RESSALVA

Atendendo solicitação do autor, o texto completo desta dissertação será disponibilizado somente a partir de 13/05/2027.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

PROGRAMA DE PÓS-GRADUAÇÃO
EM CIÊNCIA DA COMPUTAÇÃO

FUSÃO BASEADA EM RANQUEAMENTO PARA RECUPERAÇÃO AUMENTADA POR MÚLTIPLOS MODELOS DE LINGUAGEM

MURILO MISSANO BELL

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

MURILO MISSANO BELL

FUSÃO BASEADA EM RANQUEAMENTO PARA RECUPERAÇÃO AUMENTADA POR MÚLTIPLOS MODELOS DE LINGUAGEM

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Daniel Carlos Guimarães Pedronette
Coorientador: Prof. Dr. Ivan Rizzo Guilherme

Rio Claro - SP
2026

B434f

Bell, Murilo Missano
Fusão baseada em ranqueamento para recuperação aumentada por múltiplos modelos de linguagem / Murilo Missano Bell. -- Rio Claro, 2026
82 f. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Daniel Carlos Guimarães Pedronette
Coorientador: Ivan Rizzo Guilherme

1. Ciência da computação. 2. Recuperação da informação. 3. Processamento de linguagem natural (Computação). I. Título.

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

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa contribui para o avanço da recuperação de informação ao investigar múltiplas perspectivas na representação da necessidade de informação. A abordagem tem potencial de tornar sistemas de busca mais eficazes, facilitando a localização de informações relevantes e apoiando aplicações de inteligência artificial que dependem de recuperação.

POTENTIAL IMPACT OF THIS RESEARCH

This research contributes to the advancement of information retrieval by investigating multiple perspectives in the representation of information needs. The approach has the potential to make search systems more effective, improving the retrieval of relevant information and supporting artificial intelligence applications that rely on retrieval.

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

MURILO MISSANO BELL

FUSÃO BASEADA EM RANQUEAMENTO PARA RECUPERAÇÃO AUMENTADA POR MÚLTIPLOS MODELOS DE LINGUAGEM

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora

Prof. Dr. DANIEL CARLOS GUIMARÃES PEDRONETTE
IGCE / UNESP / Rio Claro (SP)

Prof. Dr. LUCAS PASCOTTI VALEM
ICMC / USP / São Carlos (SP)

Prof. Dr. ARNALDO CANDIDO JUNIOR
IBILCE / UNESP / São José do Rio Preto (SP)

Conceito: Aprovado.

Rio Claro (SP), 13 de maio de 2026.

AGRADECIMENTOS

Agradeço à Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), ao Instituto de Geociências e Ciências Exatas (IGCE), ao Centro de Ciências Naturais Aplicadas (UNESPetro) e ao Departamento de Estatística, Matemática Aplicada e Computação (DEMAC) pela infraestrutura e ambiente acadêmico que possibilitaram o desenvolvimento desta pesquisa.

Aos meus pais, meus avós, minha irmã e demais familiares pelo amor incondicional, pelo incentivo e pelo o suporte durante o desenvolvimento deste trabalho.

Ao meu orientador, Prof. Dr. Daniel Carlos Guimarães Pedronette e ao meu co-orientador, Prof. Dr. Ivan Rizzo Guilherme, pela orientação, pelos ensinamentos e pela disponibilidade em contribuir com minha formação como pesquisador.

Aos amigos e colegas da UNESPetro e do DEMAC pelas conversas, pela troca de experiências, pelas sugestões e pelos momentos de descontração.

Aos professores do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da UNESP, pelos conhecimentos transmitidos ao longo do curso, pelas discussões e por contribuírem para minha formação acadêmica.

Aos membros da banca examinadora, por aceitarem o convite para avaliar o presente trabalho e pelas contribuições e sugestões para o aprimoramento da dissertação.

À Fundação para o Desenvolvimento da UNESP (FUNDUNESP) e à Petróleo Brasileiro S.A. (Petrobras) pelo apoio financeiro concedido por meio do Processo nº 3070/2019.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processos nº 313193/2023-1 e nº 2024/04890-5, respectivamente.

Por fim, agradeço a todos que, de alguma forma, contribuíram para a realização desta dissertação, mesmo que não tenham sido mencionados nominalmente. Cada gesto de apoio foi importante para que este trabalho se tornasse realidade.

“Os limites da minha linguagem significam os limites do meu mundo.“
(Ludwig Wittgenstein) [1]

RESUMO

Os Grandes Modelos de Linguagem (LLMs – *Large Language Models*) apresentam potencial em várias áreas da computação dada sua capacidade de geração de textos coerentes e semanticamente relevantes. Apesar de seu sucesso, a alucinação permanece como um dos principais problemas enfrentados por esses modelos. Métodos de Geração Aumentada por Recuperação (RAG – *Retrieval-Augmented Generation*) têm se demonstrado abordagens efetivas para mitigação do problema. Nas abordagens de RAG, bases de conhecimento e sistemas de recuperação de informação são utilizados para proporcionar informações contextuais aos LLMs. Com o intuito de melhorar as informações contextuais retornadas aos modelos de linguagem, o método RAG-Fusion explora o uso de LLMs para expansão de consultas, por meio da geração de variações da consulta original. As variações são submetidas a um sistema de recuperação, que produz um conjunto de *rankings*. Os *rankings* são combinados por meio do método de Fusão do Ranqueamento Recíproco (RRF – *Reciprocal Rank Fusion*), produzindo o *ranking* final. Apesar do RAG-Fusion e demais métodos de expansão de consultas serem amplamente discutidos, pelo que se tem conhecimento, nenhuma abordagem investiga a complementaridade nos níveis de geração e recuperação de forma conjunta; tampouco analisa diferentes métodos de fusão e seus impactos no contexto de expansão de consultas. O presente trabalho foca em analisar essa complementaridade, propondo a abordagem de **R**ecuperação **A**umentada por **M**últiplos **MO**delos de li**N**guagem (RAMMON). O RAMMON integra múltiplos LLMs para expansão de consultas, o que expande o espectro de busca ao considerar diferentes interpretações semânticas que os modelos podem possuir; e um sistema de recuperação híbrido, ampliando as representações das consultas. Além disso, o presente trabalho propõe três novos métodos de fusão de *rankings*, comparando-os com o RRF no cenário de expansão de consultas definido no RAMMON. Os resultados experimentais demonstram que o RAMMON obtém resultados comparáveis aos métodos de estado da arte em expansão de consultas, superando-os na métrica de *Recall*, em média. Um dos métodos de fusão propostos supera o RRF na métrica de *Precision*. O presente trabalho explora a complementaridade como principal conceito, contribuindo para a expandir a análise do conceito na área de expansão de consultas e com potencial de embasar novas pesquisas em complementaridade de métodos e modelos.

Palavras-Chave: expansão de consultas; fusão baseada em ranqueamento; recuperação de informação; grandes modelos de linguagem.

ABSTRACT

Large Language Models (LLMs) have demonstrated potential across several areas of computer science due to their ability to generate coherent and semantically relevant text. Despite their success, hallucination remains one of the main challenges faced by these models. Retrieval-Augmented Generation (RAG) methods have proven to be effective approaches for mitigating this problem. In RAG approaches, knowledge bases and information retrieval systems are employed to provide contextual information to LLMs. Aiming to improve the contextual information returned to language models, the RAG-Fusion method explores the use of LLMs for query augmentation by generating variations of the original query. These variations are submitted to an information retrieval system, which produces a set of rankings. The rankings are then combined using the Reciprocal Rank Fusion (RRF) method, producing the final ranking. Although RAG-Fusion and other query expansion methods have been widely discussed, to the best of our knowledge, no existing approach jointly investigates complementarity at both the generation and retrieval levels; nor does any study analyze different ranking fusion methods and their impact in the context of query expansion. This work focuses on analyzing such complementarity by proposing the Retrieval Augmented by Multiple Language Models (RAMMON – **R**ecuperação **A**umentada por **M**últiplos **MO**delos de li**N**guagem) approach. RAMMON integrates multiple LLMs for query expansion, expanding the search spectrum by considering the different semantic interpretations that language models may provide; and a hybrid retrieval system, enriching query representations. Furthermore, our work proposes three novel rank fusion methods and compares them with RRF within the query augmentation scenario defined by RAMMON. The experimental results show that RAMMON achieves comparable results to state-of-the-art query augmentation methods, outperforming them in terms of Recall on average. One of the proposed fusion methods surpasses RRF in terms of Precision. Our work explores complementary as their main concept, contributing to the expansion of this concept analysis in the field of query augmentation and has the potential to support future research on the complementarity of methods and models.

Keywords: query augmentation; ranking-based fusion; information retrieval; large language models.

Title in english Ranking-based Fusion for Retrieval Augmented by Multiple Language Models.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AP	*Average Precision*
API	*Application Programming Interface*
BC	*Borda Count*
BEIR	*Benchmarking IR*
BMβ	*Best Match* β
BERT	*Bidirectional Encoder Representations from Transformers*
CCE	*Core Content Extraction*
CNN	*Convolutional Neural Network*
DCG	*Discounted Cumulative Gain*
DMQR-RAG	*Diverse Multi-Query Rewriting Framework*
E5	*Embeddings from Bidirectional Encoder Representations*
ERRF	*Exponential Reciprocal Rank Fusion*
GQR	*General Query Rewriting*
HNSW	*Hierarchical Navigable Small Worlds*
IDCG	*Ideal Discounted Cumulative Gain*
IPG-SW	*Instructional Permutation Generation - Sliding Window*
KG	*Knowledge Graph*
KWR	*Keyword Rewriting*
LLM	*Large Language Model*
LRRF	*Logarithmic Reciprocal Rank Fusion*
MAP	*Mean Average Precision*
MILL	*Mutual VerIfication method with Large Language model*
MMLF	*Multi-query Multi-passage Late Fusion*
MQRF-RAG	*Multi-Query Rewriting Framework*

MRR	*Mean Reciprocal Rank*
NDCG	*Normalized Discounted Cumulative Gain*
PAR	*Pseudo-Answer Rewriting*
PARADE	*Passage Representation Aggregation for Document Reranking*
PLM	*Pre-trained Language Model*
PRP	*Pairwise Ranking Prompting*
RAG	*Retrieval-Augmented Generation*
RAGF	*RAG-Fusion*
RAMMON	Recuperação Aumentada por Múltiplos MOdelos de liNguagem
RI	Recuperação de Informação
RRF	*Reciprocal Rank Fusion*
SRRF	*Sigmoid Reciprocal Rank Fusion*
TF-IDF	*Term Frequency - Inverse Document Frequency*
T5	*Text-to-Text Transfer Transformer*

LISTA DE SÍMBOLOS

d_i	*i-ésimo* documento da coleção de documentos.
$\mathcal{D}$	Coleção de documentos.
q	Consulta da coleção de consultas.
N	Total de documentos de $\mathcal{D}$, ou seja $N = \|\mathcal{D}\|$.
$\mathcal{D}_L$	Subconjunto de $\mathcal{D}$ contendo L documentos.
L	Tamanho do subconjunto $\mathcal{D}_L$, ou seja, $L = \|\mathcal{D}_L\|$.
τ_q	Representação do *ranking* obtido para a consulta q.
s	Função para o cálculo do *score* de relevância.
j	Valor do julgamento de relevância.
V	Vocabulário da coleção de documentos.
t	Termo do vocabulário V.
$w_{i,t}$	Peso do termo t no documento d_i.
$\text{freq}_{i,t}$	Frequência do termo t no documento d_i.
df_t	Quantidade de documentos que contém o termo t.
$w_{q,t}$	Peso do termo t na consulta q.
n_t	Total de documentos de $\mathcal{D}$ que contém o termo t.
K_1	Constante do BMβ, definida empiricamente.
$len(d_i)$	Tamanho do documento d_i.
avg_doclen	Média do tamanho dos documentos de $\mathcal{D}$.
$B_{i,t}$	Fator de frequência do termo t no documento d_i no método BM25.
b	Contante no intervalo $[0, 1]$ do método BM25.
Rel_q	Conjunto de relevantes para a consulta q.
$\mathcal{Q}$	Coleção de consultas.
K	Limite de corte do *ranking*.

$\mathcal{D}_{K,q}$	Subconjunto de $\mathcal{D}$, contendo K documentos, obtidos com base na consulta q.
$rank_q$	Posição do primeiro relevante no *ranking* gerado pela consulta q.
$rel(k)$	Valor de relevância atribuído ao documento na k-ésima posição do *ranking*.
$\mathcal{T}_q$	Conjunto de *rankings* obtidos com base na consulta q, ou seja, $\mathcal{T}_q = \{\tau_1, \tau_2, ...\tau_W\}$.
W	Quantidade de listas ranqueadas de $\mathcal{T}_q$, ou seja, $W = \lvert\mathcal{T}_q\rvert$.
S^i_{Comb}	Função para o cálculo do novo *score* de relevância do documento com base em sua posição no *ranking* $\tau_i \in \mathcal{T}_q$, para os métodos da família Comb.
S_{method}	Função para o cálculo do novo *score* dos documentos utilizando o método $method$, em que $method = \{BC, RRF, PRP, LRRF, ERRF, SRRF\}$.
λ	Constante do método de Fusão do Ranqueamento Recíproco.
$p_{i,j}$	Probabilidade do documento d_i ser mais relevante que o documento d_j.
psg_j	j-ésima passagem do conjunto de passagens que representa um determinado documento, por exemplo, $d_i = \{\text{psg}_1, \text{psg}_2, \ldots \text{psg}_j, \ldots, \text{psg}_{np}\}$.
np	Quantidade de passagens de um determinado documento.
d_i^{cls}	Representação densa de D_i^{cls}.
D_i^{cls}	Conjunto das representações de relevância das passagens de documento um determinado documento.
n_{wind}	Tamanho da janela deslizante.
n_s	Valor do deslize da janela deslizante a cada iteração.
$\mathcal{X}$	Espaço de textos.
$\mathcal{M}$	Espaço de modelos codificadores.
μ	Modelo codificador, em que $\mu \in \mathcal{M}$.
δ	Quantidade de dimensões do vetor gerado pelo modelo codificador.
σ	Função de medida de similaridade.
$r_{\mathcal{D}}$	Função de recuperação baseado no conjunto de documentos $\mathcal{D}$.
ϵ	Função de expansão de consultas.

w	Número de reescritas de uma determinada consulta.
τ_q^*	Lista ranqueada obtida após o processo de combinação de *rankings*.
f	Função de fusão baseada em ranqueamento.
$\mathcal{R}_q$	Conjunto de *rankings* obtidos para a consulta q após aplicação da expansão de consultas e da recuperação.
ρ	Hiperparâmetro do método de Fusão de Ranqueamento Recíproco Exponencial (ERRF).
α	Hiperparâmetro do método de Fusão de Ranqueamento Recíproco Sigmoide (SRRF).
$n_{\mathcal{Q}}$	Quantidade de consultas de um conjunto de dados.
n_r	Quantidade de *rankings* em que um determinado documento aparece.

SUMÁRIO

1 INTRODUÇÃO

O fundamento de um sistema de Recuperação de Informação (RI) é recuperar e ranquear documentos de acordo com sua relevância a uma consulta fornecida por um usuário. O termo "documentos" pode ser entendido como qualquer tipo de informação: textos, imagens, vídeos, áudio, entre outros [2]. O objetivo desses sistemas é garantir que os documentos recuperados atendam à necessidade de informação do usuário, ou seja, que a recuperação seja eficaz; e que o sistema responda ao usuário em tempo hábil, ou seja, que a recuperação seja eficiente. Com o intuito de proporcionar a eficácia e a eficiência dos sistemas de RI, a área de pesquisa de recuperação de informação dedica-se ao estudo de novos algoritmos e técnicas. Atualmente, os métodos de domínio textual estudados nessa área podem ser classificados como métodos de recuperação esparsa e métodos de recuperação densa [2, 3].

Na abordagem de recuperação esparsa, o *score* de relevância dos documentos a uma determinada consulta é obtido a partir da análise de informações léxicas intrínsecas ao texto. O modelo vetorial e o modelo probabilístico são dois dos métodos mais conhecidos nessa abordagem [4]. Apesar de serem muito adotados dentro da área de RI, os métodos de recuperação esparsa sofrem com os problemas de *lexical gap* [5] e *vocabulary mismatch* [6]. Ambos os termos são frequentemente utilizados de forma intercambiável e referem-se à dificuldade dos sistemas em retornar documentos que atendam à necessidade de informação do usuário. A dificuldade ocorre quando consultas e documentos apresentam similaridade semântica, mas não similaridade léxica [2, 7, 8].

Como forma de reduzir o impacto do problema apresentado pela recuperação esparsa, foram propostos os métodos de recuperação densa. A abordagem de recuperação densa adota, principalmente, modelos baseados em aprendizado profundo, os quais se popularizaram na última década devido aos avanços no estado da arte obtidos. Dentre os avanços, podem-se citar aqueles alcançados pela arquitetura de Transformers [9], que atingiu o estado da arte em áreas como processamento de linguagem natural [10, 11, 12], visão computacional [13, 14, 15] e processamento de áudio [16, 17].

A proposta da recuperação densa é que os modelos de aprendizado profundo produzam representações vetoriais de baixa dimensionalidade do conteúdo dos documentos, denominadas de *embeddings* [8]. Os *embeddings* são compostos pelas informações semânticas dos dados, em que, com base no cálculo de similaridade entre os vetores, é possível obter o *score* de relevância dos documentos. As abordagens de recuperação densa podem ser subdivididas em duas, de acordo com a arquitetura do modelo utilizada para a geração dos *embeddings*. São estes os modelos com arquitetura *cross-encoder* e os modelos com arquitetura *bi-encoder* [3]. Os modelos *cross-encoders* apresentam a dependência consulta-documento no processo de geração de *embeddings*, em que ambos são enviados em conjunto para definir a relevância do documento para a respectiva consulta. Já os *bi-encoders* codificam as consultas e os documentos de forma

independente, em que é aplicado um algoritmo de similaridade entre os *embeddings* gerados para calcular a relevância do documento [2].

Além de tópicos de estudo para a melhoria no processo de recuperação, a pesquisa na área de RI também aborda formas de aprimorar técnicas que ocorrem no entorno da recuperação. As técnicas são denominadas: (*i*) pré-recuperação, que ocorre antes da busca ser realizada; e (*ii*) pós-recuperação, que acontece depois da execução da busca [18]. Métodos de expansão de consulta [19, 20], reescrita de consulta [21], re-ranqueamento contextual [22, 23, 24] e fusão de *rankings* [25, 26, 27] são exemplos dessas técnicas. Os métodos de re-ranqueamento contextual têm como objetivo gerar um novo *ranking* a partir do *ranking* original retornado pelo processo de busca e de informações contextuais adicionais presentes nos dados. Os métodos re-ranqueiam a lista ranqueada original com o objetivo de melhorar a eficácia dos resultados.

A pesquisa na área de RI ganha ainda mais importância com o surgimento e a popularização dos Grandes Modelos de Linguagem (LLMs – *Large Language Models*). Os sistemas de RI se tornam responsáveis por recuperar documentos de contexto para embasar as respostas geradas por esses modelos e mitigar a geração de textos imprecisos [28]. Os LLMs advêm dos Modelos de Linguagem Pré-treinados (PLMs – *Pretrained Language Models*) e são baseados, principalmente, na arquitetura de Transformers do tipo *decoder-only* (somente decodificador). Diferente dos PLMs, os LLMs são treinados em uma quantidade massiva e diversa de dados, podendo atingir centenas de bilhões de parâmetros [29]. A escala de treinamento proporciona aos modelos uma enorme capacidade de compreensão e geração de textos, com habilidades de aprendizado *zero-shot* e *few-shot*, em que nenhum ou poucos exemplos são passados no *prompt* do modelo como base para realizar uma tarefa, e não há necessidade de tunagem de parâmetros. LLMs como DeepSeek-R1 [30], GPT-4 [31], Llama 3 [32], Mistral-7B [33] e Phi4 [34], possuem grande capacidade em uma variedade de tarefas [35], como tradução [36], classificação de textos [37] e perguntas e respostas [38].

Os sistemas de RI auxiliam os LLMs a mitigar um de seus principais problemas: a alucinação, ou seja, a produção de respostas imprecisas ou irrelevantes [35]. Na abordagem de Geração Aumentada por Recuperação (RAG – *Retrieval-Augmented Generation*), os documentos recuperados pelos sistemas de RI passam a integrar o *prompt* dos LLMs, o que fundamenta e contextualiza o modelo para atender a solicitação do usuário [28]. A técnica de RAG-Fusion [39] introduz a expansão de consultas utilizando LLMs e a fusão baseada em ranqueamento com o método de Fusão de Ranqueamento Recíproco (RRF – *Reciprocal Rank Fusion*) nos sistemas de RAG. A partir da consulta original, o modelo de linguagem gera diversas reformulações da consulta, ampliando o escopo de busca. Cada variação é submetida ao sistema de recuperação, o qual produz um conjunto de listas ranqueadas. As listas são posteriormente combinadas pelo RRF para produzir o *ranking* final. Ao gerar múltiplas consultas e combinar os resultados obtidos da recuperação, o RAG-Fusion obtém um *ranking* mais alinhado com as necessidades de informação do usuário e com maior relevância de contexto para LLMs em sistemas de RAG.

Considerando premissas similares ao RAG-Fusion, trabalhos recentes [40, 41, 42, 43] focam

na estratégia de expansão de consultas para aumentar a eficácia dos sistemas de RI e ampliar a representação da consulta original. Os métodos DMQR-RAG [42] e MQRF-RAG [43] implementam quatro estratégias de reescrita de consultas diferentes e um método adaptativo de seleção utilizando LLM. O MILL [40] aplica um sistema de verificação mútua baseado na similaridade por cosseno entre os documentos recuperados e os documentos gerados pelo modelo de linguagem. O MMLF [41] adota um fluxo em três estágios. O primeiro estágio aplica a geração de sub-consultas por meio de LLM, seguido de um segundo estágio que implementa a expansão de consulta para passagem (*query-to-passage expansion*). O terceiro estágio realiza a recuperação e a combinação dos *rankings*.

Apesar dos avanços na área de expansão de consultas, até onde se sabe, as abordagens existentes se fundamentam na utilização de um único LLM para a execução dos métodos propostos; ou seja, não há análises que avaliem a capacidade de complementaridade entre LLMs voltados para a produção de múltiplas perspectivas de uma mesma consulta. Tal fato resulta em métodos que não exploram o potencial da variedade de interpretações semânticas da combinação de múltiplos LLMs. Além disso, os métodos de expansão de consultas geralmente adotam a abordagem RRF como estratégia de fusão padrão, enquanto métodos alternativos de fusão permanecem ainda pouco explorados na área. Baseado no escopo dos trabalhos, destaca-se uma potencial oportunidade de pesquisa com o intuito de aprimorar a eficácia dos resultados obtidos. A melhora na eficácia visa advir por meio da análise de como a complementaridade de LLMs pode auxiliar na geração de múltiplas consultas, e como os *rankings* produzidos são combinados.

A complementaridade é um importante tópico na área de RI, em que diferentes paradigmas de recuperação visam capturar aspectos distintos de relevância, o que motiva a produção de modelos de recuperação híbridos [44]. De forma similar, a complementaridade também é relevante na combinação de múltiplos LLMs [45]. Esses modelos podem variar em decorrência das diferenças em arquitetura, quantidade de parâmetros, método de tokenização, vocabulário, dados de treinamento e metodologia; o que resulta em respostas distintas.

Com foco na exploração da complementaridade nos níveis de geração e recuperação, é proposto o método de **R**ecuperação **A**umentada por **M**últiplos **MO**delos de li**N**guagem (RAMMON). O RAMMON é um método que explora o uso de múltiplos LLMs e múltiplos modelos de recuperação. A partir da combinação de ambos, visa-se enriquecer a representação da necessidade de informação do usuário ao gerar variações da consulta original e considerar sinais léxicos e semânticos de relevância. Através da exploração das capacidades de complementaridade das consultas produzidas por diferentes modelos e dos documentos retornados por diferentes abordagens de recuperação, o método proposto tem como intuito capturar uma maior variedade de interpretações e representações contextuais da consulta. Os *rankings* de cada consulta gerada, para cada recuperador, são combinados por meio de métodos de fusão.

Considerando a área pouco explorada da combinação dos *rankings* de múltiplas consultas, são propostos três novos métodos de fusão de listas ranqueadas, que são analisados em relação à sua eficácia e integrados no RAMMON.

6 CONCLUSÃO

O presente trabalho propôs a implementação da abordagem de Recuperação Aumentada por Múltiplos MOdelos de liNguagem (RAMMON). Através da abordagem, foram exploradas, de forma conjunta, as capacidades de complementaridade na expansão de consultas por meio de múltiplos LLMs e na recuperação ao utilizar um sistema de busca híbrido. A complementaridade na expansão de consultas e na recuperação considerou múltiplas interpretações semânticas e representações da consulta. Adicionalmente, foram propostos três novos métodos de fusão baseados em ranqueamento (LRRF, ERRF e SRRF) com o intuito de avaliar como novas abordagens de fusão se comparam com o RRF.

Os resultados evidenciam que a complementaridade entre LLMs tende a ser eficaz ao considerar diferentes perspectivas de uma mesma consulta. A aplicação de múltiplos LLMs possibilitou o aumento de documentos relevantes na lista ranqueada, destacado pelo aumento no *Recall@100* na avaliação experimental. A combinação do método de fusão LRRF com o RAMMON apresentou resultados competitivos, superando o RRF em *Precision@10* e *Recall@100* na maioria dos conjuntos de dados. Tal fato sugere que há possibilidade de melhora nos resultados de recuperação ao considerar métodos alternativos de fusão de *rankings* na área de expansão de consultas.

Na comparação com métodos SOTA, observou-se que o RAMMON apresentou a maior média de *Recall@100*, e o maior valor de *NDCG@10* em 3 dos 8 conjuntos de dados avaliados. A análise da abordagem proposta em relação aos métodos SOTA indicou que, como os métodos podem ser aplicados de forma conjunta com o RAMMON, sua combinação em um único *pipeline* pode melhorar os valores de *NDCG@10* para os demais conjuntos. O fato embasa investigações futuras que ampliem a análise da complementaridade entre geração e recuperação, incluindo a complementaridade entre diferentes métodos de expansão de consultas.

O trabalho contribuiu para expandir a análise na área de expansão de consultas, ao considerar a complementaridade nos níveis de geração e recuperação como elementos relevantes. Por meio da avaliação da abordagem proposta, forneceu-se uma perspectiva mais abrangente sobre o impacto da complementaridade na recuperação de informação. A perspectiva mais abrangente serve de base para pesquisas futuras sobre a aplicação do conceito de complementaridade entre diferentes métodos de expansão. Em conjunto, a análise dos métodos de fusão evidenciou oportunidades associadas a integração dos métodos com abordagens de expansão de consultas, indicando potencial para o desenvolvimento de novos métodos de fusão na área. Como resultado das contribuições do trabalho, produziu-se um artigo científico, o qual foi submetido à revista acadêmica *Information Sciences* [107] e encontra-se em processo de revisão.

Relacionado às limitações do RAMMON, destaca-se que o método de expansão de consultas se limitou a dois LLMs como reescritores e dois modelos de recuperação na avaliação experimental. O aumento da quantidade de modelos pode levar à obtenção de melhores resultados.

Outro ponto a ser considerado é a variabilidade das respostas dos LLMs, que pode influenciar os resultados das expansões. A quantidade de consultas geradas se limitou a 3 consultas por modelo. O impacto da variação do valor de w_{Mistral} e w_{Phi4} pode ser avaliado em trabalhos futuros. O custo computacional da abordagem pode ser considerado uma das limitações para sua execução em ambientes com baixo recurso computacional, como em sistemas embarcados.

Já as limitações dos métodos de fusão, têm-se que os métodos não consideram a ponderação das listas ranqueadas de acordo com o método de recuperação ou o tipo da consulta (original ou expandida). Tal fato pode limitar a capacidade de explorar a complementaridade entre os *rankings*. Adicionalmente, os métodos não incorporam informações contextuais presentes nos documentos, o que limita as contribuições potenciais para a melhoria da eficácia do ranqueamento.

Como trabalhos futuros, visa-se ampliar a avaliação do método de expansão de consultas proposto (RAMMON), aumentando a variabilidade e a quantidade dos LLMs e modelos de recuperação aplicados. Como possíveis expansões do RAMMON pode-se citar a adição de métodos DSI [101], modelos *cross-encoders*, LLMs baseados em Difusão e LLMs de arquitetura híbrida. Visa-se também avaliar o impacto da variação na quantidade de consultas geradas pelos LLMs. Além disso, pretende-se expandir o nível de complementaridade do RAMMON para considerar a complementaridade entre diferentes métodos de expansão de consultas. Têm-se também como intuito investigar a integração dos métodos de fusão baseada em ranqueamento com informações contextuais dos documentos, propondo métodos de fusão que se alinhem a esse contexto. Por fim, visa-se analisar os impactos da ponderação dos *rankings* de acordo com o tipo de busca e a consulta, com o objetivo de melhorar o re-ranqueamento das listas ranqueadas.

REFERÊNCIAS

1 WITTGENSTEIN, L. **Tractatus Logico-Philosophicus**. London: Routledge & Kegan Paul, 1922.

2 LIN, J.; NOGUEIRA, R.; YATES, A. **Pretrained Transformers for Text Ranking: BERT and Beyond**. Cham: Springer Cham, 2022. (Synthesis Lectures on Human Language Technologies). ISBN 978-3-031-01053-8.

3 ZHAO, W. X. et al. Dense text retrieval based on pretrained language models: A survey. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 42, n. 4, fev. 2024. ISSN 1046-8188. Disponível em: https://doi.org/10.1145/3637870.

4 BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca**. 2. ed. [S.l.]: Bookman, 2013. ISBN 978-8582600481.

5 BERGER, A. et al. Bridging the lexical chasm: statistical approaches to answer-finding. In: **Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2000. (SIGIR '00), p. 192–199. ISBN 1581132263. Disponível em: https://doi.org/10.1145/345508.345576.

6 FURNAS, G. W. et al. The vocabulary problem in human-system communication. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 30, n. 11, p. 964–971, nov 1987. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/32206.32212.

7 THAKUR, N. et al. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)**. [s.n.], 2021. Disponível em: https://openreview.net/forum?id=wCu6T5xFjeJ.

8 ZHAN, J. et al. **Learning To Retrieve: How to Train a Dense Retrieval Model Effectively and Efficiently**. 2020. Disponível em: https://arxiv.org/abs/2010.10469.

9 VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

10 DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: https://aclanthology.org/N19-1423.

11 SCHNEIDER, E. T. R. et al. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In: **Proceedings of the 3rd Clinical Natural Language Processing Workshop**. Online: Association for Computational Linguistics, 2020. p. 65–72. Disponível em: https://www.aclweb.org/anthology/2020.clinicalnlp-1.7.

12 SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: ____. [S.l.: s.n.], 2020. p. 403–417. ISBN 978-3-030-61376-1.

13 DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. **ICLR**, 2021.

14 LI, J. et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: **Proceedings of the 40th International Conference on Machine Learning**. [S.l.]: JMLR.org, 2023. (ICML'23).

15 OQUAB, M. et al. Dinov2: Learning robust visual features without supervision. 04 2023.

16 CHEN, X. et al. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In: **ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2021. p. 5904–5908.

17 GULATI, A. et al. Conformer: Convolution-augmented transformer for speech recognition. In: . [S.l.: s.n.], 2020. p. 5036–5040.

18 GAO, Y. et al. **Retrieval-Augmented Generation for Large Language Models: A Survey**. 2024.

19 GAO, L. et al. Precise zero-shot dense retrieval without relevance labels. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 1762–1777. Disponível em: https://aclanthology.org/2023.acl-long.99.

20 WANG, L.; YANG, N.; WEI, F. Query2doc: Query expansion with large language models. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 9414–9423. Disponível em: https://aclanthology.org/2023.emnlp-main.585.

21 MA, X. et al. Query rewriting in retrieval-augmented large language models. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 5303–5315. Disponível em: https://aclanthology.org/2023.emnlp-main.322.

22 DONOSER, M.; BISCHOF, H. Diffusion processes for retrieval revisited. In: **2013 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2013. p. 1320–1327.

23 QIN, D. et al. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: **CVPR 2011**. [S.l.: s.n.], 2011. p. 777–784.

24 ZHAO, Y. et al. Modelling diffusion process by deep neural networks for image retrieval. In: **British Machine Vision Conference 2018, BMVC 2018**. Newcastle, UK: BMVA, 2019. Disponível em: https://ro.uow.edu.au/eispapers1/3498/.

25 PEDRONETTE, D. C. G.; VALEM, L. P.; LATECKI, L. J. Efficient rank-based diffusion process with assured convergence. **Journal of Imaging**, v. 7, n. 3, 2021. ISSN 2313-433X. Disponível em: https://www.mdpi.com/2313-433X/7/3/49.

26 PEDRONETTE, D. C. G. et al. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. **IEEE Transactions on Image Processing**, v. 28, n. 12, p. 5824–5838, 2019.

27 VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. In: **2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2016. p. 249–256.

28 ARSLAN, M. et al. A survey on rag with llms. **Procedia Computer Science**, v. 246, p. 3781–3790, 2024. ISSN 1877-0509. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024). Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050924021860.

29 ZHAO, W. X. et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023. Disponível em: http://arxiv.org/abs/2303.18223.

30 DEEPSEEK-AI. **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**. 2025. Disponível em: https://arxiv.org/abs/2501.12948.

31 OPENAI. **GPT-4 Technical Report**. 2024. Disponível em: https://arxiv.org/abs/2303.08774.

32 GRATTAFIORI, A. et al. **The Llama 3 Herd of Models**. 2024. Disponível em: https://arxiv.org/abs/2407.21783.

33 JIANG, A. Q. et al. **Mistral 7B**. 2023. Disponível em: https://arxiv.org/abs/2310.06825.

34 ABDIN, M. et al. **Phi-4 Technical Report**. 2024. Disponível em: https://arxiv.org/abs/2412.08905.

35 MINAEE, S. et al. **Large Language Models: A Survey**. 2025. Disponível em: https://arxiv.org/abs/2402.06196.

36 ZHU, W. et al. Multilingual machine translation with large language models: Empirical results and analysis. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Findings of the Association for Computational Linguistics: NAACL 2024**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 2765–2781. Disponível em: https://aclanthology.org/2024.findings-naacl.176/.

37 KOSTINA, A. et al. **Large Language Models For Text Classification: Case Study And Comprehensive Review**. 2025. Disponível em: https://arxiv.org/abs/2501.08457.

38 SCHIMANSKI, T. et al. Towards faithful and robust LLM specialists for evidence-based question-answering. In: KU, L.-W.; MARTINS, A.; SRIKUMAR, V. (Ed.). **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 1913–1931. Disponível em: https://aclanthology.org/2024.acl-long.105/.

39 RACKAUCKAS, Z. Rag-fusion: A new take on retrieval augmented generation. **International Journal on Natural Language Computing**, Academy and Industry Research Collaboration Center (AIRCC), v. 13, n. 1, p. 37–47, fev. 2024. ISSN 2319-4111. Disponível em: http://dx.doi.org/10.5121/ijnlc.2024.13103.

40 JIA, P. et al. MILL: Mutual verification with large language models for zero-shot query expansion. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 2498–2518. Disponível em: https://aclanthology.org/2024.naacl-long.138/.

41 KUO, Y.-C. et al. MMLF: Multi-query multi-passage late fusion retrieval. In: CHIRUZZO, L.; RITTER, A.; WANG, L. (Ed.). **Findings of the Association for Computational Linguistics: NAACL 2025**. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. p. 6602–6613. ISBN 979-8-89176-195-7. Disponível em: https://aclanthology.org/2025.findings-naacl.367/.

42 LI, Z. et al. **DMQR-RAG: Diverse Multi-Query Rewriting for RAG**. 2024. Disponível em: https://arxiv.org/abs/2411.13154.

43 YANG, J. et al. Optimization of rag multi query rewrite generation strategy based on markov decision process. In: **Proceedings of the 2025 3rd International Conference on Communication Networks and Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2025. (CNML '25), p. 142–148. ISBN 9798400713231. Disponível em: https://doi.org/10.1145/3728199.3728221.

44 LEE, D. et al. On complementarity objectives for hybrid retrieval. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 13357–13368. Disponível em: https://aclanthology.org/2023.acl-long.746/.

45 CHEN, Z. et al. **Harnessing Multiple Large Language Models: A Survey on LLM Ensemble**. 2025. Disponível em: https://arxiv.org/abs/2502.18036.

46 NOGUEIRA, R. et al. Multi-stage document ranking with bert. **arXiv preprint arXiv:1910.14424**, 2019.

47 REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: INUI, K. et al. (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. 3982–3992. Disponível em: https://aclanthology.org/D19-1410/.

48 ZHU, Y. et al. Large language models for information retrieval: A survey. **CoRR**, abs/2308.07107, 2023. Disponível em: https://arxiv.org/abs/2308.07107.

49 SHANAHAN, M. Talking about large language models. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 67, n. 2, p. 68–79, jan 2024. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/3624724.

50 BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? 🦜 In: **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2021. (FAccT '21), p. 610–623. ISBN 9781450383097. Disponível em: https://doi.org/10.1145/3442188.3445922.

51 XU, Z.; JAIN, S.; KANKANHALLI, M. S. Hallucination is inevitable: An innate limitation of large language models. **ArXiv**, abs/2401.11817, 2024. Disponível em: https://api.semanticscholar.org/CorpusID:267069207.

52 BROWN, T. et al. Language models are few-shot learners. In: LAROCHELLE, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

53 WEI, J. et al. Finetuned language models are zero-shot learners. **ArXiv**, abs/2109.01652, 2021. Disponível em: https://api.semanticscholar.org/CorpusID:237416585.

54 WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In: **Proceedings of the 36th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2024. (NIPS '22). ISBN 9781713871088.

55 ZHANG, Y. et al. Siren's song in the ai ocean: A survey on hallucination in large language models. **ArXiv**, abs/2309.01219, 2023. Disponível em: https://arxiv.org/abs/2309.01219.

56 MIENYE, I. D.; SWART, T. G. Ensemble large language models: A survey. **Information**, v. 16, n. 8, 2025. ISSN 2078-2489. Disponível em: https://www.mdpi.com/2078-2489/16/8/688.

57 WANG, S. et al. A survey on rank aggregation. In: LARSON, K. (Ed.). **Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24**. International Joint Conferences on Artificial Intelligence Organization, 2024. p. 8281–8289. Survey Track. Disponível em: https://doi.org/10.24963/ijcai.2024/915.

58 FOX, E. A.; SHAW, J. A. Combination of multiple searches. **NIST special publication SP**, NATIONAL INSTIUTE OF STANDARDS & TECHNOLOGY, v. 243, 1994.

59 LEE, J. H. Analyses of multiple evidence combination. In: **Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 1997. (SIGIR '97), p. 267–276. ISBN 0897918363. Disponível em: https://doi.org/10.1145/258525.258587.

60 BORDA, J.-C. de. Mémoire sur les élections au scrutin. **Histoire de l'Académie Royale des Sciences**, Paris, France, v. 12, 1781.

61 CORMACK, G. V.; CLARKE, C. L. A.; BUETTCHER, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: **Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2009. (SIGIR '09), p. 758–759. ISBN 9781605584836. Disponível em: https://doi.org/10.1145/1571941.1572114.

62 LI, C. et al. Parade: Passage representation aggregation for document reranking. **arXiv preprint arXiv:2008.09093**, 2020.

63 SUN, W. et al. Is ChatGPT good at search? investigating large language models as re-ranking agents. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 14918–14937. Disponível em: https://aclanthology.org/2023.emnlp-main.923.

64 QIN, Z. et al. Large language models are effective text rankers with pairwise ranking prompting. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Findings of the Association for Computational Linguistics: NAACL 2024**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 1504–1518. Disponível em: https://aclanthology.org/2024.findings-naacl.97.

65 XIAO, S. et al. C-pack: Packed resources for general chinese embeddings. In: **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2024. (SIGIR '24), p. 641–649. ISBN 9798400704314. Disponível em: https://doi.org/10.1145/3626772.3657878.

66 TEAM, A. G. Q. **Qwen2 Technical Report**. 2024. Disponível em: https://arxiv.org/abs/2407.10671.

67 MIN, S. et al. AmbigQA: Answering ambiguous open-domain questions. In: WEBBER, B. et al. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 5783–5797. Disponível em: https://aclanthology.org/2020.emnlp-main.466/.

68 YANG, Z. et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: RILOFF, E. et al. (Ed.). **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 2369–2380. Disponível em: https://aclanthology.org/D18-1259.

69 VU, T. et al. FreshLLMs: Refreshing large language models with search engine augmentation. In: KU, L.-W.; MARTINS, A.; SRIKUMAR, V. (Ed.). **Findings of the Association for Computational Linguistics: ACL 2024**. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 13697–13720. Disponível em: https://aclanthology.org/2024.findings-acl.813/.

70 RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, JMLR.org, v. 21, n. 1, jan. 2020. ISSN 1532-4435.

71 RACKAUCKAS, Z.; CÂMARA, A.; ZAVREL, J. Evaluating rag-fusion with ragelo: an automated elo-based framework. In: **Proceedings of the First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024)**. Washington D.C., USA: CEUR-WS.org, 2024. v. 3752, p. 92–112. Workshop co-located with SIGIR 2024. Disponível em: https://ceur-ws.org/Vol-3752/.

72 WANG, L. et al. **Multilingual E5 Text Embeddings: A Technical Report**. 2024. Disponível em: https://arxiv.org/abs/2402.05672.

73 CRASWELL, N. et al. Overview of the trec 2019 deep learning track. **ArXiv**, abs/2003.07820, 2020.

74 CRASWELL, N. et al. Overview of the trec 2020 deep learning track. **ArXiv**, abs/2102.07662, 2020.

75 KAMALLOO, E. et al. **Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard**. 2023.

76 WANG, L. et al. **Text Embeddings by Weakly-Supervised Contrastive Pre-training**. 2024. Disponível em: https://arxiv.org/abs/2212.03533.

77 IZACARD, G. et al. **Unsupervised Dense Information Retrieval with Contrastive Learning**. 2022. Disponível em: https://arxiv.org/abs/2112.09118.

78 HE, P. et al. Deberta: Decoding-enhanced bert with disentangled attention. In: **International Conference on Learning Representations**. [s.n.], 2021. Disponível em: https://openreview.net/forum?id=XPZIaotutsD.

79 CHUNG, H. W. et al. Scaling instruction-finetuned language models. **Journal of Machine Learning Research**, v. 25, n. 70, p. 1–53, 2024. Disponível em: http://jmlr.org/papers/v25/23-0870.html.

80 TOUVRON, H. et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

81 TAY, Y. et al. Ul2: Unifying language learning paradigms. In: **International Conference on Learning Representations**. [s.n.], 2022. Disponível em: https://api.semanticscholar.org/CorpusID:252780443.

82 WACHSMUTH, H.; SYED, S.; STEIN, B. Retrieval of the best counterargument without prior topic knowledge. In: GUREVYCH, I.; MIYAO, Y. (Ed.). **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 241–251. Disponível em: https://aclanthology.org/P18-1023/.

83 DIGGELMANN, T. et al. Climate-fever: A dataset for verification of real-world climate claims. In: **Tackling Climate Change with Machine Learning Workshop at NeurIPS 2020**. Online: [s.n.], 2020.

84 HASIBI, F. et al. Dbpedia-entity v2: A test collection for entity search. In: **Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2017. (SIGIR '17), p. 1265–1268. ISBN 9781450350228. Disponível em: https://doi.org/10.1145/3077136.3080751.

85 THORNE, J. et al. FEVER: a large-scale dataset for fact extraction and VERification. In: WALKER, M.; JI, H.; STENT, A. (Ed.). **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 809–819. Disponível em: https://aclanthology.org/N18-1074/.

86 MAIA, M. et al. Www'18 open challenge: Financial opinion mining and question answering. In: **Companion Proceedings of the The Web Conference 2018**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 1941–1942. ISBN 9781450356404. Disponível em: https://doi.org/10.1145/3184558.3192301.

87 ZHANG, X. et al. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In: ATAMAN, D. et al. (Ed.). **Proceedings of the 1st Workshop on Multilingual Representation Learning**.

Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 127–137. Disponível em: https://aclanthology.org/2021.mrl-1.12.

88 NGUYEN, T. et al. Ms marco: A human generated machine reading comprehension dataset. November 2016. Disponível em: https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/.

89 KWIATKOWSKI, T. et al. Natural questions: A benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 7, p. 452–466, 2019.

90 BOTEVA, V. et al. A full-text learning to rank dataset for medical information retrieval. In: FERRO, N. et al. (Ed.). **Advances in Information Retrieval**. Cham: Springer International Publishing, 2016. p. 716–722. ISBN 978-3-319-30671-1.

91 VOORHEES, E. Overview of the trec 2004 robust retrieval track. In: **Text REtrieval Conference (TREC)**. [S.l.: s.n.], 2005. p. 4–19.

92 WADDEN, D. et al. Fact or fiction: Verifying scientific claims. In: WEBBER, B. et al. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 7534–7550. Disponível em: https://aclanthology.org/2020.emnlp-main.609/.

93 COHAN, A. et al. SPECTER: Document-level representation learning using citation-informed transformers. In: JURAFSKY, D. et al. (Ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 2270–2282. Disponível em: https://aclanthology.org/2020.acl-main.207/.

94 SUAREZ, A. et al. A data collection for evaluating the retrieval of related tweets to news articles. In: ____. [S.l.: s.n.], 2018. p. 780–786. ISBN 978-3-319-76940-0.

95 BONDARENKO, A. et al. Overview of touché 2020: Argument retrieval. In: ARAMPATZIS, A. et al. (Ed.). **Experimental IR Meets Multilinguality, Multimodality, and Interaction**. Cham: Springer International Publishing, 2020. p. 384–395. ISBN 978-3-030-58219-7.

96 DIETZ, L. et al. Trec complex answer retrieval overview. In: **Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017)**. [S.l.: s.n.], 2017.

97 VOORHEES, E. et al. Trec-covid: constructing a pandemic information retrieval test collection. **SIGIR Forum**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 1, fev. 2021. ISSN 0163-5840. Disponível em: https://doi.org/10.1145/3451964.3451965.

98 SOBOROFF, I.; HUANG, S.; HARMAN, D. Trec 2019 news track overview. In: NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). **Text REtrieval Conference (TREC)**. Gaithersburg, Maryland, USA, 2019. p. 4–18.

99 VALEM, L. P.; PEDRONETTE, D. C. G.; LATECKI, L. J. Rank flow embedding for unsupervised and semi-supervised manifold learning. **IEEE Transactions on Image Processing**, v. 32, p. 2811–2826, 2023.

100 SHAO, R. et al. **ReasonIR: Training Retrievers for Reasoning Tasks**. 2025. Disponível em: https://arxiv.org/abs/2504.20595.

101 TAY, Y. et al. Transformer memory as a differentiable search index. In: **Proceedings of the 36th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2022. (NIPS '22). ISBN 9781713871088.

102 ROBERTSON, S. et al. Okapi at trec-3. In: **Proceedings of the 3rd Text REtrieval Conference (TREC-3)**. Gaithersburg, Maryland: [s.n.], 1994. p. 109–126.

103 ROBERTSON, S.; ZARAGOZA, H. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Now Publishers Inc., Hanover, MA, USA, v. 3, n. 4, p. 333–389, abr. 2009. ISSN 1554-0669. Disponível em: https://doi.org/10.1561/1500000019.

104 NI, J. et al. Large dual encoders are generalizable retrievers. In: GOLDBERG, Y.; KOZAREVA, Z.; ZHANG, Y. (Ed.). **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 9844–9855. Disponível em: https://aclanthology.org/2022.emnlp-main.669/.

105 WANG, L. L. et al. CORD-19: The COVID-19 open research dataset. In: VERSPOOR, K. et al. (Ed.). **Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020**. Online: Association for Computational Linguistics, 2020. Disponível em: https://aclanthology.org/2020.nlpcovid19-acl.1/.

106 AJJOUR, Y. et al. Data acquisition for argument search: The args.me corpus. In: BENZMÜLLER, C.; STUCKENSCHMIDT, H. (Ed.). **KI 2019: Advances in Artificial Intelligence**. Cham: Springer International Publishing, 2019. p. 48–59. ISBN 978-3-030-30179-8.

107 BELL, M. M. et al. Multi-llm query augmentation for rag-fusion. **Information Sciences**. In Revision.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Murilo Missano Bell 04/07/2001
Nacionalidade	Brasileiro
Nome em citações bibliográficas	BELL, Murilo Missano BELL, M. M.
Currículo Lattes	http://lattes.cnpq.br/1576097622593183
ORCID	https://orcid.org/0009-0006-6212-8984
FORMAÇÃO ACADÊMICA	
2020/2024	Bacharelado em Ciência da Computação Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) – Instituto de Geociências e Ciências Exatas (IGCE) – Câmpus de Rio Claro
2017/2019	Técnico em Informática Escola Técnica Estadual (ETEC) "Professor Armando Bayeux da Silva"
PRODUÇÃO BIBLIOGRÁFICA	
CORREIA, J. V. M.; BELL, M. M.; AMORIM, J. V. R.; QUEIROZ, J.; PEDRONETTE, D.; GUILHERME, I. R.; OLIVEIRA, F. L. Analysis of automated document relevance annotation for information retrieval in oil and gas industry. *In:* Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Suzhou (China): Association for Computational Linguistics, 2025. p. 1878–1889. ISBN 979-8-89176-333-3.	
BELL, M. M.; AMORIM, J. V. R.; GUILHERME, I. R. Uma análise de métodos para busca semântica na área de processamento de linguagem natural. *In:* Anais do XXXV Congresso de Iniciação Científica da UNESP. Atibaia (Brasil), 2023.	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	
Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025, Suzhou, China. Analysis of automated document relevance annotation for information retrieval in oil and gas industry. 2025. (Apresentação de Trabalho). XV Escola Regional de Alto Desempenho de São Paulo (ERAD-SP), 2024, Rio Claro (SP), Brasil. (Ouvinte). XXXV Congresso de Iniciação Científica da Unesp (CIC UNESP), 2023, Atibaia (SP), Brasil. Uma análise de métodos para busca semântica na área de processamento de linguagem natural. 2023. (Apresentação de Trabalho).	