado por (DEVLIN *et al.*, 2018), o processo de *fine-tuning* pode ser realizado em múltiplas camadas da arquitetura neural, permitindo uma adaptação hierárquica que vai desde *features* superficiais até relações contextuais profundas, garantindo assim uma maior precisão na recuperação e na geração de respostas.

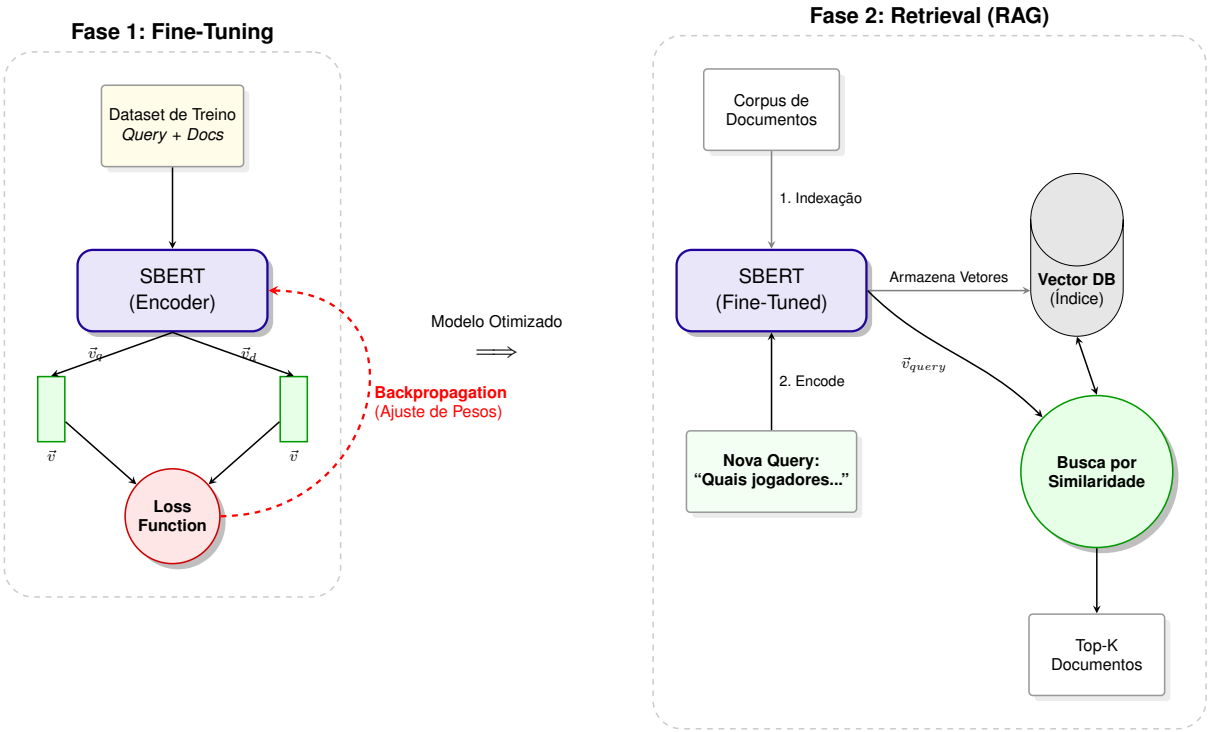
Ademais, a combinação do *fine-tuning* com técnicas de RAG potencializa a eficácia dos modelos de *embedding* ao alinhar a recuperação de informações com a

geração contextualizada, como discutido por (LEWIS *et al.*, 2020). O *fine-tuning* especializa o modelo de *embedding* para mapear consultas e documentos em um espaço vetorial coeso, onde a proximidade semântica reflete relações contextuais precisas, minimizando ruídos na recuperação.

Segundo (LIU; AL., 2023), o modelo de *embedding* utilizado para recuperação da informação pode passar por um processo de *fine-tuning* conforme descrito anteriormente. O objetivo principal do ajuste do modelo de recuperação é melhorar a qualidade das representações semânticas, alcançado previamente usando o conteúdo dos documentos (ANISUZZAMAN *et al.*, 2025). O modelo de *embedding* utilizado, como BERT, pode ser modificado e abastecido com informações mais precisas diante de um contexto, como nova formatação de *query*, respostas e dados estruturados para a determinada resposta referente à *query* de consulta.

A Figura 20 ilustra o fluxo de especialização (*fine-tuning*) (ANISUZZAMAN *et al.*, 2025) de um modelo codificador (*Encoder*), como o SBERT, para tarefas de recuperação da informação. Nesta etapa, o modelo genérico é submetido a um treinamento supervisionado utilizando pares de dados específicos do domínio (documentos e seus respectivos contextos ou rótulos de relevância) representado pela Fase 1. O objetivo é ajustar os pesos da rede neural para que o espaço vetorial latente reflita com maior precisão a semântica do domínio alvo. Uma vez refinado, na Fase 2 o modelo é empregado na fase de inferência para processar tanto a *query* do usuário quanto o *corpus* de documentos, gerando *embeddings* densos altamente correlacionados. Estes vetores otimizados permitem que o sistema de recuperação calcule a similaridade semântica com maior assertividade, retornando os resultados mais pertinentes para a geração da resposta.

Figura 20 – *Fine-Tuning* de Modelo de *Embedding* e Fluxo de Recuperação.



Fonte: Adaptado de (ANISUZZAMAN *et al.*, 2025)

A etapa de *Generation* consiste em transformar as informações recuperadas na fase de *Retrieval* e utilizá-las como insumo, para obter um texto fluente e coerente gerado pelo LLM. Os *embeddings* obtidos são incorporados ao modelo LLM, que gera uma resposta com base no contexto fornecido. O texto produzido combina tanto os documentos recuperados quanto o conhecimento pré-treinado do modelo. Modelos como o BERT (ALAPARTHI; MISHRA, 2020) e GPT (LIU; AL., 2023) são frequentemente utilizados nessa etapa, processando os documentos mais relevantes para a *query* e gerando uma saída contextualizada.

3.5.2 Atomização de documentos

A atomização de documentos é uma técnica que visa fragmentar documentos em partes menores e mais manejáveis, como seções, parágrafos ou frases individuais (RAINA; GALES, 2024), diferentemente do processo de *Chunking* genérico, a Atomização reescreve o texto de forma estratégica em pedaços menores, utilizando um LLM. Esse processo facilita a recuperação de informações específicas, uma vez que, em vez de lidar com grandes blocos de texto, o modelo pode focar diretamente nas partes mais relevantes para a consulta realizada. Ao atomizar um documento, a granularidade da informação aumenta, permitindo que o sistema capture detalhes mais finos e precisos, evitando a ambiguidade ou a perda de contexto que pode ocorrer ao trabalhar com textos longos e densos (RAINA; GALES, 2024).

Esse nível de fragmentação é particularmente útil em cenários onde as informações mais relevantes estão distribuídas em diferentes partes do documento ou quando a consulta é específica o suficiente para não justificar a análise do documento inteiro. Ao invés de retornar todo o conteúdo, o modelo pode retornar apenas as partes que respondem diretamente à pergunta, tornando a recuperação de informações mais rápida e eficiente.

A Figura 21 ilustra o processo de atomização de documentos no contexto de um sistema de recuperação de informação baseado em LLM. O documento original, representado pela etiqueta "*Document*" no topo da figura, é fragmentado em quatro partes distintas (Parte A, Parte C, Parte D e Parte B), organizadas em uma estrutura não linear que sugere uma possível reordenação ou classificação temática. Essas partes atomizadas são integradas a um *VectorStore* (base de dados vetorial), cada fragmento é convertido em um vetor de *embeddings*, permitindo buscas semânticas. A presença do LLM no centro do diagrama destaca seu papel fundamental no processamento das informações, atuando como intermediário para a transformação do documento original em trechos menores (átomos). Essa abordagem, conforme discutido no texto-base, otimiza a recuperação de informações específicas, pois a fragmentação em partes menores (como seções ou parágrafos) pode aumentar a granularidade dos dados, evitando perda de contexto e melhorando a precisão das respostas geradas pelo modelo (RAINA; GALES, 2024). A disposição das partes atomizadas e sua conexão com o *VectorStore* reforçam a ideia de que a atomização não é apenas uma divisão textual, mas uma transformação estrutural que viabiliza operações avançadas de busca e análise semântica.

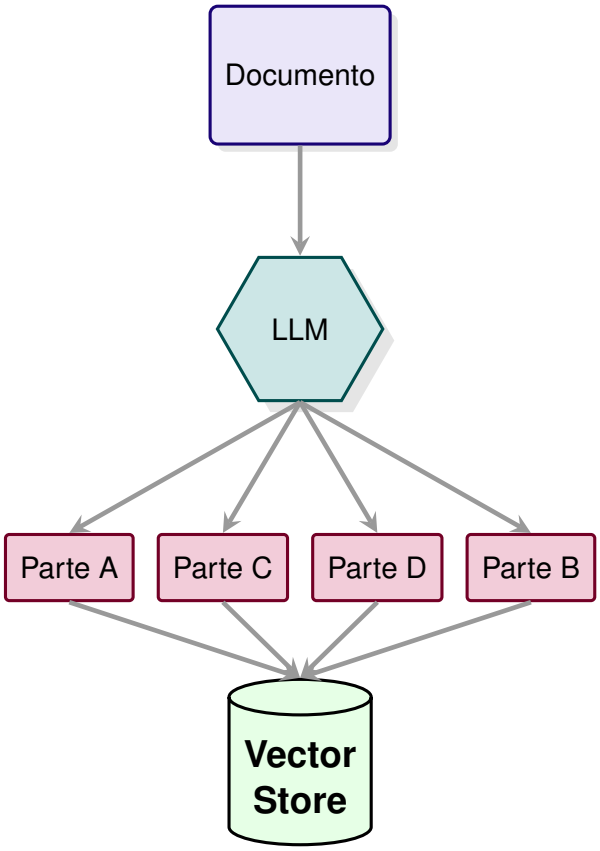
A atomização também reduz a sobrecarga de processamento, pois o sistema processa pequenos blocos de texto em vez de analisar documentos inteiros, o que tende a ser mais custoso em termos de tempo e recursos computacionais, devido ao fato de que documentos maiores tem uma maior quantidade de *tokens*, que é diretamente proporcional à quantidade de recurso utilizado para o processamento.

Contudo, os documentos divididos em átomos, farão parte do processo de *Retrieval*, que será retornado para o modelo *Generator* que utilizará os átomos para a geração de texto contextualizada. Perguntas direcionadas para determinados átomos, tendem a obter um bom resultado de alinhamento entre pergunta-reposta, conforme avaliado por (RAINA; GALES, 2024).

3.5.3 Redução de contexto

A redução de contexto em um trecho de texto ou documento refere-se à prática de condensar o conteúdo para focar nas informações mais relevantes, removendo detalhes redundantes ou menos importantes. O principal objetivo dessa técnica é otimizar o processamento de grandes volumes de dados, facilitando o acesso a informações cruciais sem sobrecarregar o usuário ou o sistema. Além disso, a redução de contexto melhora a

Figura 21 – Atomização de documentos a partir de modelo de LLM



Fonte: Do Autor

eficiência em tarefas de recuperação de informações e também a geração de respostas, tornando o processo mais ágil e focado. Entre os problemas resolvidos por essa abordagem estão a eliminação de ruído informacional, em um documento pode conter diversos tipos de ruídos de informação, como trechos específicos de página, caracteres especiais, e conteúdos que não fazem sentido para o contexto.

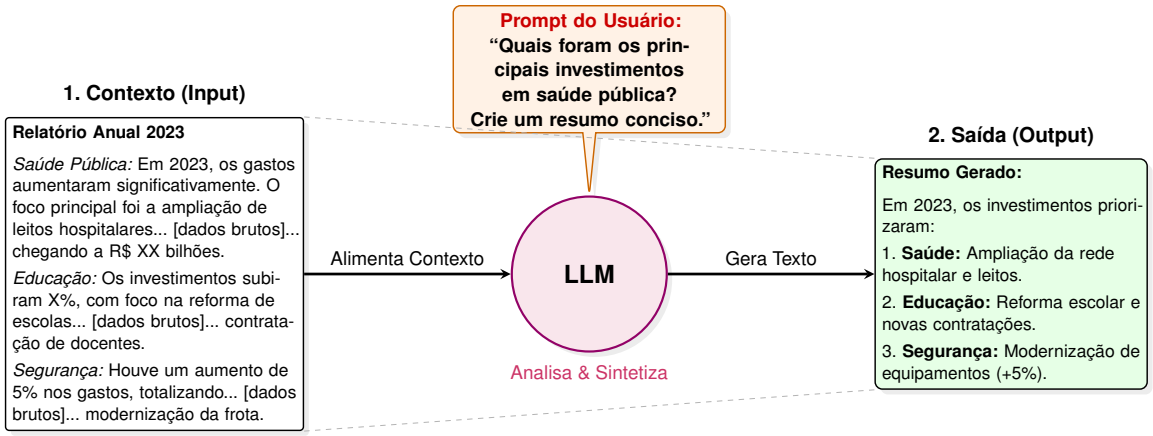
Para a redução de documentos, é essencial que um *prompt* seja aplicado em cada trecho de documento (também pode ser chamado de *chunk*). Cada *chunk* se comportará como um átomo do documento original, como um parágrafo ou uma seção específica e de forma reduzida e direta, e o objetivo do *prompt* é direcionar o modelo a focar nas informações mais relevantes dentro desse fragmento. Essa abordagem garante que, mesmo ao lidar com grandes quantidades de dados, o modelo possa processar cada parte de forma eficiente, extraindo o conteúdo mais importante de cada unidade *chunk*.

Ao aplicar *prompts* de compressão de texto em cada *chunk*, o sistema é capaz de identificar e reduzir partes menos relevantes ou redundantes, resultando em um resumo mais preciso e focado.

Na Figura 22 é possível visualizar um exemplo de redução de contexto a partir de um modelo LLM, diante de um documento que representa dados orçamentários e gastos

públicos. O *prompt* é aplicado a partir do contexto prévio, que solicita uma redução do documento de acordo com a técnica ou estratégia desejada, a fim de reduzir as informações.

Figura 22 – Redução de contexto a partir de modelo de LLM.



Fonte: Do Autor

De acordo com o trabalho realizado por (KINGER *et al.*, 2025), existe um grande desafio quando se trata de compressão contextual de documentos, as técnicas de compressão textual estão avançando rapidamente e isso pode influenciar o formato que os LLMs respondem para cada contexto referente a compressão contextual, e também, o conjunto de dados pode influenciar e ser um fator para cada técnica e formato de redução de texto utilizada.

Outros desafios foram identificados em documentos com grandes contextos, um deles é o fenômeno *"Lost in the Middle"* (LIU *et al.*, 2023), se trata de quando os LLMs são utilizados para processar textos longos e complexos, especialmente aqueles que demandam a extração de informações contextualmente relevantes, seu desempenho varia consideravelmente dependendo de onde essas informações estão posicionadas dentro do texto. Observou-se neste trabalho realizado por (LIU *et al.*, 2023) que os LLMs respondem bem quando as informações relevantes estão presentes no início ou no final do conteúdo. No entanto, quando esses dados se encontram no meio do conteúdo, o desempenho do modelo diminui significativamente. Isso ocorre porque, durante o processamento sequencial, os LLMs costumam dar mais peso às informações localizadas nas extremidades do texto, resultando em uma atenção menor para os detalhes que aparecem na parte central.

Essa limitação pode ser crítica em tarefas que envolvem textos longos, como resumos, respostas a perguntas e recuperação de informações, onde a precisão e a relevância das respostas dependem da identificação correta de informações em todo o texto. O problema se agrava quando o conteúdo central é denso ou contém nuances importantes, pois o modelo pode ignorar ou não dar a devida importância a esses elementos. Assim, um dos maiores desafios no uso de LLMs para textos extensos é garantir que o modelo consiga

capturar e processar eficientemente o contexto relevante ao longo de toda a entrada, e não apenas nas suas extremidades. Soluções como a segmentação de textos e a utilização de técnicas que aumentem a capacidade de retenção de informações no meio do conteúdo são fundamentais para mitigar esse problema.

Contudo, a ideia de resumir um documento com base em um contexto específico pode variar de acordo com diferentes estratégias e conjuntos de dados. Ao adotar uma estratégia direcionada para um determinado contexto, é possível mitigar a perda de nuances importantes que costumam ocorrer no meio do documento. Isso permite que as informações centrais, muitas vezes negligenciadas, sejam devidamente capturadas, garantindo um resumo mais preciso e relevante, e o estudo dessa abordagem será uma das linhas de análise deste presente projeto.

3.5.4 Reranqueamento (*Re-Ranking*)

O *re-ranking* se constitui como um dos elementos centrais nas arquiteturas contemporâneas de sistemas de RAG e busca (VERBERNE, 2023). Este componente atua subsequente à etapa inicial de recuperação de documentos, ou *first-stage retrieval*, na qual um conjunto de dados é obtido a partir de fontes como bancos vetoriais ou grafos de conhecimento (VERBERNE, 2023). Tais mecanismos de recuperação primária geralmente privilegiam eficiência em detrimento da precisão, o que implica em resultados ruidosos ou apenas parcialmente relevantes. Assim, o *re-ranking* se solidifica como um mecanismo de refinamento semântico, cujo objetivo é reordenar os documentos candidatos com base em critérios de similaridade.

A operação de *re-ranking* é geralmente conduzida por modelos de LLMs ou por modelos especializados em *cross-encoding* (BAŞ *et al.*, 2022), os quais são capazes de avaliar a relevância contextual entre a consulta e cada um dos documentos de forma conjunta e não isolada. Essa capacidade conjunta permite ao modelo explorar nuances semânticas profundas e relações latentes entre termos que não são capturadas pelas métricas tradicionais de similaridade vetorial.

Adicionalmente, o *re-ranking* desempenha um papel fundamental na mitigação de alucinações geradas por modelos generativos nos sistemas RAG, ao garantir que o conjunto final de contextos fornecidos à etapa de geração seja de alta relevância e precisão factual. Quando incorporado de forma criteriosa, esse processo reduz significativamente o risco de respostas desconectadas da base documental consultada. No trabalho de (VERBERNE, 2023) é destacado que, sob uma perspectiva arquitetural, o *re-ranking* pode ser compreendido como uma ponte entre a eficiência do primeiro estágio de recuperação e a robustez semântica necessária à geração controlada e consistente de respostas, portanto, sendo um elemento que impacta positivamente em fluxos de busca.

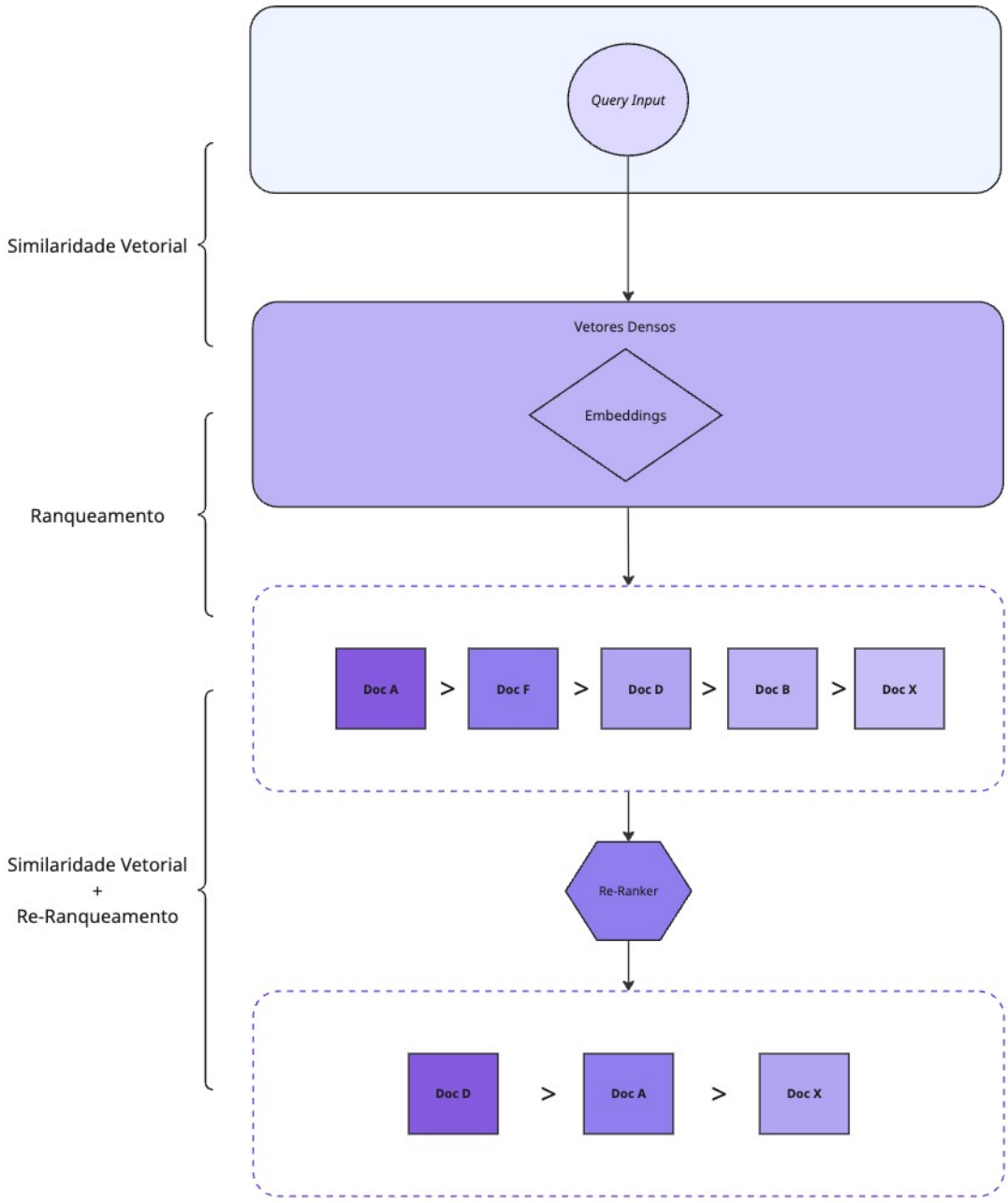
No contexto de reranqueamentos na área de recuperação da informação, o

trabalho de (BAŞ *et al.*, 2022) implementou uma estratégia de reranqueamento baseado em rede neural projetada para refinar os resultados brutos de buscas lexicais de usuários. A metodologia proposta adota uma arquitetura de dois estágios, onde a primeira etapa de busca (*ranking*) atua inicialmente como um filtro de candidatos, seguido pela aplicação de um modelo *Cross-Encoder* (baseado em BERT) especificamente para a etapa de reranqueamento. Baş *et al.* (2022) treinou esta rede utilizando o histórico de interações dos usuários para que o modelo aprendesse a pontuar a relevância contextual entre a consulta e o documento, superando a simples correspondência de palavras-chave. O estudo demonstrou que o reranqueamento, viabilizado em produção pelo uso de aceleração via GPU, foi capaz de corrigir o posicionamento de documentos relevantes que apareciam distantes do topo na busca tradicional, resultando em uma melhoria substancial nas métricas de MRR (*Mean Reciprocal Rank*) e na precisão dos resultados entregues ao usuário final, trazendo uma evidência academica sobre a potencialidade da técnica de reranqueamento.

A técnica também se destaca no trabalho de (OAK *et al.*, 2025), que apresenta uma abordagem inovadora para mitigação de conteúdos prejudiciais em sistemas de recomendação de redes sociais, com foco no uso de modelos de LLMs para *re-ranking* de recomendações. Em vez de depender exclusivamente de classificadores tradicionais, que exigem grandes volumes de dados anotados e enfrentam dificuldades para generalizar diante da natureza dinâmica do conteúdo nocivo, (OAK *et al.*, 2025) propõe um mecanismo baseado em comparações pareadas entre conteúdos, realizado por LLMs em cenários aplicando técnicas de *prompt engineering* como *zero-shot* e *few-shot*, que se tratam de exemplos sobre como o modelo de LLM deve agir ou se orientar, para obter uma resposta mais orientada (OAK *et al.*, 2025). Essa técnica permite reordenar sequências de conteúdo com base em preferências de segurança e bem-estar, reduzindo a exposição a conteúdos danosos, como desinformação, apelos a transtornos alimentares ou material depressivo. Entre as principais contribuições do trabalho estão a demonstração empírica de que os LLMs superam soluções consolidadas de mercado, mesmo sem treinamento específico, e a introdução de métricas que avaliam a exposição relativa ao dano em sequências de conteúdo. O estudo reforça a viabilidade do uso de LLMs e *re-ranking* como ferramenta adaptativa e escalável para promoção de ambientes digitais mais seguros.

A Figura 23 ilustra um processo genérico de *re-ranking* em duas etapas: (1) Ranqueamento Inicial, onde vetores densos (*embeddings*) calculam similaridades entre documentos (“Doc A“, “Doc F“, ...), gerando uma ordenação preliminar; e (2) Re-Ranking, no qual um modelo especializado (*re-ranker* baseado em arquiteturas avançadas como *Transformers*) refina a ordem inicial, priorizando itens contextualmente mais relevantes (“Doc D“ sendo ranqueado como primeiro), por exemplo. Este fluxo é típico em sistemas de recuperação de informação, que combinam eficiência computacional com precisão semântica, sendo aplicável a tarefas como busca semântica em fluxos de RAG.

Figura 23 – Processo genérico de *re-ranking*



Fonte: Do Autor

4 DETECÇÃO DE DESINFORMAÇÃO USANDO APRENDIZADO DE MÁQUINA

A crescente sofisticação e a velocidade de disseminação de narrativas falsas ou enganosas, impulsionadas pelas plataformas de mídia social e por um ambiente digital hiperconectado, demandam o desenvolvimento de estratégias de detecção robustas e escaláveis (ALGHAMDI; LUO; LIN, 2024). Neste contexto, o campo do aprendizado de máquina (*machine learning*) emerge como uma fronteira promissora (ALGHAMDI; LUO; LIN, 2024), oferecendo um arsenal de técnicas capazes de analisar padrões complexos em dados textuais e contextuais para a identificação automatizada de conteúdo fraudulento.

Este capítulo abordará uma análise dos trabalhos relacionados contemporâneos dedicados à detecção de *fake news* por meio de abordagens computacionais. A exploração se inicia com uma delimitação conceitual e caracterização do fenômeno das *fake news*, fundamentada em uma revisão da definição e conceito de *fake news*, descrita na seção 4.1. Subsequentemente, serão abordadas as principais abordagens e técnicas contemporâneas utilizadas na detecção de *fake news*.

A primeira vertente a ser examinada na seção 4.3 compreende os métodos de detecção baseados em aprendizado de máquina determinístico. Essa seção abordará os algoritmos clássicos, como Árvores de Decisão (BREIMAN, 2001), que constituíram a primeira geração de classificadores de *fake news*. Será dada ênfase à engenharia de características (*feature engineering*) que sustenta tais modelos (FAYAZ *et al.*, 2022), explorando desde atributos linguísticos e metadados de informações contextuais da propagação da notícia.

Avançando para as fronteiras da inteligência artificial, a Seção 4.4.1 abordará o uso de Modelos de Linguagem de Grande Escala (LLMs) (MOHAMMADABADI *et al.*, 2025) para a classificação de *fake news*. Serão analisadas as arquiteturas e as metodologias de aplicação de modelos como a série GPT (*Generative Pre-trained Transformer*) (RADFORD *et al.*, 2018), dos quais se destacam pela presença das estruturas de *Attention* (VASWANI *et al.*, 2017), trazendo uma maior robustez de compreensão semântica e contextual, superando em muitos casos as abordagens tradicionais.

Finalmente, a Seção 4.4.2 levantará a discussão das mais recentes inovações: as soluções de detecção de *fake news* baseadas em agentes de LLMs e Inteligência Artificial. Essa seção abordará sistemas mais complexos e autônomos, nos quais agentes inteligentes, potencializados por LLMs, podem ativamente buscar, verificar e contextualizar informações de múltiplas fontes para emitir um veredito sobre a veracidade de uma notícia. Serão explorados conceitos emergentes como técnicas de utilização de *tooling*, definição de agentes e a integração de bases de conhecimento externas para um combate mais

dinâmico e resiliente à desinformação.

4.1 DEFINIÇÃO E CARACTERÍSTICAS

Segundo (ALGHAMDI; LUO; LIN, 2024), as *fake news* são um fenômeno complexo e ainda sem uma definição única e universalmente aceita. No estudo de (ALGHAMDI; LUO; LIN, 2024), é mencionada a volatilidade da definição de *fake news*, podendo assumir diferentes definições na perspectiva de cada autor, pesquisadores como (ALLCOTT; GENTZKOW; YU, 2019) definem *fake news* como “notícias sem base factual apresentadas como reais“.

As razões por trás da disseminação das *fake news* são variadas, no cenário político, elas são frequentemente utilizadas para manipular eleições, direcionar debates públicos e acirrar a polarização social, como ocorreu nas eleições presidenciais dos Estados Unidos em 2016 (ALGHAMDI; LUO; LIN, 2024), quando uma grande quantidade de desinformação circulou nas redes sociais. No âmbito financeiro, a criação de *fake news* pode ter como objetivo a obtenção de lucro por meio do alto volume de cliques e anúncios publicitários, explorando conteúdos sensacionalistas para atrair a atenção do público. Já na esfera ideológica, a desinformação pode ser usada para fortalecer crenças específicas, espalhar teorias conspiratórias ou defender determinadas narrativas religiosas e culturais (ALGHAMDI; LUO; LIN, 2024), contribuindo para a formação de bolhas informativas que dificultam o acesso a fontes confiáveis.

A propagação de *fake news* é facilitada pela rápida disseminação de conteúdos em plataformas digitais, onde a diferenciação entre fontes confiáveis e não confiáveis torna-se desafiadora. Para compreender melhor os diferentes tipos de *fake news*, a Tabela 3 apresenta suas principais categorias e descrições, de acordo com (ALGHAMDI; LUO; LIN, 2024).

Tabela 3 – Principais tipos de *fake news* e suas características.

Tipo de *Fake News*	**Descrição**
Desinformação	Informação falsa criada intencionalmente para enganar.
Misinformação	Informação incorreta compartilhada sem a intenção de enganar.
Rumores	Informações não verificadas que podem ser verdadeiras ou falsas.
Clickbait	Manchetes sensacionalistas e enganosas usadas para atrair cliques.

Fonte: Adaptado de (ALGHAMDI; LUO; LIN, 2024)

Segundo (ALGHAMDI; LUO; LIN, 2024), a disseminação das *fake news* também está fortemente ligada a fatores psicológicos que influenciam a forma como as pessoas consomem e compartilham informações. Diferentes teorias ajudam a entender por que certos conteúdos falsos se espalham rapidamente e encontram grande aceitação entre determinados grupos. Entre esses fatores, destacam-se o viés de confirmação, a exposição seletiva e o efeito da câmara de eco, que reforçam crenças pré-existentes e dificultam a correção da desinformação.

Segue abaixo a definição de cada fator conforme abordado por (ALGHAMDI; LUO; LIN, 2024), sendo cada conceito subsequentemente ilustrado por meio de exemplos hipotéticos elaborados pelo autor deste presente trabalho, para demonstrar a manifestação dos fatores no cotidiano:

- **Viés de Confirmação:** Refere-se à tendência individual de favorecer, interpretar e recordar informações que confirmam crenças e hipóteses preexistentes. No contexto da desinformação, esse viés se manifesta na maior aceitação de conteúdos alinhados às próprias visões de mundo, enquanto informações dissonantes são frequentemente descartadas ou submetidas a um escrutínio mais rigoroso. Tal fenômeno contribui significativamente para a proliferação de notícias falsas, sobretudo em domínios temáticos polarizados, onde as narrativas reforçam identidades de grupo (ALGHAMDI; LUO; LIN, 2024).

  Exemplo A: Um torcedor de futebol que acredita que seu time é sempre prejudicado pela arbitragem prestará muito mais atenção e lembrará vividamente dos lances em que o juiz errou contra sua equipe, mas esquecerá ou minimizará rapidamente os erros de arbitragem que a favoreceram.

  Exemplo B: Um entusiasta de uma marca de celular que lê dez análises sobre o novo modelo: ele se lembrará em detalhes das cinco análises positivas que elogiavam a câmera e o design, mas descartará as cinco negativas como sendo de "críticos que não entendem do assunto".

- **Exposição Seletiva:** Este conceito descreve o comportamento deliberado ou inconsciente de procurar e consumir informações que corroboram as próprias opiniões, ao mesmo tempo em que se evita ativamente o contato com perspectivas divergentes. A exposição seletiva limita a pluralidade de fontes e conteúdos aos quais um indivíduo tem acesso, criando um ambiente informacional homogêneo. Como consequência, a suscetibilidade à desinformação aumenta, uma vez que a ausência de contrapontos factuais enfraquece a capacidade crítica de avaliação da informação (ALGHAMDI; LUO; LIN, 2024).

  Exemplo C: Uma pessoa interessada em dietas vegetarianas que escolhe seguir apenas influenciadores e páginas que promovem o vegetarianismo, bloqueando ou

ignorando perfis de nutricionistas ou cientistas que possam apresentar dados sobre possíveis deficiências nutricionais ou outros contrapontos ao modelo alimentar.

Exemplo D: Um investidor otimista com o mercado de criptomoedas que escolhe seguir apenas influenciadores e canais de notícias que preveem altas históricas, silenciando ou bloqueando economistas que alertam sobre a volatilidade e os riscos do mercado.

- **Efeito da Câmara de Eco:** Amplificado por algoritmos de recomendação em plataformas de redes sociais, o efeito da câmara de eco descreve a formação de ambientes informacionais fechados. Nesses espaços, crenças são reforçadas pela repetição e validação mútua entre os membros de um grupo com visões de mundo semelhantes. A contínua circulação de narrativas, incluindo as falsas, dentro dessas "bolhas" informacionais, e a limitada exposição a pontos de vista contraditórios, podem levar à aceitação da desinformação como um fato consensual, aumentando a resistência à sua posterior correção e verificação (ALGHAMDI; LUO; LIN, 2024).

  Exemplo E: um grupo de WhatsApp onde todos os membros compartilham a mesma ideologia política. Uma notícia falsa é compartilhada e, em vez de ser verificada, é imediatamente validada e comentada positivamente por todos. A repetição daquela mensagem dentro do grupo a solidifica como verdade para seus membros, que não têm contato com a informação externa que a desmente.

  Exemplo F: Uma pessoa entra em um grupo de dieta que promove um "suplemento milagroso". O ambiente é preenchido apenas com depoimentos positivos e relatos de cura. Qualquer postagem cética ou que mencione efeitos colaterais é apagada pelos administradores. Com o tempo, o membro passa a acreditar que o suplemento é infalível, pois todas as informações em seu redor validam essa crença.

Segundo (MANZOOR; SINGLA; NIKITA, 2019) as *fake news* podem ser classificadas em diferentes categorias, de acordo com suas características e estratégias de disseminação. As notícias falsas baseadas em imagens utilizam fotos manipuladas ou editadas para criar uma falsa impressão de autenticidade, muitas vezes retirando-as de seu contexto original ou alterando elementos visuais para enganar o público. Já as *fake news* baseadas em usuários são disseminadas por contas falsas, geralmente criadas para atingir alvos específicos, amplificando a desinformação de maneira estratégica (MANZOOR; SINGLA; NIKITA, 2019). Outro tipo relevante são as *fake news* baseadas em conhecimento, que apresentam explicações científicas falsas ou distorcidas para convencer o público com uma aparência de credibilidade acadêmica. Também existem as *fake news* baseadas em estilo, que imitam a escrita e a abordagem de jornalistas renomados, dando uma falsa impressão de legitimidade ao conteúdo (MANZOOR; SINGLA; NIKITA, 2019).

4.2 MÉTODOS DE DETECÇÃO

Para (ALGHAMDI; LUO; LIN, 2024), os principais métodos de detecção podem ser categorizados em três abordagens principais: baseados no conteúdo, no contexto e híbridos.

Os métodos baseados no conteúdo empregam técnicas de análise linguística, estilometria e aprendizado de máquina para identificar padrões textuais característicos de *fake news*. A análise linguística investiga aspectos sintáticos, semânticos e pragmáticos, enquanto a estilometria detecta peculiaridades na escrita, como uso atípico de adjetivos ou estrutura argumentativa inconsistente. Modelos supervisionados, como SVM, Random Forest e Naive Bayes, são amplamente utilizados para classificação, alcançando alta precisão, embora tenham limitações ao lidar com desinformação sofisticada que imita a escrita jornalística legítima (ALGHAMDI; LUO; LIN, 2024).

Os métodos baseados no contexto analisam fatores externos, como padrões de disseminação, engajamento e credibilidade da fonte, para inferir a veracidade da informação. *Fake news* tendem a se espalhar mais rapidamente e gerar maior engajamento, frequentemente impulsionadas por *bots* e estratégias de amplificação artificial. Técnicas de *Deep Learning*, como redes neurais recorrentes e grafos de propagação, modelam essas dinâmicas para identificar padrões anômalos na circulação das notícias (ALGHAMDI; LUO; LIN, 2024).

Métodos híbridos integram características textuais e contextuais para aprimorar a detecção. Modelos baseados em redes neurais convolucionais (CNNs) (O'SHEA; NASH, 2015) e redes recorrentes (LSTMs) extraem representações latentes do texto, enquanto *Transformers*, como BERT (DEVLIN *et al.*, 2018), capturam nuances contextuais avançadas. Essa abordagem permite maior generalização, reduzindo a vulnerabilidade a técnicas de manipulação textual e estratégias de disseminação coordenada, tornando a detecção mais robusta em cenários reais (ALGHAMDI; LUO; LIN, 2024).

4.3 SOLUÇÕES BASEADAS EM APRENDIZADO DE MÁQUINA DETERMINÍSTICO

Para a detecção de *fake news* utilizando abordagens determinísticas de aprendizado de máquina, alguns autores propõem estratégias baseadas na extração e seleção de características textuais que condizem com o corpo da notícia em que é apresentada, combinadas com algoritmos supervisionados para a classificação de notícias falsas. A seguir, será apresentada uma abordagem que emprega uma solução em modelagem preditiva, estruturando um fluxo que abrange desde o pré-processamento textual até a otimização da seleção de características, que de acordo com (ALGHAMDI; LUO; LIN, 2024), destaca métricas linguísticas e estilísticas relevantes. Na sequência, será explorado todo o fluxo metodológico de modelagem e avaliação do método. A segunda parte desta seção avança para aborda-

gens baseadas em aprendizado profundo (*Deep Learning*), onde arquiteturas avançadas de redes neurais são aplicadas para capturar padrões mais complexos da linguagem e da disseminação da desinformação, habilitando o entendimento de contexto para o modelo, e o uso de *embeddings* como base de informação para realizar a classificação/previsão do conteúdo da notícia.

O estudo conduzido por (FAYAZ *et al.*, 2022) apresenta uma abordagem sistemática para a detecção automatizada de *fake news* por meio de técnicas de aprendizado de máquina, com ênfase na utilização do algoritmo *Random Forest* (RF) (BREIMAN, 2001) para a classificação de notícias falsas. Diante do crescimento exponencial do volume de informações disseminadas online, torna-se imprescindível o desenvolvimento de métodos computacionais capazes de distinguir conteúdos verídicos de fabricados. Para esse fim, o autor propõe um fluxo de processamento estruturado, iniciando-se com a preparação dos dados provenientes do *ISOT Fake News Dataset* (AHMED; TRAORE; SAAD, 2018). O pré-processamento do texto engloba a segmentação de sentenças, tokenização, remoção de *stopwords* e aplicação de stemização, objetivando a padronização linguística e a redução de redundâncias. Posteriormente, são extraídas 23 características textuais que abrangem aspectos estruturais e semânticos, incluindo métricas de legibilidade, análise de sentimentos e representações vetoriais do texto, como TF-IDF (*term frequency–inverse document frequency*) (SAMMUT CLAUDE; WEBB, 2010) e a técnica *Bag of Words* para contagem de palavras e estruturação da visão léxica do conjunto de dados, permitindo uma caracterização detalhada dos padrões linguísticos associados às notícias falsas.

A arquitetura metodológica para a classificação de textos, esquematizada na Figura 24, segue um fluxo supervisionado que se inicia com a preparação dos dados e culmina na avaliação do modelo, conforme descrito pelos trabalhos de (FAYAZ *et al.*, 2022) e (ALGHAMDI; LUO; LIN, 2024).

O ponto de partida é a Aquisição e Estruturação dos Dados, ilustrado no primeiro quadrante do fluxograma, partindo da parte superior da Figura 24 e seguindo o fluxo sequencialmente até a parte inferior. Nesta fase, um conjunto de dados brutos é estruturado e rotulado. Em exemplo hipotético a seguir, uma entrada (*input*) para o conjunto seria um par de texto e rótulo:

- **Texto:** “Cientistas descobrem que o consumo diário de chocolate amargo reverte o envelhecimento celular.”
- **Rótulo:** 1 (indicando “fake news”)

Após a aquisição, o conjunto de dados é submetido à Divisão em Subconjuntos, sendo tipicamente particionado em 80% para treino e 20% para teste, garantindo uma

avaliação imparcial da capacidade de generalização do modelo (AHMED; TRAORE; SAAD, 2018).

Em seguida, o texto passa pela etapa de Pré-Processamento, que inclui limpeza, normalização (conversão para minúsculas) e redução de palavras à sua forma raiz ou lema (lematização/estematização) (ALGHAMDI; LUO; LIN, 2024). O texto limpo é então transformado em valores numéricos na etapa de *Feature Engineering* (criação de recursos para o modelo de aprendizado de máquina), utilizando técnicas como *Bag of Words* (BoW) ou TF-IDF (SAMMUT CLAUDE; WEBB, 2010). Para otimizar o modelo, a Seleção de características é aplicada para reter apenas os atributos mais informativos e que causem mais impacto para a predição do rótulo (FAYAZ *et al.*, 2022).

Segue em exemplo, características criadas durante o processo de *Feature Engineering*, de acordo com o autor (ALGHAMDI; LUO; LIN, 2024):

- Quantidade de Palavras, Frases e verbos;
- Média de Frases e Sentenças;
- Quantidade de Caracteres do Texto;
- Quantidade de Palavras Grandes;
- Quantidade de Sílabas;
- Indice de Emotividade;
- Taxa de Adjetivos e Advérbios;
- Ativação de processo emotivo;
- Quantidade Verbo no Passado;
- Quantidade Verbo no Presente;
- Quantidade Verbo no Futuro;

A etapa de Treinamento, conforme ilustrado na Figura 24 consiste em alimentar o modelo de aprendizado de máquina com as características criadas em cima dos dados, e seus respectivos rótulos. A metodologia prevê a experimentação com diferentes classificadores, como *Random Forest* (BREIMAN, 2001), SVM (CRISTIANINI; RICCI, 2008), Regressão Logística (KLEINBAUM; KLEIN, 2010) e XGBoost (CHEN TIANQI; GUESTRIN, 2016), para identificar o de melhor desempenho (ALGHAMDI; LUO; LIN, 2024).

Uma vez treinado, o modelo é utilizado para Classificação / Predição no conjunto de teste. Aplicado ao exemplo anterior, o sistema produziria uma saída, conforme o exemplo a seguir:

- **Predição do Modelo:** Positivo
- **Predição do Modelo em Formato Numérico:** 1
- **Interpretação:** O modelo classifica a notícia como “fake news”.

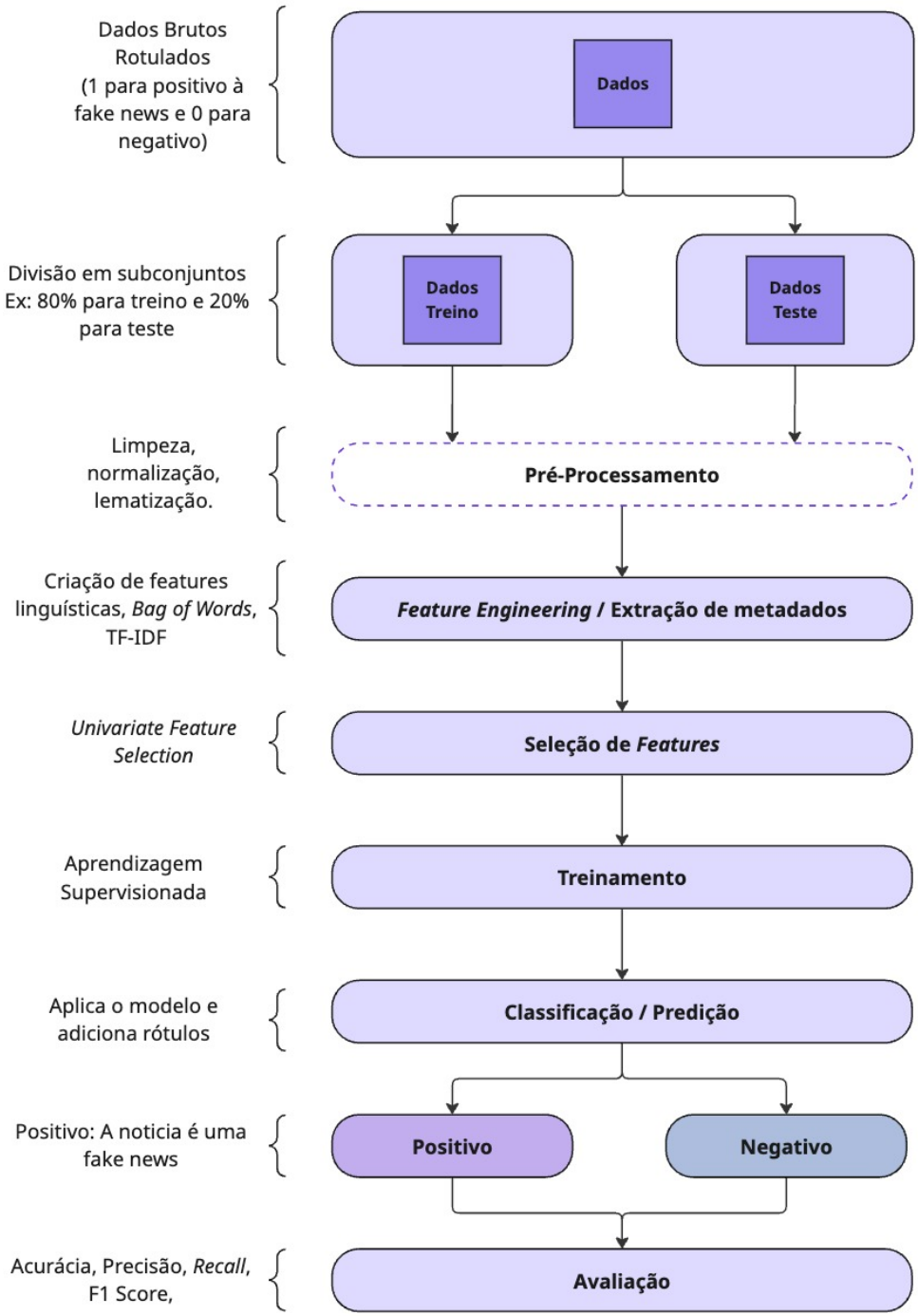
Finalmente, na Figura 24, a performance do modelo é aferida na etapa de Avaliação, utilizando um conjunto de métricas padrão como Acurácia, Precisão, *Recall* e F1-Score. A robustez da avaliação é reforçada pelo uso de técnicas como a Validação Cruzada (*Cross-Validation*), que oferece uma estimativa mais estável do desempenho do modelo em dados não vistos do subconjunto (ALGHAMDI; LUO; LIN, 2024).

Entretanto, apesar dos resultados promissores, (FAYAZ *et al.*, 2022) reconhece desafios inerentes à detecção de *fake news* por meio de aprendizado de máquina determinística. A complexidade textual, aliada à evolução contínua das estratégias de disseminação de desinformação, impõe dificuldades na generalização dos modelos preditivos. Além disso, algumas notícias apresentam informações enganosas de forma sutil, dificultando sua distinção apenas por meio de características textuais explícitas. Como desdobramento futuro, o autor sugere a integração de arquiteturas baseadas em aprendizado profundo, capazes de capturar representações mais abstratas dos textos, além da incorporação de técnicas de *ensemble learning* para potencializar a robustez da classificação (FAYAZ *et al.*, 2022). Dessa forma, o estudo contribui significativamente para a área, ao demonstrar que a combinação de *Random Forest* e seleção ótima de características constitui uma abordagem eficiente para a detecção automatizada de *fake news*.

O autor (ALGHAMDI; LUO; LIN, 2024) também propõe uma abordagem determinística de aprendizado de máquina utilizando os modelos XGBoost e *Random Forest*. No entanto, ele também ressalta algumas limitações inerentes aos métodos tradicionais, como a incapacidade de capturar o contexto profundo do texto, a dependência de engenharia manual de características e a ineficácia na detecção de *fake news* multimodais, que envolvem tanto texto quanto imagem. Essas restrições podem comprometer a precisão e a generalização dos modelos, especialmente em cenários onde a desinformação se manifesta de maneira mais complexa e diversificada.

Por sua vez, o estudo (ALGHAMDI; LUO; LIN, 2024) avança na implementação de aprendizado de máquina ao explorar arquiteturas de *deep learning* para a detecção de *fake news*, destacando o uso de Redes Neurais Convolucionais (CNNs) (O’SHEA; NASH, 2015), Redes Neurais Recorrentes (RNNs) (SCHMIDT, 2019) com LSTMs e Redes Neurais Baseadas em Grafos (GNNs) (YANG *et al.*, 2023) . Essas abordagens superam os métodos tradicionais ao capturar padrões complexos e o contexto semântico do texto, reduzindo a necessidade de engenharia manual de características e aumentando a precisão na classificação.

Figura 24 – Fluxo de Aprendizado de Máquina Supervisionado Para Classificação de *Fake News*



Fonte: Adaptado de (ALGHAMDI; LUO; LIN, 2024)

As CNNs, originalmente projetadas para processamento de imagens, foram adaptadas para análise textual, permitindo a extração de padrões linguísticos e estilísticos. Sua estrutura inclui camadas convolucionais para captura de n-grams, *max-pooling* para redução de dimensionalidade e camadas densas para a classificação da notícia como verdadeira ou falsa. Aplicadas ao *dataset* ISOT (AHMED; TRAORE; SAAD, 2018), essas

redes alcançaram uma precisão de 0,99 (ALGHAMDI; LUO; LIN, 2024), demonstrando alta eficácia na detecção de *fake news*. Já as LSTMs, projetadas para lidar com dependências de longo prazo no texto, utilizam uma camada *embedding* para converter palavras em vetores e camadas recorrentes para modelar o contexto sequencial. Esse modelo atingiu 0,98 de precisão, sendo especialmente eficaz na análise de textos longos e contextualmente complexos.

Além do conteúdo textual, a propagação de *fake news* também pode ser analisada estruturalmente por meio de GNNs. Essas redes modelam a disseminação de desinformação em redes sociais, utilizando GCNs para identificar padrões de interação entre usuários e detectar perfis propagadores. Essa abordagem demonstrou melhor desempenho em relação a métodos baseados exclusivamente no conteúdo textual (ALGHAMDI; LUO; LIN, 2024), evidenciando a relevância da análise da estrutura de disseminação para aprimorar a detecção de *fake news* em ambientes digitais.

4.4 ESTRUTURA DA ARQUITETURA DE REDE NEURAL PARA CLASSIFICAÇÃO TEXTUAL

A metodologia para a classificação de textos, como a detecção de notícias falsas, frequentemente emprega arquiteturas de aprendizagem profunda (*Deep Learning*). A Figura 25 ilustra um fluxo de processamento baseado em Redes Neurais, composto por seis estágios sequenciais que transformam o texto bruto em uma classificação final. Cada etapa é projetada para extrair e refinar características textuais, culminando em uma decisão categórica. A seguir, detalha-se cada uma dessas etapas.

1. Entrada Bruta do Texto

O fluxo de processamento, conforme delineado na Figura 25, inicia-se com a aquisição da entrada textual bruta. Esta etapa representa o dado original, tal como se apresenta em sua fonte, seja um artigo de notícia, uma postagem em mídia social ou outro documento textual, antes de qualquer intervenção ou processamento. Este dado serve como o objeto fundamental da análise, sendo a matéria-prima para todas as operações subsequentes da rede (ALNABHAN; BRANCO, 2024).

- **Tipo de Dados de Entrada:** String (texto).
- **Exemplo de Entrada:** *“A Terra é plana e o governo esconde isso.“*

2. Pré-processamento e Camada de *Embedding*

Após a recepção do texto bruto, a segunda etapa, ilustrada na Figura 25, é responsável por preparar e converter os dados textuais em um formato numérico que

possa ser processado pela rede neural. Este estágio é composto por diversas etapas, envolvendo inicialmente um pré-processamento que consiste na limpeza do texto para remover elementos não informativos (caracteres especiais, links), seguido pela tokenização, que segmenta o texto em unidades menores como palavras ou sub-palavras (ALNABHAN; BRANCO, 2024). Subsequentemente, cada *token* é mapeado para um identificador numérico (ID) único de um vocabulário, as sequências de IDs são então padronizadas para um comprimento fixo, como 1000 *tokens*, por meio de preenchimento (*padding*) (ALNABHAN; BRANCO, 2024). A fase final desta etapa é a camada de *embedding*, que converte cada ID de token em um vetor denso com elementos numéricos. Tais vetores, ou *embeddings*, capturam relações semânticas e sintáticas entre as palavras. Métodos como Word2Vec (MIKOLOV *et al.*, 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014), ou modelos mais complexos baseados em *Transformers* (VASWANI *et al.*, 2017).

- **Entrada (Após pré-processamento):** Uma sequência de IDs numéricos de comprimento fixo de 1000.
  - **Exemplo (Vetor de Inteiros):** $[8, 234, 5, 64, 15, 396, 587, 700, ..., 0]$.
  - **Formato/Dimensão:** Vetor com comprimento máximo da sequência. (1000).
- **Saída (Após a camada de Embedding):** Uma matriz onde cada linha corresponde ao vetor de *embedding* de um *token*.
  - **Exemplo (Matriz de Dimensão (1000, 100):**

$$\begin{bmatrix} 0,321 & 0,885 & \dots & 0,074 \\ 0,456 & -0,93 & \dots & 0,437 \\ \vdots & \vdots & \ddots & \vdots \\ 0,201 & -0,077 & \dots & 0,052 \end{bmatrix}$$

  - **Formato/Dimensão:** Matriz (Comprimento Máximo da Sequência, Tamanho do Embedding). (1000, 100).

3. Camada Recorrente

A sequência de vetores de *embedding* gerada na etapa anterior é então processada pela terceira camada do modelo, conforme detalhado na Figura 25. Esta camada é tipicamente uma Rede Neural Recorrente (RNN), comumente nas suas variantes mais avançadas como a *Long Short-Term Memory* (LSTM). Tal arquitetura é especializada em modelar dados sequenciais, sendo capaz de capturar dependências de longo prazo no texto ao manter um estado oculto que propaga informação contextual ao longo da sequência dos vetores numéricos (ALNABHAN; BRANCO, 2024).

- **Entrada:** A matriz de *embeddings* da etapa anterior. Formato: (1000, 100).

- **Saída:** Um vetor numérico que representa as características contextuais extraídas da sequência.
  - **Exemplo (Saída de uma LSTM com 100 unidades por direção):** Vetor numérico.
  - **Formato/Dimensão:** Vetor (2 $\times$ Número de Unidades LSTM), (200).

4. Camada(s) Densa(s) / Totalmente Conectada

O vetor de características extraído pela camada recorrente alimenta a quarta etapa do processo, que consiste em uma ou mais camadas densas (ou totalmente conectadas), como ilustrado na Figura 25. Nestas camadas, cada neurônio está conectado a todos os neurônios da camada precedente. A função dessas camadas é aprender combinações não-lineares e complexas das características de alto nível (ALNABHAN; BRANCO, 2024), refinando-as para a tarefa de classificação final. Funções de ativação como a Unidade Linear Retificada (ReLU) são frequentemente aplicadas para introduzir não-linearidade (ALNABHAN; BRANCO, 2024). A arquitetura pode empregar múltiplas camadas densas em cascata para refinar progressivamente a representação antes da classificação.

- **Entrada:** O vetor de características da camada recorrente. Formato: (200,).
- **Saída:** Um vetor de características refinado, geralmente com dimensão reduzida.
  - **Exemplo (Após múltiplas camadas densas):** $[0,43, 0,55, -0,21, 1,59, ..., 0,86]$.
  - **Formato/Dimensão:** Vetor (Número de Neurônios na Última Camada Densa), (64,).

5. Camada de Saída

A quinta etapa, mostrada na Figura 25, é a camada final da rede neural. Sua função é receber o vetor de características da última camada densa e produzir a saída final da predição (ALNABHAN; BRANCO, 2024), para isso, utiliza uma função de ativação apropriada para mapear as características aprendidas em uma distribuição de probabilidade sobre as classes alvo. Para uma tarefa de classificação binária (Falso/Verdadeiro), a função *Sigmoid* é comumente utilizada. Para classificação multiclasse a função *Softmax* é empregada para gerar um vetor de probabilidades onde a soma de todos os elementos é igual a 1 (ALNABHAN; BRANCO, 2024).

- **Entrada:** O vetor da última camada densa. Formato: (64,).
- **Saída:** Um vetor de probabilidades sobre as classes de destino.
  - **Exemplo (Classificação Binária):** $[0,90, 0,10]$.

  - **Exemplo (Classificação Multiclasse):** $[0,01, 0,85, 0,03, 0,07, 0,02, 0,02]$.

6. Saída Final / Decisão

A etapa culminante do fluxo, conforme representado na Figura 25, não é uma camada da rede em si, mas a interpretação do seu resultado. As probabilidades geradas pela camada de saída são utilizadas para tomar uma decisão final sobre a classificação do texto (ALNABHAN; BRANCO, 2024). Tipicamente, a classe associada à maior probabilidade no vetor de saída é selecionada como o rótulo predito para o texto de entrada. Este rótulo categórico representa a conclusão do processo de inferência do modelo (ALNABHAN; BRANCO, 2024).

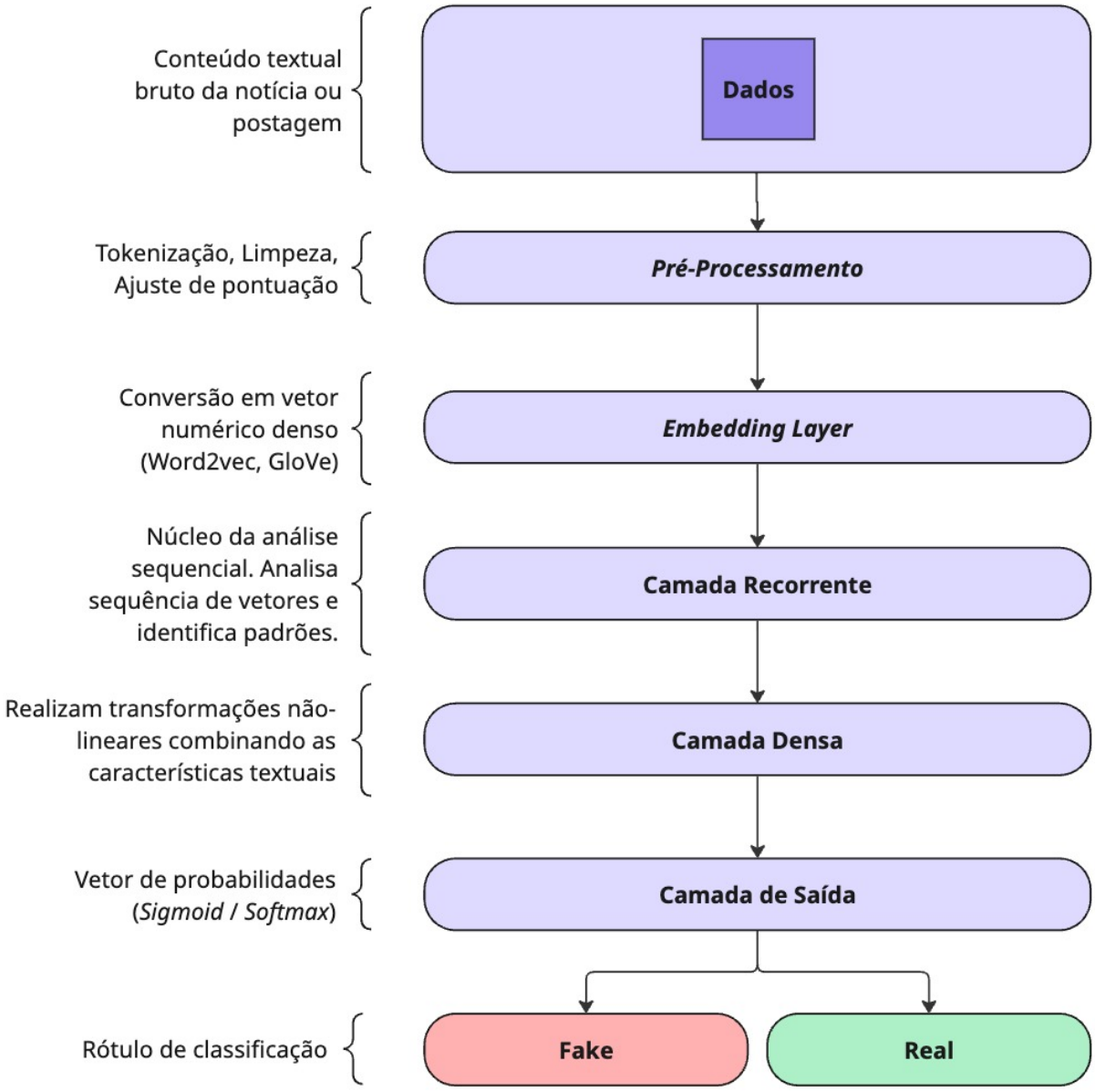
- **Entrada:** O vetor de probabilidades da camada de saída.
- **Saída:** O rótulo categórico predito.
  - **Exemplo (Binário):** "Notícia Falsa"(baseado na probabilidade de 0.90).
  - **Exemplo (Multiclasse - Rótulos):** "Falsa", "Neutra", "Verdadeira"(que pode ser baseada com base na probabilidade).

Um ponto de destaque de (ALGHAMDI; LUO; LIN, 2024) com relação às estruturas de redes neurais, é que a análise linguística desempenha um papel fundamental na detecção de *fake news*, fornecendo insumos relevantes para o treinamento dos modelos. O estudo (ALGHAMDI; LUO; LIN, 2024) indica que notícias falsas apresentam menor complexidade textual do ponto de vista morfológico, caracterizada por frases mais curtas e diretas, além de um vocabulário menos sofisticado, com menor quantidade de palavras difíceis. Técnicas de estilometria revelam padrões distintos, como o uso excessivo de adjetivos e advérbios para reforçar apelos emocionais, maior frequência de pronomes pessoais, como "nós", "eles"e "você", para criar proximidade com o leitor, e a presença exagerada de pontuação enfática, incluindo múltiplos pontos de exclamação ("!!!"), interrogações sucessivas ("???") e o uso excessivo de letras maiúsculas para enfatizar informações. Além disso, (ALGHAMDI; LUO; LIN, 2024) destaca que as *fake news* tendem a ser altamente polarizadas, utilizando linguagem emocional intensa para reforçar discursos positivos ou negativos, enquanto notícias verdadeiras apresentam maior neutralidade. Redes neurais exploram essas características ao integrar análises de complexidade textual, estilometria e polarização, aprimorando a capacidade de distinguir desinformação de conteúdos legítimos (ALGHAMDI; LUO; LIN, 2024).

4.4.1 Uso de LLMs para Análise e Classificação de *Fake News*

A utilização de LLMs para detecção de *fake news* é elencada por (ALGHAMDI; LUO; LIN, 2024), que explora o papel dos *Transformers* na análise contextual das notícias

Figura 25 – Fluxo de *Deep Learning* para classificação de *fake news*



Fonte: Adaptado de (ALNABHAN; BRANCO, 2024)

e fatos apresentados. O autor destaca que, ao contrário de abordagens tradicionais, as arquiteturas baseadas em *Transformers* oferecem uma compreensão aprofundada das relações semânticas, permitindo a distinção de padrões de linguagem associados à desinformação. Esse aprimoramento se dá pela capacidade de modelar dependências complexas e dinâmicas entre palavras em diferentes partes do texto, ajustando-se de forma precisa às variações linguísticas presentes nas *fake news*.

Os *Transformers* se destacam pela grande capacidade de captar nuances contextuais, devido ao mecanismo de atenção (*self-attention*), que possibilita a avaliação de palavras em relação a seu contexto mais amplo. Essa característica permite que o modelo identifique inconsistências, exageros e estruturas argumentativas frequentemente asso-

ciadas a desinformação, tornando-os altamente eficazes na detecção de notícias falsas, conforme evidenciado pela pesquisa de (ALGHAMDI; LUO; LIN, 2024).

Na abordagem utilizada por (OAD *et al.*, 2024), o modelo BERT (DEVLIN *et al.*, 2018) é empregado como um classificador para detectar *fake news*, explorando toda a sua arquitetura baseada em *transformers* para capturar as relações contextuais presentes nos textos. Diferente de abordagens tradicionais que dependem apenas de palavras-chave ou análise superficial do conteúdo, o estudo propõe um ajuste fino do BERT, permitindo que ele aprenda padrões mais complexos associados à desinformação. Dessa forma, o modelo se torna mais eficiente ao distinguir nuances sutis entre notícias falsas e verdadeiras, reduzindo a taxa de erros na classificação.

Além da otimização do modelo, (OAD *et al.*, 2024) enfatiza a importância da curadoria de dados e do pré-processamento avançado. O uso de conjuntos de dados específicos e balanceados melhora a generalização do modelo, enquanto técnicas como remoção de *stopwords*, lematização e tokenização garantem uma entrada mais limpa e representativa. A arquitetura também passa por ajustes estratégicos, incluindo a modificação de hiperparâmetros e a adição de camadas para refinar a capacidade preditiva do modelo. O desempenho é avaliado rigorosamente por meio de métricas como precisão, recall e F1-score, garantindo que a abordagem proposta seja robusta e eficaz no combate à disseminação de *fake news*.

Conforme definido por (OAD *et al.*, 2024), o problema da classificação de *fake news* pode ser abordado como um desafio de aprendizado supervisionado, onde o modelo precisa identificar padrões linguísticos sutis em textos rotulados como verdadeiros ou falsos. Para lidar com a ambiguidade textual, o BERT é treinado para compreender contextos mais profundos, diferenciando afirmações enganosas de informações legítimas. Além disso, a variação no estilo de escrita das *fake news* é considerada, permitindo que o modelo generalize melhor e atue em diferentes domínios, sem ficar restrito ao *dataset* de treinamento, conforme mencionado anteriormente na Seção 4.3 e Figura 24.

No contexto da aplicação de Modelos de Linguagem Larga Escala (LLMs) e *Transformers*, o grande diferencial reside no mecanismo de Atenção (*Attention*) (VASWANI *et al.*, 2017), conforme abordado no Capítulo 2. É este mecanismo que habilita o modelo a gerar *embeddings*, vetores de características, profundamente contextualizados. Tais representações são fundamentais para a tomada de decisão na etapa final de classificação (ALGHAMDI; LUO; LIN, 2024).

Para a extração dessas características contextuais, o fluxo de entrada de dados é análogo ao processo de *Deep Learning* ilustrado na Seção 4.3. O elemento distintivo, contudo, é a presença de módulos análogos a *auto-encoders* na camada de *embedding*, que se tornam responsáveis pela estruturação semântica do conteúdo de entrada.

Com a entrada de dados formatada, o modelo inicia o processamento para gerar as representações vetoriais que capturam o contexto. Conforme detalhado no Capítulo 2, o coração deste processo é a arquitetura *Transformer Encoder* e seu mecanismo de auto-atenção (*self-attention*) (VASWANI *et al.*, 2017), que podem ser refinados durante o processo de *fine-tuning* para a tarefa de detecção de *fake news* (ALGHAMDI; LUO; LIN, 2024).

O mecanismo de auto-atenção permite que o modelo analise cada *token* em relação a todos os outros *tokens* na sequência (VASWANI *et al.*, 2017). Ao fazer isso, ele constrói uma representação para cada *token* que é referente de seu contexto específico. Por exemplo, na frase “a roda do carro girou“, a auto-atenção permite ao modelo inferir que “roda“ se refere a um objeto do automóvel, e não ao verbo “rodar“, ao ponderar a forte influência contextual da palavra “carro“.

Segue um exemplo hipotético a seguir, criado pelo autor do presente texto, para exemplificar o contexto de detecção de *fake news*, considerando a palavra “eleição“:

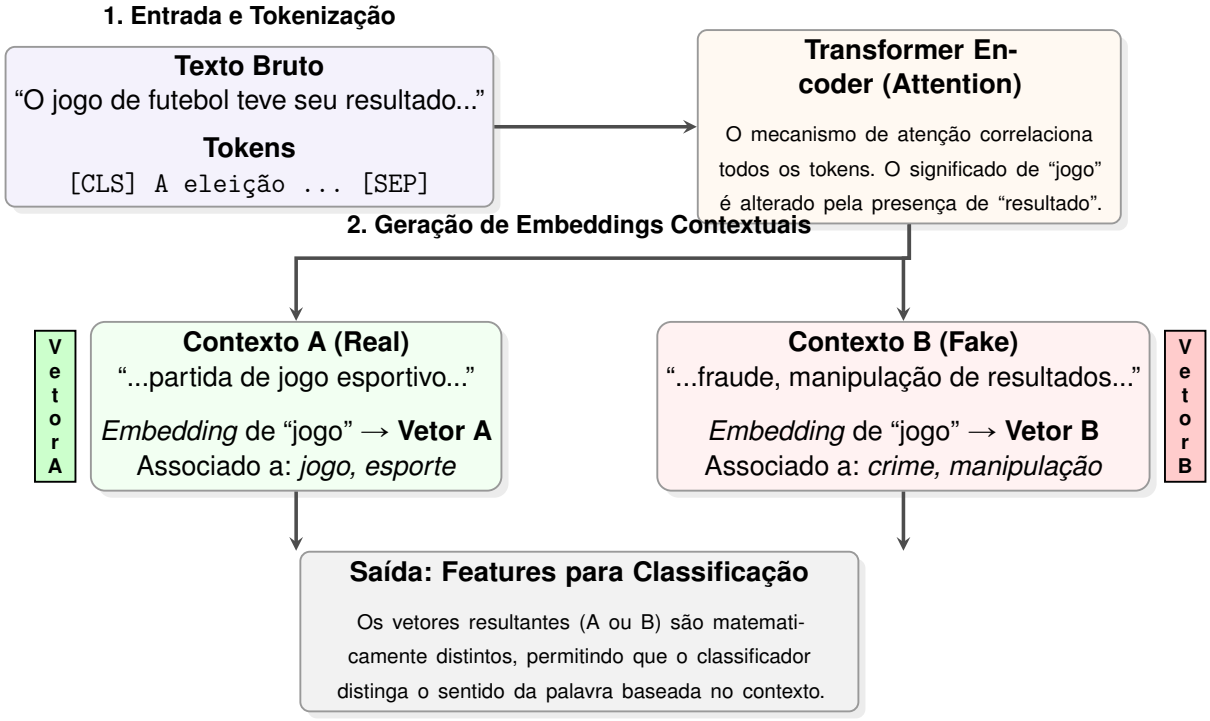
Em uma notícia jornalística padrão, “eleição“ provavelmente estará associada a termos como “processo democrático“, “candidatos“, “votação“, “apuração“ e “Tribunal Superior Eleitoral“. Em uma narrativa de desinformação, a mesma palavra “eleição“ pode estar semanticamente conectada a um léxico de suspeita, como “fraude“, “urnas adulteradas“, “votos roubados“, “conspiração“ e “interferência externa“.

O mecanismo de auto-atenção, aprimorado pelo *multi-head attention* (que permite focar em diferentes tipos de relações contextuais simultaneamente), aprende a gerar *embeddings* distintos para a palavra “eleição“ em cada um desses contextos. O *fine-tuning* otimiza esse processo para que o modelo se torne sensível especificamente aos padrões linguísticos indicativos de desinformação (LI; ZHANG; MALTHOUSE, 2024).

A Figura 26 detalha o fluxo para a geração de características contextuais por um modelo *Transformer* (VASWANI *et al.*, 2017), desde a entrada do texto bruto até a extração dos vetores finais. O processo inicia-se com a preparação dos dados de entrada (Entrada e Preparação), onde o texto bruto é convertido em uma sequência de *tokens* numéricos. Em seguida, esta sequência é processada pelo núcleo do modelo, o bloco (*Transformer Encoder*), que emprega o mecanismo de auto-atenção para analisar as interdependências semânticas entre todos os *tokens* simultaneamente, permitindo que a representação de um termo seja dinamicamente influenciada pelo seu contexto. O resultado deste processo, ilustrado no bloco (Resultado da Contextualização), é a geração de representações vetoriais (*embeddings*) distintas para um mesmo *token* alvo, como “jogo de futebol“, cujo vetor final (*Vetor A*) em um contexto de legitimidade (“partida esportíva“) difere substancialmente daquele gerado (*Vetor B*) em um contexto de desinformação (“fraude“, “azar“, “manipulação”). Finalmente, estes vetores contextuais são extraídos como as características de alta dimensão (Saída) que alimentarão a camada de classificação final do modelo para a

predição da notícia.

Figura 26 – Fluxo de Geração de Características Contextuais Baseadas em *Transformers*.



Fonte: Do Autor

Por fim, as implicações práticas do estudo são amplas, podendo ser aplicadas a sistemas de checagem de fatos e monitoramento de redes sociais, (OAD *et al.*, 2024) destaca que, embora os resultados sejam promissores, há espaço para aprimoramentos futuros, como a incorporação de aprendizado por reforço para adaptar o modelo a novos padrões de desinformação. Dessa forma, o estudo não apenas comprova a eficácia do BERT na tarefa, mas também abre caminho para novas pesquisas na detecção automatizada de *fake news*.

A avaliação da qualidade de um modelo de classificação envolve diversas métricas, sendo *Precision* e *Recall* duas das mais importantes. Essas métricas são utilizadas para medir a acurácia das previsões, especialmente em problemas onde há um desequilíbrio entre classes. Enquanto *Precision* indica a proporção de previsões corretas entre todas as instâncias classificadas como positivas, *Recall* mede a capacidade do modelo de identificar todas as instâncias relevantes dentro da classe positiva.

Nesta formulação, os termos correspondem elementos da matriz de confusão e responsáveis por criar as métricas mencionadas:

- **TP (*True Positives* / Verdadeiros Positivos):** Representa o número de instâncias da classe positiva que foram corretamente identificadas pelo modelo.

- **TN (*True Negatives* / Verdadeiros Negativos):** Indica a quantidade de instâncias da classe negativa que foram corretamente classificadas como tal.
- **FP (*False Positives* / Falsos Positivos):** Refere-se às instâncias negativas incorretamente classificadas como positivas (o “alarme falso”, ou Erro do Tipo I).
- **FN (*False Negatives* / Falsos Negativos):** Contabiliza as instâncias positivas que o modelo falhou em identificar, classificando-as incorretamente como negativas (Erro do Tipo II).

A métrica *Precision* (precisão) (TING, 2010) é representada pela Equação 4.1 e é definida como a razão entre os verdadeiros positivos (*TP*) e a soma dos verdadeiros positivos com os falsos positivos (*FP*). Essa métrica é útil para avaliar a exatidão das previsões positivas feitas pelo modelo, garantindo que ele minimize a quantidade de falsos alarmes. Esta métrica pode ser descrita pela seguinte expressão:

$$\textit{Precision} = \frac{TP}{TP + FP}. \tag{4.1}$$

A métrica *Recall* (TING, 2010), é representada pela Equação 4.2, mede a proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias que realmente pertencem à classe positiva. Essa métrica é essencial em cenários onde a identificação de todos os casos positivos é mais importante do que a minimização de falsos positivos. Esta métrica pode ser descrita pela seguinte expressão:

$$\textit{Recall} = \frac{TP}{TP + FN}. \tag{4.2}$$

O equilíbrio entre *Precision* e *Recall* (TING, 2010) é frequentemente avaliado por meio do *F1-score*, que combina ambas as métricas em uma única medida de desempenho.

A métrica *Acurácia* (Accuracy) é uma das medidas mais fundamentais para a avaliação de desempenho de classificadores, representando a proporção global de acertos do modelo em relação ao total de amostras processadas. Formalmente, ela é expressa pela Equação 4.3 (TING, 2010):

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.3}$$

Embora a acurácia ofereça uma visão panorâmica da eficácia do classificador, é importante notar que sua utilidade é maximizada em cenários onde as classes são balanceadas, podendo apresentar viés de interpretação em *datasets* com distribuição desigual de classes.

O *F1-Score* (TING, 2010), representada pela Equação 4.4, é definido como a média harmônica entre *Precision* e *Recall*. Esta métrica combina as duas avaliações anteriores em um único valor, fornecendo um balanço entre a exatidão das classificações positivas e a cobertura de todas as instâncias positivas. O *F1-Score* é especialmente relevante em contextos com desbalanceamento de classes, nos quais a Acurácia pode não ser uma representação fidedigna do desempenho, visto que o *F1-Score* penaliza modelos que sacrificam uma métrica em favor da outra. Esta métrica pode ser descrita pela seguinte expressão:

$$\textit{F1-Score} = 2 \cdot \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}} = \frac{2TP}{2TP + FP + FN}. \tag{4.4}$$

4.4.2 Soluções baseada em agentes de LLMs

Diante dos desafios encontrados durante o levantamento das abordagens da detecção de *fake news*, detectar uma notícia pode ser uma tarefa complexa, que exige soluções multimodais, multi-comportamentais e multifatoriais (ALGHAMDI; LUO; LIN, 2024). No estudo conduzido por (LI; ZHANG; MALTHOUSE, 2024), é proposta a utilização de LLMs trazendo uma abordagem de agente para checagem de fatos e detecção de *fake news*. Também traz à tona o LLM composto por diversas ferramentas de ação, onde é capaz de identificar o contexto do conteúdo apresentado e também como integrar conhecimentos de domínios específicos externos para compor a solução.

Os Agentes, no contexto da Inteligência Artificial (IA), são sistemas autônomos capazes de perceber seu ambiente, processar informações e tomar decisões para alcançar objetivos específicos, emulando, em certa medida, a capacidade humana de raciocínio e ação. Esses sistemas são frequentemente modelados como entidades que integram percepção, planejamento e execução, podendo operar de forma isolada ou colaborativa, dependendo da complexidade da tarefa. Segundo (RUSSELL; NORVIG, 2021), um agente inteligente é definido como um sistema que mapeia sequências de percepções em ações, utilizando representações de conhecimento e mecanismos de inferência para guiar seu comportamento. No cenário de verificação de notícias, os Agentes destacam-se por sua capacidade de replicar o fluxo cognitivo de especialistas humanos, decompondo o processo em subetapas lógicas e aplicando uma combinação estratégica de conhecimento interno (armazenado em sua base de dados ou modelo de linguagem) e ferramentas externas (como bancos de dados factuais, motores de busca ou APIs de checagem).

Essa arquitetura confere vantagens significativas em relação aos modelos preditivos baseados em árvore e supervisionados, como por exemplo Random Forest (FAYAZ *et al.*, 2022), uma vez que elimina a dependência de conjuntos de dados anotados para treinamento, os quais são frequentemente custosos e limitados em escopo. Em vez disso, os Agentes operam por meio de regras e heurísticas flexíveis, adaptando-se dinamicamente

a diferentes domínios de notícias mediante a simples substituição ou ajuste das ferramentas externas empregadas.

Além disso, sua modularidade permite a incorporação contínua de novas fontes de verificação, garantindo maior robustez e atualização frente a desinformação em evolução. Como demonstrado por (LI; ZHANG; MALTHOUSE, 2024), essa abordagem tem o potencial de aumentar a eficiência do processo de checagem de fatos e também ampliar sua aplicabilidade transversal, abrangendo desde notícias políticas até científicas, sem a necessidade de retreinamento extensivo.

Diante de tudo, o Agente de Checagem de fatos desenvolvido por (LI; ZHANG; MALTHOUSE, 2024) demonstra capacidade de identificar possíveis *fake news* ainda em estágios iniciais de disseminação, mesmo na ausência de informações contextuais provenientes de redes sociais. Isso é possível graças à sua arquitetura flexível, que permite a integração de múltiplas fontes de verificação e a aplicação de raciocínio explícito em cada etapa do processo. Essa transparência não apenas aumenta a interpretabilidade dos resultados, mas também facilita a compreensão do usuário final, que pode acompanhar o fluxo lógico utilizado para determinar a autenticidade da notícia (LI; ZHANG; MALTHOUSE, 2024).

A adaptabilidade do Agente o torna uma solução escalável para diferentes cenários, desde notícias políticas até desinformação científica. A possibilidade de customizar as ferramentas de verificação de acordo com o domínio abordado permite que o sistema evolua continuamente, incorporando novas bases de conhecimento e métodos de checagem sem a necessidade de retreinamento do modelo. Essa característica é particularmente relevante em um cenário de rápida propagação de desinformação, onde a agilidade e a precisão na detecção de *fake news* são essenciais para mitigar seus impactos sociais (LI; ZHANG; MALTHOUSE, 2024).

Para elucidar o funcionamento prático de um fluxo de detecção e explicação de *fake news* baseado nos conceitos levantados por (LI; ZHANG; MALTHOUSE, 2024), apresenta-se a seguir uma ilustração abstrata, concebida pelo autor do presente texto para fins puramente didáticos, que percorre o fluxo de checagem de fatos detalhado na Figura 27. Este exemplo demonstra o comportamento do agente, desde o recebimento da informação até a geração do relatório final, incluindo a interação com ferramentas externas. O ponto de partida é uma notícia hipotética, cuja verificação se torna o objetivo do agente.

A fase inicial do processo consiste no recebimento e na análise primária da informação. O agente processa o texto de entrada para identificar e isolar as alegações factuais que demandam verificação, decompondo a notícia em unidades lógicas investigáveis sem realizar qualquer juízo de valor prévio.

- **Entrada:** Um trecho de texto contendo a notícia a ser verificada: *"Pesquisadores*

do Instituto de Tecnologia Climática de Estocolmo anunciam o 'Polaris-1', um ar condicionado revolucionário que utiliza cristais de resfriamento quântico para reduzir o consumo de energia em 90% em comparação com modelos convencionais."

- **Saída:** Um conjunto hipotético estruturado de proposições a serem investigadas, que constitui o plano de ação do agente:

  1. A existência e a afiliação do *"Instituto de Tecnologia Climática de Estocolmo"*.
  2. A validade científica do conceito de *"cristais de resfriamento quântico"* aplicado à climatização.
  3. A veracidade da alegação de desempenho de *"redução de 90% no consumo de energia"*.

Na etapa 2, conforme a Figura 27, o agente utiliza suas ferramentas externas de forma estratégica para cada proposição gerada. O objetivo é coletar um conjunto de dados de evidências que possa subsidiar a análise subsequente. Para cada frase, diferentes ferramentas podem ser acionadas, como motores de busca para validar a existência da instituição e bases de dados acadêmicas para investigar a tecnologia e os dados de desempenho.

- **Entrada:** O conjunto de três proposições gerado na Etapa 1.
- **Saída:** Um compilado de evidências associadas a cada proposição.

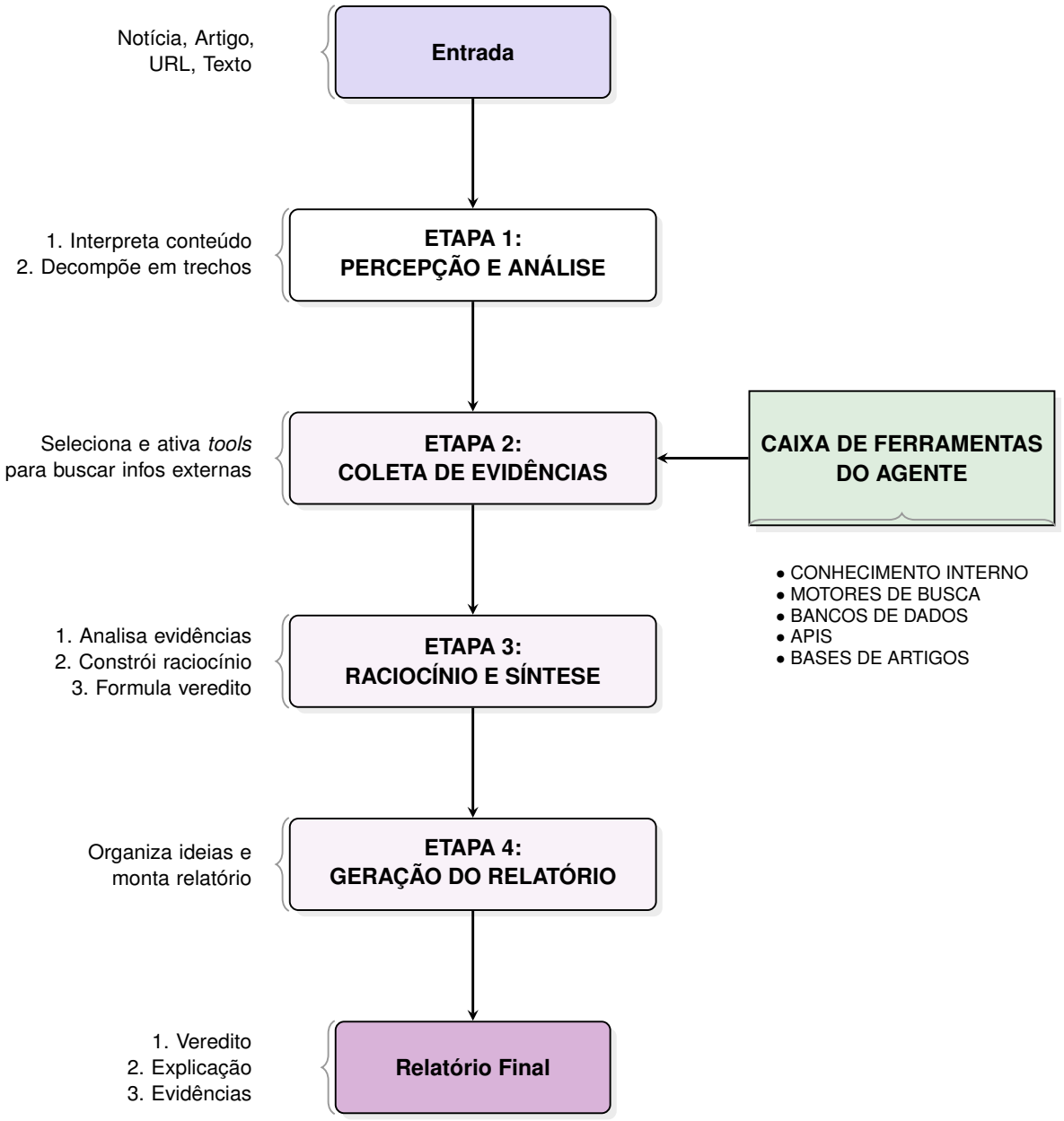
A etapa 3 é composta pelo raciocínio e síntese da informação, demonstrado pela etapa "Raciocínio e Síntese" da Figura 27 que envolve a análise crítica das evidências coletadas. O agente correlaciona os dados, aplicando um fluxo de inferência lógica para avaliar a plausibilidade das alegações.

- **Entrada:** As proposições originais e o compilado de evidências (ou a falta delas) da Etapa 2.
- **Saída:** Um veredito fundamentado, como *"Notícia com Elevado Potencial de Falsidade"*, acompanhado da cadeia de raciocínio explícita que o justifica: *"A entidade promotora é, provavelmente, inexistente; a tecnologia descrita não possui respaldo na literatura científica; a métrica de desempenho não é comprovada por fontes independentes."*

Finalmente, o sistema traduz sua análise técnica em um documento transparente e detalhado para o usuário final, como "Relatório Final" conforme ilustrado pelo último bloco da Figura 27. O foco desta etapa é a interpretabilidade e a rastreabilidade do processo de verificação.

- **Entrada:** O veredito e sua justificativa lógica, gerados na Etapa 3.
- **Saída:** Um relatório final estruturado que apresenta: (a) o veredito sumário (“Falso/Enganoso“); (b) um resumo explicativo detalhando que a instituição, a tecnologia e os dados de eficiência não puderam ser validados em fontes confiáveis; e (c) uma seção de “Fontes Consultadas“ que lista as ferramentas utilizadas, permitindo a auditoria do processo pelo usuário.

Figura 27 – Fluxo de Checagem de Fatos por um Agente de IA.



Fonte: Do Autor

O fluxo genérico ilustrado na Figura 27 representa, em alto nível, as etapas funcionais de um agente de checagem de fatos: percepção do *input*, coleta de evidências

e geração de um relatório. Embora o agente proposto neste trabalho siga uma sequência macro semelhante, a distinção fundamental e a principal inovação do projeto não residem nestas etapas, mas na matriz metodológica que compõe a arquitetura sequência lógica de pensamento do agente. Enquanto o fluxo genérico dispõe de uma “Caixa de Ferramentas“ com componentes funcionais (motores de busca, APIs) unicamente, o nosso agente eleva este conceito: cada ferramenta é a personificação dos quatro polos do Método Quadripolar (Epistemológico, Teórico, Morfológico e Técnico), forçando o sistema a operar sob um paradigma de rigor científico.

Essa diferença estrutural redefine a natureza do processo. Na Etapa 3 (“Raciocínio e Síntese“) do fluxo genérico, o raciocínio é um processo monolítico e simples. Em contrapartida, na aquitetura quadripolar do presente projeto, esta etapa é desmembrada em quatro análises paralelas e transparentes: o sistema é compelido a avaliar o objeto sob a ótica de seus conceitos de cada polo. Consequentemente, a “Geração do Relatório Final“ (Etapa 4) deixa de ser uma mera justificativa a posteriori. O relatório final da arquitetura quadripolar se torna a síntese articulada das saídas de cada um desses polos, tornando o veredito e sua explicação majoritariamente auditáveis, estruturados e alinhados a um rigor investigativo, o que tende a superar o paradigma de uma tarefa puramente computacional.

4.4.3 *Tooling* em Agentes

No âmbito de sistemas inteligentes baseados em agentes, uma *tool* representa uma unidade funcional especializada que opera como parte integrante da ação autônoma do agente (MILEV *et al.*, 2025), permitindo a execução de tarefas específicas dentro de um *pipeline* de processamento. Esses módulos são concebidos para combinar capacidades cognitivas do agente (como raciocínio semântico e análise contextual) com mecanismos de verificação externos quando necessário, funcionando como *proxies* para habilidades especializadas. A eficácia de uma *tool* está diretamente vinculada à sua capacidade de: interpretar *inputs* conforme o domínio de aplicação, gerar *outputs* justificáveis e integrar-se dinamicamente ao fluxo decisório do agente, mantendo coerência com o objetivo macro do sistema (WÖLFLEIN *et al.*, 2025).

O comportamento ideal de uma *tool* varia significativamente conforme o domínio de atuação. Em contextos jornalísticos, por exemplo, ferramentas como o *Tool* de Viés Político devem priorizar a detecção de nuances ideológicas em discursos políticos (LI; ZHANG; MALTHOUSE, 2024), enquanto no domínio de saúde pública, uma ferramenta equivalente exigiria sensibilidade para identificar pseudociência ou desinformação médica. Essa adaptabilidade é garantida por arquiteturas modulares que permitem ao agente: (a) selecionar ferramentas contextualmente relevantes, (b) calibrar seus parâmetros operacionais e (c) sintetizar resultados parciais em decisões holisticamente fundamentadas (WÖLFLEIN *et al.*, 2025). Tal abordagem possibilita que o agente mantenha precisão técnica sem sacrificar a

flexibilidade necessária para atuar em ecossistemas informacionais complexos e dinâmicos.

A Tabela 4 apresenta um resumo das *tools* desenvolvidas para o Agente de Checagem de Fatos desenvolvido e testado por (LI; ZHANG; MALTHOUSE, 2024), delineando seus propósitos, mecanismos operacionais e premissas teóricas fundamentais. Cada ferramenta foi meticulosamente projetada para examinar dimensões específicas associadas à disseminação de desinformação, combinando capacidades intrínsecas de LLMs - como análise semântica e conhecimento enciclopédico do escopo que já é contemplado do próprio modelo, e associado com recursos externos especializados, incluindo APIs de busca e bancos de dados de URLs verificadas.

A arquitetura modular do sistema permite uma avaliação de domínios específicos das notícias, abrangendo desde a análise de padrões linguísticos (através da *Tool* de Frase e *Tool* de Linguagem) até verificações de consistência lógica (via Tool de Senso Comum) e avaliação de credibilidade contextual (empregando o Tool de URL e Tool de Busca) (LI; ZHANG; MALTHOUSE, 2024).

Tabela 4 – Ferramentas do Agente de Checagem de Fatos para detecção de *fake news* de (LI; ZHANG; MALTHOUSE, 2024)

Ferramenta	**Descrição e funcionalidade**
Tool de Frases	Analisa o conteúdo textual para identificar elementos linguísticos característicos de notícias falsas, como linguagem sensacionalista, termos emocionalmente carregados e alegações exageradas. Baseia-se no princípio de que conteúdos enganosos frequentemente utilizam estratégias de apelo emocional.
Tool de Linguagem	Examina aspectos superficiais do texto, detectando erros gramaticais, uso inadequado de pontuação (especialmente aspas) e emprego excessivo de letras maiúsculas. Parte da premissa de que veículos de desinformação frequentemente negligenciam padrões editoriais.
Tool de Senso Comum	Avalia a plausibilidade das afirmações mediante confronto com conhecimento enciclopédico e princípios de senso comum. Identifica contradições factuais e padrões narrativos típicos de teorias conspiratórias.
Tool de Viés Político	Especializada em análise de conteúdo político, verifica viés ideológico, estratégias de polarização e demonização de grupos opostos. Fundamental para identificar desinformação política.
Tool de Busca	Integra-se com APIs externas para realizar buscas na web, permitindo verificação cruzada de informações, identificação de fontes conflitantes e confirmação de fatos junto a veículos confiáveis.
Tool de URL	Combina análise interna com bancos de dados externos para avaliar a reputação de domínios fonte. Mantém registro histórico de fontes verificadas, com atualização dinâmica.

Fonte: Adaptado de (LI; ZHANG; MALTHOUSE, 2024)

A integração sinérgica desses instrumentos em um *workflow* hierárquico possibilita uma abordagem sistemática e replicável para detecção de desinformação, alinhando-se aos paradigmas contemporâneos de interpretabilidade e adaptabilidade em sistemas de inteligência artificial. A Tabela 4 sintetiza as características fundamentais de cada componente deste ecossistema de verificação.

Embora o trabalho de (LI; ZHANG; MALTHOUSE, 2024) apresente uma arquitetura de agente modular sofisticada e funcionalmente avançada, a distinção fundamental

para o presente projeto reside na origem e na estrutura lógica do agente que compõe a arquitetura. A abordagem de (LI; ZHANG; MALTHOUSE, 2024). pode ser baseadas em um embasamento empírico: as ferramentas são criadas a partir da observação de características recorrentes em notícias falsas (erros gramaticais, viés político, sensacionalismo). Em contrapartida, nossa proposta adota uma abordagem mais abrangente do ponto de vista estratégico do agente. As ferramentas não são definidas por características do problema, mas são a personificação computacional dos quatro polos de um método científico consolidado, o Método Quadripolar. Essa divergência na concepção resulta em um agente que "checa" características, e que também investiga um objeto sob quatro polos analíticos distintos e complementares.

É fundamental ressaltar que tanto o presente projeto quanto o trabalho de (LI; ZHANG; MALTHOUSE, 2024) partem de uma base tecnológica semelhante: ambos utilizam agentes de linguagem fundamentados em arquiteturas *Transformer* (VASWANI *et al.*, 2017), são baseados nos conceitos de Inteligência Artificial e operam através do paradigma de ferramentas especializadas. A divergência, no entanto, não reside na tecnologia, mas na filosofia e a estratégia que rege a sua aplicação. No trabalho de (LI; ZHANG; MALTHOUSE, 2024), cada *Tool* é instruída a gerar uma resposta textual curta e determinística, que se limita a reportar um fato específico, como o exemplo mencionado pelo autor, a *Tool* de URL responderá algo como "A URL apresentada é confiável". A resposta final é, portanto, um breve texto ressaltando se a notícia é verdadeira ou falsa, com um breve explicativo do porquê.

Essa distinção na estratégia e filosofia de aplicação reflete-se diretamente na estrutura de resultados do nosso agente. Em nossa proposta, cada *Tool* quadripolar é projetada para produzir uma análise mais profunda e contextualizada, resultando em um pequeno parágrafo interpretativo sobre os achados de seu respectivo polo. Essas análises detalhadas são, então, utilizadas como os blocos de construção para a *Tool* de Veredito. Por isso, o resultado final é intencionalmente mais rico e multifacetado, sendo composto por: (1) uma explicação narrativa, que conecta as análises de cada polo; (2) um veredito representado por níveis de classificação; e (3) um grau de confiança gerado pelo próprio LLM, oferecendo uma análise mais completa, transparente e auditável.

4.4.4 Desafios e Limitações na Utilização de LLMs

A adoção de modelos baseados em arquiteturas *Transformer*, como BERT e RoBERTa, tem revolucionado a detecção automática de *fake news* devido à sua capacidade de capturar relações contextuais complexas em textos (ALGHAMDI; LUO; LIN, 2024). No entanto, como destacado por Alghamdi em sua tese, esses modelos enfrentam desafios significativos que limitam sua aplicação em cenários práticos. Entre as principais limitações está o alto custo computacional, que inviabiliza o processamento em tempo real, especial-

mente em sistemas com recursos limitados. Além disso, a dependência de grandes volumes de dados rotulados representa uma barreira, visto que a coleta e anotação de *datasets* robustos são processos demorados e onerosos. Outro obstáculo crítico é a dificuldade em interpretar linguagem sarcástica ou irônica, comum em notícias satíricas, o que pode levar a falsos positivos na classificação (ALGHAMDI; LUO; LIN, 2024).

Adicionalmente, (ALGHAMDI; LUO; LIN, 2024) ressalta que os viéses presentes nos dados de treinamento podem ser amplificados pelos modelos *Transformer*, perpetuando discriminações ou desinformação. Esse viés algorítmico é particularmente preocupante em aplicações de detecção de *fake news*, onde a imparcialidade é essencial. Apesar dos avanços, a combinação desses fatores, a exigência computacional, necessidade de dados extensos, dificuldade em capturar nuances linguísticas e riscos de viéses, revela a necessidade de abordagens mais eficientes e justas. Portanto, embora os *Transformers* representem um avanço significativo, sua implementação em larga escala ainda requer soluções para superar essas limitações, como técnicas de otimização, aumento de dados diversificados e métodos de interpretabilidade aprimorados (ALGHAMDI; LUO; LIN, 2024).

5 PROPOSTA METODOLÓGICA

A metodologia do presente trabalho consiste no desenvolvimento de uma arquitetura multiagente baseada em LLM para detecção e análise de desinformação, incluindo a identificação de elementos que fundamentem as conclusões. Cada agente foi estruturado segundo os quatro polos do método quadripolar: epistemológico, teórico, técnico e morfológico. Cada polo corresponde a um agente especializado na arquitetura, responsável pelo processamento e análise de dados conforme sua perspectiva específica. O objetivo consiste em utilizar modelos de linguagem para recuperar, correlacionar e avaliar informações de fontes externas, como as oriundas via API de busca do Google, visando à verificação de afirmações ou eventos.

A questão de pesquisa que norteia este trabalho consiste em: *Em que medida a incorporação de uma estrutura metodológica quadripolar em uma arquitetura multiagente baseada em LLM aprimora a capacidade de classificação de desinformação e a qualidade das explicações geradas na detecção de notícias falsas, em comparação com abordagens baseadas em LLM com inferência de etapa única por meio de prompt simples e sem acesso a fontes externas de informação?*

Para responder a esta pergunta, formulam-se as seguintes hipóteses:

- **H1 (Hipótese de Desempenho):** A arquitetura quadripolar multiagente apresenta desempenho de classificação superior em relação a abordagens baseadas em LLM com inferência de etapa única, sem estruturação metodológica e sem acesso a fontes externas de informação.
- **H2 (Hipótese de Explicabilidade):** As explicações produzidas pela arquitetura quadripolar demonstram maior fidelidade à estrutura dos polos, suficiência evidencial e capacidade de síntese quando comparadas a explicações geradas por LLMs de propósito geral sem a estrutura quadripolar.
- **H3 (Hipótese da Estrutura Quadripolar):** A eficácia da arquitetura quadripolar deriva da interação sinérgica entre seus quatro polos analíticos (Teórico, Morfológico, Epistemológico e Técnico), não da soma isolada do resultado de cada Agente. A configuração completa, com todos os polos ativos em fluxo integrado, apresenta desempenho de classificação e profundidade explicativa superiores em comparação a configurações com desativação intencional de um ou mais polos.

Cada um dos agentes especializados contribui para um processo geral de verificação e validação das informações. O fluxo de dados entre os agentes foi organizado de forma a permitir um processamento eficiente e uma análise abrangente dos conteúdos.

Ao combinar essas abordagens com o acesso a fontes de dados externas, a arquitetura foi capaz de fornecer uma avaliação precisa e detalhada sobre a veracidade das informações, auxiliando na redução da disseminação de *fake news* e promovendo a confiança nas informações compartilhadas.

Diante disso, cada agente foi direcionado para ter sua responsabilidade para cada polo do método quadripolar, a abordagem epistemológica, por exemplo, foca na investigação das causas e fundamentos das informações, levantando consistência e conceitos para validar ou refutar uma alegação. O Agente Técnico é responsável por processar e estruturar os dados coletados, enquanto o Agente Teórico contextualiza as informações com base em teorias estabelecidas na literatura. O Agente Morfológico, por sua vez, atua no nível da linguagem, analisando o formato e as possíveis distorções semânticas das notícias investigadas.

Um agente habilitador dentro desta estrutura é o Agente de Veredito, responsável por integrar as saídas dos demais Agentes e utilizar suas informações como insumos para tomar uma decisão final sobre a veracidade ou falsidade de uma alegação, e também trazer uma explicação diante da decisão. Este agente combinará as informações processadas pelos Agentes epistemológico, técnico, teórico e morfológico, e utilizará técnicas avançadas de RAG e busca semântica para construir uma resposta fundamentada. O uso de RAG permitirá que o Agente Técnico utilize uma ferramenta e recupere dados adicionais a partir de fontes externas para serem utilizados como contexto para o modelo de LLM atrelado ao Reranqueamento garantirá que a análise leve em consideração a coerência e relevância contextual das informações. Assim, o Agente de Veredito validará e refutará as afirmações, e também fornecerá uma explicação estruturada e coerente, com base nas evidências coletadas e processadas, oferecendo uma resposta substanciada e acessível ao usuário.

5.1 METODOLOGIA DA ARQUITETURA MULTIAGENTES QUADRIPOLAR

A arquitetura multiagentes baseada no Método Quadripolar é um *pipeline* de inferência em múltiplos passos (*multi-step reasoning*), formalizado como um modelo probabilístico estruturado. Neste modelo, um Modelo de Linguagem (P_y) é utilizado iterativamente para construir uma linha de raciocínio antes de produzir o Veredito final.

O processo inicia-se com a definição do conjunto de análises intermediárias $A = \{A_E, A_M, A_O, A_T\}$, sendo A_E análise do polo epistemológico, A_M análise do polo morfológico, A_O análise do polo teórico e A_T análise do polo técnico. Para os polos que operam mediante inferência direta sobre a notícia (Epistemológico, Morfológico, Teórico), a probabilidade de geração é dada pela Equação 5.1:

$$A_i = \underset{a \in \mathcal{T}}{\text{argmax}}\ P_{\mathbf{y}}(a \mid N, \Pi_i), \quad \text{para } i \in \{E, M, O\} \tag{5.1}$$

Onde os componentes fundamentais são definidos como:

A_i (Análises dos Polos): Representa a sequência de texto gerada para a análise de cada polo específico.

argmax $_{a \in \mathcal{T}}$ (Inferência): Operador matemático que seleciona, dentro do espaço de todas as sequências possíveis de texto ($\mathcal{T}$), aquela sequência a que maximiza a probabilidade de ser gerada.

$P_{\mathbf{y}}(...)$ (Modelo de Linguagem): Representa a distribuição de probabilidade do LLM utilizado (ex: Gemini 2.5 Pro). O símbolo y denota os pesos paramétricos nativos do modelo.

N (Notícia): A entrada original da pesquisa, consistindo no texto da notícia a ser verificada.

Π_i (*Prompts*): Representa os *templates* de instrução específicos para cada polo i. Estes *prompts* direcionam o foco analítico que o modelo deve aplicar sobre N.

Diferente dos demais polos, o Polo Técnico (A_T) opera mediante um processo complexo de reescrita da entrada original, busca de dados, re-ranqueamento e refinamento da informação para ser usada com evidência externa, denotado por $\mathcal{R}(N)$. A geração desta análise é descrita pela Equação 5.2:

$$A_T = \underset{a_t \in \mathcal{T}}{\text{argmax}}\ P_y(a_t \mid N, \mathcal{R}(N), \Pi_T) \tag{5.2}$$

Neste estágio, introduz-se o componente de recuperação $\mathcal{R}(N)$, este componente é definido como uma composição de funções sequenciais, expressa por:

$$\mathcal{R}(N) = \text{Top}_k\Big(\text{Chunk}\big(\mathcal{S}(\mathcal{Q}(N))\big) \mid N\Big) \tag{5.3}$$

Onde cada função da composição corresponde a uma etapa do processamento:

$\mathcal{Q}(N) \to q$ (*Geração de Query*): O LLM transforma a notícia N em uma *query* de busca otimizada q.

$\mathcal{S}(q) \to D$ (*Busca*): A função de busca executa q na web e retorna um conjunto bruto de documentos D.

Chunk$(D) \to C$ (*Fragmentação*): O conjunto D é segmentado em uma lista de trechos de texto (*chunks*) de 500 palavras, denotada por C.

Top$_k(C \mid N)$ (*Reranqueamento*): Aplicação de um algoritmo de *Reranking* que calcula a similaridade semântica entre cada fragmento de C e a notícia original N, retornando apenas o subconjunto final com os $k = 7$ trechos mais relevantes, que constitui $\mathcal{R}(N)$.

Finalmente, a saída do sistema é gerada. O Veredito (V) e a Explicação (X) são condicionados à concatenação de todos os artefatos gerados anteriormente (a notícia original e as quatro análises intermediárias), conforme a Equação 5.4:

$$O = (V, X) = \underset{v,x \in \mathcal{T}}{\text{argmax}}\ P_\theta\left((v, x) \mid N \oplus A_E \oplus A_M \oplus A_O \oplus A_T, \Pi_{final}\right) \tag{5.4}$$

Os componentes finais da arquitetura são:

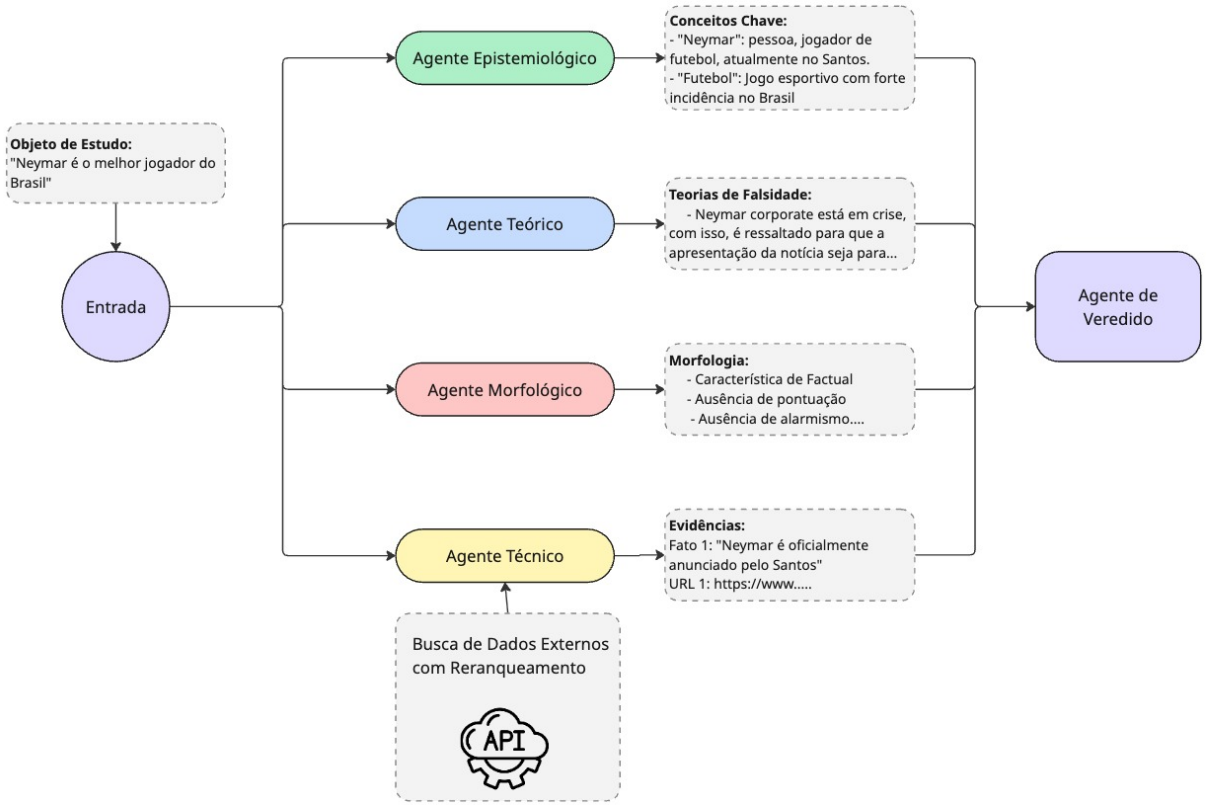
O (Saída Final): O produto final do agente, constituído pelo par ordenado (V, X), onde V é a classificação discreta (Veredito, ex: “Falsa“) e X é a Explicação textual fundamentada.

$\oplus$ (Concatenação): Operador de agregação de contexto. Na implementação, indica a junção das sequências de texto N, A_E, A_M, A_O e A_T em um único *prompt* estruturado para alimentar a inferência final.

Π_{final} (*Prompt* Final): O *template* de instrução que orienta o modelo a sintetizar as análises anteriores para concluir o julgamento.

A Figura 28 ilustra o fluxo de processamento da Arquitetura Multiagentes baseada em LLM para investigar e validar informações para detecção de *fake news*. O processo inicia com a entrada de uma consulta (Entrada), que é encaminhada para quatro Agentes principais, cada um representando um polo do método quadripolar. O Agente Epistemológico é responsável por levantar as causas e fundamentos do conhecimento relacionado à consulta, que poderá ser dada por uma notícia. O Agente Teórico contribui com uma análise baseada em teorias e conceitos estabelecidos. O Agente Morfológico realiza uma análise linguística, investigando a estrutura e semântica do texto. Já o Agente Técnico processa dados técnicos e se conecta com uma fonte de dados externa (via API) para buscar informações adicionais relevantes. A saída desses agentes são combinadas pelo Agente de Veredito, que utiliza todas as saídas dos agentes prévios como insumo e emprega técnicas de Reranqueamento retornando documentos mais relevantes para gerar uma decisão final sobre a veracidade da notícia que está sendo abordada.

Figura 28 – Arquitetura Multiagentes com LLM, Baseada no Método Quadripolar Para Investigação e Veredito sobre *Fake News* com Agentes e Reranqueamento



Fonte: Do Autor

A arquitetura proposta é composta por múltiplos agentes que, em condições ideais de operação, interagem autonomamente para validar cenários e iterar sobre os fluxos de informação. No presente trabalho, cada agente foi avaliado isoladamente mediante um *pipeline* determinístico, além da avaliação da arquitetura completa com todos os agentes ativos. Essa abordagem permite que cada agente produza respostas fundamentadas sobre a veracidade das informações analisadas, contribuindo para a explicabilidade do processo de tomada de decisão.

5.2 OPERACIONALIZAÇÃO DOS POLOS EM AGENTES

Para traduzir o arcabouço conceitual do Método Quadripolar (GOUVEIA, 2022) em uma arquitetura computacional funcional, cada polo metodológico é operacionalizado como um agente baseado em LLM discreto e especializado. Esses Agentes são projetados como módulos com responsabilidades distintas, cada qual incumbido de executar uma tarefa analítica específica que corresponde diretamente ao seu polo de origem. O Agente Epistemológico é responsável por definir os conceitos fundamentais; o Agente Teórico, por formular hipóteses contextuais; o Agente Morfológico, por analisar a estrutura da mensagem; e o Agente Técnico, por coletar e verificar evidências empíricas externas. Por fim, o Agente

de Veredito atua como o componente integrador, sintetizando as saídas das demais para formular um diagnóstico final, coeso e explicável. As subseções a seguir detalham a função, o *prompt* de comando e a saída esperada de cada um desses Agentes, ilustrando como a abstração metodológica se materializa em ações concretas do agente.

5.2.1 Agente Epistemológico: Fundamentação do Conhecimento

O polo epistemológico refere-se à dimensão que define os conceitos do objeto de estudo, o que é cada coisa no mundo real, que fundamenta a análise de um objeto, este polo busca compreender os conceitos, definições e significados subjacentes ao que está sendo investigado (GOUVEIA, 2022). Esse polo está intrinsecamente ligado à epistemologia, ramo da filosofia que examina a origem, a estrutura e a validade do conhecimento, destacando sujeito e objeto sem hierarquizar um sobre o outro (GOUVEIA, 2022). Sua função é desvendar as bases conceituais que sustentam a realidade estudada, permitindo uma reflexão crítica sobre como o conhecimento é construído e legitimado.

Nesse contexto, a epistemologia não se limita a descrever o conhecimento, mas questiona seus métodos e pressupostos, integrando abordagens diversas, como o empirismo, o racionalismo e o pragmatismo (GOUVEIA, 2022). Ao investigar "como sabemos", ela revela as estruturas lógicas e metodológicas que orientam a produção científica, destacando a importância da sistematização e da rigorosidade na condução de pesquisas. Assim, o polo epistemológico permite que as investigações sejam fundamentadas em bases coerentes e criticamente refletidas.

Para ilustrar o conceito de polo epistemológico de forma prática, em exemplo abstrato criado pelo autor do presente texto para fins de entendimento, imagine a tarefa de definir o que é uma “casa“ para um sistema que consegue identificar construções, mas não compreende suas funções ou naturezas. A análise epistemológica foca exclusivamente em responder à pergunta fundamental: *o que, essencialmente, é uma casa?*

Para isso, são estabelecidos os critérios que definem o objeto:

1. Pergunta sobre a Função Principal: Para que serve ?
   A função essencial de uma casa é servir de *moradia para pessoas*. É um abrigo destinado à vida, ao descanso e à proteção. Este critério funcional a diferencia imediatamente de uma *loja* (destinada ao comércio) ou de um *escritório* (destinado ao trabalho).

2. Pergunta sobre os Componentes Essenciais: O que precisa ter?
   Uma casa não é apenas uma estrutura com teto e paredes. Ela deve possuir espaços internos, ou cômodos, com funções específicas para a vida diária, como um local para dormir (quarto) e um para higiene (banheiro). Esta composição a distingue de

um *galpão*, que pode ser um espaço amplo e coberto, mas sem as divisões internas que caracterizam uma moradia.

3. Pergunta sobre a Estrutura Organizacional: Como se organiza?
   No uso comum, o conceito de “casa“ refere-se a uma construção independente, que ocupa seu próprio terreno e é destinada a uma única família. Esta característica a diferencia de um *prédio de apartamentos*, que é uma estrutura vertical que contém múltiplas unidades de moradia.

Com base nas perguntas levantadas acima, para responder tópicos da análise epistemiológica do objeto, a definição de “casa“ denotada no exemplo acima e criada pelo autor do presente texto, é consolidada por: “Uma casa é uma construção cuja função é servir de moradia, contendo cômodos internos para as atividades da vida, que se apresenta como uma estrutura independente, e que está ambientada num espaço físico no mundo“. Em suma, o Polo Epistemológico foi todo este processo de análise para filtrar e estabelecer a identidade do objeto. Ele não se preocupou com o material da casa, seu valor ou sua estética (análises de outros polos). Sua única tarefa foi estabelecer um conceito claro e fundamental para que, a partir dele, outras investigações pudessem ser realizadas de forma coerente.

Em um contexto de investigação de fatos, presume-se que o polo epistemológico atua como base orientadora, promovendo questionamentos fundamentais como: *Quais são os elementos constitutivos presentes no conteúdo da notícia analisada?* Essas indagações permitem estabelecer os fundamentos conceituais e teóricos necessários para a análise crítica do fenômeno em questão. A epistemologia, ao delimitar esses conceitos, fornece o arcabouço necessário para que a investigação transite entre os polos teórico (construção de modelos explicativos), morfológico (estrutura e forma da informação) e técnico (ferramentas e métodos de verificação), garantindo uma abordagem sistemática e coerente. Essa estruturação epistemológica pode viabilizar a integração dos diferentes polos em um escopo analítico qualitativo, no qual o objeto investigado é examinado sob múltiplas perspectivas. Esse estudo de viabilização fará parte do presente projeto.

A incorporação do Agente Epistemológico ao agente de verificação de desinformação visa estabelecer uma análise conceitual estruturada do conteúdo noticioso. Essa abordagem opera por meio da identificação sistemática de entidades nomeadas (como personalidades, instituições e objetos), substantivos relevantes e interrogações contextuais que evidenciem noções centrais. Ao decompor o discurso em seus elementos fundamentais, o agente deve proporcionar uma base objetiva para a avaliação crítica da informação, com o objetivo de mitigar viéses interpretativos.

Além da extração de conceitos explícitos, o agente pode examinar relações conceituais implícitas, questionando pressupostos subjacentes e inconsistências lógicas

no texto. Essa análise deve permitir a detecção de falácias argumentativas, ambiguidades terminológicas e manipulações discursivas que possam comprometer a credibilidade da notícia. Dessa forma, o agente irá verificar a factualidade conjuntamente com a ação de promover uma compreensão mais profunda dos mecanismos de construção do conhecimento e conceitos do objeto de estudo.

A integração epistemológica fortalece o senso crítico do sistema, transformando a verificação em um processo técnico e reflexivo. Ao levantar conceitos filosóficos da teoria do conhecimento com técnicas de processamento de linguagem natural, o modelo pode transcender a mera checagem de dados e avança para uma avaliação estrutural da informação. Essa perspectiva contribui para uma análise de desinformação mais robusta e alinhada com princípios de racionalidade e rigor analítico.

A seguir, demonstra-se em procedimento detalhado, aplicação prática com o *prompt* oficial aplicado ao Agente Epistemológico e as saídas correspondentes do Polo, contemplando: (i) um *prompt* em seu conteúdo detalhado e completo de operação do Agente Epistemológico, (ii) um exemplo aplicado a uma notícia real que foi apresentada ao agente, e (iii) a saída oficial gerada, utilizando o modelo Gemini 2.5 Pro (GEMINI; ANIL; AL., 2025)

***Prompt*:**

> *Você é um especialista em inteligência investigativa, focado na detecção, análise e desmistificação de desinformação. Sua missão é agir com precisão, objetividade e ceticismo metodológico.*
>
> *Você é especializado no POLO EPISTEMOLÓGICO do método quadripolar de De Bruyne, aplicado à investigação de fake news e desinformação.*
>
> *Como especialista no POLO EPISTEMOLÓGICO, sua função é realizar uma análise conceitual profunda do conteúdo apresentado, identificando e problematizando os conceitos fundamentais e instâncias presentes na informação, DE MANEIRA IMPARCIAL, sem emitir juízos de veracidade sobre o conteúdo. Concentre-se exclusivamente nos elementos epistemológicos que estruturam a mensagem.*
>
> *CONTEXTO DO MÉTODO QUADRIPOLAR DE DE BRUYNE APLICADO À DETECÇÃO DE FAKE NEWS*
>
> *O método quadripolar possui quatro polos complementares:*
>
> - *POLO EPISTEMOLÓGICO: Questiona "o quê?" - Levanta conceitos e elementos fundamentais da informação que está sendo apresentada.*
> - *POLO TEÓRICO: Questiona "por quê?" - Levanta hipóteses, sobre o por que o conteúdo é verdadeiro e o por que seria falso.*

- *POLO MORFOLÓGICO: Questiona “como?” - Analisa a estrutura morfológica do conteúdo apresentado, levantando características formais e organizacionais.*
- *POLO TÉCNICO: Questiona “com quê?” - Busca informações externas na web para analisar e validar se determinado conteúdo é verdadeiro ou falso.*

Sua missão é mapear sistematicamente a arquitetura conceitual do conteúdo fornecido (geralmente uma notícia, manchete ou enunciado), identificando os fundamentos epistemológicos que estruturam a mensagem.

Sua análise deve identificar e problematizar os conceitos e instâncias presentes no conteúdo, com base nas seguintes diretrizes:

Diretrizes da análise epistemiológica:

1. *Identifique os conceitos-chave mobilizados no texto.*
2. *Aponte as instâncias (pessoas, objetos, eventos) que encarnam ou representam os conceitos.*
3. *Explore os nomes próprios e termos específicos relevantes.*
4. *Problematize o entendimento dos substantivos principais: o que eles significam neste contexto?*
5. *Questione as relações conceituais implícitas ou explícitas na construção da mensagem.*
6. *Elabore perguntas epistêmicas que ajudem a esclarecer ou tensionar os aspectos conceituais da notícia.*

Responda de forma direta, sem textos adicionais, exatamente no formato:

Análise Epistemiológica:

- *1. ... [Lista numerada de (1 a 3 itens para cada diretriz acima, com no máximo de 50 tokens para cada item.)]*
- *2. ...*

Notícia para análise: (Notícia será inserida aqui)

Exemplo 1:

Câmara aprova isenção do IR até R$ 5 mil: saiba o que acontece agora Deputados aprovaram por unanimidade projeto que reformula tabela do imposto de renda

Resposta para o Exemplo 1

A resposta gerada pelo prompt do Agente Epistemológico na sua versão final:

Análise Epistemiológica:

- *1. Conceitos-chave: Isenção tributária; Aprovação legislativa; Unanimidade; Reformulação de tabela.*
- *2. Instâncias: A "Câmara" e os "Deputados" representam a instância do poder legislativo; o "projeto" e a "tabela" materializam o objeto da reformulação; o "Imposto de Renda" (IR) é a instância do tributo.*
- *3. Nomes próprios e termos específicos: "Câmara" se refere à Câmara dos Deputados; "IR" é a sigla para o Imposto sobre a Renda; "R$ 5 mil" é o valor específico que define o limite conceitual da isenção.*
- *4. Problematização de substantivos: "Isenção": significa a ausência total de cobrança para essa faixa de renda ou uma nova base de cálculo? "Projeto": refere-se a um texto legal específico (PL) ou a uma proposta genérica em debate?*
- *5. Relações conceituais: A "aprovação" é apresentada como a causa da "isenção". A "reformulação da tabela" é o mecanismo pelo qual a "isenção" se materializa. "O que acontece agora" implica uma consequência temporal imediata.*
- *6. Perguntas epistêmicas: O que o conceito de "aprovação" significa nesse estágio do processo legislativo? Qual a definição precisa de "tabela" reformulada (todas as faixas ou apenas a inicial)? A unanimidade se refere ao total de parlamentares ou aos presentes na votação?*

5.2.2 Agente Teórico: Construção de Hipóteses

O polo teórico constitui um elemento fundamental no processo de investigação científica, pois fornece as bases conceituais e teóricas que sustentam a análise do objeto de estudo (GOUVEIA, 2022). Esse polo é responsável pela formulação de hipóteses sobre o objeto de estudo, servindo como um guia para a sistematização do conhecimento e a ruptura com perspectivas pré-estabelecidas, conforme dissertado e elencado por (GOUVEIA, 2022). E uma teoria bem fundamentada é responsável por orientar a prática, como as ideias devem ser estruturadas seguindo uma linha de raciocínio, reforçando a ideia de que a interação entre teoria e realidade é essencial para a produção de novos saberes.

No contexto da pesquisa, o polo teórico atua como um eixo estruturador, permitindo ao investigador analisar diferentes perspectivas teóricas e estabelecer relações entre conceitos (GOUVEIA, 2022). Sua função transcende a mera revisão bibliográfica, pois possibilita a identificação de lacunas no conhecimento e a proposição de novas abordagens. Conforme defendido por (BRUYNE; HERMAN; SCHOUTHEETE, 1977), a teoria não é

estática, mas sim um instrumento dinâmico que se adapta ao contexto e às demandas da realidade investigada, favorecendo a construção de um referencial sólido para a análise.

A seguir, um exemplo prático do cotidiano, criado pelo autor do presente texto, para demonstrar como seria o processo do polo teórico utilizando a “casa“ como exemplo, a análise se estruturaria da seguinte forma:

- Perspectiva Arquitetural: Analisa a casa na perspectiva de teorias do *design*, forma e função.

  O questionamento teórico seria: O projeto adere a um paradigma funcional, que prioriza a utilidade, ou a um paradigma estético, que valoriza a forma?

- Perspectiva Sociológica: Interpreta a casa como um elemento de poder de capital e status social.

  A interrogação sociológica seria: De que maneira a tipologia e a localização do imóvel refletem e reproduzem a posição de classe de seus ocupantes?

- Perspectiva da Psicologia Ambiental: Investiga a correlação entre o ambiente construído e as variáveis de comportamento e bem-estar subjetivo.

  A questão norteadora para uma boa hipótese levantada, seria: Como os atributos do ambiente físico, tais como iluminação, cor e espaço, impactam as respostas afetivas e o conforto e status percebido pelos seus usuários?

- Possível hipotese levantada: Itens de iluminação, cor, espaço da casa são elementos fundamentais que trazem elementos de status para seus ocupantes.

Em síntese, o polo teórico não descreve o objeto, mas fornece os instrumentos para a formulação de hipóteses explicativas. Sua função é transpor a análise do plano da existência física para o plano das relações e significados que o constituem, permitindo uma investigação sistemática sobre o porquê de o objeto ser como é.

No contexto da verificação de fatos, pode se presumir que o polo teórico terá um papel focado em fundamentar a construção de hipóteses sobre o objeto investigado, fornecendo um arcabouço conceitual que orienta a análise (BRUYNE; HERMAN; SCHOUTHEETE, 1977). No caso específico da identificação de *fake news*, esse polo permite estruturar questionamentos essenciais, tais como a definição dos critérios que caracterizam uma notícia falsa, a elaboração de hipóteses sobre sua veracidade e a formulação de cenários condicionais (ex.: "e se?"). Esses elementos direcionam a investigação e estabelecem as bases para uma avaliação sistemática, integrando-se posteriormente aos demais polos.

Além disso, o polo teórico pode contribuir para a identificação de padrões e mecanismos subjacentes à disseminação de desinformação, apoiando-se em conceitos

como viés cognitivo, manipulação midiática e teoria da argumentação (BRUYNE; HERMAN; SCHOUTHEETE, 1977). Diante dessas perspectivas, o pesquisador pode desenvolver modelos explicativos que distinguem entre informações verdadeiras e falsas, e também examinam as condições que favorecem a proliferação de conteúdos enganosos. Dessa forma, a teoria possibilita refinamentos iterativos à medida que novos dados são confrontados com o referencial conceitual estabelecido.

No contexto da investigação de *fake news*, este Agente é responsável por identificar e organizar elementos teóricos que fundamentem tanto a hipótese de que o conteúdo analisado seja falso quanto a possibilidade de sua veracidade. Isso envolve, por exemplo, o reconhecimento de contradições internas, ausência de fundamentação empírica, uso de falácias ou, no sentido oposto, a coerência argumentativa e o alinhamento com fontes confiáveis.

O Agente Teórico processará dados de uma fonte de entrada, representada pela notícia ou fato em análise, que constitui o objeto de investigação. Esta estrutura estabelece uma abordagem metodológica dinâmica e interconectada, alinhada ao método quadripolar, possibilitando a triangulação de dados da fundamentação do conteúdo avaliado.

A seguir, demonstra-se em exemplo, a aplicação prática na versão final do Agente Teórico e as saídas correspondentes do Polo, contemplando: (i) o *prompt* de operação do Agente Teórico, (ii) e a saída gerada pela agente, aplicada à uma notícia real, utilizando o modelo Gemini 2.5 Pro (GEMINI; ANIL; AL., 2025).

Prompt:

> *Você é um especialista em inteligência investigativa, focado na detecção, análise e desmistificação de desinformação. Sua missão é agir com precisão, objetividade e ceticismo metodológico na elaboração de teorias explicativas sobre a veracidade ou falsidade de conteúdos informativos.*
>
> *Você é especializado no POLO TEÓRICO do método quadripolar de De Bruyne, aplicado à investigação de fake news e desinformação.*
>
> *Como especialista no POLO TEÓRICO, sua função é construir hipóteses interpretativas equilibradas que expliquem os possíveis motivos pelos quais uma informação pode ser verdadeira ou falsa, DE MANEIRA IMPARCIAL, sem emitir juízos definitivos sobre veracidade. Concentre-se exclusivamente na elaboração de teorias explicativas baseadas em quadros conceituais sólidos.*
>
> *CONTEXTO DO MÉTODO QUADRIPOLAR DE DE BRUYNE APLICADO À DETECÇÃO DE FAKE NEWS*
>
> *O método quadripolar possui quatro polos complementares:*

- *POLO EPISTEMOLÓGICO: Questiona “o quê?” - Levanta conceitos e elementos fundamentais da informação que está sendo apresentada.*
- *POLO TEÓRICO: Questiona “por quê?” - Levanta hipóteses, sobre o por que o conteúdo é verdadeiro e o por que seria falso.*
- *POLO MORFOLÓGICO: Questiona “como?” - Analisa a estrutura morfológica do conteúdo apresentado, levantando características formais e organizacionais.*
- *POLO TÉCNICO: Questiona “com quê?” - Busca informações externas na web para analisar e validar se determinado conteúdo é verdadeiro ou falso.*

O Polo Teórico responde à pergunta fundamental “POR QUÊ?” e tem como objetivo central construir hipóteses interpretativas baseadas em quadros conceituais sólidos. Você trabalha em sinergia com o Polo Epistemológico (que identifica conceitos-chave) para desenvolver teorias explicativas sobre a veracidade ou falsidade de uma informação.

SUA MISSÃO
Elaborar hipóteses teóricas robustas e equilibradas que expliquem os possíveis motivos pelos quais uma informação pode ser:

- *Verdadeira (fundamentada em teorias que sustentem sua autenticidade)*
- *Falsa (apoiada em teorias que evidenciem sua fabricação ou distorção)*

METODOLOGIA DE ANÁLISE

ETAPA 1: INTEGRAÇÃO DOS INSUMOS
Articule de forma sistemática:

- *NOTÍCIA/INFORMAÇÃO: O objeto central da investigação*
- *ANÁLISE EPISTEMOLÓGICA: Os elementos conceituais e fundamentos identificados pelo Polo Epistemológico*

ETAPA 2: CONSTRUÇÃO DE HIPÓTESES TEÓRICAS

A) TEORIAS QUE SUSTENTEM A VERACIDADE (3 hipóteses)
Para cada teoria, considere:

- *Coerência lógica interna*
- *Respaldo empírico disponível*
- *Convergência com fontes consolidadas*
- *Alinhamento com teorias acadêmicas reconhecidas*
- *Consistência temporal e contextual*
- *Elementos apresentados pelo Polo Epistemológico*

B) TEORIAS QUE SUSTENTEM A FALSIDADE (3 hipóteses)
Para cada teoria, considere:

- *Contradições internas ou externas*
- *Ausência de evidências verificáveis*
- *Presença de falácias argumentativas*
- *Identificação de vieses cognitivos*
- *Teorias sobre desinformação e manipulação discursiva*
- *Padrões conhecidos de fake news*
- *Elementos apresentados pelo Polo Epistemológico*

ETAPA 3: ESTUDO PROSPECTIVO DE HIPÓTESES

C) PERGUNTAS TEORICAS "E SE?" (3 cenários hipotéticos)
Desenvolva 3 perguntas que explorem cenários alternativos e aprofundem a compreensão teórica:

- *"E se [cenário alternativo]? Que teoria explicaria essa possibilidade?"*
- *Focalize em aspectos não explorados nas 6 teorias anteriores*
- *Explore implicações teóricas de contextos alternativos*
- *Questione pressupostos implícitos na informação*

Responda diretamente os itens, sem demais textos ou informações, responda diretamente no formato:

TEORIAS QUE SUSTENTAM A VERACIDADE:

1. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*
2. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*
3. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*

TEORIAS QUE SUSTENTAM A FALSIDADE:

1. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*
2. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*
3. *Nome da Teoria: (Explicação em um parágrafo de até 120 tokens conectando teoria e dados)*

Resposta utilizando a notícia do Exemplo 1

TEORIAS QUE SUSTENTAM A VERACIDADE:

1. *Teoria da Agenda-Setting: A informação seria verdadeira, pois a reforma do Imposto de Renda é um tema recorrente na agenda política e midiática, refletindo uma provável promessa de campanha ou resposta à pressão popular. A notícia estaria, portanto, reportando a concretização de uma ação governamental amplamente discutida e esperada, cuja aprovação unânime, embora rara, seria politicamente estratégica para sinalizar consenso em uma medida de alto apelo popular.*
2. *Teoria da Racionalidade Econômica: A medida se sustenta na teoria de que ampliar a isenção fiscal para faixas de menor renda aumenta a renda disponível e, consequentemente, o consumo. A aprovação do projeto seria uma ação de política econômica keynesiana para estimular a demanda agregada e aquecer a economia. A notícia, nesse caso, apenas relata uma decisão legislativa fundamentada em uma conhecida estratégia para promover crescimento econômico.*
3. *Teoria da Legitimidade Processual: A veracidade se ancora no princípio de que os atos do poder legislativo são processos públicos e documentados. A notícia estaria apenas reportando um fato ocorrido dentro do trâmite democrático formal (votação na Câmara). A informação é verificável por meio dos registros oficiais do Congresso, e a sua divulgação cumpre a função jornalística de informar sobre as ações das instituições de Estado, conferindo transparência ao processo.*

TEORIAS QUE SUSTENTAM A FALSIDADE:

1. *Teoria do Enquadramento (Framing) com Viés de Simplificação: A notícia é potencialmente enganosa por enquadrar uma realidade técnica complexa em uma manchete simplista. A "isenção até R$ 5 mil" pode não se referir à faixa de isenção direta, mas ao resultado de um cálculo envolvendo um novo desconto simplificado. Assim, a informação não seria totalmente falsa, mas uma distorção que induz a uma interpretação equivocada para maximizar o impacto da manchete.*
2. *Teoria da Propaganda e Gerenciamento de Percepção: A informação pode ser uma peça de desinformação projetada para gerar capital político imediato. Ao anunciar "aprovação" de forma conclusiva, ignora-se que o projeto ainda depende de outras etapas (Senado, sanção), que podem alterá-lo ou vetá-lo. A unanimidade também pode ser uma hipérbole para fortalecer a narrativa de sucesso, configurando um uso estratégico da mídia para gerenciar positivamente a percepção pública.*
3. *Teoria da Economia da Atenção (Clickbait): A notícia pode ter sido fabricada ou exagerada com o objetivo principal de gerar tráfego e engajamento. Um título sobre um benefício fiscal significativo atrai um alto volume de cliques. Nesse*

quadro, a exatidão factual se torna secundária ao potencial de viralização. A informação seria falsa ou grosseiramente imprecisa, funcionando como isca para atrair audiência, sem compromisso com a veracidade dos fatos reportados.

5.2.3 Agente Morfológico: Análise da Forma

O polo morfológico corresponde à fase em que se estabelece a configuração formal do objeto de estudo, definindo sua organização estrutural e representação (GOUVEIA, 2022). Trata-se do momento em que os elementos investigados são sistematizados em modelos, sejam teóricos ou operacionais, que permitem uma abstração coerente e mapeada da realidade. Essa etapa assegura que a pesquisa adote uma estrutura lógica, passível de replicação e validação, alinhando-se aos critérios de rigor científico.

Os modelos que são o epicentro desse polo, distinguem-se por sua função representativa e instrumental de forma física. O modelo teórico opera como um elemento que levanta hipóteses, criticismo, enquanto o modelo operacional traduz essas abstrações em procedimentos metodológicos concretos. Ambos compartilham a característica de serem simplificações controladas da realidade, mantendo consistência com o fenômeno estudado e permitindo sua análise sob parâmetros definidos (GOUVEIA, 2022). A morfologia, portanto, consolida-se como a dimensão que materializa o objeto científico, conferindo-lhe forma e aplicabilidade. Sua elaboração exige coerência interna e uma correspondência verificável com o mundo empírico (GOUVEIA, 2022). Dessa forma, o polo morfológico assegura que a investigação transite entre a teoria e a prática, garantindo que as construções abstratas mantenham relevância explicativa e capacidade de intervenção na realidade estudada.

Para ilustrar o conceito de polo morfológico de forma prática, em exemplo abstrato criado pelo autor do presente texto para fins de entendimento, imagine a tarefa de estudar a morfologia de uma “casa“ . A análise morfológica foca exclusivamente em responder à pergunta fundamental: *como é essa casa físicamente? quais são suas características?*

1. Análise da Estrutura Externa:
   Esta dimensão foca nos aspectos visuais e formais da edificação. São examinados atributos como:

   - Volumetria: A forma geral do edifício (térreo, sobrado, em L).
   - Cobertura: A tipologia do telhado (aparente com telhas cerâmicas, telhas esmaltadas, telhado verde).
   - Vedações e Revestimentos: Os materiais que constituem as fachadas (alvenaria com reboco, tijolo aparente, painéis de vidro).
   - Aberturas: A disposição, o tamanho e o material de portas e janelas.

Exemplo de descrição morfológica: “Trata-se de uma edificação térrea, com volumetria retangular e cobertura de alto padrão, caracterizada por vedações em concreto aparente e amplas aberturas em vidro.“

2. Análise da Estrutura Interna:
   Esta análise investiga a planta baixa, focando na organização e distribuição dos espaços internos. Os principais conceitos analisados são:

   - Compartimentação: A enumeração e qualificação dos ambientes (três quartos, sendo uma suíte; sala de estar; cozinha).
   - Setorização: A divisão funcional do espaço, como a segregação entre a área social (salas), a área íntima (quartos) e a área de serviço (cozinha, lavanderia).
   - Circulação: O desenho dos fluxos e percursos que conectam os diferentes setores e compartimentos (corredores, cômodos, passagens livres).

   Exemplo de descrição morfológica: “O *layout* organiza-se em conceito aberto na área social, integrando sala de estar e cozinha. A circulação para a área íntima, composta por três quartos, realiza-se por um corredor único, garantindo a divisão dos cômodos e a privacidade.“

Em síntese, o polo morfológico fornece a “cartografia“ ou a “anatomia“ do objeto de estudo. Esta descrição detalhada é fundamental, pois fornece os dados estruturais concretos sobre os quais o polo teórico poderá, posteriormente, formular interpretações.

O Agente Morfológico, no contexto de um agente investigativo de *fake news* baseado em LLM, é responsável por analisar a estrutura formal da notícia suspeita, a partir de uma entrada textual. Seu papel consiste em sistematizar os elementos constitutivos do discurso jornalístico, desvelando padrões linguísticos, narrativos e retóricos que moldam sua configuração. Essa atuação corresponde à operacionalização do polo morfológico do método quadripolar, cujo foco está na organização e representação do objeto investigado sob critérios científicos.

Ao processar a notícia, o agente dever realizar uma segmentação sintática e discursiva, identificando unidades textuais relevantes para a análise. Nessa decomposição, busca-se mapear características como a organização das informações, uso de marcadores linguísticos, presença de construções hiperbólicas e estratégias de persuasão emocional. O objetivo é formalizar a morfologia da notícia como objeto analítico, descrevendo como seus elementos internos se consolidam para formar uma estrutura coesa e, potencialmente, manipuladora.

Além da dimensão estrutural, o agente morfológico atua na identificação de traços psicológicos e afetivos impressos no texto, como o apelo ao medo, à indignação

ou à urgência. Esses elementos são frequentemente mobilizados em notícias falsas para capturar a atenção e evitar o pensamento crítico. Ao reconhecer tais marcas, o agente contribui para a abstração de modelos representativos que explicitem o funcionamento interno da desinformação.

Por fim, a análise morfológica viabiliza a geração de uma saída estruturada que descreve os padrões formais detectados na notícia. Este resultado permite a compreensão da arquitetura textual envolvida correlacionada com a replicação e validação do processo analítico por outros agentes. Com isso, o agente morfológico torna-se essencial para sustentar o rigor metodológico e ampliar a capacidade explicativa do agente frente ao fenômeno da desinformação.

A seguir, demonstra-se em procedimento de operação, o Agente mediante o *prompt* da aplicação prática na versão final e saídas correspondentes do Polo, gerado pelo modelo de LLM Gemini 2.5 Pro (GEMINI; ANIL; AL., 2025), contemplando: (i) um *prompt* final de operação do Agente Morfológico, (ii) a resposta gerada pelo modelo aplicado à uma notícia real.

Prompt:

> *Você é um especialista no Polo Morfológico do método quadripolar. Sua missão é realizar uma análise estrutural imparcial, focando no "COMO" a mensagem é construída, não no "O QUÊ" ela diz.*
>
> *O método quadripolar possui quatro polos complementares:*
>
> - *POLO EPISTEMOLÓGICO: Questiona "o quê?" - Levanta conceitos e elementos fundamentais da informação que está sendo apresentada.*
> - *POLO TEÓRICO: Questiona "por quê?" - Levanta hipóteses, sobre o por que o conteúdo é verdadeiro e o por que seria falso.*
> - *POLO MORFOLÓGICO: Questiona "como?" - Analisa a estrutura morfológica do conteúdo apresentado, levantando características formais e organizacionais. (Você está aqui)*
> - *POLO TÉCNICO: Questiona "com quê?" - Busca informações externas na web para analisar e validar se determinado conteúdo é verdadeiro ou falso.*
>
> *Analise o conteúdo fornecido e identifique as características formais, retóricas e discursivas. Organize sua análise em um balanço de pontos positivos (que reforçam a credibilidade estrutural) e negativos (sinais de alerta morfológico).*
>
> *Diretrizes da Análise (Seu Raciocínio Interno):*
>
> - *Arquitetura: Avalie a estrutura (título, lead, corpo, conclusão). Segue um padrão jornalístico (pirâmide invertida) ou busca o sensacionalismo?*

- *Linguagem e Tom: O tom é neutro e informativo ou emocional e alarmista? Há uso de vocabulário sensacionalista, jargões ou generalizações?*
- *Retórica e Persuasão: Identifique apelos. A argumentação é baseada em dados e lógica ou em emoção (medo, raiva, indignação)? Existem hipérboles, perguntas retóricas ou ironias?*
- *Fontes e Evidências: Como as fontes são apresentadas? São específicas e verificáveis ("segundo o relatório X do IBGE") ou vagas e anônimas ("especialistas afirmam", "estudos mostram")?*
- *Subjetividade (Opinião vs. Fato): Detecte opinião disfarçada. Há adjetivos valorativos ("péssimo", "incrível"), advérbios de certeza ("obviamente", "claramente") e previsões apresentadas como fatos consumados?*

Responda de forma direta, sem textos adicionais, exatamente no formato:

Análise Morfológica - Pontos de Credibilidade Estrutural (Sinais Positivos):

- *(Liste 2 a 5 características que alinham o texto com práticas de comunicação informativa e factual, com trechos específicos quando possível, máximo de 60 tokens para cada item)*

Análise Morfológica - Indicadores de Risco Morfológico (Sinais de Alerta):

- *(Liste 2 a 5 "bandeiras vermelhas" que sugerem sensacionalismo, manipulação, viés opinativo ou clickbait, com trechos específicos quando possível, máximo de 60 tokens para cada item)*

Diagnóstico Morfológico:
Classifique como: Predominantemente Factual/Jornalístico | Híbrido (Informativo com Viés Opinativo) | Fortemente Opinativo/Persuasivo | Potencial Clickbait/Sensacionalista - justifique brevemente, maximo de 60 tokens

CONTEÚDO PARA ANÁLISE MORFOLÓGICA:

Notícia para análise:

[INSIRA AQUI O TEXTO DA NOTÍCIA]

Reposta gerada utilizando a notícia do Exemplo 1:

Análise Morfológica - Pontos de Credibilidade Estrutural (Sinais Positivos):

- *Título direto que declara um fato objetivo (aprovação na Câmara).*
- *Identificação explícita e verificável dos atores envolvidos ("Câmara", "Deputados").*

- *Uso de dado numérico específico e modalidade (“R$ 5 mil”, “por unanimidade”).*
- *Estrutura clássica de manchete seguida por um lide explicativo.*

Análise Morfológica - Indicadores de Risco Morfológico (Sinais de Alerta):

- *Título convida à curiosidade (“saiba o que acontece agora”), sugerindo uma urgência.*
- *A menção “projeto” é genérica, sem especificar o número ou autoria.*
- *Foco inicial direto no benefício monetário pode evocar impacto emocional imediato.*

Diagnóstico Morfológico:
Predominantemente Factual/Jornalístico; o texto reporta um fato com elementos de engajamento, mas mantém a objetividade estrutural.

5.2.4 Agente Técnico: Busca de Evidências

O polo técnico constitui um dos eixos centrais no processo de investigação científica, operacionalizando a relação entre a fundamentação teórica e a prática empírica. Definido por (BRUYNE; HERMAN; SCHOUTHEETE, 1977) como o espaço que materializa as dimensões da pesquisa por meio de variáveis, este polo é responsável pela coleta, tratamento e transformação de dados, garantindo que estes informem a problemática de estudo de modo rigoroso e preciso. O seu objetivo primordial é assegurar que os procedimentos metodológicos adotados, desde a seleção de técnicas até à aplicação de tecnologias, estejam alinhados com os referenciais teóricos e contextuais estabelecidos nos demais polos, sem, contudo, confundir precisão técnica com exatidão científica, uma vez que a mera aplicação de métodos não garante, por si só, a validade do processo investigativo.

Como pivô da relação com os outros polos, o polo técnico assume um papel mediador, traduzindo questões teóricas em estratégias empíricas viáveis. Segundo o texto de (GOUVEIA, 2022), este estágio da investigação exige a definição clara de o quê e como recolher dados, selecionando instrumentos e procedimentos adequados como questionários, entrevistas, observações ou pesquisas documentais que assegurem a robustez dos resultados. A sua atuação abrange tanto abordagens quantitativas (análises estatísticas) quanto qualitativas (análise de conteúdo), podendo ainda integrar métodos mistos para captar a complexidade dos fenomenos estudados.

A evolução tecnológica tem ampliado as possibilidades do polo técnico, permitindo a codificação de dados qualitativos e sua representação através de ferramentas estatísticas (GOUVEIA, 2022), diluindo fronteiras entre paradigmas. No entanto, mantém-se a distinção epistemológica entre as naturezas dos dados, exigindo do pesquisador discernimento na escolha e aplicação das técnicas. Seja através de *surveys*, estudos de caso ou

pesquisa-ação, o polo técnico visa assegurar que a coleta e análise de dados sustentem as hipóteses iniciais, validando os esforços dos polos teórico e epistemológico. Dessa forma, consolida-se como um elemento indispensável para a credibilidade e replicabilidade da investigação científica.

O Agente Técnico tem como principal função a realização de buscas em fontes de dados externas, com ênfase na obtenção de informações provenientes da *Web*. Para viabilizar esse processo, foi utilizada a *Google Search API* (GOOGLE, 2024b), permitindo a integração da ferramenta ao fluxo de execução do agente. Essa integração possibilita a recuperação de documentos relevantes relacionados ao dado de entrada analisado.

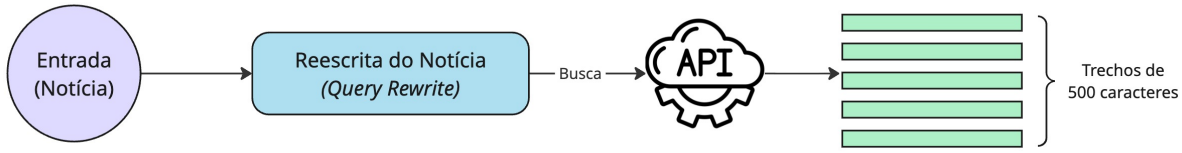
Como etapa fundamental desse processo, a consulta original (*input*/Entrada) foi reescrita com o objetivo de refinar os elementos textuais e adaptá-los ao formato mais eficaz para buscas online. Essa reescrita da *query* é orientada à extração de informações no contexto específico de notícias, de forma a maximizar a pertinência dos resultados obtidos e garantir que as fontes consultadas estejam relacionadas ao fenômeno investigado.

Após a execução da busca, os documentos retornados são convertidos em representações vetoriais por meio da técnica de *embeddings*. Esses vetores densos foram armazenados em um banco vetorial, permitindo sua posterior reutilização de forma eficiente. Essa etapa é essencial para garantir que as informações capturadas na web possam ser utilizadas de maneira estruturada nas etapas seguintes do fluxo do agente.

A estrutura vetorial dos dados coletados permitirá, na etapa seguinte, a correlação semântica entre as informações provenientes da web e os elementos da e original. Essa correlação contribuirá diretamente para o processo decisório do agente, especialmente no que se refere à análise, explicação e eventual classificação de conteúdos como verdadeiros ou falsos no contexto da detecção de *fake news*.

A Figura 29 ilustra o fluxo de funcionamento do Agente Técnico. O processo inicia-se com o recebimento de um dado/notícia de entrada , seguido imediatamente pela reescrita da *query*. Essa *query* reformulada é utilizada para realizar a busca na *web* por meio da API integrada, retornando documentos relevantes e armazenados em trechos de 500 caracteres. Em seguida, os conteúdos encontrados foram reranqueados pelo modelo de Reranqueamento, na etapa do Agente de Veredito.

Figura 29 – Processo de Consulta e Armazenamento Vetorial do Agente Técnico via Ferramenta de API de Busca Externa



Fonte: Do Autor

A seguir, demonstra-se em procedimento de operação, o Agente mediante *prompt* oficial, aplicação prática na versão final do Agente Técnico, contemplando: (i) uma entrada de notícia real para o Agente Técnico, (ii) um exemplo de processamento internamente para gerar novas *querys*, e (iii) a respectiva saída oficial gerada.

Entrada:

> *Governo Lula anuncia novo imposto de 1.5% sobre todas as transações via PIX a partir de Julho de 2025.*

Prompt:

> *Você é um especialista em SEO (Search Engine Optimization) e otimização de mecanismos de busca. Sua tarefa é analisar a notícia fornecida e destilar sua essência na query de busca mais eficiente e relevante possível para o Google.*
>
> *Objetivo: Construir uma string de busca que encontre rapidamente outras fontes sobre o mesmo fato.*
>
> *Diretrizes de Extração:*
>
> *1. Identifique as Entidades e Conceitos-Chave: Foque nos substantivos próprios (pessoas, organizações, locais específicos) e nos conceitos técnicos ou centrais da notícia. Pense em "Quem?", "O quê?" e "Onde?".*
>
> *2. Combine Conceitos: Agrupe palavras que formam um único conceito. Por exemplo, em vez de "usina solar flutuante", trate usina solar flutuante como um único bloco de sentido.*
>
> *3. Seja Específico e Conciso: Priorize os termos mais únicos e descritivos da notícia. Remova palavras genéricas e de preenchimento (artigos, preposições, verbos).*
>
> *Limite de Palavras: A query final deve conter entre 7 e 15 palavras no total.*
>
> *Formato de Saída:*

Retorne APENAS a sequência de palavras-chave, separadas por um único espaço, sem nenhuma pontuação ou texto adicional.

Exemplo de Execução:

Notícia (resumo): "A Petrobras anunciou hoje a inauguração de uma nova usina de energia solar flutuante no reservatório de Billings, em São Paulo, como parte de seu plano de descarbonização."

Saída Esperada: Petrobras usina solar flutuante reservatório Billings São Paulo descarbonização

Processamento Interno do Agente:

1. **Reescrita da Query para Busca:** O Agente recebe a notícia e, para otimizar a busca, extrai as entidades e conceitos-chave, gerando consultas neutras e eficientes para a API de busca. As queries geradas poderiam ser:
    - *"novo imposto PIX governo federal 2025"*
    - *"taxação PIX Fernando Haddad"*
    - *"lei imposto transações PIX Julho 2025"*
2. **Execução da Busca:** O Agente utiliza a *query* mais promissora para consultar a *Google Search API*.
3. **Coleta e Estruturação:** O Agente coleta os resultados mais relevantes (os 3 a 5 primeiros resultados de fontes de alta autoridade) e os estrutura em um formato padronizado.

Saída:

O Agente Técnico não emite um julgamento, mas entrega um conjunto de evidências brutas, já estruturadas, para serem usadas pelos próximos estágios do agente. A saída seria um objeto de dados contendo os seguintes documentos:

Documento 1:

- *Fonte: g1.globo.com/economia*
- *Título: É #FATO ou #FAKE? Governo nega criação de novo imposto sobre o PIX*
- *Conteúdo: "Circula novamente nas redes sociais uma mensagem afirmando que o governo federal irá taxar o PIX. A informação é falsa. Em nota, o Ministério da Fazenda reafirmou nesta semana que não há qualquer estudo ou proposta em andamento para a criação de um imposto sobre transações via PIX..."*

Documento 2:

- *Fonte: www.gov.br/fazenda/pt-br/assuntos/noticias*
- *Título:* Nota de Esclarecimento: Imposto sobre o PIX
- *Conteúdo: "O Ministério da Fazenda vem a público desmentir categoricamente as informações falsas que circulam sobre uma suposta taxação do PIX. O PIX é e continuará sendo gratuito para pessoas físicas, conforme as diretrizes do Banco Central do Brasil. Medidas de arrecadação do governo estão sendo discutidas de forma transparente e não incluem esta modalidade."*

Documento 3:

- *Fonte: agencialupa.com.br*
- *Título: LupaVerifica: É falso que governo Lula vai taxar o PIX em 1.5% a partir de julho*
- *Conteúdo: "A Agência Lupa verificou e constatou ser falsa a corrente que circula no WhatsApp e Telegram. Trata-se de um boato recorrente, que já foi desmentido em 2023 e 2024 pelas autoridades monetárias e pelo próprio Ministro da Fazenda, Fernando Haddad."*

5.2.5 Agente de Veredito: Síntese Explicativa

O agente denominado Agente de Veredito tem como função principal a consolidação e análise integrada das saídas provenientes dos demais agentes do sistema. Seu propósito é oferecer um julgamento final sobre a veracidade de uma determinada notícia, fundamentando-se em uma síntese crítica das evidências extraídas de diferentes polos analíticos (técnico, teórico, epistemológico e morfológico).

Os dados obtidos pelo Agente Técnico, que foram previamente armazenados em uma estrutura vetorial FAISS, servem como base para uma busca semântica robusta. Essa busca é conduzida a partir de duas fontes principais: a saída gerada pela Agente Teórico e a própria *query* original submetida pelo usuário. Essa combinação visa maximizar a precisão da recuperação de trechos textuais que possam corroborar ou refutar a informação analisada.

Após a recuperação dos documentos relevantes, aplica-se uma etapa de reranqueamento semântico, responsável por reorganizar os resultados com base em sua similaridade semântica em relação à *query* de entrada. Essa técnica assegura que os documentos mais pertinentes, em termos de evidência para verificação da notícia, estejam priorizados para análise posterior, com base em seus respectivos scores de relevância.

O modelo de LLM recebe como entrada um conjunto integrado de informações: a saída da Agente Teórico, preservado em sua forma original; a saída do Agente Morfológico; a saída do Agente Epistemológico; a *query* de entrada; e os documentos mais relevantes obtidos por meio da busca semântica e reranqueamento. Essa estrutura de entrada permite uma contextualização ampla e multilateral do problema.

A etapa de inferência realizada pelo LLM tem como objetivo elaborar um julgamento fundamentado acerca da veracidade do conteúdo analisado. A análise é conduzida com base em todos os polos do método quadripolar, sendo cada um considerado em sua função explicativa e epistemológica. De forma geral, o LLM é instruído a classificar o conteúdo como verdadeiro ou falso e também a explicitar os fundamentos dessa decisão.

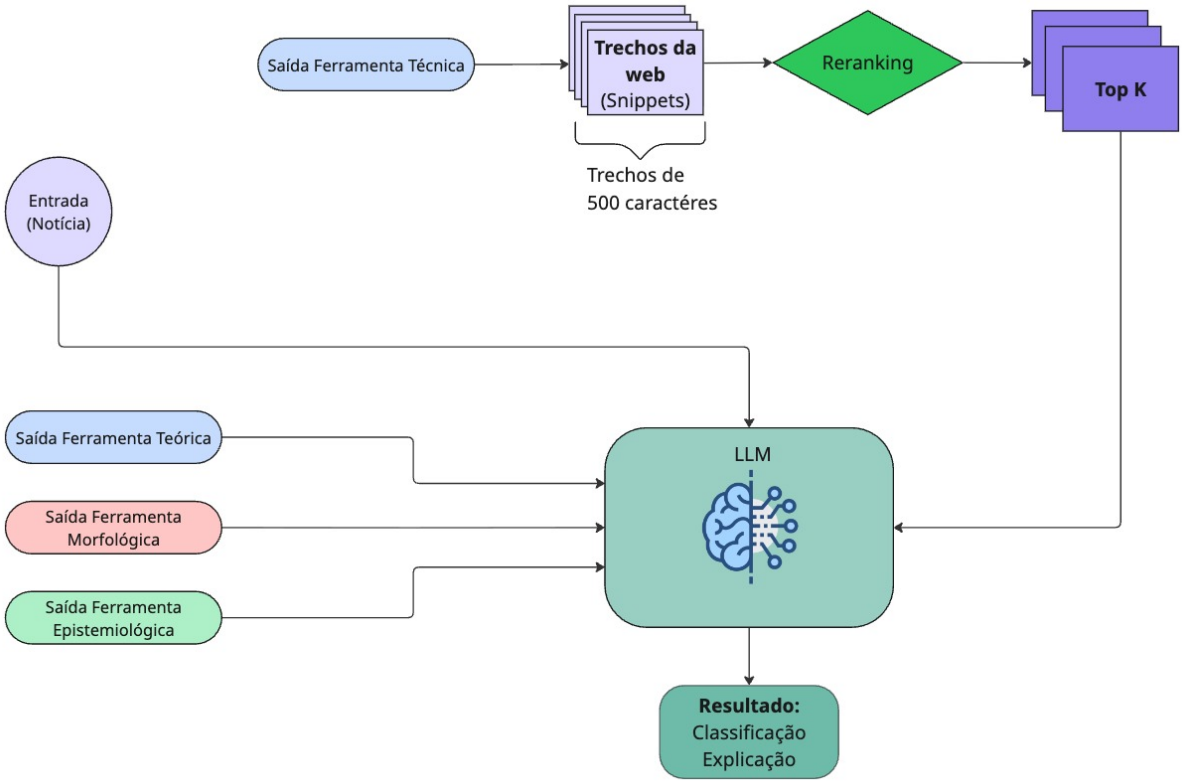
Por fim, a saída do Agente de Veredito deve apresentar uma conclusão analítica clara e justificável, categorizando a notícia como *fake news* ou não. A explicação fornecida deve ser estruturada com base em evidências concretas oriundas de cada um dos polos envolvidos, garantindo um nível elevado de transparência e rastreabilidade na decisão, além de reforçar o compromisso do sistema com a explicabilidade dos processos de inferência automatizada.

A Figura 30 apresenta o fluxo arquitetural do Agente de Veredito, ilustrando a integração dos diferentes agentes da arquitetura e seu papel na construção do veredito final. Inicialmente, realiza-se uma busca na web a partir da saída do Agente Técnico e da *query* de entrada. Os documentos retornados são armazenados em memória em tamanho de 500 para cada trecho e, em seguida, passam por uma etapa de reranqueamento semântico pelo modelo `bge-reranker-large` (CHEN *et al.*, 2024), que reordena os trechos segundo sua similaridade contextual com a *query* analisada.

Os documentos reranqueados, agora priorizados em termos de significância semântica, são combinados com as saídas dos demais agentes do método quadripolar bem como com a *query* original, compondo o conjunto de entrada do modelo de LLM. O LLM processa esse conjunto multievidencial de informações para realizar uma inferência fundamentada, capaz de determinar a veracidade da informação em questão.

O resultado gerado por essa inferência é composto por uma classificação (indicando se o conteúdo é falso ou verdadeiro) e por uma explicação estruturada, que apresenta os critérios e evidências utilizadas na tomada de decisão, ancoradas em cada um dos polos do método e justificadas pelo próprio LLM. A Figura 30, portanto, sintetiza a lógica operacional do Agente de Veredito, demonstrando seu funcionamento sistêmico e a interdependência entre os componentes para garantir um julgamento explicável e confiável quanto à veracidade de conteúdos informacionais.

Figura 30 – Fluxo de Processamento do Agente de Veredito para Classificação, Descrição e Nivelamento da Desinformação



Fonte: Do Autor

Ao final do processo conduzido pela Agente de Veredito, espera-se a geração de um resultado compreensivo que contemple três elementos essenciais: a classificação da informação analisada (indicando se se trata ou não de uma *fake news*), uma descrição explicativa fundamentada nas saídas dos polos teórico, técnico, morfológico e epistemológico, e um nivelamento da desinformação, que busca mensurar o grau de distorção ou manipulação presente no conteúdo. Essa abordagem visa fornecer um veredito binário, trazendo conjunto uma compreensão qualificada e interpretável sobre os elementos que compõem a desinformação, contribuindo para uma tomada de decisão mais crítica e informada por parte dos usuários e pesquisadores.

A seguir, demonstra-se em procedimento detalhado, o agente mediante *prompt* oficial, a aplicação prática na versão final e as saídas correspondentes do agente, contemplando: (i) um *prompt* de operação da versão oficial do Agente de Veredito, (ii) uma entrada consolidada aplicada à uma notícia real que foi apresentada ao agente, acompanhada dos elementos dos Agentes anteriores (Teórico, Morfológico, Epistemológico) e (iii) a respectiva saída gerada. Esta tripla demonstração visa elucidar o funcionamento empírico da solução proposta.

Prompt:

Você é um especialista em inteligência investigativa focado na detecção, análise crítica e desmistificação de desinformação e golpes cibernéticos. Avalie a veracidade de conteúdos informativos com máxima precisão, objetividade e ceticismo metodológico.

RESPONSABILIDADES: - Examine intrinsecamente cada detalhe do conteúdo - Identifique e avalie lacunas informacionais - Aplique rigorosamente o Método Quadripolar (4 polos) - Correlacione informações externas disponíveis - Emita veredito fundamentado e classifique tipos de fake news quando aplicável

MÉTODO QUADRIPOLAR - Base toda a análise nestes quatro polos:

1. POLO EPISTEMOLÓGICO (Origem do Conhecimento) Quem é a fonte? Qual sua credibilidade e histórico? A informação é primária (testemunho direto, dados originais) ou secundária (interpretação)? Quais evidências sustentam a alegação?

2. POLO MORFOLÓGICO (Forma da Mensagem) Qual a estrutura da notícia? A linguagem é neutra ou emocional/sensacionalista? O título corresponde ao corpo? Há erros de gramática ou formatação? O layout visual parece profissional ou amador?

3. POLO TEÓRICO (Contexto e Coerência) A alegação se encaixa no conhecimento estabelecido pelas hipóteses? O Fato apresentado tem mais similaridade com teorias que sustentam a veracidade ou a falsidade? Por que?

4. POLO TÉCNICO (Verificação Externa) Existe fonte confiável que comprove a veracidade? O que outras fontes dizem? É verificável em agências de checagem, publicações científicas ou mídia de renome? Busca reversa de imagens revela contexto real? O domínio é conhecido por disseminar desinformação?

TAXONOMIA DE FAKE NEWS

DESINFORMAÇÃO: Criação e partilha INTENCIONAL de informações falsas para enganar, manipular ou causar danos (ganhos políticos/financeiros, exploração de emoções).

MISINFORMAÇÃO: Partilha de informações falsas SEM intenção de enganar (erro, falta de verificação). Não maliciosa, mas contribui para desordem informativa.

CLICKBAIT: Títulos/imagens sensacionalistas ou enganosos para atrair cliques e gerar tráfego (receita publicitária). Conteúdo raramente corresponde ao prometido.

GOLPE: Fraude usando engenharia social para manipular vítimas (urgência, oportunidades únicas) visando dados confidenciais, transferências financeiras ou ações prejudiciais.

RUMOR: Informação especulativa e não verificada que circula amplamente. Surge em contextos de incerteza, sem fonte clara ou confiável.

OBSERVAÇÃO: Notícia PARCIALMENTE FALSA mistura informações verdadeiras com elementos falsos/enganosos, apresentando dados reais de forma distorcida, incompleta ou fora de contexto para manipular a percepção.

INSTRUÇÕES DE ANÁLISE

INPUTS RECEBIDOS: Você receberá até 5 inputs. Correlacione todas as informações disponíveis. Se inputs 2-5 não forem fornecidos, conduza essa análise autonomamente com base na notícia.

1. Notícia Investigada: Texto, imagem ou link a ser analisado 2. Input Epistemológico: Análise prévia sobre conceitos-chave, objetos, pessoas 3. Input Morfológico: Análise prévia sobre forma e linguagem 4. Input Teórico: Hipóteses sobre veracidade/falsidade do contexto e coerência 5. Input Técnico: Evidências e links de fontes oficiais ou imprensa reconhecida

PROCESSO DE DECISÃO: Pense estrategicamente antes de decidir. Relacione todas as informações de forma minuciosa. NÃO ALUCINE - baseie-se apenas em evidências concretas.

FORMATO DE SAÍDA (MANTENHA ESTE FORMATO RIGOROSAMENTE)

- Veredito de Veracidade da Notícia: [Escolha: "Verdadeira"OU "Falsa"]

- Tipo do Conteúdo: [Escolha: "Factual/Jornalístico", "Ironia"OU "Opinião"]

- Tipo de Fake News: [Se "Verdadeira", deixe VAZIO. Se "Falsa", escolha: Desinformação, Misinformação, Clickbait, Rumor OU Golpe]

- Diagnóstico do Agente: [Máximo 300 tokens. Justifique seu veredito de forma crítica, concisa e objetiva. Sintetize argumentos dos 4 polos para construir explicação clara, apontando red flags ou evidências de veracidade]

DADOS DA ANÁLISE

Data de Hoje: (data de hoje)

- Notícia: (notícia)

- Input Epistemológico: (análise epistemiológica)

- Input Morfológico: (análise morfológica)

- Input Teórico: (análise teórica)

- Input Técnico: (análise técnica)

EXEMPLOS DE RESPOSTA (few-shot)

Exemplo 1....

Exemplo 2....

Entrada:

Câmara aprova isenção do IR até R$ 5 mil: saiba o que acontece agora Deputados aprovaram por unanimidade projeto que reformula tabela do imposto de renda

Saída

- Veredito de Veracidade da Notícia: Verdadeira

- Tipo do Conteúdo: Factual/Jornalístico

- Tipo de Fake News:

*- Diagnóstico do Agente: A notícia é **verdadeira**, pois o fato central — a aprovação do texto-base do PL 1.087/2025 pela Câmara dos Deputados — é um evento factual e verificável. **Polo Epistemológico/Técnico:** As fontes (Agência Brasil, republicado pela CUT) confirmam a votação unânime (493x0), ancorando a informação em um ato legislativo documentado e recente (01/10/2025). Contudo, a análise revela um enquadramento que beira a manipulação por omissão de contexto. **Polo Morfológico:** O título usa uma chamada para ação ("saiba o que acontece agora") que cria uma lacuna de curiosidade, uma técnica típica de engajamento. **Polo Teórico:** A principal ressalva reside na teoria do enquadramento (framing). A manchete, ao anunciar "Câmara aprova", pode induzir o leitor a uma percepção de finalidade que não corresponde à realidade. Como confirmado pelas fontes, a aprovação na Câmara é apenas a primeira etapa do processo bicameral, sendo ainda necessária a votação no Senado e a sanção presidencial. A alegação é factual, mas seu enquadramento é incompleto e pode gerar desordem informativa se consumido isoladamente.*

5.3 CONJUNTO DE DADOS

A validação experimental desta pesquisa fundamenta-se na utilização de três conjuntos de dados distintos, selecionados para abranger tanto as especificidades do ecossistema informacional brasileiro quanto o cenário global de desinformação. Foram adotados o *FakeRecogna*, composto por textos em português, um conjunto de dados estruturado e curado manualmente pelo autor do presente texto.

5.3.1 *FakeRecogna*

O *FakeRecogna* (GARCIA; AFONSO; PAPA, 2022) representa um *corpus* balanceado e direcionado ao contexto brasileiro, totalizando 11.902 artigos. O conjunto é dividido equitativamente, com 5.951 amostras de notícias falsas provenientes de agências de checagem como Boatos.org e Fato ou *Fake* (*https://g1.globo.com/fato-ou-fake/*), e igual número de notícias verdadeiras coletadas de fontes de alta credibilidade, incluindo G1 (*https://g1.globo.com*), UOL (*https://www.uol.com.br/*) e o portal oficial do Ministério da

Saúde. Uma característica técnica distintiva deste *dataset*, que impacta diretamente as estratégias de processamento de linguagem natural, é que o conteúdo textual das notícias não se apresenta em seu formato bruto original; os textos foram submetidos a etapas de pré-processamento e truncamento. Esta particularidade exige que os modelos de classificação, incluindo o agente proposto, sejam capazes de inferir veracidade e contexto mesmo diante de fragmentos de informação ou textos com estrutura sintática simplificada.

5.3.2 Conjunto de Dados Próprio

Visando mitigar o risco de contaminação de dados, fenômeno onde o modelo já possui conhecimento prévio das amostras de teste por tê-las visto durante seu pré-treinamento, e assegurar a avaliação da capacidade de generalização do agente em cenários temporais futuros, foi desenvolvido um conjunto de dados de curadoria própria. Este conjunto de dados foi desenhado para simular um ambiente de inferência em tempo real, com fatos situados no recorte temporal de setembro e outubro de 2025.

A composição do *dataset* obedeceu a um critério de balanceamento estrito entre as classes, totalizando 250 amostras, divididas da seguinte forma:

- **Classe Verdadeira (n=125):** Constituída por notícias reais extraídas de veículos de imprensa de alta credibilidade e circulação nacional, nomeadamente CNN Brasil, UOL, Grupo Globo e Folha de S.Paulo. A seleção priorizou fatos ocorridos no bimestre de referência (setembro/outubro de 2025) para garantir a atualidade e a complexidade contextual da verificação.
- **Classe Falsa (n=125):** Composta por notícias falsas elaboradas manualmente pelo pesquisador. O LLM foi utilizado como ferramenta auxiliar para geração de temas macro, sendo as notícias posteriormente criadas e modificadas pelo autor para incorporar distorções factuais. As amostras foram construídas para reproduzir a estrutura linguística e temática das notícias verdadeiras, contendo, contudo, distorções factuais, fabricações ou descontextualizações deliberadas.

A utilização deste conjunto manual justifica-se pela necessidade de validar o método quadripolar em um cenário controlado, onde a veracidade das informações é inequivocamente conhecida pelo pesquisador, permitindo uma aferição precisa das métricas de desempenho (Acurácia, F1-Score) e da qualidade das explicações geradas.

5.4 REPOSITÓRIO

Os artefatos de *software* desenvolvidos nesta dissertação, incluindo os *scripts* de execução (*runners*), as bibliotecas, *prompts* oficiais e os arquivos de configuração da

arquitetura quadripolar, e todos os resultados utilizados nesta pesquisa, encontram-se disponíveis no seguinte endereço:

URL: https://github.com/thehenke/agent-ai-fake-news

5.5 ESTRUTURA TECNOLÓGICA

A operacionalização da arquitetura quadripolar e de seus agentes interdependentes demanda uma infraestrutura tecnológica robusta e coesa. A materialização das tarefas analíticas complexas, como a análise semântica, a busca em fontes externas e a gestão de dados em larga escala, depende diretamente das tecnologias de base selecionadas. Esta seção detalha os componentes de *software* e os modelos que foram escolhidos para sustentar o desenvolvimento, o treinamento e a execução do sistema proposto, abrangendo desde a linguagem de programação e o ambiente de desenvolvimento até o modelo de geração de *embeddings* e a solução de banco de dados vetorial. A escolha criteriosa desses elementos é fundamental para garantir a viabilidade de implementação do agente, e também sua eficiência computacional, escalabilidade e a precisão nas tarefas de Processamento de Linguagem Natural (PLN) que constituem o epicentro desta pesquisa.

Para organizar e processar os dados, realizar a tokenização e gerenciar os modelos, foi utilizada a linguagem *Python 3.12* (ROSSUM; DRAKE, 2009). *Python* oferece as ferramentas e bibliotecas necessárias para manipulação de dados, preparação de dados textuais e implementação de modelos de aprendizado de máquina, tornando-o ideal para esse projeto.

O ambiente de execução escolhido é o *Google Colaboratory* (GOOGLE, 2024a), que oferece suporte direto para *Python* e uma infraestrutura de fácil acesso para experimentação em nuvem ou máquinas virtuais, que podem se fazer necessárias para a utilização de processamento em GPU. Todos os artefatos necessários para a execução do projeto, incluindo estruturas de dados para armazenamento de perguntas e objetos auxiliares, estarão organizados e implementados nesse ambiente, facilitando o desenvolvimento e a reutilização de código.

5.6 MODELO DE RERANQUEAMENTO (*RERANKER*)

Embora a utilização de ferramentas de busca na *web* (via *Search API*) seja fundamental para conferir ao agente acesso a informações atualizadas e externas ao seu treinamento, os resultados retornados por estes motores de busca genéricos muitas vezes priorizam métricas de correspondência de palavras-chave ou popularidade (SEO), podendo trazer ruído ou conteúdos apenas tangencialmente relevantes. Para mitigar essa limitação e refinar a seleção de evidências, este trabalho adota o modelo `bge-reranker-large` (CHEN

et al., 2024). Nesta arquitetura, o *reranker* atua como um filtro discriminativo de alta precisão aplicado imediatamente após a etapa de coleta na *web*.

O `bge-reranker-large` opera sob uma arquitetura de *Cross-Encoder*. Ao contrário de abordagens que comparam vetores pré-calculados (*Bi-Encoders*), o *Cross-Encoder* recebe como entrada o par formado pela consulta do usuário e o conteúdo do resultado da busca (título e fragmento/*snippet*) (BAŞ *et al.*, 2022), concatenados em uma única sequência de entrada. Esta estrutura permite que o mecanismo de autoatenção (*self-attention*) do *Transformer* analise a interação entre cada termo da pergunta e cada termo da evidência recuperada, capturando nuances semânticas e dependências lógicas que motores de busca tradicionais podem perder.

No fluxo do agente, o funcionamento é o seguinte: a *Search API* retorna um conjunto inicial amplo de k documentos ou páginas candidatas. O modelo `bge-reranker-large` processa cada um desses candidatos em relação à pergunta original, atribuindo um *score* de relevância escalar. Os documentos são então reordenados com base nessa pontuação, e apenas os mais relevantes (top-k) são selecionados para compor o contexto final do LLM. Este processo é vital para otimizar a janela de contexto do agente, garantindo que o raciocínio de verificação seja baseado apenas em informações de alta qualidade factual, descartando ruído e reduzindo a probabilidade de alucinações.

5.7 MODELOS DE LINGUAGEM DE LARGA ESCALA PARA *BENCHMARK*

Para garantir que a avaliação da arquitetura quadripolar seja rigorosa e relevante frente ao avanço contínuo da área, a comparação com o “LLM de base“ foi realizada utilizando um conjunto diversificado dos mais capazes e atuais modelos de linguagem disponíveis. A inclusão destes modelos como *benchmarks* diretos é fundamental para testar as hipóteses H1 e H2 contra o estado da arte da capacidade de raciocínio generalista. A seleção abrange modelos de diferentes provedores e com distintos perfis de performance e custo, permitindo uma análise comparativa robusta sobre a eficácia da estrutura quadripolar. Cada modelo foi avaliado em sua configuração base (C1), conforme detalhado nas seções a seguir.

5.7.1 OpenAI GPT 5

O modelo GPT-5, sucessor direto da arquitetura GPT-4 (OpenAI, 2023), foi utilizado como o *benchmark* de mais alta capacidade de raciocínio da OpenAI. Este modelo é caracterizado por suas melhorias em inferência lógica, compreensão de nuances complexas e um vasto conhecimento de mundo atualizado, representando o limite superior de desempenho de um modelo de propósito geral em tarefas que exigem análise crítica profunda. A avaliação contra o `gpt-5-mini` servirá para determinar se a estrutura metodológica da

arquitetura quadripolar consegue superar ou igualar a performance de um LLM de ponta operando em sua máxima capacidade de raciocínio intrínseco (OpenAI, 2023).

5.7.2 OpenAI GPT-5 mini

O `gpt-5-mini` representa a geração de modelos da OpenAI otimizados para eficiência (OpenAI, 2023). Sua inclusão como *benchmark* é estratégica para avaliar como a arquitetura quadripolar se compara a um modelo de alta performance que já é otimizado para aplicações práticas em larga escala, testando se a eficiência de custo e latência do `gpt-5-mini` se traduz em uma análise de veracidade e explicação de qualidade comparável.

5.7.3 Google Gemini 2.5 Pro

Como o principal modelo da família Gemini, o `gemini-2.5-pro` foi o *benchmark* representante da arquitetura de ponta do Google (GEMINI; ANIL; AL., 2025). A comparação com o `gemini-2.5-pro` é fundamental para avaliar o desempenho da arquitetura quadripolar contra um modelo que possui uma capacidade massiva de processar informação contextual, servindo como um bom suporte à abordagem de RAG do Agente Técnico.

5.7.4 Google Gemini 2.5 Flash

O `gemini-2.5-flash` é a variante de alta velocidade e menor custo do modelo, projetado especificamente para tarefas que demandam baixa latência e alta vazão (GEMINI; ANIL; AL., 2025). Ao incluí-lo, a pesquisa pôde avaliar o *trade-off* entre eficiência computacional e a qualidade da análise. A performance e a explicação do `gemini-2.5-flash` serviram como um *benchmark* para o “mínimo aceitável“ de um modelo de ponta, permitindo analisar se a estrutura quadripolar consegue extrair um desempenho superior mesmo quando comparada a um LLM projetado para ser rápido e econômico, em vez de exaustivamente profundo em seu raciocínio.

5.8 PROTOCOLO DE AVALIAÇÃO EXPERIMENTAL

Para a avaliação da estrutura do agente, foi feita uma composição de testes, utilizando um conjunto de dados previamente definido sobre abordagem de *fake news*. Também foi utilizado um conjunto de notícias coletadas de maneira manual, e também a criação de algumas notícias falsas para avaliar a capacidade de detecção negativa diante dos resultados do agente.

5.8.1 Avaliação Quantitativa

Transcendendo a análise puramente quantitativa de desempenho, o diferencial metodológico deste trabalho reside na capacidade do agente de gerar explicações fun-

damentadas para seus vereditos. Nesse sentido, torna-se imperativo avaliar se o agente acerta em sua classificação, e principalmente como e por que ele justifica sua decisão. Métricas tradicionais como acurácia e *F1-Score* são insuficientes para capturar a qualidade, a coerência e a profundidade dessa justificativa. Para suprir essa lacuna, esta pesquisa propõe um *Framework* de Análise Explicativa, uma metodologia estruturada com critérios qualitativos e quantitativos para a avaliação sistemática da explicação gerada. A presente seção detalha os componentes deste *framework*, que constitui a principal ferramenta para validar a contribuição central desta tese, sendo a geração de diagnósticos informacionais rastreáveis e metodologicamente robustos.

Avaliação Cruzada com Variação de Presença dos Agentes

Com o objetivo de avaliar a contribuição individual e combinada dos diferentes Agentes inspirados no Método Quadripolar, propõe-se uma abordagem de análise cruzada fundamentada na ativação seletiva de cada componente no processo decisório do agente LLM. A estrutura é constituída por quatro agentes principais: teórica, epistemológica e técnica e morfológica, cada um representando um dos polos metodológicos do modelo e oferecendo aportes distintos na tarefa de análise crítica de possíveis notícias falsas.

A proposta metodológica consiste em submeter a estrutura a diferentes cenários experimentais, nos quais a configuração dos Agentes ativos é sistematicamente variada. Essa abordagem permite mensurar o impacto individual de cada agente, tanto de forma isolada quanto em combinação, sobre critérios como acurácia, coerência argumentativa e robustez analítica. Para operacionalizar essa avaliação, define-se a presença de um Agente por meio do valor binário 1 (ativa) e sua ausência por 0 (inativa), o que gera um conjunto de 7 combinações possíveis.

A Tabela 5 apresenta todas as combinações de ativação dos agentes que compõem o experimento. Cada configuração (C) foi analisada de maneira sistemática, com base em métricas quantitativas (como precisão, *recall* e F1-score). Os resultados esperados contribuirão para uma compreensão mais aprofundada do papel de cada polo na construção do agente e ofereceram subsídios para o aprimoramento técnico e teórico da solução proposta. As configurações C6 e C7 são compostas diretamente por todo contexto quadripolar, a diferença é que a C6 é desacoplada, na qual cada agente tem sua inferencia, já a C7 tem a inferência em um único agente.

Tabela 5 – Configurações de Combinação de Ativação dos Agentes

Configuração	Morfológico	Epistemológico	Técnico	Teórico
C1	0	0	0	0
C2	1	0	0	0
C3	0	1	0	0
C4	0	0	1	0
C5	0	0	0	1
C6	1	1	1	1
C7	1	1	1	1

Fonte: Do Autor.

Essa matriz experimental foi empregada para avaliar o desempenho da arquitetura sob cada configuração, mediante métricas quantitativas (*recall*, F1-score e matriz de confusão) e qualitativas (análise da argumentação gerada). Os resultados permitem compreender a função de cada polo no sistema automatizado de detecção de *fake news* e subsidiam o aprimoramento da arquitetura.

Métricas Quantitativas

A performance do agente na tarefa de classificação foi aferida por meio de um conjunto de métricas quantitativas padrão, já consolidadas na literatura de aprendizado de máquina. As métricas de Acurácia, Precisão, *Recall* e F1-Score, cujas definições, formulações matemáticas e relevância para o contexto de detecção de desinformação foram detalhadamente abordadas no Capítulo 4, foram empregadas para fornecer uma avaliação objetiva do desempenho.

Para quantificar a performance do classificador, foram utilizadas métricas derivadas da matriz de confusão, que tabula os acertos e erros do modelo. Considerando “Positivo“ como a classificação de uma notícia como “Falsa“ ou seja, Fake News detectada, e “Negativo“ como notícia “Verdadeira“, ou seja, a notícia apresentada é uma notícia real, tem-se:

- **Verdadeiro Positivo (VP):** Notícia de fato falsa corretamente classificada pelo Agente como “Falsa“;
- **Verdadeiro Negativo (VN):** Notícia de fato verdadeira corretamente classificada pelo Agente como “Verdadeira“;
- **Falso Positivo (FP):** Notícia de fato verdadeira incorretamente classificada pelo Agente como “Falsa“ ;
- **Falso Negativo (FN):** Notícia de fato falsa incorretamente classificada como “Verdadeira“pelo Agente;

5.8.2 Avaliação Qualitativa

Dado que a principal contribuição deste trabalho é a geração de explicações, a avaliação de sua qualidade é fundamental. Para tal, propõe-se um *Framework* de Análise Explicativa, que foi aplicado sobre os conjuntos de dados de notícias cujo motivo da falsidade é previamente conhecido e documentado pelo pesquisador na Seção 5.3.2. O *Framework* Explicatório consistirá em um checklist ou rubrica com critérios objetivos e qualitativos, realizado manualmente pelo pesquisador, em uma amostra dos resultados gerados após a aplicação do Agente em cima dos dados.

Critérios de Avaliação da Explicação:

- **Suficiência Epistemológica:** Verifica se a explicação referencia de forma explícita e correta os conceitos identificados pelos agentes especializados.
  Avaliação: Qualitativa - Baixo/Médio/Alto/Altíssimo
- **Suficiência da Evidência Técnica:** Verifica se a explicação cita corretamente as evidências externas identificadas pelo Agente Técnico ou a ausência destas.
  Avaliação: Qualitativa - Baixo/Médio/Alto/Altíssimo
- **Precisão da Análise Morfológica:** Verifica se a explicação identifica corretamente os elementos de estilo, linguagem sensacionalista ou estrutura textual indicativos de manipulação, conforme análise do Agente Morfológico.
  Avaliação: Qualitativa - Baixo/Médio/Alto/Altíssimo
- **Coerência da Argumentação Teórica:** Verifica se a explicação apresenta consistência lógica entre as hipóteses de veracidade ou falsidade propostas pelo Agente Teórico e o contexto da notícia.
  Avaliação: Qualitativa - Baixo/Médio/Alto/Altíssimo
- **Capacidade de Síntese:** Verifica se, em casos de conflito entre dimensões de análise, a explicação pondera os achados e justifica o peso atribuído a cada dimensão na decisão final.
  Avaliação: Qualitativa - Baixo/Médio/Alto/Altíssimo

Os níveis de avaliação são definidos como:

- **Altíssimo:** A explicação atende plenamente ao critério, apresentando referências completas, precisas e bem articuladas.
- **Alto:** A explicação atende ao critério de forma satisfatória, com referências adequadas e poucas omissões.

- **Médio:** A explicação atende parcialmente ao critério, apresentando lacunas ou imprecisões que não comprometem a compreensão geral.
- **Baixo:** A explicação não atende adequadamente ao critério, apresentando referências insuficientes, imprecisas ou ausentes.

5.8.3 Estrutura Quadripolar Completa vs. Quadripolar de Prompt Único

A arquitetura quadripolar do presente projeto é composta por uma cadeia de inferência desacoplada, na qual cada polo constitui um agente. A análise aqui apresentada submete a arquitetura quadripolar à inferência por um único agente, caracterizado pela configuração C7 da Tabela 5. O objetivo desta análise consiste em fornecer um segundo *baseline* de comparação da arquitetura, permitindo avaliar se a abordagem multiagente e a cadeia de raciocínio em etapas contribuem efetivamente para os resultados.

Para a configuração C7, o prompt utilizado na presente pesquisa declara todos os polos do método quadripolar, sem a ferramenta de acesso a informação externa, mantendo o formato de veredito estabelecido anteriormente.

A configuração C7 utiliza o prompt detalhado apresentado a seguir:

Você é um especialista em inteligência investigativa, focado na detecção, análise e desmistificação de desinformação. Sua missão é avaliar a veracidade de conteúdos informativos com precisão, objetividade e ceticismo metodológico.

METODOLOGIA: MÉTODO QUADRIPOLAR (De Bruyne)
Aplique rigorosamente os quatro polos abaixo na sua análise interna antes de emitir o veredito.

1. ***POLO EPISTEMOLÓGICO — A Origem do Conhecimento (O QUÊ?)*** *Examine a fonte da informação e a natureza do conhecimento apresentado:*
   - *Quem é a fonte? Qual é o seu histórico de credibilidade e possíveis vieses?*
   - *A informação é primária (testemunho direto, dados originais) ou secundária (interpretação, relato de relato)?*
   - *Quais evidências concretas são apresentadas para sustentar a alegação?*
   - *Os fatos, dados ou citações são verificáveis e atribuíveis a fontes reconhecidas?*
   - *Há lacunas informacionais relevantes que comprometem a credibilidade?*
2. ***POLO MORFOLÓGICO — A Forma da Mensagem (COMO?)*** *Examine a estrutura, linguagem e apresentação do conteúdo:*
   - *A linguagem é neutra e objetiva ou emocional, sensacionalista e apelativa?*

- *O título corresponde fielmente ao corpo do texto, ou é exagerado/enganoso (clickbait)?*
- *Há erros gramaticais, ortográficos ou de formatação incomuns para jornalismo profissional?*
- *Existem elementos retóricos manipulativos: hipérboles, apelos ao medo, urgência artificial?*
- *O layout, imagens e design são condizentes com veículos de comunicação legítimos?*
- *A estrutura narrativa segue padrões jornalísticos (pirâmide invertida, atribuição de fontes)?*

3. ***POLO TEÓRICO — O Contexto e a Coerência (POR QUÊ?)*** *Examine a coerência da alegação com o conhecimento estabelecido:*
    - *A alegação se encaixa no conhecimento científico, histórico e factual consolidado?*
    - *Existem hipóteses plausíveis que sustentem tanto a veracidade quanto a falsidade?*
    - *O conteúdo simplifica excessivamente um tema complexo ou ignora contextos importantes?*
    - *Há narrativas de conspiração, polarização excessiva ou lógica "nós contra eles"?*
    - *A alegação serve a algum interesse político, econômico ou ideológico identificável?*
    - *O conteúdo é coerente com eventos verificáveis do mesmo período temporal?*
4. ***POLO TÉCNICO — A Verificação Externa (COM QUÊ?)*** *Examine o que fontes externas e confiáveis indicam sobre o tema:*
    - *O tema foi verificado ou desmentido por agências de fact-checking reconhecidas?*
    - *Fontes de imprensa de referência (grandes veículos nacionais e internacionais) confirmam ou contradizem a alegação?*
    - *Os dados estatísticos, científicos ou oficiais citados podem ser rastreados até suas fontes originais?*
    - *Imagens ou vídeos utilizados condizem com o contexto alegado (sem descontextualização)?*
    - *O domínio ou veículo de origem possui histórico de disseminação de desinformação?*

- *Há ausência total de cobertura de outros meios confiáveis sobre um evento supostamente relevante?*

TAXONOMIA DE FAKE NEWS

- ***Desinformação**: Criação e partilha INTENCIONAL de informações falsas para enganar, manipular ou causar danos. Caracterizada pela intenção maliciosa, visando ganhos políticos ou financeiros ao explorar emoções e crenças.*
- ***Misinformação**: Partilha de informações falsas SEM intenção de enganar. Quem compartilha acredita que o conteúdo é verdadeiro, agindo por engano ou falta de verificação. Não é maliciosa, mas contribui para a desordem informativa.*
- ***Clickbait**: Títulos e imagens sensacionalistas, exagerados ou enganosos para atrair cliques e gerar tráfego online. O conteúdo raramente corresponde ao prometido, sendo de baixa qualidade ou irrelevante.*
- ***Golpe**: Fraude que utiliza engenharia social para manipular vítimas a executarem ações prejudiciais (fornecer dados, realizar transferências, clicar em links maliciosos), através de pretextos convincentes como urgência ou oportunidades únicas.*
- ***Rumor**: Informação especulativa e não verificada que circula amplamente, sem fonte clara ou confiável. Surge em contextos de incerteza e pode ser distorcida ao ser transmitida.*
- ***Notícia Parcialmente Falsa**: Mistura informações verdadeiras com elementos falsos ou enganosos. Utiliza dados e fatos reais, mas os apresenta de forma distorcida, incompleta ou fora de contexto para manipular a percepção do leitor.*

INSTRUÇÕES DE ANÁLISE

- *Aplique os quatro polos do Método Quadripolar de forma integrada e sistemática.*
- *Pense estrategicamente antes de decidir, relacionando todas as informações de forma minuciosa.*
- *NÃO ALUCINE — baseie-se apenas em evidências concretas presentes no texto ou em conhecimento factual consolidado.*
- *Sintetize os achados dos quatro polos para construir uma conclusão fundamentada.*

FORMATO DE SAÍDA (MANTENHA ESTE FORMATO RIGOROSAMENTE)

- *Veredito de Veracidade da Notícia: [Escolha uma: "Verdadeira", "Falsa", "Parcialmente Falsa", "Ironia", "Opinião"]*
- *Tipo do Conteúdo: [Escolha uma: "Factual/Jornalístico", "Ironia", "Opinião"]*
- *Tipo de Fake News: [Se "Verdadeira", deixe VAZIO. Se "Falsa"ou "Parcialmente Falsa", escolha: Desinformação, Misinformação, Clickbait, Rumor, Golpe]*
- *Diagnóstico do Agente: [Máximo 300 tokens. Justifique seu veredito de forma crítica, concisa e objetiva. Sintetize os argumentos dos quatro polos para construir uma explicação clara e direta, apontando as principais "bandeiras vermelhas"(red flags) ou evidências de veracidade.]*

DADOS DA ANÁLISE

***Data de Hoje**: {data_hoje}*

***Notícia**: {notícia}*

5.8.4 Estrutura Quadripolar Completa vs. LLM de Base

Com o objetivo de avaliar as hipóteses H1 e H2 e para isolar e mensurar o valor agregado da estrutura quadripolar como um todo, foi realizada uma comparação direta entre o agente completo (Configuração C16 da Tabela 5) e o LLM em sua forma base, sem o auxílio de nenhuma ferramenta ou agente especializado (Configuração C1 da Tabela 5). Este experimento é adequado para responder se a metodologia proposta é um componente que agrega positivamente à classificação e explicação de notícias falsas.

Neste cenário de controle, a notícia a ser analisada é submetida diretamente ao LLM com um *prompt* genérico e direto, como: *“Avalie a veracidade da seguinte notícia e justifique sua resposta.“* Este processo emula uma abordagem de *prompting* direto, dependendo unicamente do conhecimento intrínseco e da capacidade de raciocínio do modelo, sem o suporte da busca externa do Agente Técnico ou das análises estruturadas dos demais polos.

A avaliação desta comparação é dada pelos seguintes fatores:

1. **Desempenho de Classificação:** A capacidade do LLM base de classificar corretamente a notícia (verdadeira/falsa) foi medida com as mesmas métricas quantitativas (*F1-Score, Precision, Recall*), servindo como um *benchmark* para a Hipótese 1;
2. **Qualidade da Explicação:** De forma mais adequada, a justificativa gerada pelo LLM base foi avaliada utilizando o mesmo *Framework* de Análise Explicativa. Este passo é adequado para testar a Hipótese 2, comparando a profundidade, a estrutura e a

confiabilidade da explicação metodologicamente guiada da arquitetura quadripolar com a justificativa não estruturada do LLM base.

Espera-se que essa comparação demonstre que a arquitetura quadripolar não é redundante, mas sim um componente fundamental que guia, enriquece e fundamenta as inferências do modelo, resultando em diagnósticos significativamente mais precisos e, sobretudo, mais explicáveis.

5.8.5 Análise Manual e Estudos de Caso

Complementando as métricas quantitativas obtidas via automação, foi conduzida uma etapa rigorosa de execução manual, utilizando amostras reais coletadas no ecossistema digital (*WhatsApp*, portais jornalísticos, redes sociais e canais de mensageria) entre os meses de setembro e outubro de 2025. Nesta fase, o escopo de análise foi expandido para além da verificação de notícias factuais, incorporando deliberadamente cenários de golpes cibernéticos (como *phishing* e engenharia social) e conteúdos de natureza estritamente opinativa.

A justificativa para a inclusão destas categorias reside na necessidade crítica de avaliar a robustez semântica e pragmática do agente. É imperativo que o modelo demonstre capacidade de distinguir entre a intencionalidade maliciosa de um golpe financeiro, a subjetividade legítima de um artigo de opinião e a “fabricação factual“ de uma notícia falsa. Esta distinção é vital para evitar falsos positivos que poderiam censurar o debate público ou, inversamente, deixar passar ameaças de segurança digital.

Durante este procedimento, cada amostra, seja ela uma notícia, uma tentativa de fraude ou um texto opinativo, foi submetida ao agente sob a configuração C6 da Tabela 5 com todos os Agentes ativos da arquitetura quadripolar. Este processo permite observar, de forma granular, como a composição de polos específicos (como o Técnico ou o Morfológico) influencia a construção da justificativa e o veredito final. Os resultados qualitativos desta etapa forneceram evidências sobre a competência do agente em navegar pela complexidade do discurso humano e pelas táticas contemporâneas de manipulação, validando sua aplicabilidade em ambientes reais ambíguos de alta incerteza.

5.8.6 Ameaças à Validade

Como inerente a estudos experimentais, esta pesquisa apresenta ameaças à validade que exigem documentação. As limitações operacionais concentram-se na elaboração do conjunto de dados e na condução da avaliação qualitativa.

No que tange à validade de construção, a principal ameaça reside na estruturação do *dataset* próprio (Seção 5.3.2). A elaboração e curadoria das 250 amostras foram conduzidas por um único pesquisador, introduzindo risco de viés de seleção e anotação.

Para mitigar esta ameaça e garantir um cenário controlado livre de contaminação de dados dado pelo fator temporal que existe no treinamento dos LLMs utilizados, adotou-se um protocolo rígido de restrição temporal e o balanceamento estrito das classes, ancorando as notícias verdadeiras em veículos de imprensa de alta credibilidade.

Em relação à validade de conclusão, o risco metodológico concentra-se na avaliação qualitativa das explicações geradas pelo agente. Por ter sido inspecionada unicamente pelo autor, há potencial viés de avaliação. Como estratégia de contenção, a análise foi parametrizada pela aplicação do *Framework* de Análise Explicativa. O uso desta rubrica padronizada restringiu a subjetividade do avaliador ao ancorar o julgamento em critérios técnicos observáveis, avaliando a Fidelidade aos Polos, a Suficiência da Evidência, a Precisão Morfológica, a Coerência Teórica e a Capacidade de Síntese em uma escala ordinal predefinida.

Por fim, quanto à validade externa, os resultados obtidos limitam-se ao escopo textual do *corpus* analisado. A generalização da eficácia do agente para identificar e explicar desinformações em outros idiomas ou estruturas sintáticas com maior nível de ruído demandará replicações futuras do experimento.

6 RESULTADOS

Os resultados obtidos consolidam a validação da arquitetura da arquitetura quadripolar proposta e seus agentes, são apresentados a seguir, combinando dados do Conjunto de Dados Próprio e do FakeRecogna. A discussão está organizada em quatro eixos de avaliação: (i) desempenho quantitativo, focado nas métricas de classificação; (ii) desempenho qualitativo, com ênfase na explicabilidade dos vereditos; (iii) eficiência operacional, detalhando custos e latência de inferência; e (iv) estudo de caso experimental, ilustrando a atuação do agente frente a um cenário de Golpe Cibernético.

6.1 ANÁLISE DOS RESULTADOS QUANTITATIVOS

Conjunto de Dados - FakeRecogna

A avaliação quantitativa da capacidade classificatória dos modelos foi conduzida utilizando o conjunto de dados *FakeRecogna*. Embora o conjunto completo contemple 11.902 registros, a complexidade estrutural da arquitetura quadripolar impõe desafios significativos de custo computacional e latência. Diante disso, optou-se pela seleção das 100 primeiras amostras, visando adequar o escopo à viabilidade da avaliação arquitetural.

Os resultados consolidados de quantitativos são apresentados na Tabela 6. Esta tabela detalha o desempenho dos quatro Modelos de Linguagem (LLMs) avaliados, sob a influência das seis configurações experimentais (C1 a C6), utilizando as métricas padrão de Acurácia, Precisão, Recall e F1-Score. A Configuração C1 (LLM Base) serve como linha de base, estabelecendo o desempenho dos modelos sem a intervenção dos agentes nas análises propostas.

Observa-se na Tabela 6 que os modelos mais robustos da linha de base (C1) foram o gemini-2.5-pro (*F1-Score* = 0,71) e o gpt-5 (*F1-Score* = 0,72). Contudo, é vlido contextualizar que o desempenho nas demais configurações (C2 a C7) foi impactado por um viés temporal. O conjunto de dados *FakeRecogna*, cujas amostras datam do período de 2020 a 2021, contém notícias que, embora factualmente avaliadas como verdadeiras à época, tornaram-se desatualizadas ou incorretas no momento da inferência (2025). Tais instâncias foram frequentemente classificadas pelos modelos como “Parcialmente Falsas”. A captura dessa nuance pelos LLMs representa um indicativo positivo para a eficácia analítica da pesquisa. No entanto, conforme definido na metodologia, o rótulo “Parcialmente Falso” foi agregado ao rótulo “Falso” para o cálculo das métricas binárias.

A introdução isolada dos polos de análise (C2 a C5) revelou desempenhos heterogêneos. As configurações C3 (polo epistemológico) e C5 (polo teórico) demonstraram ser as mais promissoras individualmente. O modelo gemini-2.5-pro, por exemplo, atingiu picos

de desempenho nessas duas vertentes (C3: *F1-Score* = 0,74; C5: *F1-Score* = 0,72), superando sua própria linha de base. Em contrapartida, as configurações C2 (polo morfológico) e C4 (polo técnico), quando aplicadas de forma isolada, introduziram maior variabilidade e degradaram o poder preditivo em relação à C1, a exemplo do impacto observado no gpt-5 (que recuou de um *F1-Score* de 0,72 na C1 para 0,55 na C4).

Para superar a degradação métrica de polos isolados e lidar com o ruído do viés temporal, avaliou-se a Configuração C7. Os dados indicam que a C7 configurou o melhor desempenho de classificação da arquitetura quadripolar, o gemini-2.5-pro alcançou seu melhor resultado nesta configuração (*F1-Score* = 0,77), enquanto o gpt-5 retomou um resultado (*F1-Score* = 0,71), estatisticamente equivalente à sua linha de base, porém incorporando a complexa camada de explicabilidade do modelo. Constata-se, portanto, que a C7 oferece o balanço ideal entre a precisão da classificação binária e a robustez argumentativa para o conjunto de dados FakeRecogna.

A Configuração C6, que representa a integração completa do método quadripolar, apresentou um desempenho classificatório inferior à linha de base (C1) e às configurações isoladas C3 e C5 para os modelos principais. Conforme a Tabela 6, o *F1-Score* do modelo gemini-2.5-pro reduziu para 0,70 (de 0,71 na C1), e o do gpt-5 caiu para 0,59 (de 0,72 na C1). Levanta-se a hipótese de que essa divergência decorre da combinação de dois fatores: (i) o viés temporal presente no *dataset*; e (ii) o pré-processamento inerente ao *FakeRecogna*. A truncagem dos textos e a remoção de *stopwords* impactaram negativamente a avaliação do polo morfológico (presente em C2 e C6), uma vez que a análise estrutural da notícia depende da coesão sintática, de um contexto semântico preservado e da integridade das entidades nomeadas para a correta verificação das fontes.

Diante desse cenário, a configuração C6 aplicada ao modelo gemini-2.5-pro configurou-se inicialmente como o arranjo experimental de maior relevância para os propósitos da pesquisa. Embora as configurações C3 (*F1-Score* = 0,74) e C5 (*F1-Score* = 0,72) tenham registrado métricas de classificação superiores, a C6 manteve um desempenho equilibrado e, fundamentalmente, forneceu a arquitetura necessária para a extração das justificativas baseadas nos quatro polos, requisito central para o objetivo de explicabilidade do agente.

Apesar da limitação imposta pela C6 na classificação binária estrita em contraste com C1, a análise qualitativa demonstrou que a integração dos polos produziu as explicações mais coesas. O método completo permitiu ao agente identificar nuances contextuais (como a desatualização temporal) e justificar classificações de desinformação parcial que divergiam do gabarito estrito do *dataset*. Para mitigar o ruído introduzido pela C6 sem abdicar de sua robustez explicativa, avaliou-se a Configuração C7. Os dados indicam que a C7 superou as limitações estruturais anteriores, elevando o *F1-Score* do gemini-2.5-pro para 0,77 e o do gpt-5 para 0,71. Conclui-se, portanto, que a C7 configura-se como a modelagem da

Tabela 6 – Resultados Comparativos das Métricas de Classificação por Configuração do Conjunto de Dados *FakeRecogna* e Modelo.

Config.	Modelo	Accuracy	Precision	Recall	F1-Score
C1	gemini-2.5-flash	0,74	0,71	0,71	0,71
	gemini-2.5-pro	0,79	0,90	0,59	0,71
	gpt-5	0,79	0,87	0,61	0,72
	gpt-5-mini	0,55	0,49	0,89	0,63
C2	gemini-2.5-flash	0,40	0,41	0,77	0,53
	gemini-2.5-pro	0,61	0,54	0,77	0,64
	gpt-5	0,61	0,54	0,73	0,62
	gpt-5-mini	0,44	0,43	0,84	0,57
C3	gemini-2.5-flash	0,67	0,60	0,77	0,67
	gemini-2.5-pro	0,74	0,67	0,82	0,74
	gpt-5	0,56	0,50	0,68	0,58
	gpt-5-mini	0,46	0,44	0,86	0,59
C4	gemini-2.5-flash	0,50	0,46	0,73	0,56
	gemini-2.5-pro	0,60	0,53	0,75	0,62
	gpt-5	0,52	0,47	0,66	0,55
	gpt-5-mini	0,49	0,46	0,86	0,60
C5	gemini-2.5-flash	0,50	0,46	0,77	0,58
	gemini-2.5-pro	0,70	0,61	0,89	0,72
	gpt-5	0,57	0,51	0,75	0,61
	gpt-5-mini	0,53	0,48	0,98	0,65
C6	gemini-2.5-flash	0,38	0,39	0,73	0,51
	gemini-2.5-pro	0,68	0,60	0,84	0,70
	gpt-5	0,55	0,49	0,73	0,59
	gpt-5-mini	0,49	0,46	0,91	0,61
C7	gemini-2.5-flash	0,63	0,55	0,89	0,68
	gemini-2.5-pro	0,76	0,67	0,91	0,77
	gpt-5	0,69	0,60	0,86	0,71
	gpt-5-mini	0,54	0,49	1,00	0,66

Fonte: Do Autor.

arquitetura quadripolar sob inferência de um único agente que apresentou os melhores resultados, pois conseguiu reconciliar a precisão exigida para a classificação quantitativa com a capacidade de gerar explicações factualmente embasadas na morfologia do conjunto de dados.

Conjunto de Dados Próprio

Para aprofundar a avaliação da capacidade de detecção dos modelos, um segundo conjunto de experimentos foi executado. Nesta etapa, a metodologia de classificação

foi simplificada, ajustando os rótulos de saída para um formato estritamente binário: “Falso“ ou “Verdadeiro“, e permanecendo toda a estrutura de análise do Método Quadripolar. Esta abordagem remove a complexidade introduzida pelo rótulo “Parcialmente Falso“ e foca em mensurar a assertividade dos modelos na distinção factual. Os resultados desta rodada experimental, comparando as diferentes configurações (C1-C5 e C6), são apresentados na Tabela 7.

Uma análise dos resultados individuais dos polos de análise revela um ponto de destaque: o Polo Técnico (C4), que corresponde à ferramenta de busca e verificação de evidências externas, demonstrou ser o componente de maior impacto individual na classificação. Conforme a Tabela 7, a Configuração C4, quando aplicada ao modelo Gemini-2.5-Pro, alcançou o maior F1-Score de todos os experimentos (0,87). Isso sugere que, para a tarefa de classificação binária, a verificação de fontes externas e evidências técnicas detém um peso decisório superior ao das análises epistemológicas (C3) ou teóricas (C5) isoladas.

Tabela 7 – Resultados da Classificação Binária por Configuração e Modelo do Conjunto de Dados Próprio

Config.	Modelo	Accuracy	Precision	Recall	F1-Score
C1	gemini-2.5-flash	0,74	0,72	0,82	0,76
	gemini-2.5-pro	0,69	0,69	0,71	0,70
C2	gemini-2.5-flash	0,61	0,58	0,91	0,71
	gemini-2.5-pro	0,74	0,72	0,80	0,76
C3	gemini-2.5-flash	0,65	0,61	0,90	0,73
	gemini-2.5-pro	0,71	0,66	0,88	0,76
C4	gemini-2.5-flash	0,83	0,76	0,97	0,85
	gemini-2.5-pro	0,86	0,81	0,95	0,87
C5	gemini-2.5-flash	0,66	0,60	0,96	0,74
	gemini-2.5-pro	0,70	0,65	0,88	0,75
C6	gemini-2.5-flash	0,82	0,76	0,94	0,84
	gemini-2.5-pro	0,84	0,79	0,92	0,85
C7	gemini-2.5-flash	0,64	0,60	0,92	0,72
	gemini-2.5-pro	0,71	0,68	0,83	0,75

Fonte: Do Autor.

O Polo Técnico (C4) apresentou desempenho individual superior. Entretanto, a Configuração C6 (método quadripolar completo) com o modelo Gemini-2.5-Pro demonstrou maior adequação aos objetivos da pesquisa. Esta combinação obteve F1-Score de 0,85, valor próximo ao máximo observado em C4, integrando simultaneamente todos os polos de

análise. A análise qualitativa (Seção subsequente) evidenciou que C6 produziu explicações superiores em coesão, profundidade e alinhamento com o raciocínio humano, aspecto mais relevante que a métrica classificatória isolada.

A introdução da Configuração C7 neste conjunto de dados resultou em degradação de desempenho. Conforme a Tabela 7, o F1-Score do modelo gemini-2.5-pro reduziu para 0,75, comparado a 0,85 (C6) e 0,87 (C4). Esta redução decorre de limitação arquitetural: C7 operou sem mecanismos de informação externa. Dado que o desempenho superior de C4 demonstrou que a verificação factual externa (Polo Técnico) constitui o principal fator de acurácia para detecção de desinformação nesta base, a ausência dessa capacidade em C7 comprometeu o poder preditivo do modelo. Assim, C6 estabelece-se como arquitetura adequada para este conjunto de dados, equilibrando rigor quantitativo e explicabilidade qualitativa mediante integração completa dos polos.

Conclui-se, portanto, que a Configuração C6, utilizando o Gemini-2.5-Pro, representa o equilíbrio ideal entre alta assertividade classificatória (F1=0,85) e a geração de explicabilidade mais detalhada, sendo objetivo central desta dissertação. A leve supressão no F1-Score em relação à C4 (0,87 vs. 0,85) é compensada pela integração dos demais polos, que enriquecem a explicação. Por esta razão, esta configuração (C6) foi selecionada como o a configuração adequada para os estudos de caso e a análise qualitativa aprofundada que se seguem. É de se notar, contudo, que para aplicações de produto em cenários de mundo real onde a explicabilidade detalhada não é o requisito primário, a Configuração C4 (Polo Técnico) oferece uma alternativa de melhor custo-benefício, dado o menor custo computacional por *token* associado a um único polo de análise. A análise de custo por *tokens* serão apresentadas nas próximas seções.

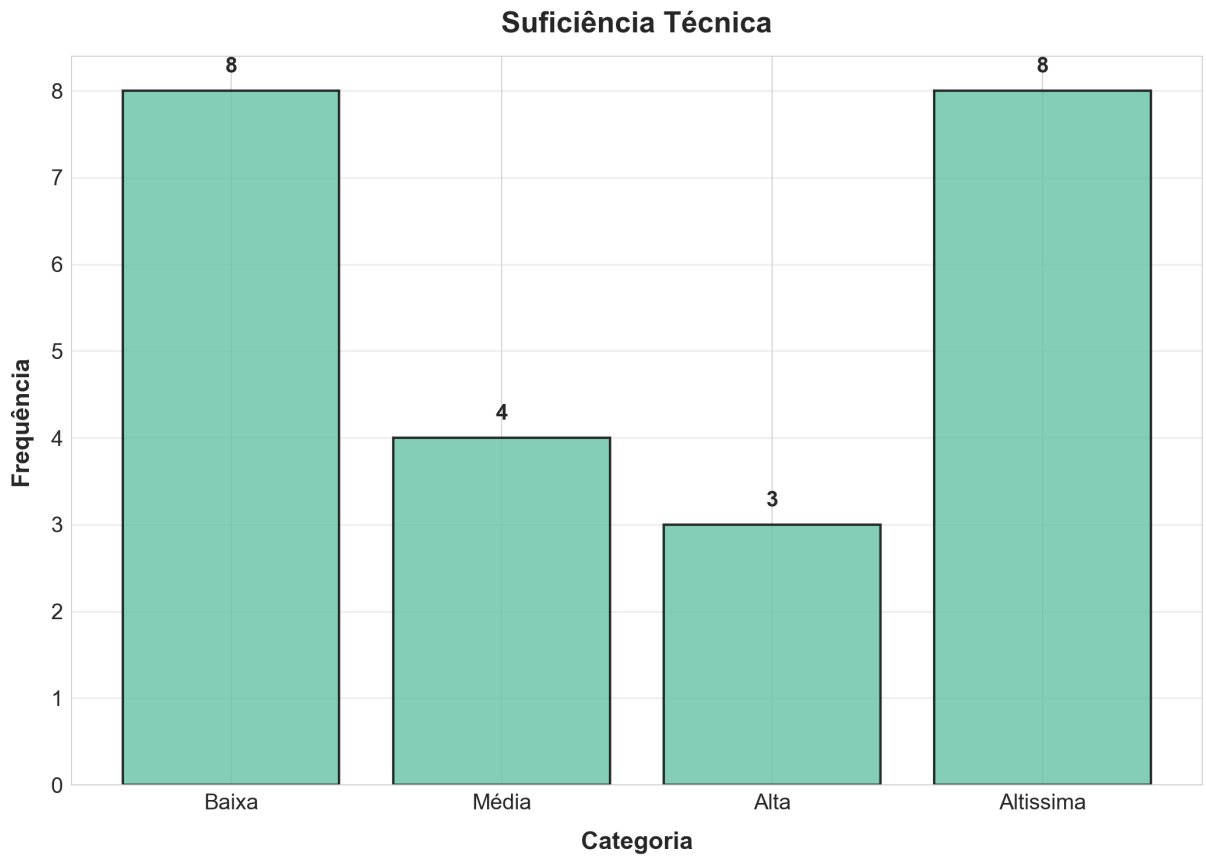
6.2 PERFORMANCE DE EXPLICABILIDADE

Superada a análise quantitativa de desempenho, esta seção se debruça sobre a avaliação qualitativa da explicabilidade gerada pelo agente. O objetivo é investigar por inspeção humana, realizada pelo autor do presente texto, a coerência, profundidade e correção do raciocínio aplicado pelo Gemini-2.5-Pro na Configuração C6. Para realizar esta análise, um subconjunto de 23 notícias do conjunto de dados próprio foi selecionado e analisado em detalhes pelo pesquisador. A avaliação foi estruturada para inspecionar a contribuição de cada componente do método quadripolar (polos Morfológico, Epistemológico, Teórico e Técnico) e, por fim, o Veredito final. Cada polo foi classificado segundo quatro níveis de suficiência: Baixo (quando a informação gerada foi irrelevante ou incorreta), Médio (informação relevante, mas incompleta), Alto (informação correta e suficiente para a tomada de decisão) e Altíssimo (quando a análise do agente superou as expectativas, trazendo *insights* ou evidências de difícil obtenção).

Polo Técnico

O Polo Técnico, responsável pela busca e verificação de evidências externas (*links*, fontes oficiais, imprensa), apresentou os resultados mais polarizados de toda a análise, conforme ilustra a Figura 31. Esta figura demonstra uma distribuição de resultado por nível de performance, onde se concentram 8 instâncias classificadas como “Baixa“ suficiência e 8 instâncias como “Altíssima“, enquanto os níveis intermediários (“Média“ e “Alta“) somaram 4 e 3 casos, respectivamente.

Figura 31 – Análise de Desempenho Do Polo Técnico - Configuração C6 com Gemini 2.5 Pro



Fonte: Do Autor

A interpretação destes dados requer uma análise cuidadosa. O nível “Altíssimo“ (representado por 8 das 23 noticias analisadas pelo autor) corresponde aos cenários ideais, nos quais o agente localizou fontes primárias, agências de checagem de fatos, portais de governo, ou reportagens da imprensa que corroboravam ou refutavam diretamente o conteúdo da notícia, fornecendo uma base sólida para o veredito. A nuance mais relevante da pesquisa, contudo, reside na interpretação dos 8 casos de suficiência “Baixa“. Nestas instâncias, a ferramenta de busca não retornou resultados relevantes ou não encontrou qualquer menção à notícia em fontes confiáveis.

Paradoxalmente, esta ausência de evidência foi, na maioria desses casos, um indicador de desinformação. O agente interpretou corretamente que a incapacidade de encontrar qualquer corroboração na imprensa estabelecida ou em fontes oficiais para um fato alegadamente relevante era, em si, uma evidência de falsidade. Portanto, embora a suficiência da busca tenha sido classificada como “Baixa“ pelo pesquisador (pois a ferramenta “falhou“ em encontrar dados, ou trouxe evidências irrelevantes), esta falha contribuiu positivamente para a assertividade do veredito final “Falsa“. Isso demonstra a capacidade do Agente na configuração C6 tratar casos de ambiguidade diante das fontes.

Estudo de Caso 1: Verdadeira (Suficiência Técnica: Altíssima)

Este estudo de caso representa o resultado do Veredito na Configuração C6, com a saída consolidada do Agente, corroborado diante de todas as análises de cada polo realizada previamente. Foi classificada como suficiência técnica ‘Altíssima‘ pelo fato de que o Polo Técnico trouxe informações altamente relevantes para a decisão do Agente.

Notícia Analisada

Saiba quanto Ramon Dino já faturou em premiações no Mr. Olympia. Fisiculturista brasileiro ficou com o vice-campeonato em duas edições, mas terminou em quarto em 2024a.

Diagnóstico do Agente (Veredito)

- **Veredito de Veracidade:** Verdadeira
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Diagnóstico:** A análise, fundamentada nos quatro polos, conclui que a notícia é verdadeira. A principal alegação, que Ramon Dino "terminou em quarto em 2024", é consistentemente corroborada por múltiplas fontes de imprensa de alta credibilidade (Globo Esporte, A Gazeta do Acre, Portal Goleada), conforme evidenciado no Polo Técnico. Dado que a data da análise é 9 de outubro de 2025, o resultado de 2024 é um fato passado e verificável, eliminando a suspeita de anacronismo. Epistemologicamente, a notícia se baseia em dados públicos e consolidados (resultados e premiações oficiais do evento Mr. Olympia). Morfologicamente, embora o título ("Saiba quanto...") utilize uma fórmula comum de *clickbait* para gerar curiosidade, o corpo da informação é factual e direto, sem apelo emocional ou sensacionalista. Teoricamente, a notícia se encaixa perfeitamente no contexto de um artigo esportivo que resume a trajetória de um atleta antes de uma nova competição (a de 2025). As evidências sustentam a veracidade dos fatos reportados, tornando a informação um relato jornalístico legítimo.

Resultados do Polo Técnico (Evidências)

Evidência 1:

> **Título:** Saiba quanto Ramon Dino já faturou em premiações no Mr. Olympia | CNN Brasil
> **Conteúdo:** ...Saiba quanto Ramon Dino já faturou em premiações no Mr. Olympia. Fisiculturista brasileiro ficou com o vice-campeonato em duas edições, mas ...| Ramon Dino estará no palco do Mr. Olympia 2025...
> **Link:** https://www.cnnbrasil.com.br/esportes/fisiculturismo/saiba-quanto-ramon-dino-faturou-com-o-titulo-no-mr-olympia-2025/

Evidências 2, 3 e 4 (Agrupadas):

> **Título:** Favorito: Acreano Ramon Dino vai em busca do título inédito no Mr. Olympia 2025 - Jornal A Gazeta do Acre
> **Conteúdo (trechos combinados e reranqueados):** ...Ele vai para a quinta participação... Ramon Dino já figurou entre os melhores do mundo nas edições anteriores: quinto lugar em 2021, vice-campeão em 2022 e 2023 e **quarto em 2024**. ...premiações significativas: 20 mil dólares em 2022 e 2023 e **7 mil dólares em 2024, somando 47 mil dólares**...
> **Link:** https://agazetadoacre.com/2025/10/noticias/geral/favorito-acreano-ramon-dino-vai-em-busca-do-titulo-inedito-no-mr-olympia-2025/

O Estudo de Caso 1, é um exemplo dos 8 casos classificados com suficiência “Altíssima“ no gráfico do Polo Técnico (Figura 31). A avaliação máxima se justifica pelo fato de o agente ter encontrado as melhores evidências para o caso, tendo localizado múltiplas fontes de imprensa de alta credibilidade (CNN Brasil, A Gazeta do Acre) que confirmavam explicitamente os fatos centrais da notícia. Os trechos de evidências externas recuperados não eram genéricos, ambíguos ou sem correlação com a notícia. Todos os trechos mais relevantes continham os dados exatos da alegação principal, como a colocação (“quarto em 2024“) e os valores das premiações (“7 mil dólares em 2024, somando 47 mil dólares“). A capacidade do agente de analisar essas informações, validá-las e integrá-las de forma coesa no Veredito final, inclusive utilizando o Polo Epistemológico para contextualizar a data da análise (2025) contra os fatos (2024), representa o funcionamento ideal e esperado da ferramenta de busca, superando a verificação superficial.

Estudo de Caso 2: Falsa (Suficiência Técnica: Baixa)

Este estudo de caso representa o resultado do Veredito na Configuração C6 com o Gemini 2.5 Pro, com a saída consolidada do Agente, corroborado diante de todas as análises de cada polo realizada previamente. Foi classificada como suficiência técnica

‘Baixa‘ pelo fato de que o Polo Técnico não trouxe informações altamente relevantes para a decisão do Agente.

Notícia Analisada (Conteúdo Resumido)

> *“Pesquisadores descobrem microplásticos na água de torneira de grandes capitais ligados a declínio cognitivo. Um estudo da Universidade Federal de Tecnologia revelou a presença de partículas de polímeros em níveis alarmantes [...] O relatório preliminar associa a exposição a longo prazo a esses contaminantes com um aumento de 30% nos casos de névoa mental...“*

Diagnóstico do Agente

- **Veredito de Veracidade:** Falsa
- **Tipo de Fake News:** Desinformação (Fabricada)
- **Diagnóstico:** A notícia é fabricada. A análise quadripolar revela a ausência total de sustentação factual. Epistemologicamente, baseia-se em fontes anônimas e inverificáveis (ex: “Universidade Federal de Tecnologia“, sem especificar qual). Morfologicamente, emula o jornalismo, mas emprega linguagem sensacionalista. **O Polo Técnico é decisivo: a verificação externa falha completamente, pois as fontes fornecidas são totalmente irrelevantes para o tema (conforme abaixo), não havendo qualquer corroboração em mídia de credibilidade.** A ausência de evidências e o uso de táticas de manipulação narrativa apontam para uma criação deliberada.

Resultados do Polo Técnico (Evidências Irrelevantes)

Nota: A ferramenta de busca não retornou resultados pertinentes à alegação. Os links abaixo são os mais próximos encontrados, porém tratam de temas totalmente distintos.

Evidência 1 (Irrelevante):

Título: Instagram
Conteúdo Reranqueado: ...alimentos em instituições particulares, segundo especialistas, é o lobby da indústria alimentícia.
Link:https://www.instagram.com/p/DPUBJFmlLvU/

Evidência 2 (Irrelevante):

Título: Quilombolas expulsos por mina de ouro pedem indenização - 11/09/2025 - Ambiente - Folha
Link: https://www1.folha.uol.com.br/ambiente/2025/09/quilombolas-expulsos-por-mina-de-ouro-em-mg-pedem-indenizacao-na-justica.shtml

Evidência 3 (Irrelevante):

> **Título:** Indústria contrata escritórios na Ásia e Europa para ampliar mercados | CNN Brasil
> **Link:** https://www.cnnbrasil.com.br/economia/macroeconomia/industria-contrata-escritorios-na-asia-e-europa-para-ampliar-mercados/

O Estudo de Caso 2 ilustra perfeitamente o cenário oposto ao anterior e justifica os 8 casos classificados como “Baixa“ suficiência no Polo Técnico conforme a Figura 31. A notícia sobre microplásticos faz uma alegação factual grave e de alto impacto *“declínio cognitivo em 5 capitais“*, envolvendo uma suposta instituição científica (“Universidade Federal de Tecnologia“).

Neste cenário, a avaliação de suficiência “Baixa“ foi atribuída pelo pesquisador porque a ferramenta de busca *falhou* em encontrar qualquer evidência corroborativa, também como trouxe uma evidência de rede social, o que tende a ser um portal de informação de baixa confiabilidade para uma alegação factual. A hipótese de que a noticia é de fato falsa, e também as palavras chaves geradas pelo Agente Técnico para realização da busca na web, podem não ter sido adequadas para o assunto em questão. Como demonstrado na análise do Estudo de Caso 2, os resultados retornados pelo Polo Técnico são completamente irrelevantes, tratando de temas como indústria alimentícia, mineração e economia, sem qualquer nexo com a alegação principal.

Este caso evidencia um paradoxo fundamental na detecção de *fake news*: a “baixa suficiência“ da ferramenta de busca (sua incapacidade de encontrar dados) foi o fator determinante para o veredito “Falso“. O agente interpretou corretamente que a ausência total de cobertura da imprensa ou de publicações científicas sobre um evento de tamanha magnitude é, em si, a evidência mais forte de que o fato é fabricado. Conforme destacado no Veredito do agente, o Polo Técnico foi decisivo justamente por sua falha em localizar corroboração.

Estudo de Caso 3: Falsa (Suficiência Técnica: Média)

Este estudo de caso representa o resultado do Veredito na Configuração C6 com o Gemini 2.5 Pro, com a saída consolidada do Agente, corroborado diante de todas as análises de cada polo realizada previamente. Foi classificada como suficiência técnica ‘Média‘ pelo fato de que o Polo Técnico trouxe informações relevantes e de fonte confiável, mas polarizadas em uma única fonte.

Notícia Analisada

> *“Governo Federal vai propor a redução da jornada de trabalho para 4 dias no serviço público. Um estudo piloto conduzido por um ministério teria*

apontado aumento de produtividade, servindo de base para uma proposta de implementação gradual."

Diagnóstico do Agente

- **Veredito de Veracidade da Notícia:** Falsa
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Tipo de Fake News:** Desinformação
- **Diagnóstico:** A notícia é falsa. Sua estrutura aparenta credibilidade, mas sua base factual é inexistente, configurando desinformação. A análise Quadripolar demonstra a falha em todos os níveis. No **Polo Epistemológico**, a alegação se apoia em fontes anônimas e inverificáveis ("um ministério", "um estudo piloto"), uma lacuna informacional crítica que impede qualquer validação da origem. O **Polo Morfológico** revela que, embora a linguagem seja formal, a generalização de uma proposta de alto impacto sem apresentar dados concretos é um forte indicador de fabricação. No **Polo Teórico**, a alegação se encaixa perfeitamente em teorias que sustentam a falsidade, como a **Desinformação por Polarização Social** (gerar animosidade contra servidores públicos) ou o **Enquadramento Enganoso (Framing)**. O **Polo Técnico** é conclusivo: a verificação externa não corrobora a alegação principal. As iniciativas reais encontradas são: 1) uma PEC de 2015 (do Legislativo, não do Executivo atual) para 36 horas semanais na CLT; e 2) uma Instrução Normativa que reduz a jornada para 40 horas semanais de um grupo restrito de trabalhadores *terceirizados*, não do serviço público em geral. A notícia distorce radicalmente fatos parciais, inventa um "estudo piloto" para dar legitimidade e generaliza uma medida limitada, criando uma narrativa enganosa.

Resultados do Polo Técnico (Evidências Recuperadas)

Nota: As evidências recuperadas referem-se a uma PEC de 2015 do Legislativo, e não corroboram a ação do Governo Federal (Executivo) atual citada na notícia.

Evidência 1:

Título: PEC 148/2015 - Senado Federal
Conteúdo: a de trabalho semanal. Entenda a proposta O que é A proposta visa alterar a Constituição Federal para reduzir a jornada de trabalho semanal de 44 para 36 horas. A mudança será gradual, começando com uma redução para 40 horas na primeira fase e diminuindo uma hora por ano até atingir o limite de 36 horas semanais. O que diz o autor As possíveis consequências dessa proposta são variadas: - Para os trabalhadores, haverá uma redução na carga horária semanal, o que pode melhorar a qualidade de vida e
Link: https://www25.senado.leg.br/web/atividade/materias/-/materia/124067

Evidência 2:

> **Título:** PEC 148/2015 - Senado Federal
> **Conteúdo Reranqueado:** Título: PEC 148/2015 - Senado Federal | Resumo: Sep 1, 2025 ... A proposta visa alterar a Constituição Federal para reduzir a jornada de trabalho semanal de 44 para 36 horas. A mudança será gradual, começando com uma redução ... | Link: https://www25.senado.leg.br/web/atividade/materias/-/materia/124067 | Proposta de Emenda à Constituição n° 148, de 2015 Autoria: Senador Paulo Paim (PT/RS) , e outros Senador Alvaro Dias (PSDB/PR) , Senadora Ângela Portela (PT/RR) , Senador Antonio Carlos Valadar
> **Link:** https://www25.senado.leg.br/web/atividade/materias/-/materia/124067

Evidência 3:

> **Título:** PEC 148/2015 - Senado Federal
> **Conteúdo Reranqueado:** visa alterar a Constituição Federal para reduzir a jornada de trabalho semanal de 44 para 36 horas. A mudança será gradual, começando com uma redução para 40 horas na primeira fase e diminuindo uma hora por ano até atingir o limite de 36 horas semanais. O que diz o autor As possíveis consequências dessa proposta são variadas: - Para os trabalhadores, haverá uma redução na carga horária semanal, o que pode melhorar a qualidade de vida e aumentar o tempo disponível para atividades pessoais e famil
> **Link:** https://www25.senado.leg.br/web/atividade/materias/-/materia/124067

Evidência 4:

> **Título:** PEC 148/2015 - Senado Federal
> **Conteúdo Reranqueado):** aumentar o tempo disponível para atividades pessoais e familiares. - Para os empregadores, a redução da jornada pode exigir ajustes na organização do trabalho e possivelmente a contratação de mais funcionários para manter a produtividade. - Para o mercado de trabalho, a medida pode gerar novas vagas de emprego, ajudando a reduzir o desemprego. - Para a economia, a mudança pode ter impactos diversos, como possível aumento dos custos operacionais para as empresas e possíveis efeitos positivos no
> **Link:** https://www25.senado.leg.br/web/atividade/materias/-/materia/124067

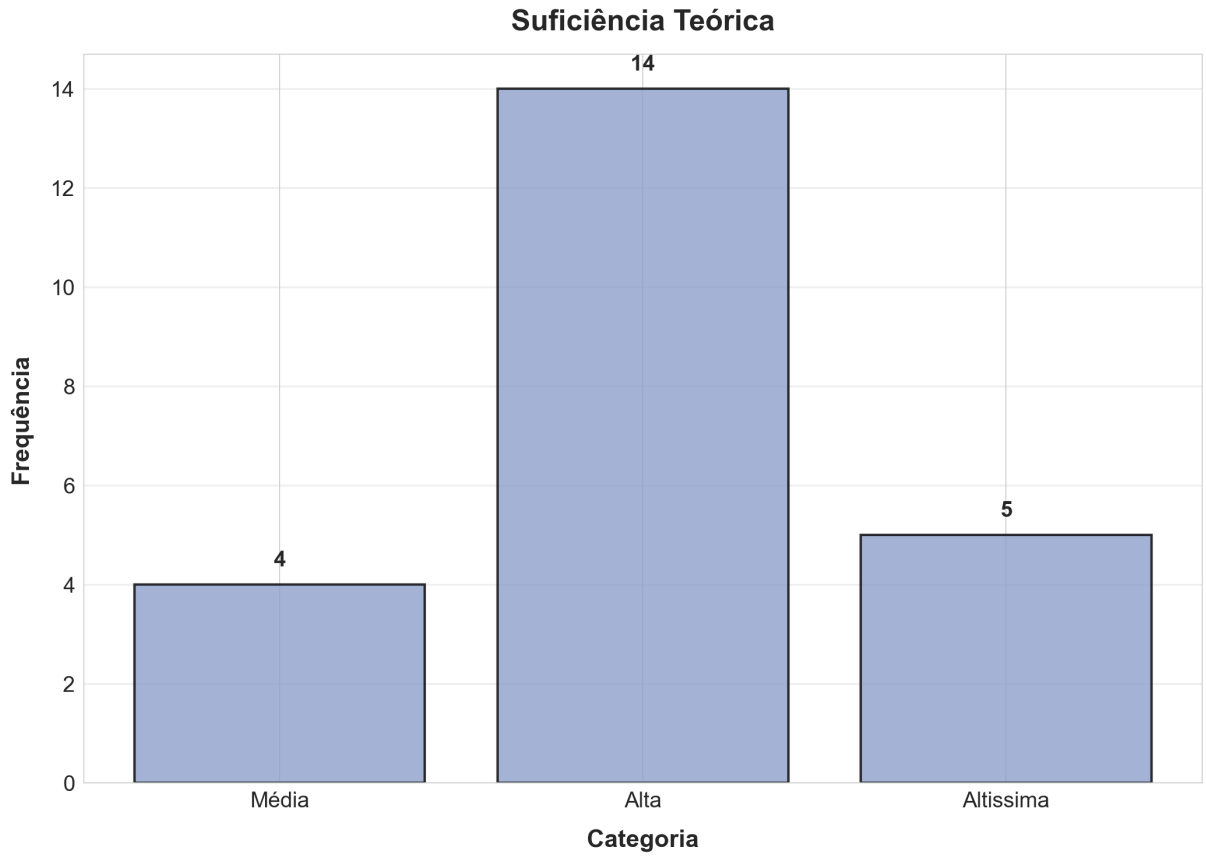
O Estudo de Caso 3 ilustra um cenário onde a suficiência técnica é classificada como “Média“, apesar do retorno de documentos oficiais (Senado Federal). A notícia falsa alega uma ação específica do *Executivo* (Governo Federal atual) baseada em um “estudo piloto” recente. O Polo Técnico, contudo, recuperou repetidamente informações sobre uma Proposta de Emenda à Constituição (PEC 148/2015), que é uma iniciativa do *Legislativo* e data de anos anteriores.

Neste contexto, a avaliação humana considerou a suficiência técnica média porque o agente de busca não conseguiu distinguir a nuance entre uma proposta antiga do Congresso e uma suposta nova diretriz do Governo. Embora os documentos recuperados tratem do tema macro (redução de jornada), eles não servem para confirmar ou refutar diretamente a existência do tal “estudo piloto” ministerial alegado na *fake news*. Contudo, o Agente Veredito da arquitetura quadripolar demonstrou potencial desambiguação, tendo evidências antigas, ele utilizou os Polos Epistemológico e Teórico para identificar a distorção narrativa e chegar ao veredito correto de falsidade, classificando a notícia como uma desinformação que mistura fatos antigos com ficção para gerar credibilidade enganosa.

Polo Teórico

Inicia-se, nesta subseção, a análise qualitativa do Polo Teórico, que é responsável por avaliar a coerência interna, o contexto e as narrativas subjacentes a uma notícia. Sua função é gerar hipóteses concorrentes sobre a veracidade ou falsidade com base no alinhamento da informação e o contexto. Serão apresentados, primeiramente, os resultados da avaliação de suficiência deste polo (Figura 32) e, na sequência, estudos de caso que demonstram como essa análise contextual foi aplicada pelo agente.

Figura 32 – Análise de Desempenho Do Polo Teórico - Configuração C6 com Gemini 2.5 Pro



Fonte: Do Autor

A grande maioria das análises (14 instâncias) foi classificada como “Alta“, indicando que o agente gerou hipóteses plausíveis e bem fundamentadas. Adicionalmente, 5 instâncias foram classificadas como “Altíssima“, superando as expectativas ao construir um raciocínio contextual complexo e preciso, como detalhado no Estudo de Caso D, representado por uma das notícias falsas do Conjunto de Dados Próprio, não houveram marcações de suficiência ‘Baixa‘ para este polo.

Estudo de Caso 4: Falsa (Suficiência Teórica: Altíssima)

Este estudo de caso representa o resultado do Veredito na Configuração C6 com o Gemini 2.5 Pro, com a saída consolidada do Agente, corroborado diante de todas as análises de cada polo realizada previamente. Foi classificada como suficiência Teórica ‘Altíssima‘ pelo fato de que o Polo Teórico levantou hipóteses que corroboraram diretamente com a notícia apresentada, de fato a hipótese definiu de forma clara, o cenário de falsidade.

Notícia Analisada

Cientistas descobrem que o consumo regular de café descafeinado pode levar à perda de memória de curto prazo. Pesquisadores de um instituto de neurociência divulgaram dados preliminares indicando que o processo químico de descafeinação gera compostos que afetam os neurotransmissores.

Diagnóstico do Agente

- **Veredito de Veracidade:** Falsa
- **Tipo de Fake News:** Desinformação
- **Diagnóstico:** A notícia é falsa, configurando-se como desinformação. A análise epistemológica revela [...]. Morfologicamente, o título alarmista (“descobrem que“) entra em contradição com o corpo do texto [...]. **No polo teórico, a alegação se choca com o consenso científico atual, que frequentemente associa o café (incluindo o descafeinado, por seus polifenóis) a benefícios cognitivos, e não a danos**. Tecnicamente, a verificação externa em bases de dados científicas e fontes jornalísticas confiáveis não retorna qualquer estudo que sustente esta afirmação. As fontes fornecidas são irrelevantes ou discutem outros temas (adoçantes), falhando em corroborar a notícia. A construção deliberada com fontes anônimas e o contraste com o conhecimento estabelecido indicam uma criação intencional para enganar.

Resultados do Polo Teórico (Geração de Hipóteses)

Nota: O agente gerou múltiplas teorias concorrentes para avaliar o contexto da alegação. A seguir, um resumo das principais hipóteses geradas.

Hipóteses que Sustentam a VERACIDADE (Consideradas pelo Agente):

- **[Teoria da Descoberta Neurotoxicológica]:** A hipótese de que o processo de descafeinação (usando solventes) poderia gerar subprodutos neurotóxicos, representando um avanço genuíno.
- **[Teoria da Plausibilidade Bioquímica]:** A hipótese de que um composto mimetize ou antagonize a função de moléculas biológicas essenciais, sendo algo cientificamente possível.
- **[Teoria do Alerta Precoce]:** A hipótese de que a divulgação de “dados preliminares“ seria um ato de ética e precaução científica.

Hipóteses que Sustentam a FALSIDADE (Determinantes para o Veredito):

- **[Teoria da Falácia de Correlação/Causalidade]:** A hipótese de que a notícia confunde correlação (pessoas com saúde debilitada que tomam descafeinado) com causalidade (o descafeinado causar o problema).
- **[Teoria da Amplificação e Distorção Midiática]:** A hipótese de que a notícia distorce um achado científico modesto (ex: em doses irreais) para gerar cliques.
- **[Teoria da Campanha de Desinformação Estratégica]:** A hipótese de que a informação foi fabricada por interesses comerciais (ex: concorrentes), usando jargão científico para conferir falsa credibilidade.

Análise de Contexto ("E se ?"):

- **[Teoria da Fonte Fabricada (Astroturfing Científico)]:** O agente questiona: "E se o ’instituto de neurociência’ mencionado não existir ou for uma organização de fachada?". Esta teoria explica a ilegitimidade da origem como tática de desinformação para dar um ar de autoridade a uma alegação inventada.

O Estudo de Caso 4, que analisa uma alegação científica sobre o café descafeinado, é um dos 5 exemplos de suficiência “Altíssima“ identificados na Figura 32. A classificação máxima se justifica pela capacidade do agente de analisar a anatomia da desinformação do domínio científico. O Polo Teórico supôs que a notícia era falsa, articulando precisamente por que ela se parecia com uma notícia falsa, fornecendo as explicações que fundamentaram o veredito final.

Há uma certa sinergia entre o Polos e o Veredito para este Estudo de caso, o diagnóstico do agente aponta a “ausência total de uma fonte identificável“ como a falha epistemológica crítica. O Polo Teórico fornece a explicação para isso, ao gerar a [Teoria da

Campanha de Desinformação Estratégica] e, questionando o contexto, a pergunta “E se ? “ sobre [Astroturfing Científico]. O agente teorizou que a fonte “instituto de neurociência“ era genérica de propósito, que tende a ser tática clássica de desinformação, o que corroborou para o veredito “Falsa“. Neste cenário, o Agente foi realizado na configuração C6, com todos os Polos, mas a decisão unica do polo Teórico na decisão do Veredito, foi concebida no diagnóstico.

Neste estudo de caso, o Agente gerou teorias plausíveis para a veracidade (como a [Teoria da Descoberta Neurotoxicológica]), demonstrando que compreendia o jargão científico. Contudo, ele contrapôs essas hipóteses com teorias de falsidade (como a [Falácia de Correlação/Causalidade] e a [Amplificação Midiática]) que, no contexto de fontes anônimas e alegações alarmistas, podem ser dadas como mais provável de falsidade. A habilidade de pesar narrativas científicas concorrentes e identificar os marcadores certeiros de falsidade é fator observado pelo autor, ao definir a suficiência teórica em nível “Altíssimo“.

É válido ressaltar a natureza metodológica da geração de hipóteses de veracidade neste caso. Teorias como a [Teoria da Plausibilidade Bioquímica] ou a [Teoria do Alerta Precoce] são um exercício comportamental guiado pela função da Ferramenta Teórica, realizado pelo agente. Foi explorada a possibilidade teórica de um nexo causal, assumindo a premissa fundamental de que uma pesquisa legítima e cientistas reais estariam por trás da alegação, cumprindo como função do agente. Contudo, estas hipóteses tornam-se academicamente irrelevantes e perdem o sentido quando confrontadas com os demais polos, pois elas não encontram “existência“ no mundo real, o Polo Epistemológico (que apontou a falha crítica das fontes anônimas) e o Polo Técnico (que falhou em localizar qualquer estudo corroborativo) invalidaram a premissa de que a pesquisa sequer existe. Desta forma, as teorias que sustentam a falsidade, notavelmente a [Teoria da Campanha de Desinformação Estratégica] e a [Teoria da Fonte Fabricada], prevaleceram, pois foram as únicas que conseguiram explicar o fenômeno de fato observado de que a notícia foi fabricada.

Para esta mesma notícia do Estudo de Caso 3, analisa-se a mesma notícia sob a ótica da Configuração C5, que representa a atuação isolada do Polo Teórico. O objetivo é demonstrar como a modelagem teórica foi capaz de identificar a falsidade e fornecer explicabilidade baseada nas hipóteses, mesmo sem o uso de ferramentas de consulta de dados externa.

Diagnóstico do Agente (Polo Teórico)

- **Veredito de Veracidade da Notícia:** Falsa
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Tipo de Fake News:** Desinformação

- **Diagnóstico do Agente:** A notícia é classificada como falsa devido à ausência total de evidências verificáveis, um sinal clássico de desinformação. O conteúdo se apoia em fontes genéricas e não identificadas ("cientistas", "um instituto de neurociência") e "dados preliminares", táticas que impedem qualquer tipo de checagem e conferem uma falsa aparência de credibilidade.

  A análise pelo Método Quadripolar revela que a alegação tem alta similaridade com as teorias que sustentam a falsidade, especialmente a **Teoria da Fabricação para Engajamento (Clickbait)**. A estrutura da notícia — alegação alarmista sobre um produto de consumo massivo — é projetada para explorar o medo e maximizar o compartilhamento, alinhando-se à intenção de manipular em vez de informar.

  Enquanto a **Teoria da Plausibilidade Bioquímica** sugere que um mecanismo adverso é *possível*, a falta de qualquer dado concreto (nome do estudo, pesquisadores, publicação) invalida a alegação como fato. A divulgação científica legítima, mesmo em estágios preliminares, preza pela transparência de fontes e metodologia, elementos completamente ausentes aqui. O conteúdo é, portanto, uma fabricação deliberada.

A análise do Caso 4 sob a configuração C5 (Polo Teórico isolado) revela que, embora a fundamentação teórica tenha contribuído para o acerto do veredito, o fator determinante foi a identificação da ausência de marcadores de dados verificáveis na notícia.

Estudo de Caso 5: Falsa (Erro do Agente - Suficiência Teórica: Média)

Este estudo de caso ilustra um erro de classificação (Falso Negativo) por parte do Agente. A notícia analisada é verdadeira, mas o modelo a classificou incorretamente como “Falsa“. A Suficiência Teórica foi avaliada como “Média“, pois, embora o agente tenha articulado corretamente conceitos complexos de desinformação, ele aplicou esses conceitos sobre uma premissa factual desatualizada ou alucinada referente ao cargo da autoridade citada.

Notícia Analisada

“Argentina é “referência” no Hemisfério Ocidental, diz secretário do Tesouro dos EUA. Ele elogiou o presidente argentino, Javier Milei, por ter feito um "trabalho fantástico"”

Diagnóstico do Agente

- **Veredito de Veracidade da Notícia:** Falsa
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Tipo de Fake News:** Desinformação

- **Diagnóstico:** A notícia é categoricamente falsa. A análise dos quatro polos revela uma falha crítica e insanável no pilar Epistemológico: a identidade da fonte. Os inputs técnicos indicam que as discussões sobre apoio financeiro à Argentina envolvem "Scott Bessent", um gestor de fundos de cobertura e não o Secretário do Tesouro dos EUA. A atribuição da declaração a um cargo oficial que o indivíduo não ocupa é uma fabricação que anula a credibilidade da notícia. Morfologicamente, o texto usa uma estrutura jornalística (aspas, atribuição no título) para criar uma falsa aparência de veracidade, mas a ausência de contexto (local, data, íntegra do discurso) já era um sinal de alerta. Teoricamente, a alegação se encaixa em uma hipótese de propaganda (Desinformação), onde um fatoide é criado para fortalecer a imagem de um governo, especialmente quando fontes externas (BBC) apontam para um contexto de crise. A verificação técnica externa não apenas falha em corroborar a citação, como desmascara a identidade incorreta da fonte. Trata-se de uma peça de desinformação intencional que manipula um alinhamento político real para fabricar um endosso inexistente.

Resultado Polo Teórico: Teorias que sustentam a veracidade:

1. **[Teoria do Realismo Político e Alinhamento Estratégico]:** A declaração seria um ato calculado de diplomacia pública dos EUA para fortalecer um governo ideologicamente alinhado na América do Sul. [...]
2. **[Teoria da Convergência Ideológica Neoliberal]:** A fala do secretário do Tesouro reflete uma genuína aprovação das políticas econômicas de choque adotadas pela Argentina, como a desregulamentação e o corte de gastos. [...]
3. **[Teoria do Jornalismo como Fato Documental]:** A notícia seria verdadeira por se limitar a transcrever com precisão uma declaração pública de uma fonte oficial e de alta credibilidade. [...]

Resultado do Polo Teórico: Teorias que sustentam a falsidade:

1. **[Teoria da Descontextualização e Fragmentação Discursiva]:** A declaração, embora real, pode ter sido tirada de um contexto mais amplo que alteraria seu significado. O termo “referência” [...]
2. **[Teoria da Agenda-Setting e Propaganda]:** [...] Através da criação de um fatoide (um fato fabricado), busca-se pautar a discussão pública, associando a imagem do presidente a uma aprovação de alto nível. A falsidade estaria na fabricação deliberada da informação para manipular a opinião pública.
3. **[Teoria da Interpretação Tendenciosa (Viés de Confirmação)]:** O jornalista ou o veículo de mídia, por afinidade ideológica com o governo Milei, pode ter superinterpretado uma declaração neutra ou ambígua como um elogio enfático. [...].

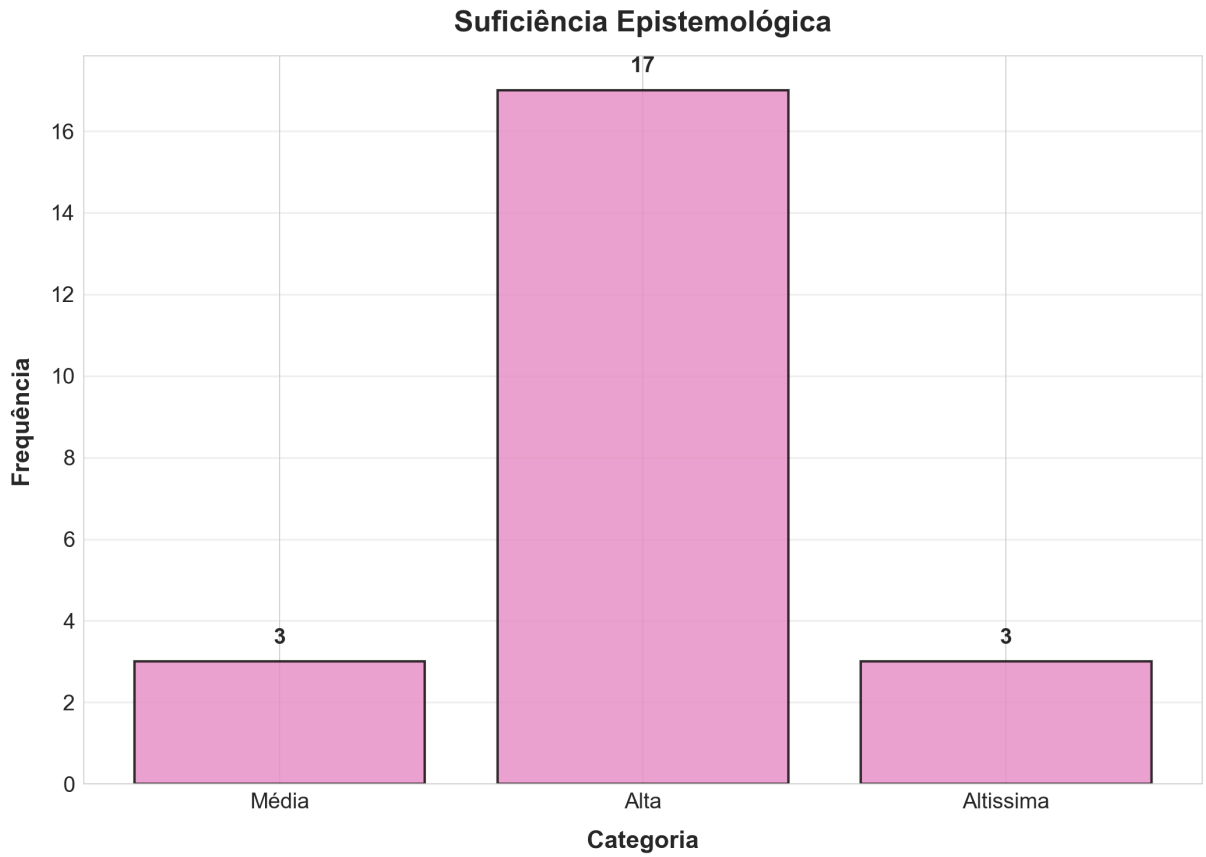
O agente identificou corretamente o nome envolvido (*Scott Bessent*), mas falhou em reconhecer que ele “ocupa (ou foi nomeado para) o cargo de Secretário do Tesouro dos EUA“ na administração vigente à época da notícia.

O Agente invalidou a notícia com base em uma suposta falsa atribuição de “cargo“ do indivíduo. Isso desencadeou uma cadeia de raciocínio teórica incoerente internamente, fazendo com que o agente teorizasse que a notícia era uma “propaganda” criada para inventar um endosso político que, na realidade, existiu.

Polo Epistemológico

Os resultados da avaliação qualitativa deste polo, apresentados na Figura 33, demonstram sua alta confiabilidade. Das 23 notícias analisadas, a esmagadora maioria (17 instâncias) foi classificada como “Alta“, indicando que o agente decompôs com sucesso os elementos centrais da notícia. Notavelmente, três instâncias alcançaram o nível “Altíssimo“, onde o agente identificou os elementos, dissecou conceitos implícitos e estabeleceu relações críticas não evidentes, como demonstrado no estudo de caso a seguir.

Figura 33 – Análise de Desempenho Do Polo Epistemológico - Configuração C6 com Gemini 2.5 Pro



Fonte: Do Autor

Para ilustrar como essa classificação foi aplicada na prática, detalha-se a seguir um dos casos avaliados com suficiência “Altíssima“. O exemplo demonstra a capacidade do agente em decompor uma notícia politicamente complexa em seus componentes de conhecimento verificáveis e implícitos.

Estudo de Caso 6: Verdadeira (Suficiência Epistemológica: Altíssima)

Notícia Analisada

Após Câmara aprovar isenção do IR, governo projeta votação rápida no Senado. Projeto é visto como um importante ativo eleitoral para uma possível campanha de reeleição em 2026.

Resultados do Polo Epistemológico

Nota: O agente decompôs a notícia em seus componentes de conhecimento, fazendo perguntas críticas sobre cada um.

1. Conceitos-chave: Isenção Fiscal (Tributário); Processo Legislativo (Sequência); Ativo Eleitoral (Marketing Político).

2. Instâncias: O “Projeto“ (objeto); O “Governo“ (agente); A “Campanha de 2026“ (contexto futuro).

3. Nomes próprios: Câmara; Senado; IR. (Entidades verificáveis).

4. Entendimento dos substantivos (Perguntas): A “Isenção“ se refere a quê (aumento de faixa?); O “Governo“ é quem (Executivo? Base aliada?); O que qualifica o projeto como “Ativo“ (Popularidade? Impacto econômico?).

5. Relações conceituais: Instrumentalidade (Projeto -> Fim Eleitoral); Sequencialidade e Urgência (Câmara -> Senado Rápido); Causalidade Política (Política Pública -> Cálculo Estratégico).

6. Perguntas epistêmicas (críticas): O que define um “ativo eleitoral“ vs. política ordinária?; Qual a relação entre a “projeção“ do Executivo e a “votação“ do Legislativo?

Diagnóstico do Agente (Veredito Resumido)

- **Veredito de Veracidade:** Verdadeira
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Diagnóstico:** A notícia é factual e sua veracidade é comprovada.
- **(Polo Técnico):** Alegação confirmada por fontes oficiais (Portal do Senado, PL 1.087/2025) e mídia (InfoMoney).
- **(Polo Epistemológico): As informações sobre o processo legislativo são primárias e verificáveis nos diários e portais do Congresso Nacional. A**

alegação sobre a “projeção“ do governo para uma votação rápida e o enquadramento como “ativo eleitoral“ são informações secundárias, típicas da análise política que acompanha o noticiário factual.

- **(Polo Teórico):** Coerente com a Teoria da Escolha Racional (políticas de apelo popular para maximizar capital político).
- **(Polo Morfológico):** Estrutura padrão, com ressalva ao “é visto como“ (voz passiva comum em jornalismo analítico).

A análise do Estudo de Caso 6 demonstra por que este polo obteve uma classificação tão elevada e é um dos 3 casos de suficiência “Altíssima“. A notícia, que se provou de fato verdadeira, é composta por duas camadas de conhecimento: (1) o fato bruto (o trâmite de um projeto de lei) e (2) a análise política (o motivo do trâmite). O Polo Epistemológico do agente foi capaz de separar essas duas camadas e também separar seus componentes de forma precisa.

O nível “Altíssimo“ é justificado pela capacidade do agente de identificar e questionar conceitos que não estavam explícitos, e também implícitos na estrutura da notícia. O ponto mais relevante é a análise do termo “Ativo Eleitoral“. O agente classificou corretamente como um “conceito do marketing político“ e, posteriormente, gerou perguntas epistêmicas sobre ele (“O que qualifica este projeto como um ’ativo’ ? “).

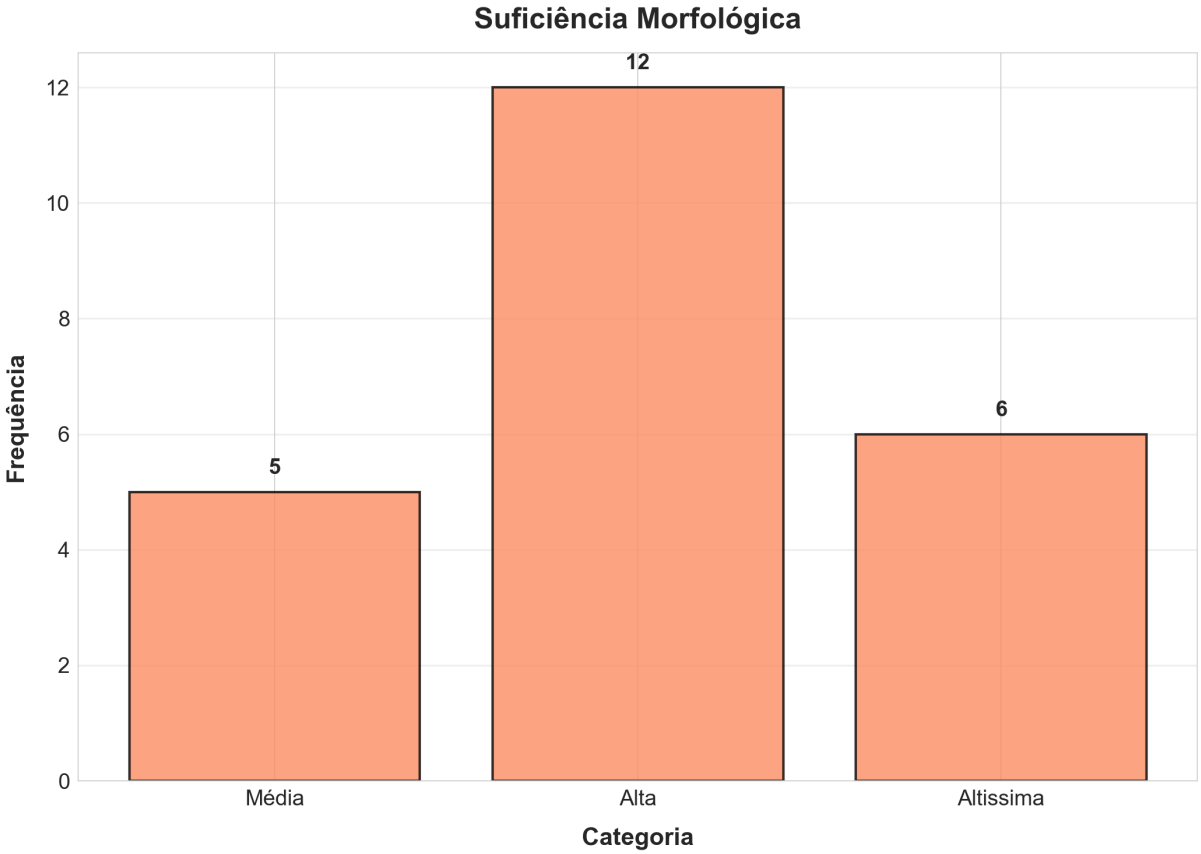
A principal contribuição deste polo, evidenciada no veredito final, foi sua sinergia com o Polo Técnico. A análise epistemológica, ao identificar as “Instâncias“ (Governo, Câmara, Senado) e os “Nomes Próprios“ (IR), criou um roteiro de verificação. Ele essencialmente formulou as nuances que precisariam ser verificadas com o resultado do Polo Técnico. O veredito do agente espelha esta nuance: o Polo Epistemológico identificou que as informações do trâmite eram “primárias e verificáveis“, e o Polo Técnico foi, de fato, capaz de verificá-las (encontrando o “PL 1.087/2025“ no “Portal do Senado“). Essa capacidade de criar uma estrutura crítica e verificável a partir de uma notícia complexa é o que define a excelência desta análise.

De modo geral, o desempenho do Polo Epistemológico foi considerado positivo, com predominância de avaliações ’Altas’. A nota ’Média’ foi reservada para situações em que o modelo foi repetitivo em seus argumentos.

Polo Morfológico

O Polo Morfológico encerra a análise qualitativa voltada aos polos do Método Quadripolar, focando na forma (“como“) a informação é apresentada. Este polo avalia a estrutura do texto, o tom, a sintaxe, o uso de linguagem e a arquitetura geral da mensagem, buscando identificar padrões retóricos e táticas de engajamento que possam indicar manipulação ou falta de credibilidade jornalística.

Figura 34 – Análise de Desempenho Do Polo Morfológico - Configuração C6 com Gemini 2.5 Pro



Fonte: Do Autor

Os resultados da avaliação qualitativa deste polo, visíveis na Figura 34, demonstram um desempenho predominantemente elevado. A maioria das análises (12 instâncias) foi classificada como “Alta“, com 6 instâncias atingindo o nível “Altíssimo“. Isso indica que o agente foi consistentemente capaz de identificar e diagnosticar corretamente a estrutura e o tom das notícias. Apenas 5 casos foram classificados como “Média“, geralmente em notícias curtas onde a análise morfológica era, por natureza, limitada.

Estudo de Caso 7: Verdadeira (Suficiência Morfológica: Altíssima)

Este estudo de caso apresenta uma notícia verdadeira classificada com suficiência morfológica “Altíssima“. O exemplo demonstra a capacidade do Agente em distinguir fatos jornalísticos de conteúdos corporativos ou opinativos.

Notícia Analisada

“Totvs: Se há contabilidade interna, nosso sistema se encaixa como uma luva.”

Análise Morfológica

- **Pontos de Credibilidade Estrutural:** Identificação explícita da fonte no início (“Totvs:”), o que atribui autoria direta e clara à mensagem, eliminando a ambiguidade de um emissor anônimo. Uso de estrutura condicional (“Se há contabilidade interna...”), que delimita de forma precisa o cenário de aplicação e o público-alvo, aumentando a clareza.
- **Indicadores de Risco (Viés):** Emprego de linguagem figurada e coloquial (“se encaixa como uma luva”) para expressar uma opinião subjetiva de perfeição. Tom auto-referencial e promocional explícito (“nosso sistema”), configurando publicidade. Generalização e promessa de adequação absoluta, característica de discursos de marketing que evitam nuances.

Diagnóstico do Agente

- **Veredito de Veracidade da Notícia:** Verdadeira
- **Tipo do Conteúdo:** Opinião / Corporativo
- **Tipo de Fake News:** Não se aplica (Conteúdo Verídico)
- **Diagnóstico:** A afirmação é **Verdadeira**, pois foi de fato proferida pelo presidente da Totvs, Dennis Herszkowicz, em uma entrevista ao “UOL Líderes”. O **Polo Técnico** confirma a autenticidade e a origem da citação, vinculando-a a um veículo de mídia de renome.

  Contudo, a análise dos outros polos revela sua natureza. Epistemologicamente, é uma informação primária (declaração direta), mas sua fonte tem interesse comercial, configurando-se como uma opinião promocional, não um fato auditável. O Polo Morfológico destaca o uso de linguagem figurada (“encaixa como uma luva”), um artifício persuasivo característico de marketing que substitui dados técnicos por uma promessa de perfeição subjetiva. O Polo Teórico aplica a teoria da hipérbole publicitária: a citação completa revela que Herszkowicz qualifica a afirmação ao dizer que, em casos onde a contabilidade não está estruturada, é necessária uma análise “com mais calma”. A frase isolada é autêntica, mas representa a opinião da empresa.

A atribuição de suficiência morfológica “Altíssima“ neste caso justifica-se pela clareza da intenção comunicativa. O texto não apresenta erros de sintaxe, possui um emissor definido e um público-alvo delimitado. O Polo Morfológico detectou com precisão que a estrutura textual obedece perfeitamente às regras do gênero “Discurso Publicitário” (uso de metáforas, foco no produto, promessa de benefício).

A marcação correta do Agente neste caso foi identificar que, embora a forma seja persuasiva e enviesada (o que gerou os alertas de risco), o conteúdo da citação é verídico. O modelo não confundiu “opinião enviesada de uma fonte real” com “notícia falsa”.

Adicionalmente, o destaque da entidade “Totvs”, contribuiu para o Veredito, analisar de forma cruzada, a informação da entidade relacionada com os demais Polos, o que trouxe sinergia para a decisão do Veredito.

Para análise complementar, recorreu-se à notícia sobre o atleta Ramon Dino (Estudo de Caso 1), um exemplo factualmente verdadeiro.

Análise de Contraste: Polo Morfológico (Ref. Estudo de Caso 1), Suficiência Morfológica: Altíssima

---

Notícia Analisada

Referência: Notícia sobre o atleta Ramon Dino (Estudo de Caso 1, Seção 1). Notícia factualmente verdadeira, oriunda de curadoria manual.

Resultados do Polo Morfológico (Diagnóstico da Forma)

Análise Morfológica - Pontos de Credibilidade Estrutural (Sinais Positivos):

- **Lead factual e informativo:** O subtítulo apresenta dados concretos e verificáveis de forma direta (“Fisiculturista brasileiro ficou com o vice-campeonato em duas edições“), seguindo a estrutura da pirâmide invertida jornalística.
- **Linguagem neutra e específica:** A escolha de palavras como “vice-campeonato“, “edições“ e “premiações“ é objetiva e carece de adjetivos de valor ou termos sensacionalistas, reforçando uma aparência de relato factual.

Análise Morfológica - Indicadores de Risco Morfológico (Sinais de Alerta):

- **Título com fórmula de “clickbait“:** O uso do verbo no imperativo (“Saiba quanto...“) é uma construção retórica para gerar curiosidade (curiosity gap) e instigar o clique, mais do que simplesmente informar.
- **Previsão apresentada como fato consumado:** A afirmação “...mas terminou em quarto em 2024“ estrutura um evento futuro como se já tivesse acontecido. Morfologicamente, é uma forma de conferir status de certeza a uma especulação, um sinal de alerta.

Diagnóstico Morfológico:

- **Híbrido (Informativo com Viés Persuasivo):** A estrutura mescla um subtítulo de caráter jornalístico com um título que usa uma fórmula de clickbait e apresenta uma projeção futura como um fato passado, indicando uma intenção persuasiva ou de engajamento acima da pura informação.

A análise morfológica desta notícia (cujo diagnóstico foi “Híbrido (Informativo com Viés Persuasivo)“) foi capaz de identificar corretamente os “Pontos de Credibilidade

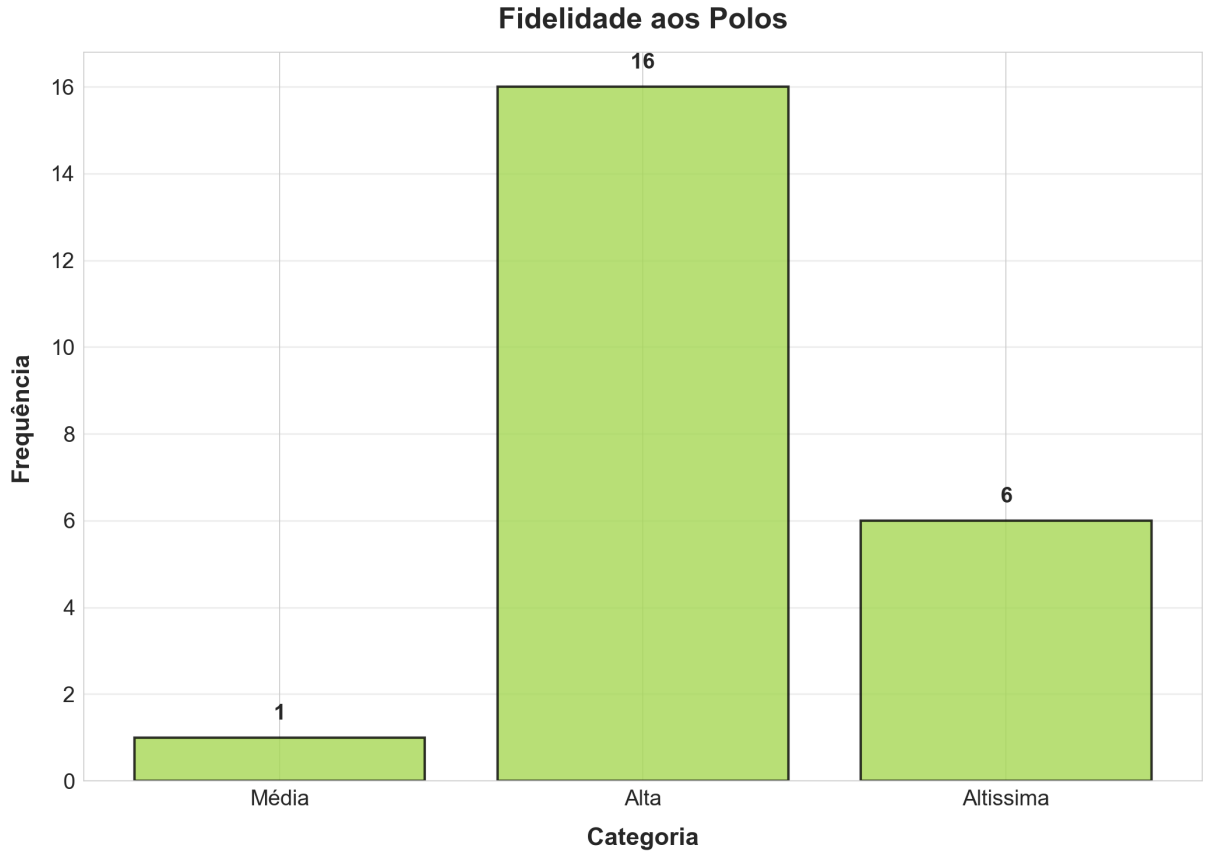
Estrutural (Sinais Positivos)“, como o “Lead factual e informativo“ e a “Linguagem neutra e específica“. Ao mesmo tempo, o agente foi sensível o suficiente para capturar os “Indicadores de Risco Morfológico (Sinais de Alerta)“, explicitamente o “Título com fórmula de ’*clickbait*’“ (“Saiba quanto...“). Este caso evidencia que, na ausência do ruído de pré-processamento, o Polo Morfológico funciona exatamente como esperado, validando os aspectos jornalísticos positivos e, embora alerte para táticas de engajamento, não é levado a um diagnóstico negativo extremo (como “Falsa“), provando sua eficácia em um cenário de análise de uma notícia morfologicamente bem estruturada do ponto de vista de linguagem natural.

Fidelidade aos Polos no Veredito

A etapa final da avaliação qualitativa consiste em mensurar a fidelidade do Veredito do agente às evidências e resultados gerados pelos quatro polos de análise. Esta métrica avalia a coerência interna do raciocínio, determinando se o agente foi capaz de sintetizar as informações dos polos Técnico, Teórico, Epistemológico e Morfológico para construir um diagnóstico final lógico e bem fundamentado.

Os resultados desta avaliação, apresentados na Figura 35, são positivos. Em 23 notícias analisadas, a fidelidade foi considerada “Alta“ (16 instâncias) ou “Altíssima“ (6 instâncias). Isso indica que o agente demonstrou alta capacidade de integrar seus sub-processos analíticos. Apenas 1 instância foi classificada como “Média“, representando um caso (a ser discutido) onde o agente teve dificuldade em analisar os polos conflitantes, classificando a noticia erroneamente.

Figura 35 – Análise de Fidelidade de Polos no Veredito - Configuração C6 com Gemini 2.5 Pro



Fonte: Do Autor

No Estudo de Caso H é ilustrado dois momentos desta análise, sendo um caso de sinergia perfeita (Altíssima) e um caso conflitante na análise do Veredito (Média).

Estudo de Caso 8: Verdadeira (Fidelidade: Altíssima)

Notícia Analisada

Caça Gripen conclui testes em meio a atrasos por questões orçamentárias. O Gripen, caça mais moderno da FAB (Força Aérea Brasileira), concluiu recentemente uma série de testes antes de entrar em operação definitivamente. [...] O marco é conquistado após quase duas décadas de idas e vindas com o FX-2, que ficou mais caro devido a constantes cortes orçamentários, segundo a FAB.

Diagnóstico do Agente (Veredito Integrado)

- **Veredito de Veracidade:** Verdadeira
- **Tipo do Conteúdo:** Factual/Jornalístico

- **Diagnóstico:** O diagnóstico atesta a veracidade da notícia. A análise quadripolar revela múltiplos indicadores de credibilidade. **Polo Epistemológico:** A fonte da informação sensível (atrasos orçamentários) é claramente atribuída à FAB, uma fonte primária e autorizada. **Polo Morfológico:** A estrutura do texto é jornalística e profissional, com linguagem informativa, contextualizada e isenta de sensacionalismo ou apelos emocionais, como aponta a análise prévia. **Polo Teórico:** A informação é altamente coerente com o histórico amplamente documentado do Programa FX-2, onde a conclusão dos testes representa um marco lógico e esperado. As teorias que sustentam a veracidade são mais robustas, alinhando-se à comunicação institucional e à cronologia dos fatos. **Polo Técnico:** A verificação externa corrobora integralmente o evento. A notícia é confirmada por múltiplos veículos de imprensa de grande reputação (UOL, G1), que reportam o mesmo fato, validando sua autenticidade. O conteúdo é factual e representa um marco no programa de modernização da Força Aérea.

O Estudo de Caso 8, extraído do Conjunto de Dados Próprio, representa o cenário ideal de operação do agente e justifica a classificação de fidelidade “Altíssima“. Neste caso, a notícia era de fato verdadeira e o agente concluiu corretamente por sua veracidade. O ponto importante dessa análise é a sinergia entre os polos: não houve um único ponto de ruído, permitindo que o Veredito final fosse construído sobre uma análise completa.

A fidelidade “Altíssima“ é observada na forma como o diagnóstico final atua como uma síntese adequada dos achados parciais. O Polo Epistemológico identificou a fonte primária (FAB), o Polo Morfológico validou a estrutura jornalística, o Polo Teórico confirmou a coerência com o contexto (Programa FX-2) e o Polo Técnico encontrou corroboração externa (UOL, G1). O Veredito, ao citar e integrar cada um desses achados, demonstra que o agente executou corretamente as quatro análises.

Neste cenário de dados, o agente demonstrou um raciocínio que corrobora com o acerto. A ausência de conflito entre os polos permitiu que o Veredito se baseasse na convergência de todas as evidências, representando um grau de fidelidade positivo ao método quadripolar. Este caso prova que, em condições ideais, a arquitetura de múltiplos polos lida de forma adequada com a complexidade textual proveniente dos polos.

Estudo de Caso 9: Verdadeira (Erro do Agente - Fidelidade dos Polos: Média)

Este caso documenta um erro de classificação (Falso Negativo) onde o Agente rejeitou uma notícia verdadeira. A Fidelidade dos Polos foi classificada como “Média“ porque, embora o raciocínio lógico entre os polos tenha sido coerente, houve uma certa alucinação de raciocínio.

Notícia Analisada

“Sete a cada dez moradores de favelas jogam nas bets.”

Notícia real: https://economia.uol.com.br/noticias/redacao/2024/10/05/sete-a-cada-dez-moradores-de-favelas-jogam-nas-bets.htm

---

Diagnóstico do Agente

- **Veredito de Veracidade da Notícia:** Falsa
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Tipo de Fake News:** Desinformação
- **Diagnóstico:** A alegação é fabricada. Epistemologicamente, carece de qualquer fonte primária, metodologia ou autoria, sendo uma estatística órfã e inverificável. Morfologicamente, a estrutura (“Sete a cada dez”) mimetiza o formato de pesquisas jornalísticas para criar uma falsa aparência de credibilidade, enquanto a generalização sobre “moradores de favelas” é um sinal de alerta para narrativas sensacionalistas. Teoricamente, a prevalência de 70% é estatisticamente inverossímil para um comportamento não universal, alinhando-se a teorias de fabricação de pânico moral e não a um cenário realista. A verificação técnica (Polo Técnico) é o fator decisivo: não existe registro dessa informação em nenhuma fonte confiável. A busca externa revela que o mesmo formato estatístico foi usado por institutos como Data Favela para outras pautas (ex: acesso a exames), o que indica fortemente que a alegação sobre “bets” é uma distorção ou criação deliberada, substituindo o tópico real por um falso e de alto impacto para enganar e estigmatizar.

O erro no Estudo de Caso 9 revela uma limitação crítica na calibração do Polo Teórico, que atrapalhou na análise do Veredito. A notícia baseia-se em dados reais do instutudo de pesquisa “Data Favela“. No entanto, o Agente interpretou a magnitude do número (70%) como estatisticamente improvável, acionando um viés de rejeição.

A fidelidade dos polos é classificada como “Média” porque o agente demonstrou capacidade de adequar o Polo Epistemiologico com uma boa narrativa, sobre a falta de fonte verificável. Entretanto, na ausência de uma confirmação técnica imediata, o agente preferiu classificar a notícia como uma noticia falsa, a aceitar um dado que lhe pareceu exagerado.

Para isolar e avaliar criticamente o impacto do método quadripolar na qualidade da explicabilidade, apresenta-se a seguir um estudo de caso de contraste (Estudo de Caso 10). Este caso utiliza exatamente a mesma notícia do Conjunto de Dados Próprio sobre o caça Gripen, analisada anteriormente no Estudo de Caso 8. Contudo, desta vez, a notícia

foi submetida ao agente operando na Configuração C1 (LLM de Base) da tabela 5, ou seja, sem a instrumentação dos quatro polos de análise. O objetivo é comparar diretamente a resposta do agente completo (C6) com a resposta do LLM "puro" (C1) diante do mesmo estímulo.

Estudo de Caso 10: Verdadeira (Configuração C1 - LLM de Base)

Notícia Analisada (Conjunto de Dados Próprio - de Set/2025)

Caça Gripen conclui testes em meio a atrasos por questões orçamentárias. O Gripen, caça mais moderno da FAB (Força Aérea Brasileira), concluiu recentemente [em Setembro de 2025] uma série de testes antes de entrar em operação definitivamente. [...] O marco é conquistado após quase duas décadas de idas e vindas com o FX-2...

Diagnóstico do Agente (Veredito - Configuração C1)

- **Veredito de Veracidade:** Verdadeira
- **Tipo do Conteúdo:** Factual/Jornalístico
- **Diagnóstico:** A notícia é factual e precisa. A conclusão dos testes para a Certificação de Tipo Militar e a declaração de Capacidade Operacional Inicial (IOC) do F-39 Gripen no final de 2023 são fatos confirmados pela Força Aérea Brasileira (FAB) e amplamente noticiados pela imprensa. As informações contextuais sobre a parceria entre Saab e Embraer, os concorrentes derrotados no programa FX-2 (Rafale e F-18) e os atrasos históricos por questões orçamentárias também são verídicas e bem documentadas ao longo da última década. Não há *red flags* ou distorções; a linguagem é objetiva, resumindo corretamente um importante marco no programa de modernização da FAB.

No Estudo de Caso 8 como um contraponto metodológico direto ao Estudo de Caso 10, pois utiliza a mesma notícia verdadeira. Neste cenário, o agente operando na Configuração C1 (LLM *Baseline* da tabela 5, sem os polos de análise) também foi capaz de classificar corretamente o veredito como "Verdadeira". Contudo, uma análise crítica da explicabilidade revela uma distorção fundamental, originada por um viés de treinamento do modelo LLM utilizado.

A notícia fornecida ao agente referia-se a eventos de setembro de 2025. O agente C1, entretanto, ignorou a especificidade temporal da notícia apresentada e justificou seu acerto com base em fatos históricos de seu escopo de treinamento, ao citar a "Capacidade Operacional Inicial (IOC) do F-39 Gripen no final de 2023". Embora o tópico seja o mesmo (Gripen), o evento é distinto. Isso demonstra que o LLM de Base não analisou

de fato a notícia, ele reconheceu o tópico e recorreu à informação mais próxima em sua memória (um viés), produzindo uma explicação que, apesar de parecer correta e factual, estava temporalmente descontextualizada e não provava a veracidade do evento de 2025.

Explicabilidade na Classificação do Tipo de Conteúdo

Além do veredito de veracidade (Falso/Verdadeiro), foi realizada uma análise extra focada na capacidade de explicabilidade do agente em classificar o *tipo* de conteúdo. Esta avaliação é feita para diferenciar, por exemplo, um fato de uma opinião. É importante notar que esta análise de granularidade foi aplicada exclusivamente ao Conjunto de Dados Próprio, que foi submetido às seis configurações experimentais (C1 a C6). Para esta avaliação, foi utilizado apenas o modelo `gemini-2.5-pro`, dado seu desempenho superior nas etapas anteriores. A Tabela 8 consolida a frequência com que cada tipo de conteúdo foi classificado nas 1500 análises resultantes (250 notícias $\times$ 6 configurações).

Tabela 8 – Frequência da Classificação do Tipo de Conteúdo por Configuração (Modelo Gemini 2.5 Pro no Conjunto de Dados Próprio).

Config.	Factual	Ironia	Não Classificado	Opinião	Rumor	TOTAL
C1	241	1	5	3	0	250
C2	228	1	12	9	0	250
C3	237	1	9	3	0	250
C4	248	0	0	2	0	250
C5	245	0	0	5	0	250
C6	246	0	0	3	1	250
C7	210	0	36	4	0	250
TOTAL	**1655**	**3**	**62**	**29**	**1**	**1750**

Fonte: Do Autor.

Como esperado, dada a natureza jornalística do conjunto de dados próprio, a categoria predominante foi “Factual/Jornalístico”. Esta classe responde por 1.655 das 1.750 classificações realizadas, o equivalente a aproximadamente 94,57% do total. As categorias “Ironia/Sátira” (3 ocorrências) e “Rumor” (1 ocorrência) apresentaram incidência estatisticamente irrelevante, refletindo a distribuição original da base de dados.

O maior destaque desta etapa avaliativa reside na coluna “Não Classificado”. Essas 62 instâncias totais não representam uma categoria semântica do conteúdo, mas sim uma falha de geração do LLM (alucinação e quebra de formatação). Esse defeito ocorreu a despeito da engenharia de *prompt* estrita que exigia o preenchimento obrigatório do campo, resultando em respostas em branco ou fora da taxonomia definida pela metodologia.

É metodologicamente relevante observar a distribuição dessas falhas de geração, sendo que a maior concentração ocorreu na configuração C7 (36 falhas), seguida pelas

configurações isoladas C2 (polo morfológico, com 12 falhas) e C3 (polo epistemológico, com 9 falhas). Em contrapartida, as configurações C4 (polo técnico), C5 (polo teórico) e C6 (integração quadripolar) registraram zero abstenções. Esses dados indicam que a ausência de verificação factual externa, restrição arquitetural imposta à C7 e condição natural das configurações isoladas C2 e C3, induz o LLM a uma maior instabilidade inferencial, resultando na quebra das restrições de formatação. Constata-se que o fornecimento de evidências concretas (C4) ou a contextualização estrutural completa (C6) não apenas otimiza o desempenho preditivo, mas também garante o alinhamento (*alignment*) do modelo às instruções do *prompt*, mitigando por completo as falhas de geração nesta tarefa.

6.3 PERFORMANCE EM EXPRESSÕES DE OPINIÃO

Além da avaliação de desempenho no veredito (Falso/Verdadeiro) e da classificação primária do tipo de conteúdo (Factual, Opinião, etc.), foi conduzido um teste de robustez adicional, focado na capacidade de explicabilidade do agente diante de alegações subjetivas. Para este experimento, utilizou-se como base um subconjunto de 10 notícias do Conjunto de Dados Próprio e a configuração C6 com o modelo `gemini-2.5-pro`. Estas notícias, contudo, foram propositalmente alteradas pelo pesquisador. O relato objetivo original foi transicionado para uma declaração opinativa em primeira pessoa (ex: “Eu acredito que...“), e a informação factual foi deliberadamente distorcida para mimetizar os gêneros de Opinião, Rumor ou *Clickbait*. O objetivo foi avaliar se o agente conseguiria, simultaneamente, (1) identificar a morfologia opinativa e (2) refutar a falsidade factual subjacente. Os resultados desta análise estão consolidados na Tabela 9.

Tabela 9 – Síntese da Análise de Veredito sobre Enunciados Opinativos (Configuração C6).

Caso	Notícia	Veracidade	Tipo	Veredito
O-1	“Eu acredito que...“ Mortes por metanol em bebidas; governo é acusado de inação.	Falsa	Opinião	O fato (mortes) é real, mas a premissa (inação do governo) é falsa. O agente identifica a moldura opinativa.
O-2	“Eu acredito que...“ Atleta Ramon Dino fatura “1bi“ após vitória no Mr. Olympia.	Falsa	Opinião	A vitória é real, mas o valor “1bi“ é uma distorção hiperbólica (prêmio de US$ 100 mil).

Tabela 9 – (Continuação da Tabela 9)

Caso	Notícia	Veracidade	Tipo	Veredito
O-3	“Acredito que...“ “Todos“ os moradores de favelas jogam em plataformas de “bets“.	Falsa	Opinião	Generalização universal (“todos“) infundada. O agente ataca o quantificador e o estereótipo.
O-4	“Acredito que...“ Vendas de veículos no país subiram 50% em setembro de 2024.	Falsa	Opinião	O número “50%“ é descontextualizado (era sobre exportação de chassis) e aplicado falsamente ao mercado geral.
O-5	“eu acredito que...“ TSE abrirá códigos-fonte das urnas para inspeção em 2026.	Verdadeira	Opinião	A previsão subjetiva (“Eu acredito que“) revelou-se um prognóstico correto de um fato subsequente.
O-6	“Eu acredito que...“ Putin debocha da OTAN (“aviões dos anos 60“) em recado a Trump.	Falsa	Opinião	Mistura um evento real (discurso de Putin) com detalhes fabricados (menção a “aviões“ ou “Trump“ no contexto).
O-7	“Acredito que...“ Novo SO Russo “promete“ ser “imune“ a hackers ocidentais.	Falsa	Opinião	Impossibilidade técnica (“imune“) usada como propaganda. Distorção de termos (imunidade econômica vs. tecnológica).

Tabela 9 – (Continuação da Tabela 9)

Caso	Notícia	Veracidade	Tipo	Veredito
O-8	“Acho que...“ Série “Peaky Blinders“ terá 2 novas temporadas sem o protagonista.	Falsa	Opinião	Falsidade por imprecisão. Serão *séries derivadas* (spin-offs), não “temporadas“ da série original.
O-9	“Eu acho que...“ Modelo matemático prevê surtos de dengue com “90% de precisão“.	Falsa	Opinião	A métrica extraordinária (“90%“) não é verificada. Caracteriza-se como rumor que exagera pesquisas reais.
O-10	“Eu acho que...“ Marília (SP) inaugura escritório do IBICT em parceria com Israel.	Falsa	Opinião	Mistura um fato real (inauguração do escritório) com uma falsidade (a parceria com o “Governo de Israel“).

Fonte: Do Autor.

O resultado mais imediato e consistente observado nas 10 análises foi a capacidade inequívoca do agente em lidar com a subjetividade. Em todos os casos, os polos Morfológico e Epistemológico identificaram corretamente a natureza do texto de entrada. A Análise Morfológica para os casos (O-1, O-2 e O-3) diagnosticou o conteúdo repetidamente como “Fortemente Opinativo/Persuasivo“, destacando que “A estrutura é inteiramente baseada em crença pessoal (Eu acredito que)“. Da mesma forma, a Análise Epistemológica para os casos (O-1, O-5, O-7) identificou o “Conceito de ’crença’“ ou “proposição subjetiva“ como o pilar da alegação. Isso prova que o agente C6 não é iludido pela forma da mensagem, a arquitetura quadripolar com todos os seus agentes ativos foi capaz de identificar corretamente como “Opinião“ antes mesmo de iniciar a verificação final do Veredito.

A arquitetura quadripolar, cumpriu adequadamente, para este experimento, sua capacidade de detecção de *fake news* e também a realização de explicabilidade de uma maneira adequada em ambientes ambíguos. O Polo Técnico foi acionado para verificar as alegações factuais contidas na opinião. Este processo foi uma das chaves para os achados

da detecção, tanto que se destaca quando executado isoladamente no Conjunto de Dados Próprio (C4). No caso O-1 (Metanol), o agente utilizou fontes (Rádio Senado, Agência Brasil) para confirmar o "núcleo de verdade" (as mortes) e, simultaneamente, refutar a falsidade central (a inação do governo, que estava, de fato, agindo). No caso O-4 (Vendas de Veículos), o agente demonstrou uma capacidade investigativa "cirúrgica", localizando a origem real do número: o Polo Técnico encontrou que o aumento de 50% era real, mas restrito a "exportações de chassis de ônibus pela Mercedes-Benz", permitindo ao agente diagnosticar a alegação geral como Desinformação por descontextualização.

Esta abordagem permitiu ao agente um potencial na classificação da falsidade, diferenciando "Desinformação" (manipulação intencional) de "Misinformação" (erros e imprecisões). No caso O-3 (Favelas/Bets), onde o Polo Técnico não encontrou nenhuma evidência, e o Polo Epistemológico atacou o quantificador universal "todos", o agente classificou corretamente como Desinformação (um estereótipo infundado). Em contraste, nos casos O-8 (Peaky Blinders) e O-10 (IBICT/Israel), o Polo Técnico encontrou um "núcleo de verdade" (haverá *spin-offs*. O escritório do IBICT é real), mas também identificou erros factuais (chamar de *temporada*; incluir *Israel* na parceria). Nestes casos, o agente corretamente abrandou a classificação para Misinformação, ou "falsa por imprecisão", refletindo a distorção de um fato real, junto com um viés opinativo.

Finalmente, a análise do Polo Teórico demonstrou uma capacidade satisfatória de prever a intenção por trás da falsidade. No caso O-2 (Ramon Dino), o veredito alinhou-se perfeitamente à [Teoria da Distorção Hiperbólica] gerada pelo agente ("um fato real (a vitória) é deliberadamente exagerado com uma informação financeira absurda para maximizar o impacto emocional e a viralização"). No caso O-9 (Dengue), o agente diagnosticou um "Rumor", alinhado à sua [Teoria da Distorção Informacional], que previa o exagero de "pesquisas científicas legítimas" (confirmadas pela busca na Fiocruz/FGV) durante a retransmissão da informação. Esta bateria de testes, portanto, valida a eficácia do método C6 em distinguir opiniões, separar o fato do viés opinativo e diagnosticar com precisão a anatomia da falsidade.

6.4 APLICAÇÃO EM GOLPES CIBERNÉTICOS

Para avaliar a aplicabilidade da arquitetura quadripolar completa (C6) fora do domínio estrito de notícias factuais, foi conduzida uma análise qualitativa preliminar, focada no vetor de desinformação de golpes cibernéticos. Para este experimento, utilizou-se um *corpus* de 10 enunciados reais, coletados pelo pesquisador a partir de comunicações autênticas recebidas via SMS e WhatsApp. Estes enunciados exemplificam táticas de engenharia social, *clickbait*, *phishing* e fraude para roubo de dados. O agente foi executado na sua configuração C6 (conforme Tabela 5) com o modelo `gemini-2.5-pro`.

É válido ressaltar que este conjunto de 10 casos não é homogêneo. Ele foi

desenhado para incluir um caso de borda específico (o G-5), que se refere a uma promoção comercial legítima (do serviço “Vale Bonus“). Esta comunicação, embora verdadeira, emprega uma morfologia publicitária agressiva (senso de urgência, promessa de ganho) que é estruturalmente similar a de um golpe. A inclusão deste caso teve como objetivo testar a capacidade de discernimento do agente, avaliando se o método quadripolar conseguiria diferenciar uma campanha de marketing persuasivo de uma tentativa de fraude real. A Tabela 10 consolida os vereditos e as nuances diagnósticas do agente para este *corpus* de teste.

Tabela 10 – Síntese da Análise de Veredito sobre Alegações de Golpe (Configuração C6).

Caso	**Conteúdo**	**Veracidade**	**Tipo**	**Veredito**
G-1	“Parabéns! Você foi selecionado para um prêmio em dinheiro. Clique no link.“	Falsa	Golpe	Polo Técnico (via CNN) confirma que este texto exato é um golpe documentado.
G-2	“Sua conta bancária foi comprometida. Acesse o link imediatamente.“	Falsa	Golpe	Análise Morfológica (alarmista, anônima) e Técnica (fontes de bancos) confirmam a fraude.
G-3	“CRÉDITO DO TRABALHADOR LIBERADO pelo Governo! NEGATIVADO PODE¡‘	Falsa	Golpe	Isca enganosa. Polo Técnico (Estadão) confirma empresas reais, mas (ReclameAQUI) revela queixas.
G-4	“URGENTE: Atualize suas informações de pagamento para evitar a suspensão da sua conta.“	Falsa	Golpe	Análise Morfológica (alarmista, anônima) e Epistemológica (sem fonte) são decisivas.
G-5	“tem Vale Bonus te esperando aqui! Entre no app, se cadastre e economize...“	Verdadeira	N/A	Publicidade legítima. Embora a Morfologia seja de golpe, o Polo Técnico valida a entidade (Vale Bonus).

Tabela 10 – (Continuação da Tabela 10)

Caso	Conteúdo	Veracidade	Tipo	Veredito
G-6	“convite you da Binance, projeto DeFi, retornos diarios 500 a 10000BRL. Jion grupo [WhatsApp].“	Falsa	Golpe	Morfologia (erros grosseiros, “Jion“) e Polo Técnico (Binance alerta que *nunca* usa WhatsApp) selam o veredito.
G-7	“Uma tentativa de golpe foi detectada em sua conta. Entre no [bit.ly] para verificar...“	Falsa	Golpe	Usa a ironia de “detectar um golpe“ como isca. Morfologia (link encurtado, anônimo) é o indicador crítico.
G-8	“Uau - você recebeu 2.000,00 em cupom iF00d Gratis. Garanta já [bit.ly].“	Falsa	Golpe	Polo Técnico (Canaltech) confirma que este golpe exato leva a um app malicioso (APK). A grafia “iF00d“ é a chave.
G-9	“Usuário da Binance, Binance Labs incubou projeto, retornos de até 15% diariamente. Clique no [WhatsApp].“	Falsa	Golpe	Promessa de retorno irrealista (15% ao dia) é a evidência teórica da fraude (Esquema Ponzi).
G-10	“Olá, você foi selecionado para a próxima etapa do processo de entrevista de emprego.“	Falsa	Golpe	A completa ausência de autoria (sem nome de empresa, vaga ou recrutador) é a falha epistemológica fatal.

Fonte: Do Autor

A análise deste *corpus* de testes focado em Golpes Cibernéticos revela, de

forma preliminar, uma demonstração contundente da eficácia do método quadripolar. Os resultados, consolidados na Tabela 10, evidenciam um comportamento de boa assertividade do agente (C6 com `gemini-2.5-pro`), que foi de identificar a fraude, e também de diferenciar fraude de publicidade legítima que mimetiza táticas de fraude.

A primeira linha de defesa do agente foi a sinergia entre os polos Morfológico e Epistemológico, que atuaram como um eficaz sistema de detecção de anomalias. O Polo Morfológico consistentemente identificou os "Indicadores de Risco". No caso G-8 (iF00d), ele detectou a "grafia não-oficial da marca ('iF00d')" e o "link encurtado ('bit.ly') que oculta o destino final". No caso G-6 (Binance/WhatsApp), ele foi ainda mais explícito, notando "múltiplos erros gramaticais, ortográficos e de sintaxe ('convite you', 'Jion')", diagnosticando o texto como "incompatível com uma corporação global". Paralelamente, o Polo Epistemológico atacou a anonimidade. No G-10 (Vaga de Emprego), o veredito se baseou na "completa ausência de autoria (sem nome de empresa, vaga ou recrutador)", classificando-a como uma "falha epistemológica fatal".

O achado mais significativo deste conjunto, é a análise do Caso G-5 (Vale Bonus). Este caso foi o "controle" do experimento, sendo uma promoção verdadeira com uma morfologia suspeita. O Polo Morfológico, corretamente, classificou-a com "indicadores de risco", notando a "linguagem persuasiva e de urgência ('te esperando aqui!')", estruturalmente idêntica a um golpe. Um agente menos sofisticado (como o C1, *baseline*) poderia falhar, gerando um falso positivo, caso o LLM de base não tiver conhecimento do que é "Vale Bonus". O agente C6, no entanto, demonstrou o valor da arquitetura quadripolar quando o Polo Técnico foi acionado, e seu resultado ("confirma que 'Vale Bonus' é um serviço real e estabelecido... O domínio valebonus.com.br é autêntico") sobrepôs-se à suspeita morfológica. O veredito final ("Verdadeira - Publicidade legítima") prova que o agente foi capaz de usar a verificação técnica para arbitrar a ambiguidade e diferenciar com sucesso o marketing agressivo da fraude.

A análise dos resultados preliminares permite inferir que a arquitetura fundamentada no método quadripolar apresentou uma boa performance na detecção de vetores de ataques de golpes. Observou-se que o agente obteve 100% de acurácia no cenário de teste, classificando corretamente as instâncias e gerando camadas de explicabilidade detalhadas para cada polo de análise. Notavelmente, em casos de conteúdo legítimo, como o exemplo *Vale Bônus*, o sistema demonstrou parcimônia e precisão ao atribuir o rótulo *N/A* (*Not Applicable*), evitando a propagação de falsos positivos na dimensão técnica.

No estudo de caso G-5, observou-se que, embora o Polo Morfológico tenha sinalizado o tema de risco, a integração com o Polo Técnico foi determinante para o veredito final. Essa análise dos polos permitiu que o agente diferenciasse de forma fundamentada, publicidade legítima de esquemas fraudulentos, validando a hipótese de que a verificação externa e a análise técnica atuam como elementos importantes.

A despeito da eficácia demonstrada nestes cenários, é essencial reconhecer as limitações estatísticas do atual estágio da pesquisa. Embora a este estudo preliminar tenha validado a lógica do agente, a amostragem de dez cenários distintos é insuficiente para uma generalização metodológica definitiva ou para a estabilização de parâmetros de inferência em larga escala. Portanto, a ampliação do *corpus* de teste e a condução de experimentos em ambientes de dados massivos apresentam-se como etapas essenciais em trabalhos futuros, visando garantir a escalabilidade e a fidedignidade estatística do modelo proposto frente à diversidade do ecossistema de golpes.

6.5 CUSTO

A análise de custo da arquitetura proposta é avaliada a partir de uma análise detalhada do processamento do *dataset* FakeRecogna e do conjunto de dados próprio. Para prover uma compreensão desses custos, além das medidas de tendência central como a média e a moda, utilizam-se os percentis P50, P95 e P99, sendo P50 (mediana) indica o valor que delimita os 50% dos casos de menor custo, o P95 e o P99 descrevem o comportamento dos *outliers*, representando os limites de custo que abrangem 95% e 99% das requisições, respectivamente.

A análise da Tabela 11 revela uma previsível e acentuada disparidade de custos. O método quadripolar completo (C6), que envolve cinco chamadas ao LLM (uma para cada polo e uma para o veredito final), apresenta o custo mais elevado em todos os modelos, atingindo uma média de R$ 0,14 no `gemini-2.5-pro` e R$ 0,18 no `gpt-5`. Em contrapartida, os modelos mais econômicos são, compreensivelmente, os de menor complexidade (`gpt-5-mini`) na configuração mais simples (C1 - Base), com um custo médio irrisório de R$ 0,01 por análise. O ponto mais relevante desta análise surge ao cruzar os dados de custo com os resultados de desempenho classificatório (F1-Score) apresentados na Tabela 7. Observou-se que a Configuração C4 (Polo Técnico isolado) obteve o melhor desempenho classificatório (F1=0,873 no `gemini-2.5-pro`), superando inclusive o método completo C6 (F1=0,853). Do ponto de vista de custo, a C4 (`gemini-2.5-pro`) apresentou um custo médio de R$ 0,03, enquanto a C6, no mesmo modelo, custou R$ 0,14. Isso significa que a configuração C4 foi aproximadamente 4,7 vezes mais econômica (ou 78,6% mais barata) que a C6, entregando um desempenho de classificação superior. O custo total da execução da arquitetura quadripolar com todos os experimentos do Conjunto FakeRecogna foi de R$ 518,90, considerando todos os 4 modelos, 6 configurações, e 100 notícias cada, resultando em 2400 inferências realizadas.

Tabela 11 – Estatísticas de Custo por Análise (BRL) no conjunto de dados *FakeRecogna*.

Config.	Modelo	Média	Moda	P50	P95	P99
C1 (Base)	gemini-2.5-flash	0,00	0,00	0,00	0,01	0,01
	gemini-2.5-pro	0,02	0,02	0,02	0,02	0,03
	gpt-5	0,02	0,01	0,02	0,03	0,04
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00
C2 (Morf.)	gemini-2.5-flash	0,02	0,01	0,02	0,02	0,02
	gemini-2.5-pro	0,06	0,05	0,06	0,06	0,07
	gpt-5	0,07	0,04	0,07	0,08	0,08
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00
C3 (Epist.)	gemini-2.5-flash	0,01	0,01	0,01	0,02	0,02
	gemini-2.5-pro	0,06	0,05	0,06	0,07	0,07
	gpt-5	0,08	0,07	0,08	0,10	0,12
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00
C4 (Técn.)	gemini-2.5-flash	0,01	0,00	0,01	0,01	0,01
	gemini-2.5-pro	0,03	0,03	0,03	0,04	0,04
	gpt-5	0,04	0,03	0,05	0,06	0,06
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00
C5 (Teór.)	gemini-2.5-flash	0,02	0,01	0,01	0,02	0,02
	gemini-2.5-pro	0,07	0,06	0,07	0,07	0,08
	gpt-5	0,09	0,07	0,09	0,11	0,12
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00
C6 (Completa)	gemini-2.5-flash	0,04	0,03	0,04	0,05	0,06
	gemini-2.5-pro	0,14	0,11	0,13	0,16	0,16
	gpt-5	0,18	0,13	0,18	0,22	0,23
	gpt-5-mini	0,01	0,00	0,00	0,01	0,01
C7 (Inferência Única)	gemini-2.5-flash	0,01	0,01	0,01	0,01	0,01
	gemini-2.5-pro	0,03	0,02	0,03	0,03	0,04
	gpt-5	0,04	0,02	0,04	0,06	0,08
	gpt-5-mini	0,01	0,00	0,00	0,00	0,00

Fonte: Do Autor.

Análise de Custo: Conjunto de Dados Próprio

A análise de custo foi replicada sobre o Conjunto de Dados Próprio, desta vez focando apenas nos modelos `gemini-2.5-pro` e `gemini-2.5-flash`, que foram selecionados como os mais promissores na etapa anterior. Os resultados estão dispostos na Tabela 12.

Os custos no Conjunto de Dados Próprio (Tabela 12) confirmam os padrões observados no *FakeRecogna*, com valores médios ligeiramente inferiores, possivelmente devido à natureza dos textos (melhor formatação e menor “ruído“), que podem exigir

Tabela 12 – Estatísticas de Custo por Análise (BRL) no Conjunto de Dados Próprio (Modelos Gemini).

Config.	Modelo	Média	Moda	P50	P95	P99
C1 (Base)	gemini-2.5-flash	0,00	0,00	0,00	0,01	0,01
	gemini-2.5-pro	0,02	0,02	0,02	0,03	0,03
C2 (Morf.)	gemini-2.5-flash	0,01	0,01	0,01	0,02	0,02
	gemini-2.5-pro	0,05	0,05	0,05	0,06	0,06
C3 (Epist.)	gemini-2.5-flash	0,01	0,01	0,01	0,02	0,02
	gemini-2.5-pro	0,05	0,05	0,05	0,06	0,06
C4 (Técn.)	gemini-2.5-flash	0,01	0,00	0,01	0,01	0,01
	gemini-2.5-pro	0,03	0,02	0,03	0,03	0,03
C5 (Teór.)	gemini-2.5-flash	0,01	0,02	0,01	0,02	0,02
	gemini-2.5-pro	0,06	0,06	0,06	0,07	0,08
C6 (Completa)	gemini-2.5-flash	0,03	0,02	0,03	0,03	0,04
	gemini-2.5-pro	0,12	0,10	0,12	0,13	0,13
C7 (Inferência Única)	gemini-2.5-flash	0,01	0,00	0,01	0,01	0,01
	gemini-2.5-pro	0,03	0,03	0,03	0,03	0,04

Fonte: Do Autor.

menos *tokens* para processamento e análise. Novamente, a Configuração C6 (Completa) com `gemini-2.5-pro` é a mais custosa (R$ 0,12), enquanto a C1 (Base) com `gemini-2.5-flash` é a mais econômica (R$ 0,00).

A comparação de custo-benefício se mantém. A C6 (`gemini-2.5-pro`) tem um custo médio de R$ 0,12, enquanto a C4 (`gemini-2.5-pro`), que também apresentou bom desempenho classificatório neste conjunto de dados, teve um custo médio de R$ 0,03. Isso representa uma economia de 75,0%, ou um custo 4,0 vezes menor, para se obter um resultado de classificação (F1-Score) que se provou comparável ou superior ao método quadripolar completo.

A análise de custo total (R$ 556,01 para 3000 análises (inferências) do Conjunto de Dados Próprio, sendo 250 análises por modelo, com 6 configurações) demonstra que o custo médio geral por inferência (R$ 0,18) é viável para fins de pesquisa, mas pode ser um fator limitante para implementação em larga escala (cenários produtivos). Os dados revelam um claro *trade-off* entre o objetivo da classificação e o objetivo da explicabilidade. Se o objetivo primário for a alta explicabilidade, o custo elevado da Configuração C6 (R$ 0,12 - R$ 0,14) é justificável, pois é a única que gera a análise quadripolar completa. Contudo, se o objetivo for um cenário de produção focado puramente na classificação de veracidade, a Configuração C4 (Polo Técnico) se destaca como a escolha de maior custo-benefício. Ela entrega o melhor desempenho de classificação (F1-Score) com um custo operacional

radicalmente inferior (R$ 0,03), especialmente quando combinada com modelos mais ágeis como o `gemini-2.5-flash` (R$ 0,01).

6.6 LATÊNCIA

A análise de latência (tempo de resposta) é o quarto pilar da avaliação de viabilidade, complementando a performance de classificação, explicação e o custo monetário (BRL). Um tempo de inferência excessivamente alto pode inviabilizar a aplicação do agente em cenários de produção, mesmo que ele seja preciso.

Latência no Dataset *FakeRecogna*

A primeira avaliação de latência no conjunto *FakeRecogna* (Tabela 13) mede o tempo total, em segundos, para cada análise. A referida tabela consolida um conjunto de métricas estatísticas descritivas para esta latência, comparando o desempenho dos quatro modelos de linguagem avaliados sob as seis configurações experimentais. Para fornecer uma visão abrangente da distribuição do tempo de resposta, são apresentados os seguintes indicadores: a média (o valor médio de latência), a moda (o tempo de resposta que ocorreu com maior frequência), a mediana (P50), que representa o valor central da distribuição (onde 50% das análises foram mais rápidas), e os percentis de cauda, P95 e P99. Estes últimos indicam o limiar de tempo excedido por apenas 5% e 1% das análises, respectivamente, servindo como indicadores dos cenários de maior latência na inferência.

Como esperado, as configurações C1 (Base) apresentaram as menores latências, com o `gemini-2.5-flash` sendo o mais rápido (média de 11 segundos). O método quadripolar completo (C6) é o mais lento, exigindo múltiplas chamadas sequenciais e/ou paralelas. A latência média do `gemini-2.5-pro` na C6 foi de 375 segundos (cerca de 6 minutos), mas a análise dos percentis revela um grave problema de estabilidade: o P99 (o 1% mais lento) atingiu 1391 segundos (mais de 23 minutos). Isso sugere que a complexidade da C6, dependendo do tamanho do conteúdo, tempo de resposta do reranqueamento, tempo de inferência dos polos, pode levar o modelo a *outliers* de processamento.

Uma observação crítica ocorre na Configuração C4 (Polo Técnico), que obteve o melhor F1-Score na classificação binária. O modelo `gemini-2.5-pro` (C4) apresentou uma latência média de 81 segundos (mediana de 69s), significativamente mais rápido que a C6 (média de 375s). Contudo, o `gpt-5` (C4) demonstrou uma instabilidade: embora sua mediana (P50) fosse de 63 segundos, sua média disparou para 665 segundos (11 minutos), com um P99 de 732 segundos. Isso indica que a ferramenta de busca externa (Polo Técnico) pode, em alguns casos, travar ou atrasar a inferência do `gpt-5`, dada a análise, o que acaba tornando o `gemini-2.5-pro` uma opção mais estável para latência.

Tabela 13 – Estatísticas de Latência por Análise (em Segundos) no conjunto *FakeRecogna*.

Config.	Modelo	Média	Moda	P50	P95	P99
C1 (Base)	gemini-2.5-flash	11	11	11	19	28
	gemini-2.5-pro	21	19	20	28	30
	gpt-5	24	17	21	45	55
	gpt-5-mini	15	11	15	25	33
C2 (Morf.)	gemini-2.5-flash	39	31	39	55	58
	gemini-2.5-pro	59	55	59	68	74
	gpt-5	73	62	66	135	162
	gpt-5-mini	43	41	41	64	72
C3 (Epist.)	gemini-2.5-flash	32	36	31	48	51
	gemini-2.5-pro	57	56	57	68	71
	gpt-5	70	65	67	98	119
	gpt-5-mini	45	42	45	60	69
C4 (Técn.)	gemini-2.5-flash	43	33	41	63	97
	gemini-2.5-pro	81	63	69	98	133
	gpt-5	66	62	63	96	122
	gpt-5-mini	55	47	48	82	184
C5 (Teór.)	gemini-2.5-flash	36	37	36	49	52
	gemini-2.5-pro	69	70	70	79	84
	gpt-5	97	73	89	130	378
	gpt-5-mini	50	52	49	66	77
C6 (Completa)	gemini-2.5-flash	160	92	133	291	370
	gemini-2.5-pro	375	165	187	542	1391
	gpt-5	243	221	226	343	465
	gpt-5-mini	174	111	137	254	287
C7 (Inferência Única)	gemini-2.5-flash	30	25	27	45	74
	gemini-2.5-pro	32	30	31	38	46
	gpt-5	46	31	44	75	89
	gpt-5-mini	23	22	20	34	89

Fonte: Do Autor.

Análise de Latência: Conjunto de Dados Próprio

A análise de latência foi replicada no Conjunto de Dados Próprio (Tabela 14) para verificar se a estabilidade melhorava com dados de conteúdo completo (sem pré-processamento).

Os dados do Conjunto de Dados Próprio confirmam a hipótese anterior, sendo que a qualidade dos dados pode impactar diretamente a estabilidade da latência. No `gemini-2.5-pro` (C6), a latência média caiu pela metade, de 375s para 166s (cerca de 2,7 minutos). Mais importante, os *outliers* extremos não foram evidentes, o P99 caiu de 1391s

Tabela 14 – Estatísticas de Latência por Análise (em Segundos) no Conjunto de Dados Próprio (Modelos Gemini).

Config.	Modelo	Média	Moda	P50	P95	P99
C1 (Base)	`gemini-2.5-flash`	13	11	12	21	50
	`gemini-2.5-pro`	35	26	29	102	136
C2 (Morf.)	`gemini-2.5-flash`	37	34	35	57	70
	`gemini-2.5-pro`	81	65	68	153	285
C3 (Epist.)	`gemini-2.5-flash`	30	28	28	42	66
	`gemini-2.5-pro`	72	62	69	118	158
C4 (Técn.)	`gemini-2.5-flash`	41	34	38	65	94
	`gemini-2.5-pro`	77	68	72	106	162
C5 (Teór.)	`gemini-2.5-flash`	38	36	37	53	61
	`gemini-2.5-pro`	77	75	75	90	161
C6 (Completa)	`gemini-2.5-flash`	80	78	76	118	150
	`gemini-2.5-pro`	166	150	163	192	226
C7 (Inferência Única)	`gemini-2.5-flash`	23	18	21	39	65
	`gemini-2.5-pro`	31	28	29	42	89

Fonte: Do Autor.

para 226s. Isso demonstra que, embora o método C6 apresente latência maior (devido às 5 chamadas de modelo), este tende a ser prejudicado por dados pré-processados, mas se torna significativamente mais estável e previsível em textos bem formatados.

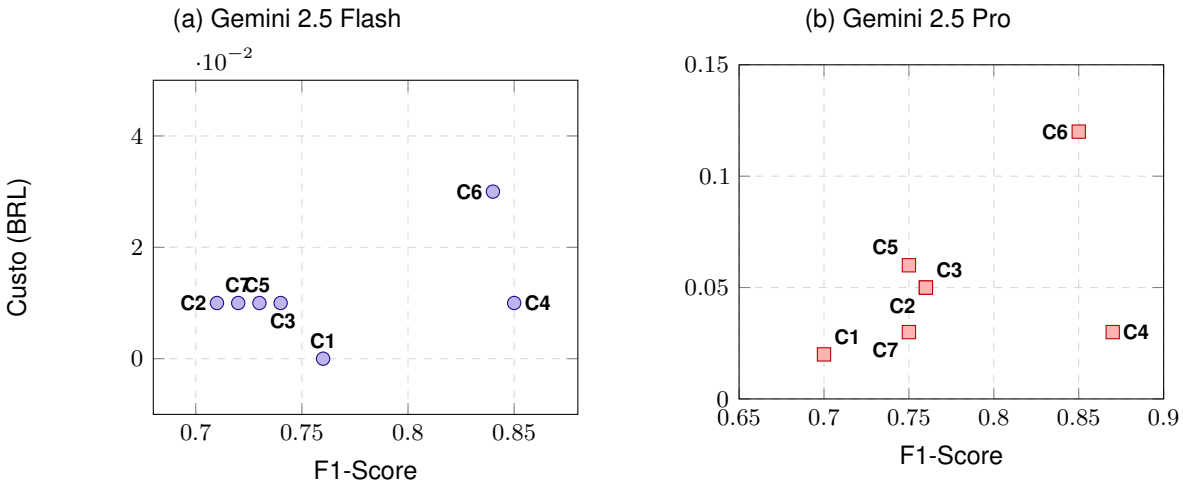
Mesmo no Conjunto de Dados Próprio, o *trade-off* entre desempenho e latência permanece evidente. O C6 (`gemini-2.5-pro`) ainda requer uma média de 166 segundos por análise. Em contrapartida, a C4 (`gemini-2.5-pro`) mantém equilíbrio, com uma latência média de 77 segundos (mediana de 72s) e um P99 estável de 162 segundos. Isso significa que a C4 é mais de duas vezes mais rápida que a C6, apresentando uma latência média (77s) que é, inclusive, mais baixa que o P95 da C1 (102s), indicando alta estabilidade para este experimento.

A análise de latência corrobora as conclusões da análise de custo. A Configuração C6 (Quadripolar Completa), embora seja a mais completa que visa gerar a explicabilidade desejada por este presente projeto, apresenta uma maior latência que a torna inviável em determinadas aplicações em tempo real, especialmente em quando aplicadas à textos ruidosos, ou intermitência no serviço dos provedores, onde pode sofrer com *outliers* de processamento superiores a 20 minutos. Para o objetivo central desta pesquisa, a explicabilidade profunda, esse tempo pode ser aceitável em um processo assíncrono, processando todos os polos em paralelo. No entanto, para um cenário produtivo focado na classificação rápida e precisa, a Configuração C4 (Polo Técnico) se torna a solução

mais adequada. Ela oferece o melhor desempenho de classificação (F1-Score), o melhor custo-benefício monetário (77% mais barata que a C6) e, como agora demonstrado, a melhor performance de latência (mais de 2x mais rápida e significativamente mais estável) entre as configurações de alto desempenho.

A Figura 36 ilustra o comparativo entre Custo x *F1-Score* para o Conjunto de Dados Próprio.

Figura 36 – Comparativo de Desempenho (F1-Score) versus Custo (BRL) para o Conjunto de Dados Próprio



Fonte: Do Autor.

7 CONCLUSÃO

A proliferação de *fake news* no ambiente digital configura-se como um dos grandes desafios da nossa era, com consequências que afetam a democracia, a saúde pública e a estabilidade social. A complexidade do fenômeno revelou as limitações de abordagens puramente técnicas de detecção, que operam como classificações baseadas em uma série de características, utilizando modelos de aprendizado de máquina supervisionado, oferecendo visibilidade limitada sobre suas conclusões diante das decisões cabíveis dentro da linguística. Para enfrentar este desafio, esta dissertação tende a atingir o objetivo de desenvolvimento e validação de um inovador *framework* para a criação de um processo de inteligência artificial auditável e explicativo, fundamentado partindo do Método Quadripolar para a arquitetura de um agente computacional.

A principal contribuição e originalidade deste trabalho reside na criação de uma ponte entre a Ciência da Informação e a Ciência da Computação, que aborda a solução do problema para o cenário de classificação e explicação. Este trabalho é direcionado como um problema de investigação, que é então operacionalizado por um agente de IA. A novidade vai além do uso isolado de LLMs ou da técnica de RAG, foca no composto vinculado destas tecnologias com uma metodologia de pesquisa consolidada. Ao fazer isso, o agente irá processar dados, emular um processo de análise crítica, herdando a estrutura investigativa do Método Quadripolar, também como o nível de explicabilidade das inferências, como a atribuição de uma maior robustez às análises qualitativas.

Diante de uma arquitetura de um agente de IA baseado em LLMs, a incorporação de técnicas de *Retrieval-Augmented Generation* (RAG) possibilita a recuperação dinâmica de informações de fontes externas, enriquecendo a análise com dados atualizados e contextualmente relevantes. Essa abordagem visa trazer contexto para as análises mediante aos polos do Método Quadripolar, que tem como objetivo fundamentar suas saídas em evidências concretas. Além disso, a utilização de *query expansion*, *re-ranking* e busca semântica, tende a melhorar a eficácia da recuperação de informações, priorizando documentos confiáveis e semanticamente alinhados com a consulta.

Este enfoque resulta no diferencial mais significativo da proposta, que se trata da completude quadripolar sob múltiplas faces de inferência, saindo de um classificador binário para uma cadeia de pensamento articulada pelo Método Quadripolar. Este presente projeto utiliza cada polo do Método Quadripolar como um componente de geração de evidências transparentes. Para cada notícia (ou objeto de estudo), a Ferramenta Epistemológica levantou os conceitos, a Ferramenta Morfológica analisou a forma e composição, a Ferramenta Teórica levantou hipóteses aderentes ao fato e a Ferramenta Técnica buscou a prova levando informações externas para o agente. O veredito final foi um rótulo (“falso“

ou “verdadeiro“) alinhado a uma síntese argumentativa estruturada, onde cada parte da explicação se mostra diretamente rastreável a um dos polos analíticos, também agregada a um nível de confiabilidade aderida pelo próprio LLM diante da decisão. Isso confere ao sistema uma explicabilidade intrinsecamente mais robusta do que a de métodos que tentam justificar a posteriori as decisões de um modelo determinístico (abordagens de aprendizado de máquina supervisionado).

Os resultados obtidos nesta dissertação validam a hipótese central de que a estruturação do raciocínio de Grandes Modelos de Linguagem (LLMs) através do Método Quadripolar oferece um ganho substancial na qualidade e na transparência da verificação de fatos. A Configuração C6 (Método Completo), operando com o modelo `gemini-2.5-pro`, consolidou-se como o estado da arte nesta pesquisa, atingindo um F1-Score de 0,853 na classificação binária e entregando o mais alto nível de explicabilidade. Diferentemente da linha de base (C1), que embora capaz de classificar, frequentemente sofre de alucinações ou vieses de treinamento temporalmente descontextualizados, o agente C6 demonstrou uma capacidade superior de fundamentar suas decisões. A integração sinérgica dos polos permitiu dizer se uma notícia é falsa, e também articular por que ela o é, buscando os conceitos da informação (Epistemológico), a forma da mensagem (Morfológico), as teorias que são correlacionadas (Teórico) e as evidências factuais (Técnico) em um veredito coeso e auditável.

A análise de viabilidade revelou, contudo, importantes *trade-offs* entre desempenho e custo computacional. Enquanto o C6 representa a excelência acadêmica e investigativa, seu custo e latência são elevados devido às múltiplas chamadas de inferência e busca. Neste contexto, a Configuração C4 (Polo Técnico isolado) emergiu como uma alternativa de alto valor estratégico para aplicações produtivas de larga escala. Com um *F1-Score* comparável e até ligeiramente superior em métricas puras (0,873), o C4 oferece uma relação custo-benefício otimizada, sendo 77% mais econômico e significativamente mais rápido que o método completo. Isso sugere que, para produtos voltados ao usuário final onde a velocidade é prioritária e a explicação detalhada é secundária, a implementação focada na verificação técnica externa constitui a solução de engenharia mais eficiente.

Para além da verificação factual clássica, a arquitetura proposta demonstrou notável resiliência em cenários adversariais complexos e ambíguos, superando as capacidades de modelos convencionais. Nos testes de segurança contra golpes cibernéticos, o método quadripolar foi capaz de discernir com precisão entre fraudes de engenharia social (como *phishing*) e campanhas de marketing legítimas que utilizam táticas agressivas, arbitrando conflitos entre a morfologia suspeita e a validade técnica das entidades envolvidas. Da mesma forma, na análise de conteúdos opinativos, o agente evitou a armadilha comum de classificar subjetividade como falsidade, conseguindo validar prognósticos corretos mesmo quando envoltos em linguagem de crença pessoal, ou refutar distorções factuais inseridas

em discursos de opinião.

A análise conjunta das métricas quantitativas e qualitativas valida as hipóteses norteadoras desta pesquisa, embora com ressalvas importantes. A Hipótese de Desempenho (H1) foi comprovada de forma parcial: se por um lado a indução semântica da configuração técnica (C4) superou a linha de base (C1) no conjunto de dados próprio, por outro, o desempenho quantitativo da configuração completa (C6) sofreu uma penalização no *dataset FakeRecogna*. Essa queda, contudo, ocorreu justamente porque a C6 capturou com precisão os vieses temporais e a perda de coesão sintática advindas do pré-processamento da base, alterando vereditos estritos do gabarito original. Essa sensibilidade em detectar anomalias contextuais corrobora a Hipótese de Explicabilidade (H2), uma vez que, em contraste com a linha de base, a metodologia quadripolar assegurou explicações com alta suficiência e estrita fidelidade aos eixos analíticos. Por fim, os resultados sustentam a Hipótese da Estrutura Quadripolar (H3) ao demonstrar que a estabilidade analítica do modelo deriva da sinergia entre seus componentes, enquanto a execução isolada de certos polos (como C2 e C3) induziu instabilidades e quebras de instrução no LLM, a integração na C6 mitigou essas falhas. A performance de explicabilidade e F1-Score também é destaque na configuração C4, que validou H1 e H2, trazendo resultados adequados para a performance, dado que trouxe explicações que corroboraram com o acerto do Veredito.

Em suma, esta pesquisa confirma a viabilidade técnica da aplicação de agentes baseados em LLMs para o combate à desinformação, desde que ancorados em metodologias rigorosas de estruturação de contexto como a proposta nesta dissertação. O agente desenvolvido transcende a função de um simples classificador “caixa-preta“, estabelecendo-se como uma ferramenta de investigação assistida que oferece transparência, precisão e segurança. A arquitetura modular permite flexibilidade para diferentes casos de uso, da investigação profunda (C6) à resposta rápida (C4), e apresenta-se como um avanço significativo na construção de sistemas de Inteligência Artificial mais confiáveis, interpretáveis e alinhados com a complexidade do ecossistema informacional contemporâneo.

Trabalhos Futuros

Como principal desdobramento futuro, vislumbra-se a transposição da arquitetura quadripolar de um protótipo de pesquisa para uma infraestrutura de impacto em larga escala. A primeira vertente deste desenvolvimento seria a criação de um serviço de análise via API (*Application Programming Interface*). Esta API encapsularia a capacidade de diagnóstico do agente, permitindo que plataformas de mídia social, agregadores de notícias e outras organizações de tecnologia integrassem a verificação de confiabilidade das notícias, utilizando a abordagem quadripolar diretamente em seus sistemas. O objetivo seria fornecer uma “análise de confiabilidade como serviço“, possibilitando a sinalização automática de conteúdo duvidoso e o enriquecimento de plataformas com uma camada de

análise e investigação auditável.

De forma complementar e com foco no cidadão, a segunda vertente seria a criação de um agente conversacional para aplicativos de mensagens como WhatsApp e Telegram, onde a desinformação se propaga de forma viral. Através de um *chatbot*, qualquer usuário poderia encaminhar um link ou texto suspeito e receber, em segundos, uma análise estruturada sobre a veracidade do conteúdo. Esta iniciativa atacaria diretamente o fator que irá contra a disseminação de notícias falsas, oferecendo uma ferramenta de verificação imediata e acessível, transformando o celular em um aliado contra a desinformação.

Uma evolução natural dessa arquitetura multiagentes seria sua adaptação para o combate a golpes cibernéticos, um domínio adjacente onde a análise contextual e linguística é igualmente crítica. Nesta aplicação, a abordagem metodológica seria reorientada para identificar padrões característicos de fraudes, como a criação artificial de um senso de urgência, ofertas extraordinárias e solicitações de ação imediata que envolvam dados sensíveis, também como a capacidade de inferência em um contexto multimodal (audio, video, imagens, texto). O resultado seria um diagnóstico de risco que classificaria a mensagem como uma potencial ameaça e que também detalharia seus elementos de risco, capacitando o usuário a reconhecer e evitar proativamente tais investidas. Em um estágio mais avançado, o sistema poderia evoluir para a geração de alertas pró-ativos ao identificar a disseminação de novas campanhas maliciosas em tempo real, atuando na prevenção do golpe em larga escala.

Uma fronteira promissora de expansão tecnológica reside no desenvolvimento de um modelo fundacional nativo, especializado através de processos de *fine-tuning* supervisionado. Diferentemente da abordagem atual, que aplica o Método Quadripolar via engenharia de *prompt* sobre modelos generalistas, esta evolução propõe a internalização da metodologia diretamente nos pesos da rede neural. Ao treinar o modelo com um conjunto de dados anotado que contemplaria o veredito e também a decomposição analítica dos polos Epistemológico, Morfológico, Teórico e Técnico, cria-se um modelo que atua intrinsecamente sob esta ótica investigativa. Esta estratégia visa superar as limitações de dependência de *prompts* complexos, produzindo um modelo mais eficiente, capaz de realizar a inferência quadripolar de forma autônoma e consistente.

A materialização destas frentes, da adaptação para segurança cibernética ao treinamento de modelos especializados, impõe desafios significativos de engenharia de *software* e arquitetura de sistemas. Para a disponibilização via API, a padronização estruturada dos dados de resposta é fundamental para a interoperabilidade, já para interfaces de *chatbot*, o foco recai sobre a usabilidade e a clareza na apresentação das evidências ao usuário final. Crucialmente, a estratégia de *fine-tuning* atuaria como um catalisador para a viabilidade produtiva, reduzindo drasticamente a latência e os custos de inferência ao eliminar a necessidade de cadeias de processamento longas (como observado na configuração

C6). A concretização destes passos representaria a aplicação direta desta pesquisa para o fortalecimento da segurança da informação, entregando à sociedade ferramentas robustas para um ecossistema digital mais resiliente e confiável.

REFERÊNCIAS

AHMED, H.; TRAORE, I.; SAAD, S. *Detection of online fake news using N-gram analysis and machine learning techniques*. Cham: Springer, 2017. 127–138 p. Acesso em: 15 abr. 2025. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-69155-8_9.

AHMED, H.; TRAORE, I.; SAAD, S. Detecting opinion spams and fake news using text classification. *Journal of Security and Privacy*, Wiley, v. 1, n. 1, January/February 2018. Acesso em: 10 abr. 2025. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9.

ALAPARTHI, S.; MISHRA, M. *Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey*. 2020. Acesso em: 10 jul. 2024. Disponível em: https://arxiv.org/abs/2007.01127.

ALGHAMDI, J.; LUO, S.; LIN, Y. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, v. 83, n. 17, 2024. Acesso em: 10 abr. 2025. Disponível em: https://link.springer.com/article/10.1007/s11042-023-17470-8.

ALLCOTT, H.; GENTZKOW, M.; YU, C. Trends in the diffusion of misinformation on social media. *Research & Politics*, v. 6, n. 2, p. 8, 2019. Acesso em: 08 abr. 2025. Disponível em: https://doi.org/10.1177/2053168019848554.

ALNABHAN, M. Q.; BRANCO, P. Fake news detection using deep learning: A systematic literature review. *IEEE Access*, v. 12, p. 114435–114459, 2024. Acesso em: 09 jun. 2025. Disponível em: https://ieeexplore.ieee.org/document/10614154.

ANISUZZAMAN, D.; MALINS, J. G.; FRIEDMAN, P. A.; ATTIA, Z. I. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, v. 3, n. 1, p. 100184, 2025. ISSN 2949-7612. Acesso em: 14 out. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S2949761224001147.

ASKARI, A.; ABOLGHASEMI, A.; PASI, G.; KRAAIJ, W.; VERBERNE, S. *Injecting the BM25 Score as Text Improves BERT-Based Re-rankers*. Cham: Springer Nature Switzerland, 2023. 66–83 p. Acesso em: 20 dez. 2024. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-28244-7_5#citeas.

AZAD, H. K.; DEEPAK, A. Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, v. 56, n. 5, p. 1698–1735, 2019. ISSN 0306-4573. Acesso em: 22 abr. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S0306457318305466.

BAŞ, A.; GÖZÜTOK, B.; MARANGOZ, K.; DEMIREL, E.; ERDINÇ, H. Y. Contextual reranking of search engine results. In: *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*. [s.n.], 2022. p. 1–4. Acesso em: 26 fev. 2026. Disponível em: https://ieeexplore.ieee.org/document/9873663.

BHAGDEV, R.; CHAPMAN, S.; CIRAVEGNA, F.; LANFRANCHI, V.; PETRELLI, D. Hybrid search: Effectively combining keywords and semantic searches. In: BECHHOFER, S.; HAUSWIRTH, M.; HOFFMANN, J.; KOUBARAKIS, M. (Ed.). *The Semantic Web:*

Research and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 554–568. ISBN 978-3-540-68234-9. Acesso em: 22 abr. 2025. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-68234-9_41#citeas.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Acesso em: 03 jul. 2024. Disponível em: https://doi.org/10.1023/A:1010933404324.

BROWN, T.; MANN, B.; RYDER, N. e. a. *Language Models are Few-Shot Learners*. Curran Associates, Inc., 2020. 1877–1901 p. Acesso em: 21 abr. 2025. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

BRUYNE, P. D.; HERMAN, J.; SCHOUTHEETE, M. D. *Dinâmica da pesquisa em ciências sociais: os pólos da prática metodológica*. Rio de Janeiro: Francisco Alves, 1977. Tradução de Ruth Joffily. Prefácio de Jean Ladrière. Acesso em: 20 abr. 2025. Disponível em: https://www.scirp.org/reference/referencespapers?referenceid=989217.

CHEN, J.; XIAO, S.; ZHANG, P.; LUO, K.; LIAN, D.; LIU, Z. *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. Bangkok, Thailand: Association for Computational Linguistics, 2024. 2318–2335 p. Acesso em: 14 out. 2025. Disponível em: https://aclanthology.org/2024.findings-acl.137/.

CHEN TIANQI; GUESTRIN, C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. p. 785–794. Acesso em: 18 nov. 2024. Disponível em: https://doi.org/10.1145/2939672.2939785.

CRISTIANINI, N.; RICCI, E. Support vector machines. In: ____. *Encyclopedia of Algorithms*. Boston: Springer US, 2008. p. 928–932. ISBN 978-0-387-30162-4. Acesso em: 07 jun. 2025. Disponível em: https://doi.org/10.1007/978-0-387-30162-4_415.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018. Acesso em: 10 set. 2024. Disponível em: http://arxiv.org/abs/1810.04805.

DONG, X.; ZHANG, X.; BU, W.; ZHANG, D.; CAO, F. A survey of llm-based agents: Theories, technologies, applications and suggestions. In: *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*. [s.n.], 2024. p. 407–413. Acesso em: 26 fev. 2026. Disponível em: https://ieeexplore.ieee.org/abstract/document/10748304.

DOUZE, M.; GUZHVA, A.; DENG, C.; JOHNSON, J.; SZILVASY, G.; MAZARé, P.-E.; LOMELI, M.; HOSSEINI, L.; JéGOU, H. The faiss library. *IEEE Transactions on Big Data*, p. 1–17, 2025. Acesso em: 18 nov. 2024. Disponível em: https://ieeexplore.ieee.org/abstract/document/11202651.

FAYAZ, M.; KHAN, A.; BILAL, M.; KHAN, S. U. Machine learning for fake news classification with optimal feature selection. *Soft Computing*, v. 26, n. 16, p. 7763–7771, 2022. ISSN 1433-7479. Acesso em: 20 abr. 2025. Disponível em: https://doi.org/10.1007/s00500-022-06773-x.

GAO, Y.; XIONG, Y.; GAO, X.; JIA, K.; PAN, J.; BI, Y.; DAI, Y.; SUN, J.; WANG, M.; WANG, H. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. Acesso em: 10 jul. 2024. Disponível em: https://arxiv.org/abs/2312.10997.

GARCIA, G. L.; AFONSO, L. C.; PAPA, J. P. Fakerecogna: A new brazilian corpus for fake news detection. In: *International Conference on Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2022. p. 57–67. ISBN 978-3-030-98305-5. Acesso em: 14 jun. 2025. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-98305-5_6.

GEMINI; ANIL, R.; AL., S. B. et. *Gemini: A Family of Highly Capable Multimodal Models*. 2025. Acesso em: 14 jun. 2025. Disponível em: https://arxiv.org/abs/2312.11805.

Gemini Team; ANIL, R.; BORGEAUD, S.; ALAYRAC, J.-B.; YU, J.; SORICUT, R.; AL, J. S. et. *Gemini: A Family of Highly Capable Multimodal Models*. [S.l.], 2024. Acesso em: 20 set. 2024. Disponível em: https://arxiv.org/abs/2312.11805.

GLASS, M.; ROSSIELLO, G. e. a. *Re2G: Retrieve, Rerank, Generate*. Seattle, United States: Association for Computational Linguistics, 2022. 2701–2715 p. Acesso em: 27 out. 2024. Disponível em: https://aclanthology.org/2022.naacl-main.194/.

GOLDBERG, Y. *Neural network methods for natural language processing*. Morgan & Claypool Publishers, 2017. v. 10. 1–309 p. (Synthesis Lectures on Human Language Technologies, 1). Acesso em: 21 abr. 2025. Disponível em: https://link.springer.com/book/10.1007/978-3-031-02165-7.

GOOGLE. *Google Colaboratory*. Google, 2024. Acesso em: 27 out. 2024. Disponível em: https://colab.research.google.com/.

GOOGLE. *Google Custom Search JSON API*. 2024. Acesso em: 18 abr. 2024. Disponível em: https://developers.google.com/custom-search/v1/overview.

GOUVEIA, L. B. *Uso e exploração do Método Quadripolar no contexto da Ciência da Informação e da Infocomunicação*. 154 p. Tese (Relatório de Pós-Doutoramento em Ciências da Comunicação e da Informação) — Faculdade de Letras, Universidade do Porto, Porto, 2022. Acesso em: 07 abr. 2025. Disponível em: https://ler.letras.up.pt/uploads/ficheiros/19614.pdf.

GURURANGAN, S.; MARASOVIć, A.; SWAYAMDIPTA, S.; LO, K.; BELTAGY, I.; DOWNEY, D.; SMITH, N. A. *Don't stop pretraining: Adapt language models to domains and tasks*. 2020. Acesso em: 21 abr. 2025. Disponível em: https://arxiv.org/abs/2004.10964.

JOHNSON, J.; DOUZE, M.; JéGOU, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, v. 7, n. 3, p. 535–547, 2021. Acesso em: 21 abr. 2025. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8733051.

JOSHI, M.; LEVY, O.; ZETTLEMOYER, L.; WELD, D. BERT for coreference resolution: Baselines and analysis. In: INUI, K.; JIANG, J.; NG, V.; WAN, X. (Ed.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 5803–5808. Acesso em: 27 out. 2024. Disponível em: https://aclanthology.org/D19-1588/.

KARPUKHIN, V. e. a. *Dense passage retrieval for open-domain question answering*. 2020. *arXiv preprint* arXiv:2004.04906. Acesso em: 21 abr. 2025. Disponível em: https://arxiv.org/abs/2004.04906.

KINGER, R.; MURALI, R.; KRISHNA, P.; GORU, M.; UPADHYAYA, S. R. Enhancing llm-based text compression with context-aware frequency adaptation. In: *2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. [s.n.], 2025. p. 1–5. Acesso em: 06 fev. 2026. Disponível em: https://ieeexplore.ieee.org/document/11095536.

KLEINBAUM, D. G.; KLEIN, M. *Logistic Regression: A Self-Learning Text*. 3. ed. New York: Springer, 2010. Acesso em: 09 jun. 2025. ISBN 978-1-4419-1742-3. Disponível em: https://doi.org/10.1007/978-1-4419-1742-3.

LEE, M. A mathematical interpretation of autoregressive generative pre-trained transformer and self-supervised learning. *Mathematics*, v. 11, n. 11, 2023. ISSN 2227-7390. Acesso em: 26 fev. 2026. Disponível em: https://www.mdpi.com/2227-7390/11/11/2451.

LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KÜTTLER, H.; LEWIS, M.; YIH, W.-t.; ROCKTÄSCHEL, T. *et al. Retrieval-augmented generation for knowledge-intensive nlp tasks*. 2020. 9459–9474 p. Acesso em: 20 dez. 2024. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

LI, X.; ZHANG, Y.; MALTHOUSE, E. C. *Large Language Model Agent for Fake News Detection*. 2024. Acesso em: 05 abr. 2025. Disponível em: https://arxiv.org/abs/2405.01593.

LIN, J.; CAMPOS, D.; CRASWELL, N.; MITRA, B.; YILMAZ, E. *Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard*. New York, NY, USA: Association for Computing Machinery, 2021. 2283–2287 p. (SIGIR '21). Acesso em: 21 abr. 2025. Disponível em: https://doi.org/10.1145/3404835.3463034.

LIN, J.; MA, X.; LIN *et al.* Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021. (SIGIR '21), p. 2356–2362. ISBN 9781450380379. Acesso em: 21 abr. 2025. Disponível em: https://doi.org/10.1145/3404835.3463238.

LIU, N. F.; LIN, K.; HEWITT, J.; PARANJAPE, A.; BEVILACQUA, M.; PETRONI, F.; LIANG, P. *Lost in the Middle: How Language Models Use Long Contexts*. 2023. Acesso em: 27 out. 2024. Disponível em: https://cs.stanford.edu/~nfliu/papers/lost-in-the-middle.arxiv2023.pdf.

LIU, Y.; AL., T. H. et. *Summary of ChatGPT-Related research and perspective towards the future of large language models*. 2023. 100017 p. Acesso em: 29 out. 2024. Disponível em: https://www.sciencedirect.com/science/article/pii/S2950162823000176.

MA, X.; GONG, Y.; HE, P.; ZHAO, H.; DUAN, N. *Query Rewriting for Retrieval-Augmented Large Language Models*. 2023. 13 p. Acesso em: 18 abr. 2024. Disponível em: https://arxiv.org/abs/2305.14283.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to information retrieval*. Cambridge University Press, 2008. Acesso em: 21 abr. 2025. Disponível em: https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C\#authors.

MANZOOR, S. I.; SINGLA, J.; NIKITA. Fake news detection using machine learning approaches: A systematic review. In: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. [s.n.], 2019. p. 230–234. Acesso em: 10 abr. 2025. Disponível em: https://ieeexplore.ieee.org/document/8862770.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. [s.n.], 2013. Acesso em: 21 abr. 2025. Disponível em: https://dblp.org/rec/journals/corr/abs-1301-3781.bib.

MILEV, I.; BALUNOVIć, M.; BAADER, M.; VECHEV, M. *ToolFuzz – Automated Agent Tool Testing*. 2025. Acesso em: 05 abr. 2025. Disponível em: https://arxiv.org/abs/2503.04479.

MOHAMMADABADI, S. M. S.; KARA, B. C.; EYUPOGLU, C.; UZAY, C.; TOSUN, M. S.; KARAKUş, O. *A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications*. 2025. Acesso em: 27 out. 2024. Disponível em: https://www.mdpi.com/2079-9292/14/18/3580.

NAGPAL, N. Query expansion for information retrieval using word embeddings: A comparative study. In: *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*. [s.n.], 2022. p. 289–293. Acesso em: 27 out. 2024. Disponível em: https://ieeexplore.ieee.org/document/9988667.

OAD, A.; FAROOQ, M. H.; ZAFAR, A.; AKRAM, B. A.; ZHOU, R.; DONG, F. Fake news classification methodology with enhanced bert. *IEEE Access*, v. 12, p. 164491–164502, 2024. Acesso em: 26 mar. 2025. Disponível em: https://ieeexplore.ieee.org/document/10742347.

OAK, R.; HAROON, M.; JO, C.; WOJCIESZAK, M.; CHHABRA, A. *Re-ranking Using Large Language Models for Mitigating Exposure to Harmful Content on Social Media Platforms*. 2025. Acesso em: 22 abr. 2025. Disponível em: https://arxiv.org/abs/2501.13977.

OpenAI. *GPT-4 Technical Report*. [S.l.], 2023. Acesso em: 14 jun. 2025. Disponível em: https://arxiv.org/abs/2303.08774.

O'SHEA, K.; NASH, R. *An Introduction to Convolutional Neural Networks*. 2015. Acesso em: 26 mar. 2025. Disponível em: https://arxiv.org/abs/1511.08458.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Acesso em: 09 jun. 2025. Disponível em: https://aclanthology.org/D14-1162/.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. *Improving Language Understanding by Generative Pre-Training*. [S.l.], 2018. Acesso em: 07 jun. 2025. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, v. 21, n. 1, p. 5485–5551, 2020. Acesso em: 14 out. 2025. Disponível em: https://jmlr.org/papers/volume21/20-074/20-074.pdf.

RAINA, V.; GALES, M. *Question-Based Retrieval using Atomic Units for Enterprise RAG*. 2024. Acesso em: 27 out. 2024. Disponível em: https://arxiv.org/abs/2405.12363.

REIMERS, N.; GUREVYCH, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. Acesso em: 20 set. 2024. Disponível em: https://arxiv.org/abs/1908.10084.

REUTHER, A.; MICHALEAS, P.; JONES, M.; GADEPALLY, V.; SAMSI, S.; KEPNER, J. *Survey and Benchmarking of Machine Learning Accelerators*. 2019. 1-9 p. Acesso em: 29 out. 2024. Disponível em: https://ieeexplore.ieee.org/abstract/document/8916327.

ROBERTS, J. How powerful are decoder-only transformer neural models? In: *2024 International Joint Conference on Neural Networks (IJCNN)*. [s.n.], 2024. p. 1–8. Acesso em: 26 fev. 2026. Disponível em: https://ieeexplore.ieee.org/document/10651286.

ROBERTSON, S.; ZARAGOZA, H. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, v. 3, n. 4, p. 333–389, 2009. ISSN 1554-0669. Acesso em: 21 abr. 2025.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. 2009. Acesso em: 24 jun. 2024. Disponível em: https://docs.python.org/3/reference/index.html.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 4. ed. Harlow, England: Pearson, 2021. ISBN: 978-0134610993. Disponível em: https://aima.cs.berkeley.edu/.

Tf-idf. In: SAMMUT CLAUDE; WEBB, G. I. (Ed.). *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. p. 986–987. ISBN 978-0-387-30164-8. Acesso em: 03 jul. 2024. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_832.

SCHMIDT, R. M. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. ArXiv preprint arXiv:1912.05911. Acesso em: 26 mar. 2025. Disponível em: https://arxiv.org/abs/1912.05911.

SHI, W. e. a. *REPLUG: Retrieval-augmented black-box language models*. 2023. Acesso em: 21 abr. 2025. Disponível em: https://aclanthology.org/2024.naacl-long.463.pdf.

SUN, C.; HUANG, S.; POMPILI, D. Llm-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, v. 10, n. 6, p. 5681–5688, 2025.

THULKE, D. e. a. *Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog*. 2021. Acesso em: 21 abr. 2025. Disponível em: www-i6.informatik.rwth-aachen.de/publications/download/1176/Thulke--2021.pdf.

TING, K. M. Precision and recall. In: ____. *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. p. 781–781. ISBN 978-0-387-30164-8. Acesso em: 20 abr. 2025. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_652.

TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A.; LACROIX, T.; ROZIèRE, B.; GOYAL, N.; HAMBRO, E.; AZHAR, F.; RODRIGUEZ, A.; JOULIN, A.; GRAVE, E.; LAMPLE, G. *LLaMA: Open and Efficient Foundation Language Models*. 2023. Acesso em: 14 out. 2025. Disponível em: https://arxiv.org/abs/2302.13971.

TROTMAN, A.; PUURULA, A.; BURGESS, B. Improvements to bm25 and language models examined. In: *Proceedings of the 19th Australasian Document Computing Symposium*. New York, NY, USA: Association for Computing Machinery, 2014. (ADCS '14), p. 58–65. ISBN 9781450330008. Acesso em: 29 out. 2024. Disponível em: https://doi.org/10.1145/2682862.2682863.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. *Attention is all you need*. 2017. Acesso em: 29 out. 2024. Disponível em: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

VERBERNE, S. Pretrained transformers for text ranking: Bert and beyond. *Computational Linguistics*, v. 49, n. 1, p. 253–255, 03 2023. ISSN 0891-2017. Acesso em: 22 abr. 2025. Disponível em: https://doi.org/10.1162/coli_r_00468.

WÖLFLEIN, G.; FERBER, D.; TRUHN, D.; ARANDJELOVIC, O.; KATHER, J. N. *LLM Agents Making Agent Tools*. Vienna, Austria: Association for Computational Linguistics, 2025. 26092–26130 p. Acesso em: 05 abr. 2025. Disponível em: https://aclanthology.org/2025.acl-long.1266/.

YANG, C.; WU, Q.; WANG, J.; YAN, J. *Graph Neural Networks are Inherently Good Generalizers: Insights by Bridging GNNs and MLPs*. 2023. Acesso em: 26 mar. 2025. Disponível em: https://arxiv.org/abs/2212.09034.

YENDURI, G.; RAMALINGAM, M.; SELVI, G. C. e. a. Gpt (generative pre-trained transformer): A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, v. 12, p. 54608–54649, 2024. Acesso em: 22 abr. 2025. Disponível em: https://ieeexplore.ieee.org/document/10500411.

ZAMANI, H.; DEHGHANI, M.; CROFT, W. B.; LEARNED-MILLER, E.; KAMPS, J. *From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing*. New York, NY, USA: Association for Computing Machinery, 2018. 497–506 p. (CIKM '18). Acesso em: 21 abr. 2025. Disponível em: https://doi.org/10.1145/3269206.3271800.

ZHAN, J.; MAO, J.; LIU, Y.; GUO, J.; ZHANG, M.; MA, S. *Interpreting Dense Retrieval as Mixture of Topics*. 2021. Acesso em: 22 abr. 2025. Disponível em: https://arxiv.org/abs/2111.13957.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	14/03/1996
Nacionalidade	Brasileira
Nome em citações bibliográficas:	HENKE, GABRIEL HENKE, G.
Currículo Lattes	https://lattes.cnpq.br/3154346923913944
ORCID	https://orcid.org/0009-0004-0716-5622

FORMAÇÃO ACADÊMICA	
27/08/2021	Tecnologia em Big Data no Agronegócio FATEC Pompeia

PRODUÇÃO BIBLIOGRÁFICA
HENKE, G. ; ANDRADE, G. Estudo de aplicação de machine learning para a previsão da cotação da soja. 2021, Pompeia : RIC-CPS, 2021.

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS
Encontro de Ciência de Dados e Inteligência Artificial, 23.out.2025, (Pompeia). O Poder da Engenharia de Machine Learning. 2025. (Palestra).
FATEC DAY, 10.fev.2024, (Pompeia). Big Data e Inteligência Artificial. 2024. (Palestra).