



UNIVERSIDADE ESTADUAL PAULISTA

"JÚLIO DE MESQUITA FILHO"

Câmpus São José do Rio Preto

João Vitor Mariano Correia

Knowledge Enhanced Dense Retrieval in Oil and Gas Domain

Bauru
2026

João Vitor Mariano Correia

Knowledge Enhanced Dense Retrieval in Oil and Gas Domain

Dissertation submitted to the to the Postgraduate Program in Computer Science of the Institute of Biosciences, Letters, and Exact Sciences of the São Paulo State University "Júlio de Mesquita Filho", Bauru Campus, as a requirement for obtaining the title of Master in Computer Science

Advisor: Prof. Dr. João Paulo Papa

Bauru
2026

C824k

Correia, João Vitor Mariano

Knowledge Enhanced Dense Retrieval in Oil and Gas Domain /
João Vitor Mariano Correia. -- Bauru, 2026

114 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP),
Faculdade de Ciências, Bauru

Orientador: João Paulo Papa

1. Recuperação da informação. 2. Processamento de linguagem
natural. 3. Grafos de conhecimento. 4. Aprendizado profundo. I.
Título.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE JOÃO VITOR MARIANO CORREIA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 25 de fevereiro de 2026, às 14h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de JOÃO VITOR MARIANO CORREIA, intitulada "Knowledge Enhanced Dense Retrieval in Oil and Gas Domain". A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. JOÃO PAULO PAPA (Orientador(a) - Participação Virtual) do Departamento de Computação/UNESP/ Câmpus de Bauru - FC, Prof. Dr. DANIEL CARLOS GUIMARÃES PEDRONETTE (Participação Virtual) do Departamento de Estatística, Matemática Aplicada e Computação/UNESP/Câmpus de Rio Claro - IGCE, Prof. Dr. JOÃO BAPTISTA CARDIA NETO (Participação Virtual) do Faculdade de Tecnologia de Catanduva. / Centro Estadual de Educação Tecnológica Paula Souza. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Documento assinado digitalmente
 JOAO PAULO PAPA
Data: 26/02/2026 06:29:40-0300
Verifique em <https://validar.itl.gov.br>

Prof. Dr. JOÃO PAULO PAPA

Acknowledgements

Agradeço primeiramente a Deus, por Sua infinita bondade e por todas as bênçãos a mim concedidas.

Em segundo lugar, à minha mãe, por ser meu maior exemplo e por me guiar e orientar em todos os caminhos e adversidades, estando sempre ao meu lado. Um exemplo de mulher e de ser humano; sonho um dia poder ser 1% do que ela é.

Agradeço também à minha família: meu irmão Gabriel, meu padrasto Gilmar, meus avós, tios e primos. À minha namorada, Eliza, meus amigos, e meus filhos de quatro patas (Ótis e Kiki): obrigado por tornarem os dias mais leves com sua companhia.

Ao meu orientador e amigo, Professor João Paulo Papa, que me orienta, auxilia e acolhe desde a graduação. Sou grato por todas as oportunidades e portas que foram abertas durante esse tempo.

A todos os professores com quem tive contato e que foram essenciais para minha formação. À Unesp, por fornecer um ensino gratuito e de qualidade e por fazer parte de toda a minha trajetória acadêmica, desde o ensino médio e técnico no Colégio Técnico Industrial, passando pela graduação e agora no Mestrado.

Agradeço também aos colegas do laboratório Recogna por todos os trabalhos desenvolvidos, pelas discussões e pela abertura para o desenvolvimento de um trabalho científico de qualidade. Por fim, agradeço aos colegas do Projeto Radiar e aos membros da Petrobras, por compartilharem esta jornada de pesquisa e desenvolvimento.

Nothing is hard in this world if you dare to scale the heights
- Mao Zedong

Resumo

A recuperação de informação em domínios técnicos especializados enfrenta desafios significativos devido à escassez de dados rotulados, à presença de vocabulário altamente específico e à limitação de modelos semânticos generalistas em capturar relações operacionais complexas. No contexto da indústria de Óleo e Gás, relatórios técnicos como os Boletins Diários de Perfuração apresentam estrutura implícita, terminologia abreviada e forte dependência de conhecimento de domínio, tornando abordagens tradicionais de busca lexical ou modelos densos pré-treinados insuficientes em cenários de baixo supervisionamento. Este trabalho investiga se a incorporação explícita de conhecimento estrutural, derivado de uma ontologia de domínio, pode melhorar o desempenho de sistemas de recuperação de informação neural em ambientes industriais. Para isso, propõe-se um método de adaptação de modelos bi-encoder baseado em *Structure-Aware Negative Sampling*, no qual um Grafo de Informação Heterogêneo é construído a partir de uma ontologia de domínio e utilizado para minerar exemplos negativos semanticamente próximos, porém estruturalmente distintos. Além disso, são exploradas estratégias de geração sintética de consultas para mitigar a escassez de dados rotulados. A abordagem proposta é avaliada em um conjunto de dados reais de perfuração *offshore*, utilizando tanto um padrão-ouro com anotações manuais quanto um padrão-prata em larga escala, gerado por classificação automática. Os resultados experimentais demonstram que o modelo adaptado supera baselines lexicais fortes, como BM25, e modelos densos generalistas, alcançando ganhos consistentes em métricas de ranking, especialmente em consultas difíceis caracterizadas por alto descompasso vocabular. Análises geométricas do espaço latente indicam que a supervisão estrutural promove um equilíbrio mais adequado entre alinhamento e uniformidade das representações. Por fim, o modelo apresenta elevada eficiência computacional, sendo significativamente mais rápido e energeticamente eficiente do que alternativas baseadas em modelos de larga escala. Os resultados indicam que a integração de conhecimento estrutural constitui um sinal de supervisão eficaz para recuperação de informação em domínios técnicos, oferecendo uma alternativa escalável e sustentável à simples ampliação do tamanho dos modelos.

Palavras-chave: Recuperação da Informação. Processamento de Linguagem Natural. Grafos de Conhecimento. Aprendizado Profundo.

Abstract

Information retrieval in specialized technical domains poses significant challenges due to limited labeled data, highly domain-specific vocabulary, and the inability of general-purpose semantic models to capture complex operational relationships. In the Oil and Gas industry, technical documents such as Daily Drilling Reports exhibit implicit structure, abbreviated terminology, and strong dependence on domain knowledge, making traditional lexical retrieval methods and off-the-shelf dense models insufficient in low-supervision settings. This dissertation investigates whether the explicit incorporation of structural domain knowledge, derived from a domain ontology, can improve neural information retrieval performance in industrial environments. To this end, we propose a bi-encoder adaptation strategy based on *Structure-Aware Negative Sampling*, in which a Heterogeneous Information Network constructed from a domain ontology is used to mine semantically similar yet structurally distinct hard negatives. In addition, synthetic query generation strategies are employed to mitigate the scarcity of labeled training data. The proposed approach is evaluated on a real-world offshore drilling dataset using both a manually annotated Gold Standard and a large-scale Silver Standard generated via automated classification. Experimental results show that the adapted model consistently outperforms strong lexical baselines such as BM25 and general-purpose dense retrievers, achieving notable gains in ranking metrics, particularly for hard queries characterized by severe vocabulary mismatch. Latent space analysis further reveals that structural supervision promotes a more effective balance between alignment and uniformity in the learned representations. Moreover, the proposed model demonstrates high computational efficiency, offering substantial reductions in indexing time, inference latency, and energy consumption compared to large-scale language models. These findings suggest that integrating structural domain knowledge provides an effective and scalable supervision signal for information retrieval in technical domains, offering a practical alternative to increasing model size in resource-constrained industrial settings.

Keywords: Information Retrieval. Natural Language Processing. Knowledge Graphs. Deep Learning.

List of Figures

Figure 1 – Basic Information Retrieval Pipeline: The process begins with an initial query that is submitted to the retriever module. The retriever accesses the corpus and retrieves a set of relevant documents. These documents are then passed to the re-ranker module, which reorders them based on relevance scores to produce a refined, final ranked response.	23
Figure 2 – Cross-Encoder Architecture for Dense Retrieval: In this architecture, the query and each candidate document are jointly encoded by concatenating them and passing them through a transformer-based model, which outputs a relevance score. Unlike the bi-encoder, the cross-encoder processes the query-document pairs simultaneously, allowing for deeper interaction between the two. This approach, while computationally more intensive, often achieves higher accuracy by capturing nuanced relationships between the query and documents.	32
Figure 3 – Bi-Encoder Architecture for Dense Retrieval: In this architecture, the query and documents are independently encoded into dense vector representations using a shared or separate encoder model. These vectors are then compared, typically using a similarity measure to identify the most relevant documents. The bi-encoder approach enables efficient retrieval by allowing pre-computation and storage of document embeddings.	33
Figure 4 – The test dataset creation pipeline, moving from raw data to both a human-adjudicated Golden Standard and a machine-annotated Silver Standard relevance set.	54
Figure 5 – Heatmap of Pairwise Cohen's Kappa (κ) Agreement.	58
Figure 6 – Conceptual visualization of the Knowledge Graph structure. The diagram illustrates how specialized ontology classes are instantiated as nodes and interconnected through semantic relationships extracted from DDR narratives.	62
Figure 7 – Architectural overview of the proposed Knowledge-Enhanced Dense Retrieval Framework, highlighting the integration of unsupervised graph representation learning and structure-aware negative sampling.	63
Figure 8 – Comparative retrieval performance across models under gold-standard and silver-standard datasets NDCG@10.	80
Figure 9 – Performance comparison of fine-tuning strategies across NDCG@10, MAP@10, and Recall@10.	82
Figure 10 – NDCG@10 performance stratified by query difficulty.	86
Figure 11 – Pareto Frontier of Retrieval Effectiveness vs. Inference Latency.	90

List of Tables

Table 1	– SOTA Datasets for Information Retrieval with domain and corpus size . . .	52
Table 2	– Retrieval models used for document pooling and candidate generation. . .	55
Table 3	– Performance comparison of top three models for each metric. Model names are abbreviated as [Classifier] + [Training Strategy] + [Embedding] + [Feature Combination]. Values are mean $\pm$ standard deviation when available. . . .	57
Table 4	– Overview of retrieval baselines used for evaluation, spanning probabilistic lexical models, domain-specific static embeddings, sparse neural retrievers, dense bi-encoder architectures, and late-interaction models. The selection reflects both widely adopted industry standards and top-performing open-weight models, enabling a comprehensive comparison across retrieval paradigms.	77
Table 5	– Normalized Discounted Cumulative Gain (nDCG@k) for all retrieval models evaluated on the Gold and Silver Standard datasets. Results are reported at multiple cutoff values to assess ranking quality under varying retrieval depths. The table enables comparison across lexical, dense, and late-interaction paradigms under different annotation granularities.	78
Table 6	– Mean Average Precision (MAP@k) results for the evaluated retrieval models on the Gold and Silver Standard datasets. MAP emphasizes precision across the full ranked list, providing insight into the consistency of relevance ordering across different retrieval approaches and supervision regimes.	78
Table 7	– Recall@k scores for all retrieval models evaluated on the Gold and Silver Standard datasets. Recall at increasing cutoff values reflects the ability of each method to retrieve a broader set of relevant documents, complementing precision-oriented metrics and highlighting trade-offs across retrieval paradigms.	79
Table 8	– Alignment and Uniformity metrics calculated on highly relevant pairs (Score=3) from the Silver Standard. Lower values indicate better performance for both metrics.	84
Table 9	– NDCG@10 performance stratified by lexical query difficulty. The proposed model demonstrates superior performance on "Hard" queries, surpassing even the largest baselines.	85
Table 10	– Qualitative comparison of retrieval outputs for hard queries. The proposed model demonstrates superior handling of implicit causality, synonymy, and domain-specific quantitative associations compared to a strong late-interaction baseline.	87

Table 11 – Computational efficiency benchmark on NVIDIA A40. *Note: VRAM usage for the smaller models reflects the peak system allocation during the batched benchmarking process rather than the minimum memory requirement for deployment, which is strictly correlated with parameter count. 90

List of abbreviations and acronyms

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BIM	Binary Independence Model
CNN	Convolutional Neural Network
CoRT	Complementary Rankings from Transformers
DDR	Daily Drilling Report
DFR	Divergence From Randomness
DSI	Differentiable Search Index
EA	Entity Alignment
FV	Fisher Vector
GAE	Graph Autoencoder
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GPT	Generative Pre-trained Transformer
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HIN	Heterogeneous Information Network
IDF	Inverse Document Frequency
IR	Information Retrieval
KEPLM	Knowledge-Enhanced Pre-trained Language Model
KG	Knowledge Graphs
KGQA	Knowledge Graph Question Answering
LDA	Latent Dirichlet Allocation

LSH	Locality Sensitive Hashing
LSI	Latent Semantic Indexing
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
NER	Named Entity Recognition
NLL	Negative Log-Likelihood
NLP	Natural Language Processing
NN	Neural Networks
NVSM	Neural Vector Space Model
OWL	Web Ontology Language
PLM	Pre-trained Language Model
PRF	Pseudo-Relevance Feedback
PV	Paragraph Vector
QA	Question Answering
QL	Query Likelihood
RDF	Resource Description Framework
RNN	Recurrent Neural Network
TDV	Term Discrimination Values
TF	Term Frequency
UMAP	Uniform Manifold Approximation and Projection
VSM	Vector Space Model

Contents

1	INTRODUCTION	16
1.1	Motivation	16
1.2	Our Proposal	19
1.2.1	Problems	19
1.2.2	Hypothesis	20
1.2.3	Research Objectives	20
1.2.3.1	General Objective	20
1.2.3.2	Specific objectives	20
1.2.4	Dissertation Structure	21
2	LITERATURE REVIEW	22
2.1	Information Retrieval	22
2.1.1	Classical Retrieval	24
2.1.1.1	Set-Theoretic Models	25
2.1.1.2	Algebraic Models	25
2.1.1.3	Probabilistic Models	25
2.1.1.4	Language Models for IR	26
2.1.2	Semantic Retrieval	26
2.1.2.1	Query Augmentation	26
2.1.2.2	Document Augmentation	27
2.1.3	Neural Methods for Retrieval	28
2.1.3.1	Sparse Retrieval	29
2.1.3.2	Dense Retrieval	31
2.2	Knowledge Graphs	38
2.2.1	Knowledge Representations	38
2.2.2	Knowledge Representation Learning	39
2.2.2.1	Translational Distance Models	39
2.2.2.2	Semantic Matching Models	41
2.2.2.3	Graph Neural Networks	41
2.2.2.3.1	Graph Convolutional Networks (GCN)	41
2.2.2.3.2	GraphSAGE	42
2.2.2.3.3	Graph Attention Networks (GAT)	43
2.2.3	Knowledge Enhanced Pre-trained Language Models	44
2.3	Synthesis and Research Gap	49
3	METHODOLOGY	51

3.1	Datasets	51
3.1.1	Test Dataset Creation	53
3.1.2	Training Dataset Creation	59
3.2	Knowledge Graph Construction	59
3.2.1	Ontology Engineering and Design	60
3.2.2	Entity Extraction and Knowledge Injection	61
3.3	Overall Solution	62
3.3.1	Problem Formulation	64
3.3.2	Graph-Based Representation Learning	65
3.3.2.1	Feature Initialization	65
3.3.2.2	Heterogeneous Graph Encoder	66
3.3.3	Unsupervised Optimization Strategy	67
3.3.4	Structure-Aware Negative Sampling	68
3.3.5	Synthetic Query Generation	69
3.3.6	Dense Retriever Training Objective	70
3.3.6.1	Model Architecture	70
3.3.6.2	Contrastive Learning Objective	71
3.3.6.3	Training Configuration	71
3.4	Experimental Design and Baseline Configuration	72
3.4.1	Baseline Methods	72
3.4.2	Synthetic Query Generation Control	73
3.4.3	Architectural and Training Controls	73
3.5	Evaluation Metrics	74
4	RESULTS	76
4.1	Comparative Performance Analysis	76
4.1.1	Analysis of Zero-Shot Performance	76
4.1.1.1	The Dense Retrieval Gap	79
4.1.1.2	Superiority of Late Interaction	80
4.1.1.3	Impact of the Silver Standard: Recall Drop and Precision Gain	80
4.1.2	Analysis of Fine-Tuned Models	81
4.1.2.1	Synthetic Generation vs. Unsupervised Adaptation	81
4.1.2.2	Efficacy of Relevance-Aware and Graph-Enhanced Strategies	82
4.1.2.3	Surpassing the Lexical Baseline	83
4.1.2.4	Generalization in the Silver Standard	83
4.2	In-Depth Analysis	83
4.2.1	Latent Space Geometry: Alignment and Uniformity	83
4.2.2	Query-Level Analysis and Robustness	85
4.2.2.1	Stratified Performance by Difficulty	85
4.2.2.2	Win/Loss Analysis	85

4.2.2.3	Qualitative Inspection	86
4.3	Generalization and Domain Specificity	87
4.3.1	Computational Efficiency and Environmental Impact	89
4.3.1.1	Computational Efficiency Analysis	89
4.3.1.2	Environmental Impact Analysis	91
4.3.2	Summary of Empirical Findings	91
5	DISCUSSION	93
6	CONCLUSION	96
	BIBLIOGRAPHY	99

1 Introduction

The exponential growth of unstructured data in high-stakes industrial environments presents a substantial challenge for contemporary Information Science. In the Oil and Gas (O&G) industry, daily operations generate large collections of technical narratives, ranging from Daily Drilling Reports (DDRs) to shift logs that contain critical knowledge regarding safety events, equipment performance, and operational anomalies. Although this information is abundant, it remains largely inaccessible to automated retrieval systems due to vocabulary mismatch, domain-specific abbreviations, and implicit contextual dependencies that are poorly captured by general-purpose language representations. As a result, operational knowledge is often underutilized, constraining data-driven decision-making and increasing reliance on manual inspection. The primary motivation of this dissertation arises from the need to address these representation limitations and enable systematic access to such data to support decision-making, enhance operational safety, and reduce redundant operational processes.

1.1 Motivation

In recent decades, the increasing volume of available data has made Information Retrieval tasks widely used in everyday applications. These tasks are used in web browsers, personal assistants, streaming apps, image galleries, digital libraries, and other tools to provide users with the information they request (HAMBARDE; PROENCA, 2023). The main goal of Information Retrieval is to locate and retrieve materials, such as unstructured documents such as text, images, and audio, based on a user's query (MANNING; RAGHAVAN; SCHÜTZE, 2008).

The methods for retrieving information can be divided and categorized in various ways, with different types of procedures and models used for the same purpose. In a classical information retrieval approach, the process can be divided into two distinct stages: (1) indexing, which involves creating a representation of the contents of documents and queries to facilitate retrieval, and (2) retrieval, which aims to find a document that matches the query based on the representations created in the indexing stage (TAMINE; GOEURLOT, 2020).

The algorithms used in the indexing process can be categorized into traditional models (boolean or probabilistic) or alternative models (neural networks, language models, learning-to-rank, etc.). Each of these models has its own advantages and disadvantages, suitable for different application scenarios. Traditional models are usually categorized into (HAMBARDE; PROENCA, 2023):

1. Boolean models, widely used in simple information retrieval applications. They represent

queries in an n -dimensional vector space where n corresponds to the vocabulary size of the documents. The presence or absence of a term in the document is represented in this document vector, and similarity measures are calculated based on these.

2. Probabilistic models use probability concepts to model the relevance of documents to each query, and are widely explored for result ranking.

While traditional models have been effective for many uses, they exhibit limitations that restrict their applicability for certain scenarios. One major limitation is their inability to understand the meaning of a word in a specific context. This means that when a word has multiple meanings, the model will treat it as the same word, disregarding the different possible meanings and affecting the accuracy of the result. Additionally, these models do not account for semantic relationships between words, such as synonyms (words with similar meanings), antonyms (words with opposite meanings), hypernyms (words with broader meanings), and hyponyms (words with narrower meanings), among others (SANTOS, 2018).

An alternative to traditional models are the alternative information retrieval models, which encompass a broad set of techniques that can be employed to build richer representations of the search space. These techniques include Artificial Intelligence (AI), Neural Networks (NN), language models, and learning-to-rank, among others. When applied to information retrieval tasks, these techniques can capture semantic representations and non-linear relationships between the terms present in the documents. The text representations learned by neural networks from training on large labeled datasets are called "embeddings." These embeddings are dense vector representations that enable more expressive representations of natural language by the machine (ZHAO et al., 2022).

Despite the strong performance of artificial intelligence techniques, particularly deep learning models, their widespread adoption was initially hindered by the large amount of data and computational cost required to achieve such performance. To address this issue, a new paradigm emerged, and the use of pre-trained language models became widely adopted as a viable option. Models like BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2018) and GPT-2 (Generative Pre-trained Transformer 2) (RADFORD et al., 2019) are among the main examples. These models follow the pretrain-then-finetune paradigm, which involves pre-training the models on a large, general-purpose dataset and then fine-tuning them to adapt the model to specific domains. Recent research in the field of information retrieval also aims to incorporate external knowledge to enhance the results obtained in the retrieval task (ZHAO et al., 2022).

However, a persistent limitation remains: general-purpose Pre-trained Language Models (PLMs) often underperform in specialized domains. Models trained on web-scale text (e.g., Wikipedia, Common Crawl) struggle to generalize to high-stakes industries like Oil & Gas. In these environments, the distribution of vocabulary differs significantly from the general domain;

technical jargon, domain-specific abbreviations (e.g., “POOH” for “Pull Out Of Hole”), and implicit causal relationships are prevalent. Standard bi-encoder architectures often fail to model these domain-specific distinctions in zero-shot settings, leading to situations where neural dense retrievers perform worse than strong lexical baselines such as BM25 when applied to unseen domains without task-specific training. This phenomenon is often attributed to domain shift, where the distribution of vocabulary and semantic relationships in the target corpus differs significantly from the data used during pre-training. Under such conditions, lexical methods remain competitive because they rely directly on term statistics observed within the target collection. (THAKUR et al., 2021; IZACARD et al., 2022; OLSEN et al., 2025).

Recent research aims to bridge this gap by incorporating external knowledge to enhance retrieval tasks (ZHAO et al., 2022). External knowledge can be integrated from various sources: (1) *linguistic knowledge* (lexical relations, syntactic trees); (2) *encyclopedic knowledge* (structured triples of entities and relationships); (3) *commonsense knowledge* (graphs like ConceptNet (SPEER; CHIN; HAVASI, 2016) and ATOMIC (SAP et al., 2018)); and (4) *domain-specific knowledge*, such as the Unified Medical Language System (UMLS) (BODENREIDER, 2004) for medicine or specific ontologies for e-commerce (ZHANG et al., 2020a). This knowledge can be categorized by granularity, ranging from text chunks and entities to complex subgraphs (WEI et al., 2021), offering a pathway to inject structured expertise into neural retrieval models.

While incorporating structured knowledge has shown promise for improving semantic representations, an important methodological question remains: how can the relational structure encoded in knowledge graphs be effectively leveraged during the training of retrieval models? Simply augmenting textual representations with entity information is often insufficient to capture the complex dependencies between domain concepts (ZHAO et al., 2022). In specialized environments such as Oil and Gas operations, the relationships between entities form structured patterns that are not explicitly represented in the textual surface form of documents.

Recent advances in graph representation learning provide mechanisms to model such relational dependencies. Graph Neural Networks (GNNs), for example, enable the propagation of information across nodes and edges in a knowledge graph, allowing the learned representations to capture higher-order structural relationships between entities (YE et al., 2022). When integrated into retrieval pipelines, these graph-based representations can serve as an inductive bias that aligns language model embeddings with domain semantics (YANG et al., 2025; ZHUKOVA et al., 2025).

Beyond representation learning, knowledge graphs can also inform the training dynamics of retrieval models. Dense retrieval systems are typically trained using contrastive learning, where the model learns to distinguish relevant documents from non-relevant ones. However, randomly sampled negatives are often too easy, providing limited learning signal. Hard negative samples (documents that are semantically similar yet irrelevant) are known to significantly improve retrieval performance (HAMBARDE; PROENCA, 2023). In domain-specific corpora,

the structural information encoded in knowledge graphs can help identify such challenging negatives by revealing documents that share related entities or operational contexts but do not directly answer the query.

Motivated by these observations, this dissertation investigates how structured domain knowledge can be used not only to enrich semantic representations but also to guide the learning process of dense retrieval models. In particular, we explore a framework that integrates knowledge graphs with neural retrieval architectures to improve representation quality and training effectiveness in specialized technical domains.

1.2 Our Proposal

1.2.1 Problems

In the field of information retrieval using semantic techniques, there are several areas that need improved approaches to provide more relevant results to users. Although semantic models have shown efficiency in many applications, they still encounter significant challenges in achieving broader adoption in operational settings. Many of these challenges arise from the complexity of analyzing large volumes of text. Notable issues include:

1. Sometimes, when formulating a query, the words used may not exactly match the words found in the documents. This is known as the “lexical gap”, where two words that have similar meanings are different. The reasons for this error can vary, from the user’s lack of experience or expertise in the document domain to suboptimal indexing processes.
2. Beyond the lexical issue, the semantic representations generated for a text may fail to capture the high-level concepts that convey relevant information within the context, focusing instead on lower-level textual components. Addressing semantic ambiguities requires developing models with a more profound understanding of both the language and the specific application domain.
3. When assessed in specific knowledge domains, correctly interpreting context can be a challenge even for advanced language models. Technical terms and acronyms are essential for grasping the context but are frequently misunderstood or used incorrectly by these models. In more specialized domains, it becomes increasingly challenging to obtain labeled data and relevant information.

Ontologies, knowledge graphs, and other forms of knowledge representation are used to capture relationships and meanings between entities. This enables a more structured representation of semantic information, making it easier to interpret and process complex queries. To tackle the challenges faced by semantic models, it is essential to combine different

techniques. Hybrid approaches that integrate statistical, symbolic, and knowledge-based methods have been explored to achieve more effective results in semantic representation and information retrieval.

1.2.2 Hypothesis

This research is guided by the following hypothesis: *Integrating structured domain knowledge, represented as a knowledge graph, into the training process of pre-trained language models improves retrieval effectiveness in specialized technical domains when compared to general-purpose dense retrieval models trained without domain-specific structural information.*

Specifically, this work hypothesizes that representation learning guided by graph-derived relational structure introduces an inductive bias that aligns embedding spaces with domain semantics, enabling models to encode technical terminology, contextual dependencies, and entity relationships more effectively. By incorporating structural signals during training, the learned representations are expected to exhibit improved semantic coherence under conditions of vocabulary mismatch and limited labeled data, which are characteristic of industrial domains.

1.2.3 Research Objectives

1.2.3.1 General Objective

To develop and evaluate a Knowledge-Enhanced Dense Retrieval Framework, that leverages external knowledge sources, such as ontologies and knowledge graphs, with pre-trained language models to improve semantic representations in specific domains for effective application in information retrieval tasks.

1.2.3.2 Specific objectives

In order to achieve the overall objective, the following tasks are defined as specific goals for this work:

- Study the field of Information Retrieval, focusing on the evolution from classical to neural approaches, and explore Natural Language Processing techniques for semantic feature extraction.
- Investigate various forms of knowledge representation and develop methodologies for integrating them with language models to improve domain-specific understanding.
- Develop a dataset in the field of oil well drilling and extraction, comprising Daily Drilling Reports.

- To design a training framework that leverages graph-based representations to identify structurally informed hard negative samples, strengthening the contrastive learning signal used to train dense retrieval models.
- To rigorously benchmark the proposed model against a wide spectrum of baselines, including sparse, dense, and late-interaction models and validating its effectiveness in zero-shot and fine-tuned scenarios.

1.2.4 Dissertation Structure

This dissertation is organized into the following chapters:

- **Chapter 2 (Literature Review):** Reviews the theoretical foundations of IR, covering classical methods, modern neural architectures, and the emerging field of Knowledge-Enhanced Pre-trained Language Models (KEPLMs). It cites relevant studies on knowledge injection and graph neural networks.
- **Chapter 3 (Methodology):** Details the methodology used in this research, including the datasets created, the construction of the knowledge graph, and the proposed overall solution. This chapter also describes the expected outcomes and the evaluation metrics used to assess the model's performance.
- **Chapter 4 (Results):** Presents the empirical findings, including a comparative analysis of the proposed model against state-of-the-art baselines and a discussion on the project timeline and completed activities.
- **Chapter 5 (Discussion):** Interprets the key findings, contextualizing the superiority of the fine-tuned model and discussing the implications of the dense retrieval gap and the trade-offs between alignment and uniformity in the vector space.
- **Chapter 6 (Conclusion):** Summarizes the research contributions and outlines future directions for integrating this retrieval framework into broader generative AI systems.

2 Literature Review

To understand the state of the art in information retrieval, several authors have conducted bibliographic research to identify the most relevant and prominent models in the field. Zhao et al. (2022) presents a comprehensive study highlighting the main information retrieval models based on pre-trained language models. Their work provides an extensive review, covering various aspects of these models, including architecture, training methods, indexing techniques, evaluation metrics, integration with other retrieval process modules, and other advanced topics. Similarly, Guo et al. (2022) offers an in-depth literature review focusing on semantic models applied at the initial stage of retrieval. The most recent study in the field, by Hambarde e Proenca (2023), explores models used in both the first and second stages of the retrieval process.

2.1 Information Retrieval

The field of Information Retrieval has evolved significantly, with modern systems focusing on efficiently retrieving and ranking relevant documents in response to user queries. To ensure conceptual clarity, we adopt the standard terminology defined by Manning, Raghavan e Schütze (2008) and Baeza-Yates e Ribeiro-Neto (2011), which segments the IR lifecycle into three distinct phases: indexing, retrieval, and ranking. Indexing is the offline process of constructing data structures, such as inverted indices, that enable efficient access to the document corpus. The subsequent online processes are typically divided into a multi-stage pipeline. In the retrieval stage (often termed candidate generation), the primary objective is to extract a subset of potentially relevant documents from a large corpus, prioritizing recall and computational speed. The ranking (or re-ranking) stage subsequently refines this subset to present the most pertinent results to the user based on precision-oriented relevance scores.

The retriever is a crucial component of Information Retrieval as it aims to narrow down the pool of documents by identifying relevant ones based on a user's query. In an IR pipeline, the retriever is responsible for initially retrieving potentially relevant documents from a large corpus, which can contain millions to billions of documents depending on the application (e.g., Wikipedia or the Web) (GUO et al., 2022). Efficiency is vital at this stage due to the scale of the corpus, and the objective is to recall the most likely relevant documents before passing them to subsequent re-ranking stages for further refinement (ZHAO et al., 2022). A variety of algorithms have been developed over the years to perform retrieval tasks, each offering different strengths in terms of efficiency and relevance. Following the taxonomy proposed by Baeza-Yates e Ribeiro-Neto (2011), classical retrieval models are generally categorized into Set-Theoretic (e.g., Boolean), Algebraic (e.g., Vector Space), and Probabilistic models. Traditional

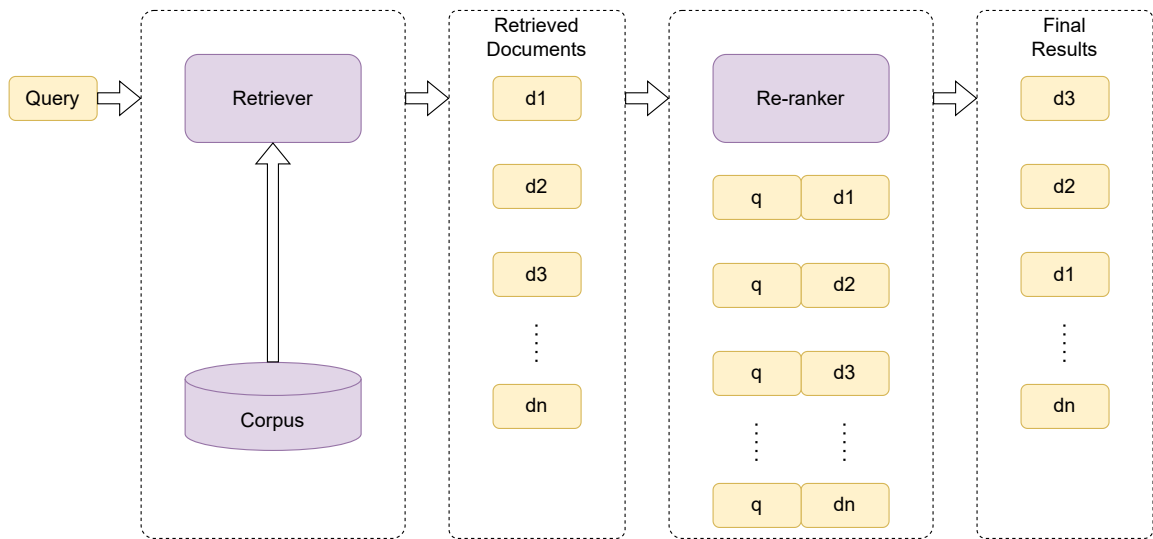


Figure 1 – Basic Information Retrieval Pipeline: The process begins with an initial query that is submitted to the retriever module. The retriever accesses the corpus and retrieves a set of relevant documents. These documents are then passed to the re-ranker module, which reorders them based on relevance scores to produce a refined, final ranked response.

models, such as the vector space model and boolean model, were foundational in early IR systems, allowing documents to be retrieved based on lexical matching with the query terms (HAMBARDE; PROENCA, 2023). These models, while efficient, are limited by their inability to capture deeper semantic relationships between words. In order to overcome these limitations, more advanced models such as Latent Semantic Indexing (LSI) (DEERWESTER et al., 1990) and Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2003) have been introduced. These models analyze latent structures in the data to reveal semantic patterns. However, the most significant recent advances have resulted from the integration of pre-trained language models, like BERT (DEVLIN et al., 2018). These models utilize deep learning techniques to create contextual embeddings for both the query and documents, thereby significantly enhancing the quality of retrieval by capturing complex semantic relationships (HAMBARDE; PROENCA, 2023).

The field of information retrieval has seen significant changes with the emergence of neural IR models. Traditional term-based models such as BM25, which rely on the bag-of-words approach, have been highly regarded for their strong performance and efficient document retrieval. While BM25 continues to serve as a strong baseline for many IR tasks, its emphasis on term frequency and lexical similarity can sometimes overlook semantic relationships between query terms and documents. In contrast, neural retrieval models map both the query and the document into high-dimensional vector spaces, and their relevance is calculated using inner product or cosine similarity (ROBERTSON et al., 1994). This approach enables the retriever to better capture semantic similarities between the query and the document content, which is particularly important as text collections become larger and more complex. These

vector-based retrieval methods have now become the primary approach, allowing for a more nuanced understanding of query-document relationships and significantly improving retrieval performance (ZHU et al., 2024).

Formally, the retriever learns a scoring function $s(q, d)$ that assigns a relevance score to a query-document pair (q, d) . Given a document collection $C = \{d_1, d_2, \dots, d_N\}$ and a query q , the retriever returns a ranked list of candidate documents $D_i = [d_{i1}, d_{i2}, \dots, d_{ik}]$. The scoring function can be expressed as

$$s(q, d) = f(\phi(q), \psi(d)), \quad (1)$$

where $q \in X$ denotes a query drawn from the query space X , and $d \in Y$ denotes a document from the document space Y . The functions $\phi : X \rightarrow \mathbb{R}^{k_1}$ and $\psi : Y \rightarrow \mathbb{R}^{k_2}$ represent encoder functions that map the query and the document to dense vector representations $\phi(q)$ and $\psi(d)$, respectively. The function $f(\cdot)$ computes a similarity score between these embeddings, which is used to rank candidate documents according to their estimated relevance (GUO et al., 2022).

The re-ranking stage is crucial for refining the initial set of documents retrieved by the retriever. While the retriever efficiently identifies potentially relevant documents from a large corpus, the re-ranking process fine-tunes the order of these documents based on more sophisticated ranking criteria, ensuring that the most relevant documents appear at the top of the list and enhancing the overall effectiveness of retrieval (ZHAO et al., 2022). This process leverages various models that differ from those used in the retrieval stage. Some of the key models employed in re-ranking include RankNet (BURGES et al., 2005), which uses a learning-to-rank approach, and neural models such as DRMM (GUO et al., 2016) and Duet (MITRA; DIAZ; CRASWELL, 2016), which apply deep learning techniques to improve ranking accuracy. These models often incorporate advanced techniques like reinforcement learning (ZHOU; AGICHTEIN, 2020), contextual embeddings (MACAVANEY et al., 2019), and attention mechanisms (LI et al., 2022), enabling the system to better understand and rank documents based on the specific criteria of the user's query (HAMBARDE; PROENCA, 2023).

2.1.1 Classical Retrieval

Traditional retrieval methods form the foundation of Information Retrieval systems, especially in their early stages. These methods, often referred to as classical or term-based retrieval, focus on matching terms between a user's query and documents in a collection. While they are efficient and simple to implement, they have limitations in handling language complexities such as synonymy, polysemy, and contextual nuances (MANNING; RAGHAVAN; SCHÜTZE, 2008; BAEZA-YATES; RIBEIRO-NETO, 2011). However, classical retrieval approaches are still

widely used in some practical applications, especially in smaller-scale systems where their speed and simplicity are advantageous (MANNING; RAGHAVAN; SCHÜTZE, 2008).

2.1.1.1 Set-Theoretic Models

The most basic of these methods is Boolean retrieval, which relies on set theory and boolean algebra. It matches exact terms in the query with terms in the document. This method considers a document relevant if it contains the specified terms without taking into account variations in meaning or context. While Boolean retrieval is fast and easy to implement, it has several limitations, especially in terms of flexibility and semantic understanding (BAEZA-YATES; RIBEIRO-NETO, 2011).

2.1.1.2 Algebraic Models

Algebraic models represent documents and queries as vectors, matrices, or tuples. The Vector Space Model (VSM) (SALTON; WONG; YANG, 1975) represents queries and documents as high-dimensional sparse vectors in a shared vector space. In this model, each dimension corresponds to a term in the vocabulary. Crucially, VSM is a framework that allows for various weighting schemes; it is not inherently probabilistic. The weight of each term in a document is determined by functions such as term frequency (TF) and inverse document frequency (IDF) (SALTON, 1991). TF-IDF is the most prominent weighting scheme within this algebraic framework, used to penalize common terms and highlight discriminative ones. The relevance between a query and a document is then computed using similarity measures, most commonly cosine similarity, which compares the angle between the query and document vectors. Documents with a higher cosine similarity score are considered more relevant (GUO et al., 2022).

2.1.1.3 Probabilistic Models

Distinct from algebraic models, Probabilistic retrieval models introduce a formal mathematical framework to estimate the probability that a document is relevant to a given query. The Binary Independence Model (BIM) (ROBERTSON; JONES, 1988) was the first to propose a probabilistic approach, representing documents as binary vectors where terms are either present or absent. While BIM made significant strides in formalizing the retrieval process, its assumptions about term independence were unrealistic. This led to the development of more sophisticated models like BM25 (ROBERTSON; ZARAGOZA, 2009), which remains one of the most widely used IR models. Unlike the standard VSM which measures geometric distance, BM25 estimates the probability of relevance based on term distribution. BM25 adjusts term weights based on document length, term frequency, and document frequency, addressing the inefficiencies and inaccuracies of earlier models.

2.1.1.4 Language Models for IR

Language models for information retrieval mark a significant advancement in traditional retrieval methods. Instead of explicitly modeling the probability of relevance, language models like the Query Likelihood Model (QL) (PONTE; CROFT, 2017) and the Divergence from Randomness (DFR) model (AMATI; RIJSBERGEN, 2002). This approach shifts the focus from binary term matches to probabilistic term occurrences, considering each document as a generative model for the query. By doing so, language models capture more nuanced information about term distributions and have been demonstrated to outperform traditional models such as TF-IDF. These models are also based on the bag-of-words assumption, similar to Vector Space Model (VSM), but they provide a richer statistical basis for modeling relevance.

2.1.2 Semantic Retrieval

2.1.2.1 Query Augmentation

Query expansion is a technique aimed at improving retrieval performance by addressing the inherent mismatch between user queries and relevant documents. The main objective of query expansion is to enhance the original query by adding related terms that are semantically or contextually relevant. This allows the system to retrieve a wider range of relevant documents that may not explicitly match the original query terms. This approach has been widely studied and applied in various information retrieval systems and can generally be categorized into two types: global methods and local methods (XU; CROFT, 1996).

Global methods for query expansion rely on external resources or the statistical co-occurrence of terms across a large corpus to identify related terms. These methods aim to enhance the query by adding terms that are likely to be relevant based on the overall structure of the corpus. One approach is to use a manually constructed thesaurus, such as WordNet (VOORHEES, 1994), which leverages lexical-semantic relations between words (e.g., synonyms, hypernyms) to expand the query. Another global method involves analyzing word co-occurrences within the corpus itself to identify terms that frequently appear together with the original query terms, providing a data-driven way to enrich the query (LESK, 1969). Although global methods can be effective in improving average retrieval performance, they tend to be less reliable across different queries due to their reliance on general patterns rather than query-specific information (GUO et al., 2022).

Local methods for query expansion take a different approach by incorporating user feedback into the retrieval process or by focusing on the top-ranked documents retrieved by the initial query. A widely used local method is the Rocchio (1971) relevance feedback algorithm, which adjusts the query representation by incorporating feedback from relevant documents. Cao et al. (2008) forms the foundation for many modern Pseudo-Relevance Feedback (PRF) (CHANG et al., 2020) techniques, which have been proven highly effective in improving the

performance of various retrieval models. The idea behind PRF is that the documents retrieved in response to the original query are likely to contain additional terms relevant to the user's information need. These terms are then used to expand the query, allowing the retrieval system to improve performance by refining the search based on the initial results.

Despite the success of PRF techniques, they are not without challenges. One of the key issues associated with PRF is the query drift problem, where the expanded query diverges too far from the original query's intent, leading to the retrieval of irrelevant documents (MITRA et al., 1997). This occurs when the terms selected for expansion are not directly aligned with the user's information need but are instead over-influenced by the top-ranked documents. Managing query drift remains an active area of research in query expansion, as it can significantly impact the effectiveness of retrieval systems.

2.1.2.2 Document Augmentation

Document augmentation, also known as document expansion, is a technique used to enhance the representation of each document by adding related terms or concepts. The goal is to improve retrieval performance by ensuring that documents more effectively match user queries, even if the query terms differ from those originally present in the document. This method is particularly useful for improving the retrieval of short or sparsely worded documents, where the limited text makes matching difficult. By expanding a document's representation, information retrieval systems can increase the likelihood of retrieving relevant documents, even when there is a difference in wording between the query and the document (GUO et al., 2022; HAMBARDE; PROENCA, 2023).

Several approaches have been explored for document augmentation, including both internal corpus-based methods and external knowledge sources. Early work by Singhal e Pereira (1999) introduced a technique where each document was treated as a query against the entire corpus, and the most relevant documents retrieved were used to enhance the original document's representation by adding a vector linearly weighted mixture of related documents into the original document. This retrieval-based expansion approach was later applied by Efron, Organisciak e Fenlon (2012), resulting in performance improvements due to the analysis of the resulting set. Additionally, clustering techniques have been used to identify related documents within the corpus, allowing for document expansion based on groupings of similar texts (KURLAND; LEE, 2004; LIU; CROFT, 2004). Significant improvements over non-expanded baselines on the TREC ad-hoc document retrieval task were reported.

The use of external resources, such as lexical databases like WordNet, has been common in document expansion. WordNet, in particular, is widely utilized to add semantically related terms to document representations, making the document more likely to match queries expressed in different ways (SHERMAN; EFRON, 2017). Studies have demonstrated that leveraging external collections like can significantly improve retrieval performance, especially when dealing

with short texts or specialized vocabularies. Document augmentation offers great potential for enhancing retrieval performance, particularly in domains with short, under-specified documents or widely varying terminologies. However, document expansion techniques have been less popular in information retrieval research due to being less amenable to rapid experiments. Every time the expansion technique changes, the corpus needs to be re-indexed, which is a costly process (HAMBARDE; PROENCA, 2023).

2.1.3 Neural Methods for Retrieval

Neural models have transformed the field of Information Retrieval by offering new ways to capture the semantic and syntactic complexities of language. In the past, IR systems relied on methods such as term-based models or simpler machine learning approaches, which treated words independently and struggled to capture deeper linguistic relationships. However, the introduction of neural models, especially those utilizing deep learning techniques, has changed this approach by introducing more advanced methods for representing and retrieving information based on continuous vector spaces and complex non-linear interactions.

One of the initial advancements in applying neural networks to retrieval was the development of models based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models allowed IR systems to represent both documents and queries in low-dimensional dense vectors, effectively capturing semantic features that traditional term-based models overlooked. By transforming text into a continuous vector space, neural models could calculate similarity measures, such as cosine similarity or dot-product, to rank documents based on their relevance to a query (ZHOU; DING; GUO, 2014; HUANG et al., 2013). This represented a significant improvement over earlier models, as it enabled the system to consider the relationships between words in context rather than treating them as independent entities. Despite these advancements, early neural methods in information retrieval were mainly used in the re-ranking stage of retrieval pipelines. The objective was to enhance the results obtained by traditional methods. Although they improved ranking accuracy, the emphasis was on refining small-scale rather than large-scale retrieval tasks (GUO et al., 2016; HUANG et al., 2013).

A significant advancement in neural models occurred with the introduction of the Transformer architecture (VASWANI et al., 2017). This architecture transformed the processing of sequence data by introducing a self-attention mechanism, unlike traditional sequence models such as RNNs. The self-attention mechanism enables each token in a sequence to consider every other token, effectively capturing dependencies and allowing efficient parallelization on modern hardware like GPUs and TPUs. This architecture is well-suited for handling large datasets and training deep neural networks, making it a fundamental component for many developments in natural language processing and information retrieval.

Based on the success of the Transformer, pretrained language models like BERT (DEVLIN et al., 2018) and GPT-2 (RADFORD et al., 2019) have further significant changes

the field of retrieval. These models are pretrained on large datasets using self-supervised learning techniques, enabling them to capture comprehensive bidirectional representations of text that can be adapted for various downstream tasks, including IR (LIU et al., 2019). BERT, in particular, introduced the concept of masking random words during pretraining to encourage the model to comprehend the context of a word based on its surrounding terms. This ability resulted in significant enhancements in IR systems, as the model could more effectively grasp the nuances and context of both queries and documents, leading to more precise retrieval (ZHAO et al., 2022).

Neural methods for semantic retrieval can be broadly categorized into three types: sparse retrieval, dense retrieval, and hybrid retrieval. Sparse retrieval methods still rely on term-based indexing, but they incorporate neural networks to better understand term importance and context. Dense retrieval models represent queries and documents entirely in continuous vector spaces, where relevance is determined by the proximity of the query and document embeddings. Hybrid methods combine elements of both, using dense embeddings to enhance the effectiveness of sparse retrieval techniques (GUO et al., 2022).

2.1.3.1 Sparse Retrieval

Sparse retrieval methods focus on representing queries and documents using sparse vectors, where only a few dimensions are active. This sparsity aligns with human memory function and offers better interpretability, leading to efficient retrieval processes. These methods aim to improve the accuracy and relevance of document retrieval while maintaining computational efficiency, which is fundamental for large-scale search tasks. Zhao et al. (2022) proposes two main categories of sparse retrieval approaches: neural weighting schemes and sparse representation learning. The former focuses on improving term weighting using neural models, while the latter aims to learn sparse representations for both queries and documents through neural networks in a latent space.

Neural weighting schemes aim to improve traditional term-based retrieval by reassessing the significance of terms using neural models, often utilizing pretrained language models. One notable method, DeepTR (ZHENG; CALLAN, 2015), introduces a mechanism to adjust the weights of terms based on their semantic relationships, using distributed representations to capture the meaning of terms in context. Another important approach involves integrating neural word embeddings into retrieval models, as discussed in (ZUCCON et al., 2015), where the effectiveness of different embeddings in improving retrieval performance is assessed. This research emphasizes the importance of embeddings for representing semantic similarity in information retrieval tasks and outlines strategies for incorporating these embeddings into existing models.

One effective approach involves predicting the importance of terms based on their meaning rather than using predefined rules. For example, DeepCT (DAI; CALLAN, 2020) uses

BERT's contextualized token representations to estimate how important terms are within specific passages. By converting these token representations into real-valued term weights, DeepCT can assign greater importance to terms that are more relevant within their context. This method has been shown to improve the initial retrieval performance by better capturing the significance of terms. HDCT (DAI; CALLAN, 2020) builds on this by extending DeepCT to handle longer documents, breaking them into passages and combining the estimated term weights across these passages. HDCT also uses various data resources, such as titles, web inlinks, and pseudo-relevance feedback, to create weak supervision signals, further refining the estimation of term importance.

Another approach, COIL (GAO; DAI; CALLAN, 2021), computes relevance by utilizing the contextualized token representations of exact matching terms. By computing the dot product between token representations from the query encoder and text encoder (considering only overlapping terms) COIL sums the term-specific similarity scores to produce a relevance score. This method allows for more precise term matching by incorporating contextual information. Lin e Ma (2021) proposed a conceptual framework to unify these methods, highlighting that models like DeepCT assign scalar weights to query terms, while models like COIL assign vector weights. They also introduced uniCOIL, which reduces the vector weights in COIL to a single dimension, retaining effectiveness while improving efficiency.

In a recent development, Frej et al. (2020) have introduced a method for learning term discrimination values (TDVs). This method offers an alternative to the traditional Inverse Document Frequency in inverted indexing. The approach, which is based on FastText (BOJANOWSKI et al., 2017), not only focuses on pairwise ranking but also applies ℓ_1 -norm minimization to reduce the memory requirements of bag-of-words representations. This optimization improves both storage and retrieval speed. Additionally, Zuccon et al. (2015) explored the integration of word embeddings in a translation language model to enhance information retrieval. Their goal is to estimate translation probabilities between terms, which helps to mitigate vocabulary mismatch by identifying semantic relationships between query terms and document content. This approach leads to more accurate relevance scoring.

Sparse representation learning focuses on directly learning sparse vectors for queries and documents in a latent space using neural networks. Unlike traditional topic models, these latent dimensions do not correspond to clear concepts but are learned to capture semantic meanings of the texts. SPLADE and its improved versions SPLADE v2 and v3 (FORMAL; PIWOWARSKI; CLINCHANT, 2021; FORMAL et al., 2021; LASSANCE et al., 2024) project each term in queries and documents to a vocabulary-sized weight vector, where each dimension represents the weight of a term in the PLM's vocabulary. The weights are estimated using the logits from masked language models, and sparsity regularization is applied to ensure the representations remain sparse. This approach allows for both term re-weighting and expansion within a unified framework.

To address the vocabulary mismatch problem, term expansion techniques enhance queries or documents by incorporating semantically related terms identified by neural models. Methods such as docTTTTTquery (NOGUEIRA; LIN; EPISTEMIC, 2019) generate potential queries that a document could be relevant to, effectively enriching the document content with these generated queries. By appending these synthesized queries to the original documents, the retrieval system can enhance recall by matching queries that may employ different vocabulary.

Semantic hashing proposed by (SALAKHUTDINOV; HINTON, 2009) converts documents into compact binary codes, allowing for efficient approximate nearest neighbor searches and achieving high performance in retrieval tasks. Another approach, SNRM (ZAMANI et al., 2018), focuses on learning sparse document representations using neural networks, resulting in improved retrieval effectiveness. Additionally, UHD-BERT (JANG et al., 2021) extends this concept by generating ultra-high-dimensional sparse representations, which have shown improvements in full-ranking tasks. Ji, Shao e Yang (2019) studied ways to improve the efficiency of interaction-focused models for first-stage retrieval by using sparse representations. Their work specifically enhanced the performance of three neural retrieval algorithms: DRMM (GUO et al., 2016), KNRM (XIONG et al., 2017), and Conv-KNRM (DAI et al., 2018). To accomplish this, they employed a Locality Sensitive Hashing (LSH) approximation, which enabled fast histogram-based kernel computation and precomputed term vectors for caching during runtime.

2.1.3.2 Dense Retrieval

Dense retrieval represents a significant shift from traditional sparse retrieval methods by focusing on the semantic interaction between queries and documents within a continuous latent space. Unlike sparse models, which rely on term matching and representations with many inactive dimensions, dense retrieval methods utilize dense, continuous embeddings that capture deeper semantic relationships between words, leading to more accurate relevance evaluation in information retrieval systems. There are two primary architectures employed in dense retrieval: the cross-encoder and the bi-encoder.

The cross-encoder model offers a distinctive approach by considering the combination of a query and document as a single sequence. To do this, the input is created by joining the query and text together, with a separator symbol ([SEP]) between them. Additionally, a special token ("[CLS]") is added at the beginning of the sequence to help generate a unified representation. Once the combined sequence is processed by the pre-trained language model, the model captures the semantic interactions between all tokens in both the query and the text (ZHAO et al., 2022).

The overall representation of the entire sequence is usually obtained from the [CLS] token, which acts as a summary of the semantic relationship between the query and the document. In some cases, alternative methods, such as averaging the token embeddings, are used to obtain the final representation. This architecture resembles earlier interaction-based

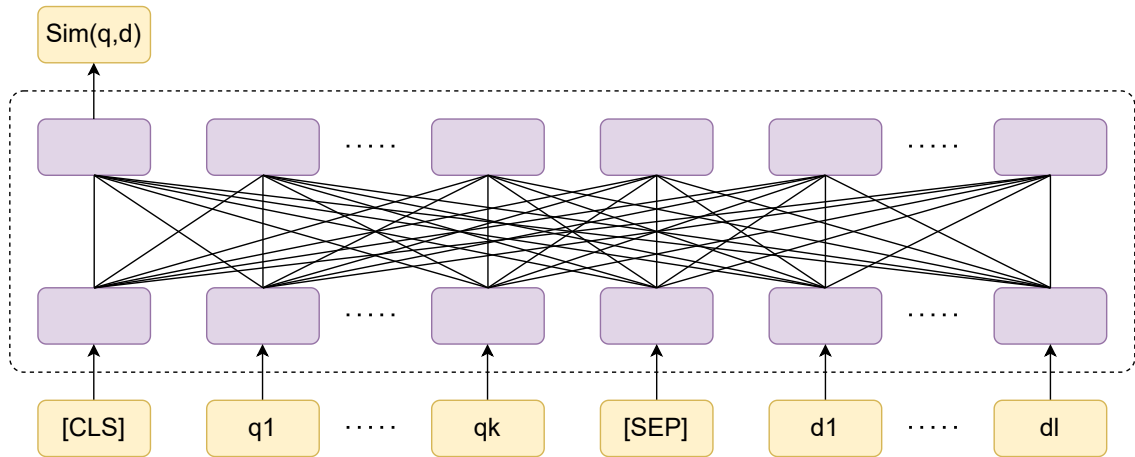


Figure 2 – Cross-Encoder Architecture for Dense Retrieval: In this architecture, the query and each candidate document are jointly encoded by concatenating them and passing them through a transformer-based model, which outputs a relevance score. Unlike the bi-encoder, the cross-encoder processes the query-document pairs simultaneously, allowing for deeper interaction between the two. This approach, while computationally more intensive, often achieves higher accuracy by capturing nuanced relationships between the query and documents.

models used before BERT, where tokens from the query and text could interact directly, allowing for a more comprehensive representation of the semantic match between the two inputs (ZHAO et al., 2022).

The bi-encoder, also known as the dual-encoder, is a two-tower architecture that is similar to earlier representation-based methods such as DSSM, but with a significant improvement: it replaces traditional neural networks with more powerful PLM encoders. This architecture is highly efficient for large-scale retrieval tasks because it can independently encode queries and documents into dense vector representations, $\phi(q)$ for the query and $\psi(d)$ for the document (GUO et al., 2022).

These dense representations are then compared using simple similarity metrics, such as cosine similarity or dot product, to calculate a relevance score. One of the major advantages of the bi-encoder setup is its ability to precompute and store document embeddings, which significantly speeds up the retrieval process when used with approximate nearest neighbor (ANN) algorithms during inference. This efficiency makes the bi-encoder a practical choice for large-scale applications where rapid retrieval is critical. Given the complexity of documents that often contain rich content and intricate structures that require careful attention while designing their representation function, dense retrieval methods are categorized based on how they handle document representation, usually categorized into term-level and document-level representation learning (GUO et al., 2022)

Term-level representation learning involves generating detailed term-level embeddings for both queries and documents, which are then compared using a similarity function to calculate

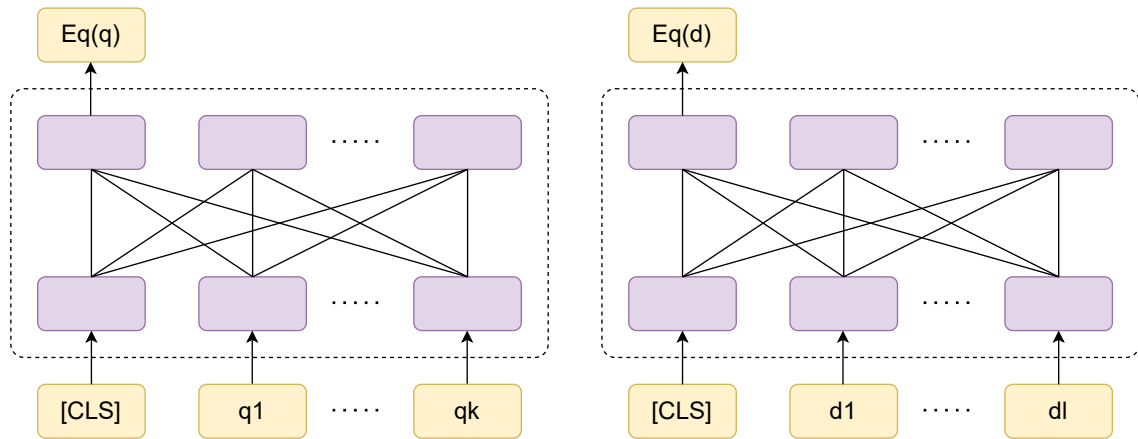


Figure 3 – Bi-Encoder Architecture for Dense Retrieval: In this architecture, the query and documents are independently encoded into dense vector representations using a shared or separate encoder model. These vectors are then compared, typically using a similarity measure to identify the most relevant documents. The bi-encoder approach enables efficient retrieval by allowing pre-computation and storage of document embeddings.

the relevance score. Traditional approaches, such as those developed by Kenter e Rijke (2015), replace term frequency in BM25 with maximum cosine similarity between word embeddings, showing improved performance over baseline methods. Other models, like DESM (MITRA et al., 2016), use Word2Vec embeddings to map query and document terms into different spaces, further improving ranking models in tasks like document re-ranking.

Recent advancements in the use of contextual word embeddings and pre-trained language models like BERT have significantly improved the quality of term-level representations. The DC-BERT model, proposed by Zhang et al. (2020), employs two BERT models to enhance the model's efficiency by decoupling the contextual encoding framework. The first BERT encodes the query online, while the second encodes documents offline and stores these representations, the dual encoders efficiently generate term representations, achieving speedups without sacrificing much retrieval performance. ColBERT (KHATTAB; ZAHARIA, 2020) takes a different approach by using a MaxSim operator to efficiently match query and document terms, yielding highly effective results while maintaining computational efficiency, especially in top- k retrieval tasks. COIL (GAO; DAI; CALLAN, 2021) simplifies this by only interacting with matching document terms, offering performance comparable to more complex models like ColBERT. Moving beyond individual terms, phrase-level representation learning extends the model to capture n-grams or sentence-level embeddings. Models like those proposed in Seo et al. (2018), Seo et al. (2019) combine dense representations with sparse contextualized embeddings, improving the retrieval quality and scalability for open-domain question answering. By precomputing phrase encodings, these models facilitate efficient retrieval, a concept further advanced by models like MUPPET (FELDMAN; EL-YANIV, 2019), which focuses on sentence-level representations in multi-hop question answering tasks.

Document-level representation learning focuses on creating dense global embeddings for both queries and documents, capturing their semantic meanings. These embeddings, $\phi(q)$ for queries and $\psi(d)$ for documents, are compared using similarity functions like dot product or cosine similarity to generate relevance scores. Initial methods aggregated word embeddings into document-level representations, such as the Fisher Vector (FV) model (CLINCHANT; PERRONNIN, 2013), which, while outperforming some earlier techniques, did not surpass models like TF-IDF in retrieval tasks. Later approaches, such as Paragraph Vector (PV) (LE; MIKOLOV, 2014), aimed to resolve the limitations of simple word embeddings by learning fixed-length representations for variable-length text but showed only moderate improvements.

To address these issues, models like Neural Vector Space Model (NVSM) (GYSEL; RIJKE; KANOULAS, 2018) and SAFIR (AGOSTI; MARCHESIN; SILVELLO, 2020) introduced unsupervised methods for learning latent representations that improved document retrieval performance. These models project query embeddings into a shared document feature space and compute relevance through cosine similarity, yielding better results than classical lexical retrieval models on benchmark datasets. Modern approaches using BERT-based models, such as DPR (KARPUKHIN et al., 2020), have further advanced document-level representation learning by utilizing dual encoders for efficient retrieval while outperforming traditional methods like BM25. Distillation techniques have also been employed to compress more complex models, into simpler architectures, maintaining effectiveness while improving inference speed and reducing computational requirements.

RepBERT proposed by Zhan et al. (2020) suggests a method for creating fixed-size contextualized embeddings to represent documents and queries. CoRT (Complementary Rankings from Transformers), introduced by Wrzalik e Krechel (2020), is a ranking method that uses textual representations from pre-trained language models to complement term-based ranking functions. Ren et al. (2021) estimates similarity relationships by comparing retrieved passages with the query, rather than working with individual tokens from each. The authors also present two formulations of similarity relationships and propose a technique for generating pseudo-labeled data through knowledge distillation. Tay et al. (2022) proposed an architecture for a sequence-to-sequence learning system that aims to map a query directly to a relevant document index. This architecture, called the Differentiable Search Index (DSI), is implemented using a pre-trained Transformer model where the document corpus is encoded directly into the model's parameters.

The cross-encoder excels at modeling fine-grained semantic interactions between query-text pairs, making it very effective for tasks that involve matching relevance. Its architecture processes the entire query and text together, allowing tokens from both inputs to interact, which captures detailed relationships. This fine-grained interaction leads to superior performance in relevance ranking tasks. However, using the cross-encoder comes with a cost: it is computationally expensive and less efficient for large-scale retrieval. In contrast, the bi-encoder is more efficient

and flexible. It encodes queries and documents separately into dense vector representations, which can then be compared using similarity functions like cosine similarity. This separation allows for the precomputation of document embeddings, making the bi-encoder suitable for large-scale retrieval tasks where speed is critical. The architecture also allows flexibility by enabling different networks for encoding queries and documents, supporting diverse retrieval needs, such as phrase indexing. However, the bi-encoder struggles to capture fine-grained interactions between queries and texts, even though methods like multi-representation embeddings have been proposed to address this issue. Despite these improvements, the bi-encoder generally performs worse than the cross-encoder in fine-grained semantic tasks (ZHAO et al., 2022).

When training dense retrievers, it's important to carefully consider several factors, especially the formulation of loss functions that guide the learning process. As described above, neural semantic retrieval models always define functions ϕ , ψ , and f in the network structure (GUO et al., 2022). These functions are usually learned from data using deep learning technology. Effective loss functions are crucial for aligning the representations of queries and documents in the vector space, ensuring that relevant documents are positioned closer to the query while irrelevant ones are pushed further away.

One commonly used objective for training dense retrievers is the Negative Log-Likelihood (NLL) loss, which maximizes the likelihood of retrieving relevant documents for a given query. Let (q_i, d_i^+) denote a query and its associated relevant document. Given a set of negative documents D^- , the loss for a single training instance can be written as

$$L(q_i, d_i^+) = -\log \frac{\exp(f(\phi(q_i), \psi(d_i^+)))}{\exp(f(\phi(q_i), \psi(d_i^+))) + \sum_{d' \in D^-} \exp(f(\phi(q_i), \psi(d')))}, \quad (2)$$

where D^- represents the set of negative documents associated with the query q_i . The function $f(\cdot)$ computes the similarity between the query embedding $\phi(q_i)$ and document embeddings $\psi(d)$. This objective encourages the model to assign higher similarity scores to relevant documents while penalizing irrelevant ones.

An alternative optimization strategy is the triplet ranking loss, which directly models the relative ordering between relevant and irrelevant documents. Given a query q_i , a relevant document d_i^+ , and an irrelevant document d_i^- , the loss is defined as

$$L(q_i, d_i^+, d_i^-) = \max\left(0, 1 - \left(f(\phi(q_i), \psi(d_i^+)) - f(\phi(q_i), \psi(d_i^-))\right)\right), \quad (3)$$

where the margin value 1 enforces a minimum separation between the similarity scores of relevant and irrelevant documents.

In order to instantiate the function $f(\cdot)$ in the loss functions, we can apply different vector similarity or distance measurements to calculate how similar the query and document

embeddings are. These measurements are important for determining how effectively the dense retriever matches relevant documents with the query. Some of the commonly used similarity metrics include:

- **Euclidean Distance:** Euclidean distance measures the straight-line distance between two vectors in a high-dimensional space. It is a dissimilarity measure, meaning smaller distances indicate higher similarity. It can be defined as

$$s(q, d) = \|\phi(q) - \psi(d)\|_2 = \sqrt{\sum_{i=1}^n (\phi_i(q) - \psi_i(d))^2}, \quad (4)$$

where n denotes the dimensionality of the embedding vectors.

- **Inner Product:** The inner product, also known as the dot product, measures the similarity between two vectors by multiplying their corresponding components and then summing the results. It is frequently employed in dense retrieval because it effectively captures the alignment between query and document embeddings. The inner product is defined as

$$s(q, d) = \langle \phi(q), \psi(d) \rangle = \sum_{i=1}^n \phi_i(q) \psi_i(d), \quad (5)$$

- **Cosine Similarity:** Cosine similarity measures the cosine of the angle between two vectors in a high-dimensional space. It focuses on the direction rather than the magnitude. This measure is widely used in information retrieval tasks as it normalizes the vectors and quantifies the similarity on a scale of -1 to 1. A value of 1 indicates that the vectors are identical, while a value of -1 indicates complete opposition.

$$s(q, d) = \frac{\langle \phi(q), \psi(d) \rangle}{\|\phi(q)\|_2 \|\psi(d)\|_2}. \quad (6)$$

When optimizing a dense retriever, it is important to select negative examples for computing the negative log-likelihood loss during training. One common strategy is random sampling, where each positive text is paired with randomly chosen negatives. However, this approach can be impractical due to memory constraints in GPU batch training when dealing with a large number of negatives. To overcome this, in-batch negatives are used, where positive texts associated with other queries in the same batch serve as negative samples for a given query.

To further improve training, cross-batch negatives have been introduced, especially in multi-GPU settings. This method involves reusing examples across GPUs, effectively increasing the number of negatives available for training. While these techniques provide more negative examples, they may not always generate hard negatives, which are irrelevant texts that still share high semantic similarity with the query. Including hard negatives is crucial for enhancing the model's ability to distinguish between relevant and irrelevant content.

Traditionally, static hard negatives are selected from a fixed set, typically derived from the top- k results of a different retriever, such as a BM25 or a pre-trained dense model. Although useful, this method is limited by the iterative nature of model convergence. Consequently, dynamic hard negatives, which are adaptively updated as the model improves, are proposed as a more robust solution for refining the retriever’s latent space.

However, the construction of high-quality training triples remains a significant bottleneck, especially in specialized domains like Oil and Gas where human-labeled data is sparse. Modern research has shifted toward unsupervised domain adaptation using the few-shot capabilities of LLMs. Early methods such as InPars (BONIFACIO et al., 2022) and Promptagator (DAI et al., 2023) demonstrate that LLMs can generate relevant synthetic queries for specific documents. To ensure reliability, Promptagator introduced round-trip consistency filtering, where a query is retained only if it retrieves its source document, thereby ensuring semantic alignment.

Building on these generation and filtering techniques, UDAPDR (SAAD-FALCON et al., 2023) proposes a comprehensive framework for unsupervised domain adaptation. This method utilizes high-performance LLMs to seed the generation of synthetic queries, which are then scaled up using more efficient models. Crucially, UDAPDR employs a distillation strategy where domain-specific rerankers, trained on the synthetic data serve as teachers to train a single, efficient bi-encoder retriever. This multi-stage process, combining synthetic query generation, consistency filtering, and knowledge distillation, represents the state-of-the-art in constructing training triples, enabling high-performance dense retrieval even in zero-shot domain adaptation scenarios.

Thakur et al. (2024) proposed the SWIM-IR framework to address the uninformative query problem often encountered in zero-shot generation. They introduced Summarize-then-Ask Prompting (SAP), which decomposes the generation into two stages. In the first stage, the LLM produces an extractive summary (e_s) from the input passage (p_s) to isolate salient technical information. In the second stage, this summary acts as an intermediate semantic signal to formulate a linguistically fluent query (q_t) in the target language. By forcing the generator to distill the passage content before asking, SAP ensures that synthetic queries are grounded in the core facts of the document rather than peripheral noise. This approach is particularly effective for industrial reports, which often cover multiple operational events in a single document, as it allows the retriever to learn from queries that precisely reflect the primary narrative of the passage.

Complementary to the focus on positive signal quality, ReadQG (JU et al., 2024) introduces a paradigm shift by generating diverse synthetic queries with continuous relevance scores. This framework utilizes soft prompt-tuning to control the relevance dynamic between a document and its generated queries. By employing learnable embedding prompts, specifically an instruction prompt (P_{inst}) and a relevance prompt ($P_{rel}(r)$), the model is conditioned to generate a query $\hat{q}(r)$ based on a target relevance score $r \in [0, 1]$. To prevent the model

from degenerating into trivial or non-informative outputs, ReadQG implements a Calibrated Sequence Likelihood (L_{cal}) objective:

$$L_{cal} = \sum_{(r,r')} \max(0, \epsilon - \log P_{fw} + \log P_{rev}). \quad (7)$$

2.2 Knowledge Graphs

Factual knowledge refers to the understanding of terminology and specific details that describe objective realities. This type of knowledge is essential for conveying precise information about particular elements or objects. Conceptual knowledge, on the other hand, involves recognizing the relationships among fundamental elements within a larger structure. This allows these components to interact and function cohesively, as seen in principles, generalizations, or theories. Procedural knowledge focuses on understanding how to carry out specific tasks or actions. It includes the use of methods, techniques, and algorithms that guide behavior or problem-solving processes. Lastly, metacognitive knowledge relates to self-awareness in cognition. It involves understanding one's own thinking processes and the ability to regulate those processes through active reflection and initiative (YANG et al., 2023).

2.2.1 Knowledge Representations

Davis, Shrobe e Szolovits (1993) define knowledge representation by highlighting five distinct roles it plays. First, it serves as a surrogate for the actual entities it represents, enabling reasoning without direct interaction. Second, it embodies ontological commitments, influencing our understanding of the world. Third, it acts as a fragmentary theory of intelligent reasoning. Fourth, it functions as a medium for computational efficiency, organizing knowledge to support inferences. Lastly, it facilitates human expression and cognitive understanding.

Several conventional methods for knowledge representation are commonly used, each with its own strengths and limitations. First-order predicate logic utilizes symbols to represent objects, relationships, and functions. While it allows for consistent inference, it struggles to handle procedural knowledge effectively. Semantic networks are represented as directed graphs, where nodes denote concepts and edges represent relationships. This approach offers simplicity and ease of use in computer systems, but it is limited to managing conceptual knowledge and cannot effectively represent dynamic knowledge like procedures. Frame representation organizes knowledge hierarchically using frames and slots for attributes, which helps avoid redundancy through inheritance. However, this method may introduce errors and lacks generality, particularly when representing procedural knowledge. Script representation illustrates sequences of actions in a temporal or causal order, making it effective for representing procedural knowledge, but it is not suitable for conceptual or factual knowledge (YANG et al., 2023).

Ontology, originally a philosophical concept, has been adapted in computing to describe structured computational models of the world. Ontologies provide organized knowledge that is valuable in areas such as information retrieval and natural language processing. Ontologies support both static and task-specific knowledge, enhancing models with reasoning capabilities. The introduction of the Semantic Web by Berners-Lee, Hendler e Lassila (2001) aimed to improve semantic understanding on a large scale by linking data through meaning rather than mere strings. This vision is realized through the Resource Description Framework (RDF), which represents data in a semantic network using triples, and the Web Ontology Language (OWL), which enables computers to reason about categories, attributes, and instances of objects.

Finally, Knowledge Graphs (KGs) are fundamental to modern knowledge representation systems. They are structured as networks, with entities represented as nodes and relationships as edges. KGs are based on Semantic Web technologies and are organized in triples denoted as $G = \{E, R, F\}$. In this structure, E represents the set of entities, where each entity is a distinct and identifiable item. The set R represents the relationships between entities, with each relationship acting as an edge connecting different entities. The set F consists of triples (h, r, t) , where h is the head entity, r is the relationship, and t is the tail entity. While KGs are effective in managing factual and conceptual knowledge, they typically do not encompass procedural knowledge due to the inherent uncertainties and complexities involved (YANG et al., 2023; YE et al., 2022).

2.2.2 Knowledge Representation Learning

Although KGs are highly effective for structuring data, the underlying symbolic nature of the triples often makes them difficult to manipulate computationally (WANG et al., 2017). To address this, Knowledge Representation Learning (KRL), also frequently referred to as Knowledge Graph Embedding (KGE), has emerged as a critical foundation for modern KG applications (YE et al., 2022). The core objective of KRL is to complete the distributed representation of entities and relations by mapping them into continuous, low-dimensional vector spaces (WANG et al., 2017; YE et al., 2022). This projection simplifies manipulation while preserving the inherent semantic structure of the graph. Compared to traditional one-hot encoding, KRL significantly improves computational efficiency and mitigates issues related to data sparseness (YE et al., 2022). These embeddings facilitate the integration of heterogeneous information and serve as a basis for advanced tasks such as link prediction, knowledge graph alignment, entity classification, and reasoning (YE et al., 2022; WANG et al., 2017).

2.2.2.1 Translational Distance Models

Translational distance models utilize distance-based scoring functions to measure the plausibility of a fact triplet (h, r, t) . The fundamental intuition is that the relationship r functions as a translation in the vector space from the head entity h to the tail entity t .

Formally, if a fact holds true, the embedding of the tail should be close to the embedding of the head plus the relation vector, i.e., $h + r \approx t$. The seminal model, TransE (BORDES et al., 2013), embeds entities and relations in a single d -dimensional vector space, scoring triples based on the distance $\|h + r - t\|$. While computationally efficient, TransE struggles with complex relations such as 1-to-N, N-to-1, and N-to-N mappings, as it forces entities involved in such relations to have identical representations.

To address these limitations, subsequent models introduced mechanisms to allow and entity to separate entity and relation spaces. TransH (WANG et al., 2014) introduces relation-specific hyperplanes, allowing an entity to have distinct representations depending on the relation involved. This is achieved with each relation r as a vector $\mathbf{r}$ on a hyperplane with w_r as the normal vector, the head h and tail t are then projected onto these relation-specific hyperplanes, enabling different roles of entities in different relations. TransR (JI et al., 2015) further generalizes this by modeling entities and relations in completely separate spaces, projecting entities into the relation space via projection matrices. Although powerful in modeling complex relations, TransR loses the efficiency of the previous methods once that it requires a matrix for each relation.

TransD (LIN et al., 2015) refines TransR by employing dynamic mapping matrices determined by both the entity and the relation. By decomposing the projection matrix into a product of vectors, TransD reduces complexity and captures the diversity of entity types more effectively. TransSparse (JI et al., 2016) further simplifies TransR by enforcing sparsity on the projection matrices to handle the heterogeneity and data imbalance. TransSparse proposes two variants: *TransSparse (share)*, which uses the same sparse projection matrix for both head and tail entities, and *TransSparse (separate)*, which utilizes distinct sparse matrices for the head and tail. This approach significantly reduces the parameter space to $\mathcal{O}(nd + (1 - u)mdk)$, where u represents the degree of sparsity.

Another line of research improves upon TransE by relaxing the strict requirement that $h + r \approx t$. TransM (FAN et al., 2014) argues that the translation assumption should not apply equally to all relations. It associates each fact with a relation-specific weight θ_r , assigning lower weights to complex mappings (1-to-N, N-to-1) to allow the tail entity to lie farther from the translated head. ManifoldE (XIAO; HUANG; ZHU, 2016) relaxes the point-wise constraint to a manifold constraint. It posits that t should lie on a hyper-sphere (manifold) centered at $h + r$ with a radius θ_r , rather than at the exact point. Similarly, TransF (FENG et al., 2016) relaxes the geometric position to a directional requirement, scoring facts based on whether t lies in the same direction as $h + r$ and h in the same direction as $t - r$. Finally, TransA (XIAO et al., 2015) introduces an adaptive metric approach, utilizing a symmetric non-negative matrix M_r for each relation to define a scoring function based on the Mahalanobis distance, offering greater flexibility for complex relation shapes.

2.2.2.2 Semantic Matching Models

Semantic matching models evaluate the likelihood of facts by examining the underlying meanings of entities and relations using similarity-based scoring functions. One model, RESCAL (NICKEL; TRESP; KRIEGEL, 2011), represents entities and relations as vectors and matrices, and scores facts through a bilinear function. This allows the model to capture the local structure of relationships in knowledge graphs. To make the model more efficient, DistMult (YANG et al., 2014) simplifies RESCAL by using diagonal matrices for relations. HolE (NICKEL; ROSASCO; POGGIO, 2016), on the other hand, combines entity representations using circular correlation and then matches them with the relation vector to score facts while SME (GLOROT et al., 2013) uses neural networks to perform semantic matching.

2.2.2.3 Graph Neural Networks

While translational and semantic matching models primarily rely on the direct structure of triplets, Graph Neural Networks (GNNs) leverage the broader topological structure of the graph. The core concept of a GNN is the *message passing* framework, where node embeddings are iteratively updated by aggregating information from their local neighborhood (YE et al., 2022). Fundamentally, a GNN aims to learn an embedded state $\mathbf{h}_v \in \mathbb{R}^d$ for each node v that encodes both its intrinsic attributes and the characteristics of its neighbors. This process is often framed as a fixed-point iteration problem based on the Banach fixed point theorem (ASIF et al., 2021). The state update for a node v at time step $t + 1$ can be formalized as:

$$\mathbf{H}^{t+1} = F(\mathbf{H}^t, \mathbf{X}) \quad (8)$$

where $\mathbf{H}$ represents the state of all nodes, $\mathbf{X}$ denotes the feature matrix (including edge and node features), and $F(\cdot)$ is a global transition function, typically a feed-forward neural network (YE et al., 2022). In a local context, the embedding of a node v is updated by aggregating features from its neighbors $\mathcal{N}(v)$ and combining them with its own previous state:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\sum_{u \in \mathcal{N}(v)} f(\mathbf{h}_u^{(l)}, \mathbf{h}_v^{(l)}) \right) \quad (9)$$

This iterative process allows the model to capture high-order dependencies and structural patterns that extend beyond immediate neighbors (YE et al., 2022; BRONSTEIN et al., 2017).

2.2.2.3.1 Graph Convolutional Networks (GCN)

Graph Convolutional Networks (GCNs) are neural architectures designed to learn representations from graph-structured data by iteratively aggregating information from neighboring nodes. Instead of operating on regular grids such as images, GCNs generalize the notion of

convolution to irregular graph domains, where each node updates its representation based on the features of its local neighborhood.

The most widely adopted formulation of GCNs was introduced by Kipf e Welling (2016), who proposed a simplified and computationally efficient propagation rule for graph convolution. Given a graph $G = (V, E)$ with adjacency matrix $\mathbf{A}$ and node feature matrix $\mathbf{X}$, a GCN layer computes new node representations according to:

$$\mathbf{H}^{(l+1)} = \sigma \left(\tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} \mathbf{H}^{(l)} \mathbf{W}^{(l)} \right) \quad (10)$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, $\mathbf{W}^{(l)}$ represents the trainable weight matrix of layer l , and $\sigma(\cdot)$ is a nonlinear activation function. The matrix $\mathbf{H}^{(l)}$ denotes the node representations at layer l , with $\mathbf{H}^{(0)} = \mathbf{X}$. Intuitively, each node updates its representation by computing a weighted combination of its own features and the features of its neighboring nodes.

This formulation performs a normalized aggregation of neighboring node features, enabling the model to propagate contextual information across the graph structure. By stacking multiple layers, GCNs capture increasingly larger receptive fields, allowing nodes to incorporate information from multi-hop neighborhoods.

However, the naive GCN architecture generally assumes a single type of edge relation. Knowledge graphs, in contrast, contain multiple relation types between entities. To address this limitation, Relational Graph Convolutional Networks (R-GCN) (SCHLICHTKRULL et al., 2018) extend the GCN framework by introducing relation-specific transformation matrices for each edge type. Because the number of parameters grows linearly with the number of relations, R-GCN employs regularization strategies such as basis decomposition to control model complexity and prevent overfitting (YE et al., 2022).

Subsequent works have proposed additional improvements for modeling relational graphs. For example, SACN integrates graph neural encoders with convolution-based decoders for knowledge graph completion, while CompGCN (VASHISHTH et al., 2020) introduces composition operators that jointly model entities and relations during message passing.

2.2.2.3.2 GraphSAGE

Most of the existing approaches to node embeddings, including the standard GCN architecture are transductive, optimizing embeddings directly for specific nodes within a single, fixed graph, making them unable to naturally generalize to unseen data without retraining. GraphSAGE (SAmple and aggreGatE) (HAMILTON; YING; LESKOVEC, 2017) proposes a framework for inductive node embeddings. Instead of training a distinct embedding vector for each node, it leverages node features, such as text attributes or node degrees to learn a set

of transferable aggregator functions. The core objective is to learn how to aggregate feature information from a node's local neighborhood. Specifically, the model trains K aggregator functions, each corresponding to a different search depth from a given node, alongside a set of weight matrices $\mathbf{W}^k$ used to propagate information between layers. A critical requirement for these aggregator functions is that they must be invariant to permutations of the input (symmetric) to effectively operate on the naturally unordered sets of neighbor vectors.

$$J_G(z_u) = -\log(\sigma(z_u^\top z_v)) - Q \cdot \mathbb{E}_{v_n \sim P_n(v)} \log(\sigma(-z_u^\top z_{v_n})) \quad (11)$$

where v is a node co-occurring near u on a fixed-length random walk, P_n is a negative sampling distribution, Q defines the number of negative samples, and σ is the sigmoid function. This framework allows GraphSAGE to simultaneously learn the topological structure of a node's neighborhood and the distribution of node features, effectively generalizing to zero-shot scenarios where new nodes are added to the graph over time.

2.2.2.3.3 Graph Attention Networks (GAT)

A significant limitation of standard GCNs is that they assign weights based largely on structural properties like node degree, effectively treating all neighbors in the receptive field with fixed or isotropic importance (ASIF et al., 2021). To address this, Veličković et al. (2018) introduced Graph Attention Networks (GATs), adapting the self-attention mechanisms that have become standard in sequence-based tasks to graph-structured data. The core idea is to compute the hidden representations of each node by attending to its neighbors, thereby allowing the model to focus on the most relevant parts of the graph structure. This attention mechanism offers advantages: it is computationally efficient due to parallelization across node-neighbor pairs, it handles nodes of varying degrees by assigning arbitrary weights, and it is directly applicable to inductive learning problems involving unseen graphs (VELIČKOVIĆ et al., 2018).

In a standard GAT layer, the input is a set of node features $\mathbf{h} = \{\vec{h}_1, \dots, \vec{h}_N\}$, where $\vec{h}_i \in \mathbb{R}^F$. To transform these features into higher-level representations, a shared linear transformation, parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$, is applied to every node. A shared attentional mechanism $a : \mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R}$ then computes attention coefficients e_{ij} that indicate the importance of node j 's features to node i :

$$e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j) \quad (12)$$

To inject graph structure, this attention is masked, meaning e_{ij} is only computed for nodes $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the first-order neighborhood of i (including i itself). In the implementation by Veličković et al. (2018), the attention mechanism a is a single-layer feedforward neural network parameterized by a weight vector $\vec{a} \in \mathbb{R}^{2F'}$, followed by a LeakyReLU

nonlinearity (with a negative input slope of $\alpha = 0.2$). The coefficients are then normalized using a softmax function to ensure comparability:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{a}^T[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{a}^T[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_k]\right)\right)} \quad (13)$$

where $\parallel$ denotes the concatenation operation. The final output features for node i , $\vec{h}'_i$, are computed as a linear combination of the features of its neighbors, weighted by the normalized attention coefficients, and passed through a nonlinearity σ :

$$\vec{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \vec{h}_j\right) \quad (14)$$

To stabilize the learning process, GAT employs *multi-head attention*, similar to the transformer architecture proposed by Vaswani et al. (2017). This involves executing K independent attention mechanisms with separate parameters $(\mathbf{W}^k, \vec{a}^k)$. For intermediate layers, the outputs from these K heads are concatenated:

$$\vec{h}'_i = \left\| \sum_{k=1}^K \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right) \right\| \quad (15)$$

However, for the final prediction layer, concatenation is often not sensible due to the potentially large output dimension. Instead, the multi-head outputs are averaged, and the final nonlinearity (such as softmax) is applied subsequently:

$$\vec{h}'_i = \sigma\left(\frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right) \quad (16)$$

2.2.3 Knowledge Enhanced Pre-trained Language Models

Knowledge-Enhanced Pre-trained Language Models (KEPLMs) integrate information at various levels of detail, making them suitable for tasks that require different types of knowledge. These models can be categorized into several types based on the kind of knowledge they incorporate: entity-fused, text-fused, syntax-tree-fused, KG-fused, and rule-fused models. Entity-fused KEPLMs incorporate linguistic and lexical knowledge by embedding entities alongside word representations. This feature supports fine-grained tasks such as named entity recognition and sentiment analysis. Text-fused enhance word embeddings by utilizing external text sources, improving performance in tasks like question answering by extracting essential information from references. Syntax-tree-fused leverage syntactic knowledge to transform or select key components of sentences, which is beneficial for tasks requiring structural awareness, such as syntactic parsing. KG-fused merge knowledge graph representations with language models,

facilitating tasks that involve detailed understanding of entity relationships. This integration enhances both language comprehension and generation tasks. Finally, rule-fused apply formal rules or logic, providing interpretability and accountability, being crucial for reasoning tasks that require transparent inferences (YANG et al., 2023).

The method of knowledge injection also plays a significant role in how PLMs integrate with and perform using external knowledge. Various approaches to knowledge injection include feature-fused, embedding-combined, data-structure-unified, knowledge-supervised, retrieval-based, and rule-guided KEPLMs. Feature-fused project specific features, such as entity spans or sentiment polarity, into embeddings using a trainable matrix. Embedding-combined transform symbolic knowledge into embeddings, aligning these with text tokens. However, this can lead to heterogeneous semantic spaces. Data-structure-unified address semantic incompatibility by converting knowledge graph triplets into sequences. This allows the same encoder to learn embeddings, although it may sacrifice some structural information. Knowledge-supervised reduce extra training costs by utilizing relational triplets as training data, which they integrate directly with input sequences. Retrieval-based consult external knowledge in real-time, selecting relevant information based on the input text, but their application is somewhat limited. Rule-guided combine PLMs for perception with symbolic rules for reasoning, offering enhanced interpretability and reliability through explicit reasoning (YANG et al., 2023).

Regarding the use of external knowledge sources to enhance pre-trained language models, Wei et al. (2021) presents a literature review outlining the main state-of-the-art techniques and practices. Among the works cited, Qin et al. (2020) employs a contrastive learning *framework* to gain a deeper understanding of entities and their relationships. They introduce two new procedures: entity discrimination, to determine which entity can be inferred from another, and relationship discrimination, to assess whether two relationships are semantically similar. In e-commerce, Zhang et al. (2020a) proposes using product association graphs derived from user consumption statistics. Additionally, Wang et al. (2020) proposes a framework that preserves the original network parameters, allowing the injection of multiple types of knowledge without overwriting or erasing the previously encoded textual knowledge.

In a more in-depth analysis of studies focused on enhancing language models through knowledge injection techniques for information retrieval tasks, Zhang et al. (2020b) propose an approach that leverages *graph neural networks* (GNNs) to estimate document relevance, aiming to enhance representations of the relationships between words in a text for document retrieval tasks. Additionally, the authors incorporate all contextual meanings of a term by using a word graph. In Selvalakshmi e Subramaniam (2019), the authors introduce a semantic information retrieval approach that utilizes feature selection and classification techniques to improve the relevance scores of documents. They achieve this by employing a feature selection algorithm based on fuzzy rough sets theory, in combination with a semantic retrieval system built on ontologies and latent Dirichlet allocation. The proposed algorithm not only enhances

the relevance of retrieved documents but also shows efficient handling of large-scale data retrieval.

The authors Lu, Du e Nie (2020) highlight the limitation of attention-based models like BERT in capturing global information about a language's vocabulary despite their strong performance in representing the contextual information of a document or sentence. In contrast, graph convolutional networks excel at representing global vocabulary information. To address this, the authors propose a model combining BERT's strengths with a GCN that incorporates domain-specific vocabulary information. Local and global information interact through the layers of the language model, enabling both types of representations to influence each other, resulting in a more comprehensive representation. In the context of medical applications, Fei et al. (2021) propose a language model with enhanced contextual information by integrating it with a large-scale biomedical knowledge graph. Additionally, a three-stage training procedure is employed to facilitate the injection of knowledge into the model. First, pre-training is conducted on a large-scale corpus (in this case, Wikipedia, PubMed, and PMC were used). This is followed by a post-training phase, where knowledge from biomedical knowledge graphs (SemMedDB and BioGrakn) is injected. Finally, fine-tuning is performed through a supervised learning process focused on information extraction tasks.

Extending the concept of fusing global structural information with contextual representations, Liu et al. (2025) address the challenge of generating coherent and informative topics in unsupervised learning. They propose a neural topic modeling framework based on Variational Autoencoders (VAE) that integrates both contextual and graph information. The model employs a Siamese encoder to fuse semantic representations from SBERT with Bag-of-Words features, enforcing constraints for topic alignment and sharpening to focus on high-quality topics. A graph decoder is introduced to reconstruct a global word relation graph derived from co-occurrences. This graph information fusion strategy adjusts word embeddings based on global relationships, significantly improving the coherence and diversity of the generated topics compared to baselines that rely solely on local context.

Yasunaga et al. (2022) propose a self-supervised pre-training method for language models that leverages text and knowledge graphs to complement textual information. The method involves creating aligned pairs in the format (text, subgraph) from a set of text documents and a large knowledge graph by sampling a text segment, extracting entities, and obtaining a subgraph based on those entities. Using these aligned pairs, the model undergoes pre-training with a cross-modal encoder to capture the relationship between the text and the graph. Two pre-training tasks are employed: masked language modeling (where some tokens in the input text are masked, and the model attempts to predict them) and link prediction (where the model attempts to predict the edges in the input graph). Seeking efficiency in passage retrieval tasks, Liu et al. (2022) propose merging query information with passage embeddings using a graph neural network called GNN-encoder. Starting with a set of queries,

the model retrieves the most relevant passages for each query and constructs a graph where the nodes represent the passages and queries, and the edges capture the correspondences and relationships between them. The GNN is then initialized with representations from pre-trained dual-encoder and cross-encoder models, followed by information propagation throughout the graph.

DrKG (NANGI et al., 2023) is a dense retrieval approach specifically designed for commonsense question answering, leveraging both language models and knowledge graphs to tackle the challenges posed by commonsense QA, which requires integrating unstructured linguistic knowledge from pre-trained language models with structured information from Knowledge Graphs such as ConceptNet. DrKG combines the nuanced contextual understanding of language models with the logical reasoning capabilities of KGs. It builds on successful frameworks like QA-GNN, which merges large language models and graph neural networks for commonsense reasoning. QA-GNN first retrieves a subgraph from ConceptNet through entity linking and then uses a Graph Attention Network to learn joint representations of the QA context and KG nodes. However, DrKG improves upon QA-GNN by addressing key limitations in subgraph extraction. QA-GNN relies on heuristic-based selection of 2-hop paths from question and answer entities, which can limit the relevance and learning capacity of the KG retrieval process. As a result, it often omits relevant nodes beyond the 2-hop network and lacks adaptability over time. DrKG incorporates dense retrieval techniques within the context of KGs. It efficiently learns dense representations that capture complex entity relationships, enhancing both retrieval and QA accuracy. Through extensive experimentation, DrKG explores various training approaches and model designs to optimize the performance of knowledge graph-based dense retrieval in end-to-end commonsense question-answering tasks.

Guo e Huang (2023) proposes to address the challenge of high time complexity in Knowledge Graph Question Answering (KGQA) systems, especially in cases involving multiple entities and relationships, where recent advancements have focused on dense relationship retrieval to improve both efficiency and accuracy. Traditional KGQA systems typically retrieve all possible entities and paths, resulting in significant computational overhead, particularly for complex, multi-hop queries. In response, a Chinese KGQA system utilizing dense relationship retrieval has been proposed to enhance performance. This system leverages the Faiss index mechanism for efficient vector similarity retrieval. By rapidly extracting the top-K relationships from a pre-built vector relationship library, it effectively narrows down the relevant paths. Additionally, by constructing optimal paths and reducing irrelevant semantic paths, the system minimizes the exponential path growth often seen with second-order and higher-order questions. This results in decreased response times and lower resource consumption. Using this method, the KGQA system can respond within one second while maintaining high efficiency and minimal accuracy loss. The integration of pre-trained language models further strengthens the system's capabilities in tasks such as text semantic similarity and entity mention recognition.

KnowledGPT (WANG et al., 2023) introduces a framework designed to enhance large language models by integrating retrieval and storage capabilities with knowledge bases. This approach addresses common challenges in LLMs, such as completeness, timeliness, faithfulness, and adaptability. While recent efforts have explored connecting LLMs to external knowledge sources, integrating KBs has been less explored and often complex. KnowledGPT fills this gap by enabling LLMs to effectively retrieve and store knowledge across various KBs. Through a method called program of thought prompting, KnowledGPT generates search queries in code format to facilitate precise operations within knowledge bases. This tool supports retrieval using pre-defined functions that are tailored to the needs of the LLM. In addition to retrieval the method allows for the storage of knowledge in a personalized KB, adapting to the specific requirements of individual users. Experiments demonstrate that this integration enables to answer a wider variety of knowledge-based questions more effectively than standalone LLMs, leveraging both widely recognized knowledge bases and personalized knowledge to create a more robust and adaptable response system.

Entity Alignment (EA) is essential for knowledge-enhanced dense retrieval, as it identifies equivalent entities across different knowledge graphs to facilitate knowledge fusion. Traditionally, EA methods relied on embedding entity structures based on relation triples. More recent approaches, however, have started to incorporate auxiliary attributes, like names. Despite this progress, the use of separate encoding techniques for structural information and attributes often hindered the mutual enhancement of these features. Wang e Chen (2024) proposes a dense entity retrieval framework for EA that utilizes language models to uniformly encode both structural and attribute information. This approach allows for efficient nearest neighbor searches for cross-KG alignment. The process begins with entity retrieval to generate alignment candidates, followed by a reranking process to refine the results and determine the final alignments.

G-Retriever (HE et al., 2024) introduces a retrieval-augmented generation approach for understanding textual graphs and improving question answering. This system allows users to interact with complex graphs through a conversational interface. Unlike existing methods that focus on basic graph tasks or simplified queries, G-Retriever is tailored for real-world textual graphs, making it suitable for various applications, including scene graph understanding, commonsense reasoning, and knowledge graph reasoning. The framework includes the development of a Graph Question Answering benchmark, which uses data from various tasks to create a robust evaluation setting. G-Retriever employs a soft prompting technique to fine-tune large language models for graph QA, enhancing graph comprehension while mitigating issues like hallucination. To effectively manage large graphs, G-Retriever utilizes Prize-Collecting Steiner Tree optimization to efficiently retrieve relevant parts of the graph, ensuring that pertinent content is prioritized. Evaluations demonstrate that G-Retriever outperforms baseline methods in textual graph tasks across different domains, showcasing its scalability and effectiveness in delivering accurate, context-aware responses.

Most recently, to address the inherent limitations of LLMs in utilizing structural signals for Knowledge Graph Completion, Yang et al. (2025) proposed the SLiNT framework. Traditional LLMs often struggle with structural sparsity and semantic ambiguity, particularly in incomplete or zero-shot scenarios. SLiNT mitigates these issues by injecting structure-aware context into a frozen LLM backbone using a three-module architecture. First, it employs Structure-Guided Neighborhood Enhancement (SGNE) to enrich sparse entities by aggregating pseudo-neighbors from the pre-trained embedding space. Second, Dynamic Hard Contrastive Learning (DHCL) is utilized to enforce fine-grained supervision; this distinguishes structurally valid entities from semantically similar but structurally divergent distractors by interpolating hard positives and negatives. Finally, the framework implements Gradient-Decoupled Dual Injection (GDDI), which performs token-level intervention to fuse these structural signals into the LLM while preserving the core parameters through efficient Low-Rank Adaptation (LoRA). This modular approach allows for robust link prediction by effectively bridging the gap between the generative capabilities of LLMs and the topological precision of KGs.

Similarly employing contrastive learning to overcome structural limitations, Zhukova et al. (2025) explore domain adaptation for the process industry, where data consists largely of irregular text logs and sparse knowledge graphs. Adapting the SciNCL framework, the authors utilize a heterogeneous graph that connects text logs with functional locations (FLs) to capture domain-specific relationships. Rather than relying on direct edges which are often sparse, the method employs graph embeddings to define a continuous similarity space. Neighborhood sampling strategies are then used to select positive and negative document triplets based on geometric proximity in this embedding space. This graph-aware contrastive fine-tuning enables the language model to effectively learn domain-specific terminology and relationships, significantly outperforming general-purpose encoders like mE5-large with fewer parameters.

2.3 Synthesis and Research Gap

Over the past decades, Information Retrieval research has evolved from classical lexical matching models toward increasingly sophisticated neural architectures capable of capturing semantic similarity beyond surface-level term overlap. Dense and late-interaction retrieval models, particularly those based on pre-trained language models, have demonstrated strong performance across a wide range of benchmarks, significantly narrowing the vocabulary mismatch problem that constrained earlier systems. However, the effectiveness of these approaches remains tightly coupled to the availability of large-scale, domain-representative training data and evaluation benchmarks, assumptions that rarely hold in specialized industrial contexts. In high-stakes technical domains such as offshore oil and gas operations, textual data exhibits characteristics that challenge contemporary retrieval models. Operational reports are highly abbreviated, context-dependent, and semantically dense, with critical distinctions encoded through domain-specific terminology and implicit procedural relationships. While neural retrieval models excel

at capturing distributional similarity, they often fail to preserve these fine-grained operational semantics, leading to retrieval outputs that are linguistically related yet operationally irrelevant.

To mitigate these limitations, recent research has explored the integration of structured knowledge through knowledge graphs and ontologies. Knowledge-enhanced approaches have shown promise in representing explicit relationships between entities and enabling more interpretable reasoning processes. Nevertheless, much of this work treats structured knowledge as an auxiliary resource, primarily leveraged during inference or downstream reasoning tasks. As a result, the representational space learned by the underlying language models remains largely shaped by unstructured text alone, limiting the influence of domain-specific structure on semantic alignment. Furthermore, existing Knowledge-Enhanced Pre-trained Language Models often rely on assumptions that are misaligned with industrial reality. These approaches typically presume the availability of dense, well-curated knowledge graphs and extensive manual annotation, conditions that are rarely met in operational environments where knowledge is fragmented, evolving, and costly to formalize. Consequently, the benefits of knowledge integration are constrained, particularly in low-supervision settings where traditional evaluation paradigms struggle to provide reliable performance estimates.

An additional and underexplored challenge lies in the construction of effective training signals for dense retrieval in specialized domains. Contrastive learning frameworks depend critically on the availability of informative negative samples to shape the geometry of the embedding space. In the absence of large-scale relevance judgments, most existing methods rely on heuristics such as random or lexical hard negatives, which inadequately reflect domain-specific semantic boundaries. This limitation directly impacts the model’s ability to distinguish between superficially similar yet operationally distinct documents. Taken together, the literature reveals a critical gap at the intersection of neural retrieval, knowledge representation, and contrastive learning. While prior work has independently advanced dense retrieval models, knowledge graph construction, and domain adaptation strategies, the potential of structured domain knowledge to actively guide representation learning remains insufficiently explored. This gap is especially pronounced in industrial domains characterized by heterogeneous entities, sparse annotations, and stringent requirements for semantic precision.

This dissertation addresses this gap by proposing a knowledge-enhanced dense retrieval framework in which a domain-specific knowledge graph is not merely an auxiliary artifact but a central component of the training process. By leveraging graph-based structural similarity to mine hard negative samples, the proposed approach aims to align the embedding space with operational semantics, enabling robust retrieval performance under low-supervision conditions. In doing so, this work positions itself as a step toward bridging the divide between the expressive power of neural language models and the structural precision required for effective information retrieval in specialized technical domains.

3 Methodology

The present chapter outlines the research design and the specific methods used to address the research questions. Section 3.1 provides a detailed overview of how data was collected, analyzed, and interpreted for the study. It starts by describing the research framework, including the theoretical underpinnings and rationale for the chosen methods. Then, it details the existing publicly available datasets commonly used in information retrieval research and also describes the process of creating a custom dataset focused on the specific objectives of this dissertation, using daily drilling reports.

3.1 Datasets

As previously discussed, language models require vast amounts of data to effectively learn textual representations. Even with the use of pre-trained models, there remains a critical need for domain-specific data to fine-tune these models for particular tasks. Building on the works of Hambarde e Proenca (2023) and Zhao et al. (2022), this section identifies several key datasets that are commonly employed for training and validating information retrieval techniques. These datasets, which are summarized in Table 1, are categorized by their respective domains and include details regarding the size of their document corpus and represent some of the most frequently utilized resources in the field due to their widespread adoption and relevance across various domains.

While the datasets listed above are foundational for general information retrieval research, they do not encapsulate the unique linguistic characteristics and specialized information needs of the Oil and Gas industry. Daily Drilling Reports are crucial documents in oil drilling operations, providing comprehensive records of significant occurrences, such as equipment issues and operational hurdles. These reports present a chronological account of the drilling process, capturing essential data needed to analyze patterns and plan maintenance. Despite their importance, DDRs are often unstructured, making it complex to extract relevant information (CORDEIRO et al., 2024). To illustrate the technical complexity and specialized syntax of these narratives, we present a representative example sourced from the work of Cinelli et al. (2021). This surrogate example mirrors the linguistic structure of actual operational logs without compromising proprietary information:

Dataset	Domain	Corpus Size
MS MARCO (NGUYEN et al., 2016)	Real queries	8.84M
BioASQ (TSATSARONIS et al., 2015)	Question answering	14.71M
Natural Questions (KWIATKOWSKI et al., 2019)	Question answering	2.68M
QASC (KHOT et al., 2019)	Question answering	17M
TREC Robust (VOORHEES, 2005)	Scientific	69.9M
Conversational Argument (RUGGERI; MESGAR; GUREVYCH, 2022)	Scientific	382K
CQADupStack (HOOGEVEEN; VERSPOOR; BALDWIN, 2015)	Question answering	457K
DBPedia (LEHMANN et al., 2015)	Entities	4.63M
Climate-FEVER (DIGGELMANN et al., 2020)	Climate changes	5.42M
NFCorpus (BOTEVA et al., 2016)	Bio-Medical	9K
REGIS (OLIVEIRA; ROMEU; MOREIRA, 2021)	Geoscience	21K
Quati (BUENO et al., 2024)	Websites	10M

Table 1 – SOTA Datasets for Information Retrieval with domain and corpus size

Daily Drilling Report Example

Open Well Pulling Drill String out of the Hole

Pulling drill string out of the hole DP 5 1/2" TT550 with 28" BHA #2 from 2411

- 2143 m in an open well without drag, except at:

- 2323-2310 m with drag of 30-50 klbf.

- 2284-2252 m with drag of 30-50 klbf.

Note: - Total CADIT 9.9 ppg used for pulling drill string out of the hole, across the open well, equal to 1100 bbl.

The selection of Daily Drilling Reports as the primary textual source is directly motivated by the central hypothesis of this research. DDRs exhibit extreme domain specificity, abbreviated syntax, and implicit operational semantics, making them an ideal stress test for evaluating whether knowledge-enhanced representation learning can overcome the limitations of purely distributional models.

Therefore, as part of a broader collaborative research and development initiative between academia and industry, a central objective was the creation of a novel dataset specifically

tailored to the oil extraction domain. This process was conducted within a comprehensive framework for technological advancement in the sector and was not limited to the specific scope of this dissertation. The data for this study was sourced from an operational database containing historical DDRs maintained by a major national oil and gas company. The selection of the 181 wells used to construct the test dataset followed explicit operational and temporal criteria rather than random sampling. These wells correspond to assets identified as operationally critical by domain managers and were actively monitored through an internal business platform, ensuring that the evaluation set reflects information needs of direct practical relevance. In addition, the selected wells correspond to relatively recent drilling operations, as older wells often exhibit outdated reporting practices and linguistic patterns that no longer align with current operational procedures. Given the extensive volume of available data, which spans numerous wells and a long historical period, it was not feasible or methodologically desirable to include the entire database in the test collection. Instead, the focus on a curated subset ensured both domain relevance and linguistic consistency with contemporary operational scenarios.

3.1.1 Test Dataset Creation

The creation of test collections for information retrieval is a difficult and expensive task. The common used method was initially proposed by (SPARCK-JONES; RIJSHERGEN, 1975), who introduces the concept of pooling to manage the otherwise infeasible task of having experts judging every document in a large corpus. Since manually evaluating the relevance of a query across the entire corpus would be impractical, it was crucial to implement this step to significantly reduce the number of passages that human annotators would need to assess. The goal was to narrow down the candidate set to a manageable number of passages per query, ensuring that human judgments could be made efficiently while maintaining the integrity of the evaluation process. This methodology, was institutionalized and popularized by the Text REtrieval Conference (TREC). The TREC paradigm, leveraging community participation to build pools for annotation, became the gold standard for IR evaluation and profoundly influenced the field (SANDERSON, 2010).

However, the success of the TREC model relies on a large, active research community and access to expert annotators. These prerequisites are often absent in specialized, high-stakes industrial domains, where the complexity of the data is high and the pool of available experts is small. Furthermore, the effectiveness of the pooling method diminishes in such contexts. The retrieval systems used to create the pool often struggle with niche terminology and complex concepts, leading to low-quality pools with a high proportion of irrelevant documents, which in turn makes the annotation process inefficient. With a reduced annotators numbers, we often struggle on creating a dataset of sufficient size, resulting in a difficult evaluation scenario where the metrics obtained for the models used in the pooling creation performs better than other models, not because they are better but because they have a 100% coverage on the

top positions, meaning that other models could retrieve relevant documents that were not assessed and are false negatives.

To address these challenges, we designed a hybrid methodology for creating our test collection, as illustrated in Figure 4. This process begins with manual expert annotation to create a high-quality Golden Standard set and then leverages this set to train a relevance classifier, enabling the cost-effective expansion to a larger Silver Standard evaluation set. The following sections describe each step of this pipeline.

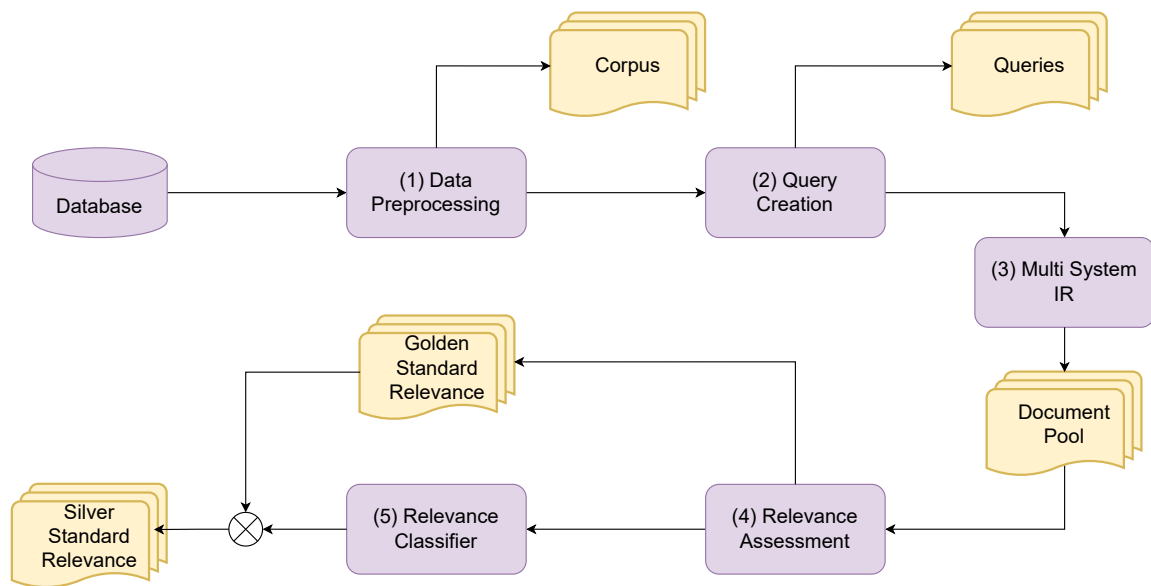


Figure 4 – The test dataset creation pipeline, moving from raw data to both a human-adjudicated Golden Standard and a machine-annotated Silver Standard relevance set.

(1) Data Preprocessing: The process begins with sourcing data from an internal operational database. A subset of 181 wells of major interest was extracted, resulting a total of 108,292 entries from 181 wells. Each entry contains 71 fields, with most fields consisting of metadata like well identification numbers, dates, operational parameters, and equipment details. The chosen DDRs span cover from March 1, 2021, to June 30, 2024, presenting a current and pertinent overview of drilling activities in the Brazilian oil and gas industry. This timeframe ensures that the dataset includes a range of operational situations, such as various drilling stages, equipment utilization patterns, and potential operational hurdles faced during the drilling operations. A detailed linguistic analysis of this corpus revealed a total of 5,737,015 tokens. The average document contains around 53 tokens, with a vocabulary of 36,430 unique tokens. This demonstrates a diverse and technical linguistic range, ideal for developing and evaluating robust, domain-specific information retrieval systems.

(2) Query Creation: The dataset queries were created through a structured process to mirror real-world well drilling scenarios. Initially, the queries were based on an Excel sheet regularly used by drilling supervisors and fiscals to record and prioritize potential risks during

drilling operations. This document provided a comprehensive set of ranked risks, forming the basis for query generation. To ensure the queries were specific to the domain and relevant to retrieval tasks, a petroleum engineering specialist was consulted. The specialist curated and refined the list, translating the ranked risks into meaningful, task-oriented queries for use in information retrieval experiments. This process ensured that the queries were aligned with the practical needs and terminology of the drilling industry, thus enhancing the relevance and applicability of the retrieval task and resulted in an total of 47 queries.

(3) Multi-System IR and Document Pooling: In order to create a pool of potential passages, for each query we used a multi-system approach by retrieving the top 10 passages from a diverse set of IR systems for each query. This pooling strategy takes advantage of the diversity in retrieval results from different IR methods, allowing for the inclusion of a wider range of potentially relevant passages. By leveraging multiple retrieval models, the pool of candidates becomes more varied, which enhances the likelihood that the final dataset will comprehensively capture the relevant information. The diversity of IR systems ensures that different aspects of relevance are considered, thereby creating a more robust evaluation dataset for the task, improving the quality of the dataset and making the annotation task more manageable and focused. The annotators evaluated the passages retrieved by the following models:

Category	Model Name / Configuration
Lexical & Sparse	BM25, DFR
Hybrid & Re-ranking	BM25 + mt5-xl, BM25 + e5-large
Domain0Specific	Petrovec-hibrido-300, Petrobev-o&g-300
Dense Retriever	e5-large, e5-base, mcontriever-base-msmarco, msmarco-distilbert-dot-v5, paraphrase-multilingual-mpnet-base-v2, bge-multilingual-gemma2
Late Interaction	colbertv2.0, jina-colbertv2.0
Commercial API	text-embedding-ada-002, text-embedding-3-small, text-embedding-3-large

Table 2 – Retrieval models used for document pooling and candidate generation.

(4) Relevance Assessment: We carefully assessed the relevance of the candidate passages retrieved earlier. The domain experts, who have specialized knowledge in well drilling, performed this task to ensure that the judgments were based on practical expertise. To ensure accuracy, each query-document pair was evaluated by at least two annotators. To maintain the integrity of the evaluation process, the documents were presented in order of doc-id rather than by their retrieval ranking. This eliminated any potential ranking bias, allowing the annotators to make relevance judgments solely based on the content of the documents. Each document was assessed on a four-point scale: not relevant, marginally relevant, fairly relevant, and very relevant, providing a nuanced understanding of its utility in relation to the query. This fine-grained relevance labeling ensured that the dataset would capture varying degrees of

relevance, enabling more sophisticated evaluation of retrieval models in future research. This set is named the Golden Standard Relevance.

(5) Relevance Classifier: While the Golden Standard is highly accurate, its size is limited by the intensive manual labor required. To scale up our evaluation resources, we utilized the Golden Standard to train a Relevance Classifier capable of automatically annotating a much larger set of documents, thereby creating a Silver Standard Relevance set. This human-in-the-loop approach empowers experts by using their initial judgments to annotate data at scale. To determine the optimal automated annotation strategy, we conducted a rigorous comparative analysis between traditional supervised machine learning models and state-of-the-art Large Language Models acting as zero-shot annotators.

- **Problem Formulation:** Due to class imbalance in the four-point scale, the labels were binarized for the classification task. Documents judged as marginally, fairly, or very relevant were consolidated into a single Relevant class, with the remainder labeled Not Relevant.
- **Supervised Classifiers:** We evaluated a suite of traditional machine learning models (e.g., Logistic Regression, SVM, Random Forest, Gradient Boosting) trained on semantic representations of query-document pairs. The experimental design tested various components, including different embedding models, query-document interaction methods, and training strategies.
- **LLMs as Zero-Shot Annotators:** As an alternative, we evaluated the zero-shot annotation capabilities of large language models (GPT-4o and GPT-4o-mini). We tested both a generic, standardized prompt and a specialist prompt enriched with domain-specific knowledge to gauge their performance against the supervised classifiers.

As shown in Table 3, the empirical results present a multifaceted view of annotation performance, where superiority depends on the specific metric of interest. The Neural Network classifier (NN+BGE+Hada) achieved the highest F1-score of **0.860**. While this figure technically surpasses the performance of the second human expert ($F1=0.793$), this result must be interpreted with caution. The classifier was explicitly trained to replicate the decision boundaries of the primary annotator, effectively learning to mimic their specific biases and labeling style. Therefore, its higher F1-score reflects a successful alignment with the training source rather than an intrinsic superiority over human judgment.

Based on the results and analysis, the classifier was deployed to annotate the complete corpus of 108,292 document for all the queries. This large-scale inference process yielded the **Silver Standard** relevance set. This resource serves the fundamental function of robustly evaluate the models explored in subsequent chapters, addressing the inherent recall limitations of the initial pooling phase. In traditional pooling, any document not retrieved by the initial

set of systems is typically unjudged and assumed irrelevant by default. This penalizes novel systems that might retrieve relevant documents missed by the pool contributors. By classifying the entire corpus, the Silver Standard identifies these hidden relevant documents that were left out of the pooling stage, ensuring a fairer and more comprehensive evaluation landscape where models are recognized for retrieving relevant content regardless of whether it appeared in the initial top-10 selection

Method	Precision	Recall	F1-Score
Human Annotator 2	0.867	0.730	0.793
GB + CQ + BGE + Hada	0.865 ± 0.028	0.696 ± 0.010	0.771 ± 0.007
NN + CQ + BGE + Hada	0.860 ± 0.017	0.860 ± 0.020	0.860 ± 0.018
SVM-RBF + CQ + BGE + Diff	0.850 ± 0.019	0.861 ± 0.012	0.856 ± 0.013
SVM-RBF + CQ + BGE + Hada	0.849 ± 0.018	0.865 ± 0.015	0.857 ± 0.014
NN + PQ + BGE + Diff	0.718 ± 0.212	0.890 ± 0.084	0.772 ± 0.161
NN + PQ + BGE + Hada	0.731 ± 0.214	0.877 ± 0.113	0.775 ± 0.174
NN + PQ + BGE + DocOnly	0.730 ± 0.209	0.876 ± 0.103	0.776 ± 0.169
GPT-4o (Baseline)	0.413	0.915	0.569
GPT-4o (Specialist)	0.522	0.940	0.671
GPT-4o-mini (Baseline)	0.494	0.877	0.632
GPT-4o-mini (Specialist)	0.407	0.983	0.575

Table 3 – Performance comparison of top three models for each metric. Model names are abbreviated as [Classifier] + [Training Strategy] + [Embedding] + [Feature Combination]. Values are mean $\pm$ standard deviation when available.

A more granular analysis reveals that the second Human Annotator achieved the highest Precision (0.867). This indicates that human experts possess superior discriminative power, effectively filtering out noise and marginally related documents that automated systems might misinterpret as useful. Conversely, the Large Language Models demonstrated the highest Recall (reaching 0.983). However, this metric is inflated by the models tendency to over-predict relevance. The LLMs adopted a highly lenient threshold, classifying a vast number of documents as positive. While this strategy ensures that nearly all relevant documents are captured, it comes at the cost of generating a noisy dataset saturated with false positives, as evidenced by their correspondingly low precision scores.

To validate the reliability of the generated labels, we computed pairwise Cohen’s Kappa (κ) coefficients between all annotation methods, shown in Figure 5. The agreement between the two human experts was substantial ($\kappa = 0.695$), establishing a strong upper-bound benchmark that confirms the feasibility of the annotation task despite the domain’s complexity. The best supervised classifier demonstrated remarkable consistency, achieving high agreement not only with its training source ($\kappa = 0.785$) but also with the held-out second expert ($\kappa = 0.669$). This indicates that the classifier successfully captured a generalizable definition of relevance rather than merely overfitting to a single annotator’s. In contrast, the agreement between humans

and zero-shot LLMs was consistently low, ranging from slight to fair ($\kappa \in [0.13, 0.40]$). This divergence suggests a fundamental mismatch between the LLMs general-purpose interpretation of relevance, and the utility-focused criteria applied by domain experts. Furthermore, the moderate agreement observed between different LLM configurations indicates a lack of internal consensus, reinforcing the conclusion that zero-shot LLMs lack the reliability required to serve as surrogate annotators in this specialized high-stakes domain.

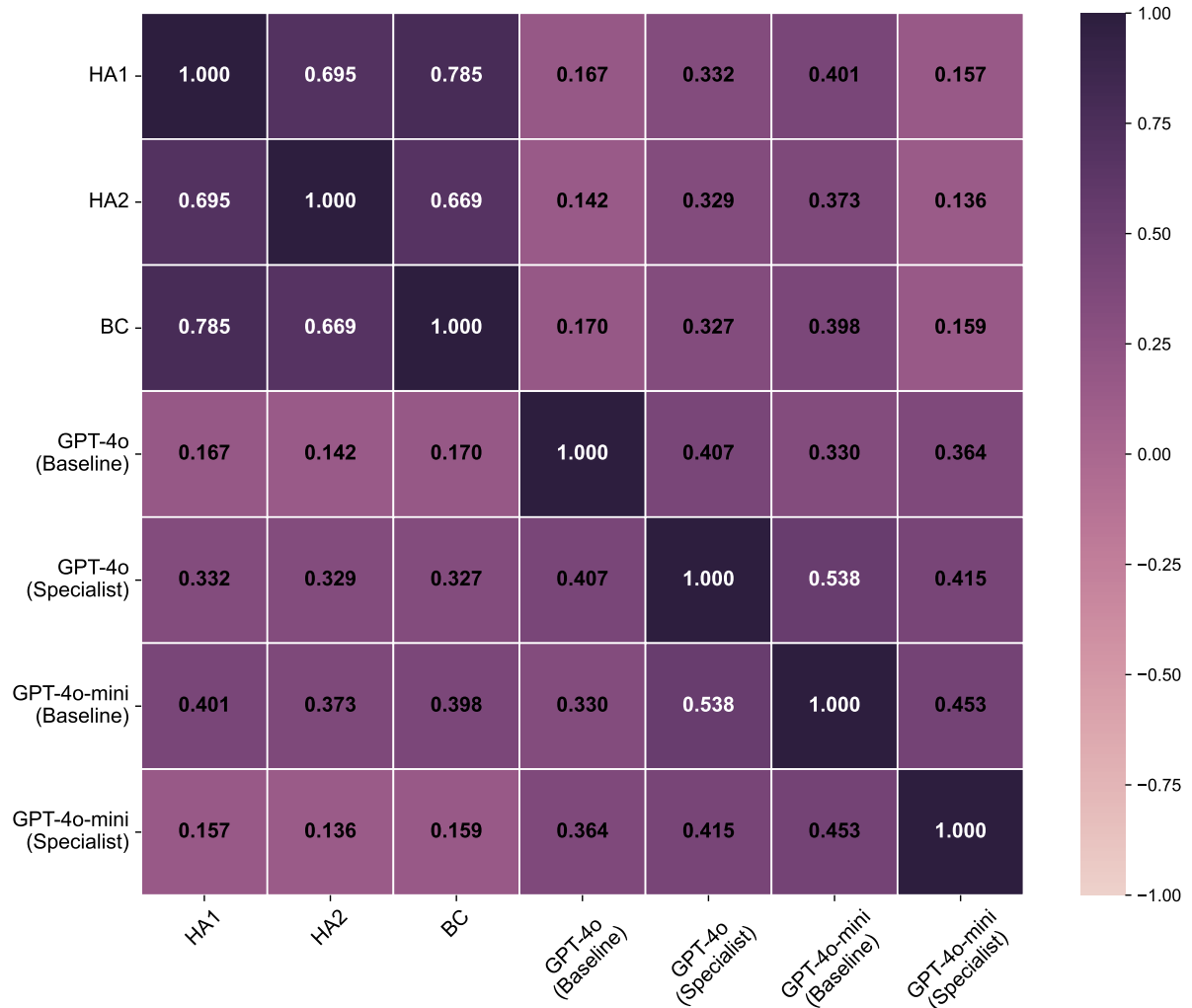


Figure 5 – Heatmap of Pairwise Cohen's Kappa (κ) Agreement.

In summary, this hybrid methodology bridges the gap between expert precision and large-scale evaluation by combining a human-curated Golden Standard with a high-fidelity automated Silver Standard. By addressing the recall limitations inherent in traditional pooling and the scarcity of domain experts, this dual-layered dataset provides a robust, domain-specific framework for benchmarking information retrieval systems within the complex context of the Oil and Gas industry.

3.1.2 Training Dataset Creation

Effective fine-tuning of bi-encoder architectures necessitates a volume of data that exceeds the capacity of manual expert annotation. To meet this requirement, we constructed a large-scale training corpus from the same operational database and temporal window (March 2021 to June 2024) used for the test set. The training set was defined by excluding the 181 wells selected for evaluation, while retaining all remaining wells that satisfied the same temporal constraints. This strategy maximizes the amount of available training data while ensuring that the linguistic style, operational context, and reporting practices remain consistent across training and evaluation. Crucially, this strict separation prevents any overlap of well-specific information between datasets, thereby eliminating data leakage and preserving the validity of the experimental results.

The resulting training corpus provides a significantly broader linguistic and operational scope than the test set. It comprises 672,775 documents sourced from 472 unique wells, totaling over 22 million words and a cleaned vocabulary of 90,948 unique tokens. This scale allows the model to capture a wide array of technical nuances and operational patterns inherent in Daily Drilling Reports.

Given the infeasibility of manual labeling at this scale, we utilized a semi-supervised approach to generate training instances in the form of (*query*, *positive_passage*, *negative_passage*) triplets. While this subsection establishes the data foundation, the specific focus of this work regarding the methodology for triplet construction and the subsequent Knowledge Graph generation process will be detailed in the following sections as part of the proposed retrieval framework.

3.2 Knowledge Graph Construction

This section details a hybrid methodology for constructing a domain-specific Knowledge Graph from heterogeneous wellbore data, integrating both structured operational records and unstructured information extracted from Daily Drilling Reports. The process follows sequential phases to transform raw data into a semantically rich graph structure, underpinned by a specialized ontology designed for offshore oil well intervention reports.

The design of the Knowledge Graph is explicitly guided by the requirements of downstream dense retrieval rather than symbolic reasoning alone. In particular, the graph is constructed to encode co-occurrence, participation, and contextual proximity between operational entities in a form that can be exploited by representation learning models. By emphasizing relational consistency over logical completeness, the resulting structure provides a controllable inductive bias for shaping embedding geometry, enabling the identification of semantically proximate yet distinct documents for structure-aware negative sampling.

3.2.1 Ontology Engineering and Design

The semantic backbone of the proposed system is the RIPPO ontology (Relatórios de Intervenção em Poços de Petróleo Offshore), which was developed to formalize the concepts found in operational descriptions (QUEIROZ et al., 2025). Unlike traditional waterfall approaches, the development of the ontology followed an agile methodology, with an iterative approach that allows for the incremental refinement of classes and relationships through a series of sprints, ensuring that the resulting schema met specific competency questions relevant to decision support, such as identifying which components were involved in specific incidents or calculating operational efficiency metrics.

The ontology organizes domain knowledge into a hierarchy comprising over 3,000 classes. These are categorized into core superclasses that map directly to the entities observed in drilling reports: *Actions* (verbs denoting operations), *Procedures* (noun-based operational tasks), *Components* (equipment and tools), *Materials* (fluids and substances), *Locations* (spatial references on the rig or well), *Organizations* (companies and teams), *Measures* (numerical values with units), *Properties* (qualities or parameters), *States* (conditions of entities), and *Incidents* (unplanned events). This structure was designed not only for data organization but also to support complex queries regarding operational sequences and equipment performance.

Complementing the class hierarchy, the ontology defines a set of semantic relationships that govern the interactions between entities. These were designed to mirror the syntactic and narrative structure of the DDRs, where operational descriptions typically link an action to specific components, materials, or parameters. Key relationships include *subActionOf* and *followedBy*, which capture the hierarchical and temporal nature of operational sequences defined in the text. The core operational logic is modeled through the *involves* relationship, which connects an *Action* to the diverse entities participating in it, such as *Components*, *Materials*, or *Procedures*. Furthermore, attribute linking is handled by specific properties like *hasProperty*, *hasMeasure*, *hasState*, and *hasUnit*, which associate entities with their qualitative and quantitative descriptors. The *reports* relationship serves as the critical bridge between the unstructured text and the structured database, linking extracted semantic actions to the main node, thereby grounding the extracted events within the specific temporal and metadata context of the daily logs.

To ensure interoperability and semantic rigor, the ontology was aligned with standard upper-level ontologies. Specifically, core concepts were mapped to the Basic Formal Ontology (BFO) and the Industrial Ontologies Foundry (IOF) specifications. For instance, *Actions* were aligned with *bfo:process*, *Components* with *iof:PieceOfEquipment*, and *Incidents* with *iof:Event*. Furthermore, domain-specific terminology was cross-referenced with the ISO 15926-4 Reference Data Library to maintain consistency with industrial standards.

3.2.2 Entity Extraction and Knowledge Injection

The process of populating the Knowledge Graph involves a transformation pipeline that converts raw text into semantic triples. The process begins by loading a Natural Language Processing (NLP) pipeline based on Spacy (specifically the *pt_core_news_lg* model). However, to address the high precision requirements of the engineering domain and the specific nomenclature of drilling reports, the system employs a hybrid Named Entity Recognition (NER) strategy. Instead of relying solely on statistical models, the system utilizes rigorous lexical patterns defined as Regular Expressions (Regex) within the ontology's annotation properties. This approach mitigates the precision issues often associated with purely stochastic models when dealing with technical jargon and acronyms unique to the oil and gas industry (QUEIROZ et al., 2025).

For each well in the dataset, a scraping process extracts semantic triples comprising a subject, predicate, and object. The extraction logic is governed by a set of heuristic rules derived from the syntactic structure of the reports. These heuristics analyze the sequence of identified entities to infer relationships; for example, if a *Component* entity follows an *Action* entity in the text, a relationship such as *<Action> involves <Component>* is generated. Similarly, if a numerical *Measure* follows a *Component*, it is linked via a *hasMeasure* relationship. These extracted triples are then instantiated as nodes and edges within the Neo4j graph database.

This heuristic extraction strategy prioritizes precision over recall, reflecting the high cost of spurious relations in the downstream graph structure. While this approach reduces noise introduced by statistical NER models in a highly technical domain, it may fail to capture entities expressed through uncommon abbreviations or syntactic variants. Extraction errors can propagate to the graph by introducing missing or incomplete relational patterns, which may affect the quality of structure-aware negative sampling. However, because the graph is treated as a weak supervisory signal rather than a source of ground truth, the framework is designed to tolerate such imperfections without assuming exhaustive correctness.

In addition to textual descriptions, structured data is integrated into the graph via explicit mapping rules. Key attributes such as well identifiers, timestamps, and operational codes are mapped to specific nodes using a "data-driven" approach. To facilitate integration, specific classes such as *TempoOperacao* were introduced to act as hubs connecting unstructured events extracted from text with the structured metadata of the daily logs. This dual representation allows the graph to support queries that require both semantic understanding of events (e.g., "Find all incidents involving a specific drill bit") and temporal precision (e.g., "within the last fiscal quarter").

The final stage involves injecting the constructed graph into the Neo4j knowledge base. Each injection is logged to facilitate performance analysis. Throughout the process, the performance of each step is monitored to optimize the bootstrapping of large datasets. The

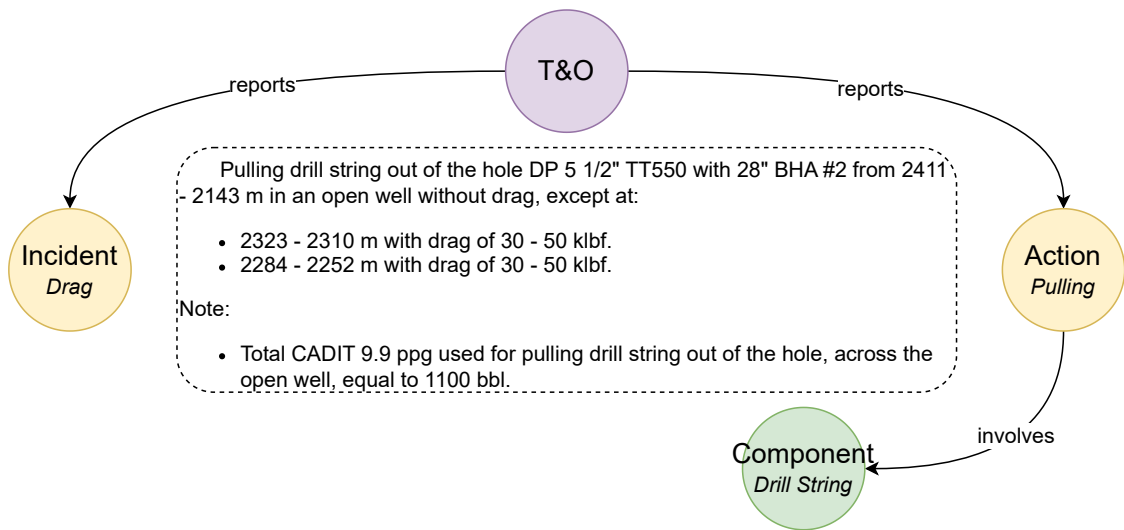


Figure 6 – Conceptual visualization of the Knowledge Graph structure. The diagram illustrates how specialized ontology classes are instantiated as nodes and interconnected through semantic relationships extracted from DDR narratives.

resulting Knowledge Graph serves as a foundational resource for downstream applications, enabling advanced reasoning tasks, calculating drilling metrics, and supporting the retrieval of complex operational information that would be difficult to retrieve reliably using keyword-based methods.

While the quality of the knowledge graph directly influences the effectiveness of structure-aware learning, this work does not assume completeness or perfect correctness of the ontology. Instead, the graph is treated as a weak but structured signal, whose utility lies in its relative relational consistency rather than exhaustive coverage. This design choice reflects the realities of industrial knowledge engineering, where partial and evolving representations are the norm.

3.3 Overall Solution

This dissertation proposes a Knowledge-Enhanced Dense Retrieval Framework designed to overcome the limitations of standard bi-encoders in domain-specific environments, such as the Oil and Gas industry. While traditional approaches rely solely on the semantic signals present in unstructured text, our solution integrates structured domain knowledge to construct a richer context for retrieval.

The core innovation of this framework lies in its training methodology. Rather than relying on external retrieval systems or Large Language Models to generate synthetic training data, as proposed in methods like InPars or UDAPDR, we leverage the intrinsic topological structure of the data itself. We formulate the problem as a learning task on a Heterogeneous Information Network (HIN), where textual reports, domain entities, and metadata are interconnected. By applying Graph Neural Networks under unsupervised paradigms, we capture

high-order structural dependencies and global semantic information.

A key limitation of purely text-driven synthetic query generation is its reliance on surface-level semantic plausibility rather than operational constraints. In preliminary experiments, synthetic queries generated without structural grounding frequently emphasized fluent descriptions while overlooking relational distinctions critical to operational relevance, such as differences in equipment state or procedural context. In contrast, topology-derived supervision constrains learning through explicit relational proximity, enabling the model to distinguish between documents that are lexically similar but operationally distinct. This structural bias provides a more discriminative training signal in environments where terminology reuse is pervasive.

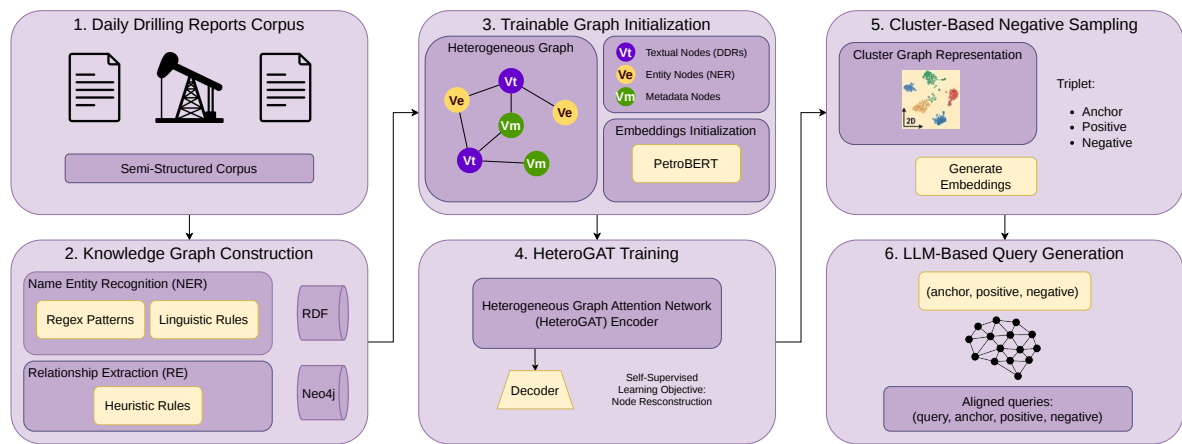


Figure 7 – Architectural overview of the proposed Knowledge-Enhanced Dense Retrieval Framework, highlighting the integration of unsupervised graph representation learning and structure-aware negative sampling.

Figure 7 summarizes the end-to-end workflow of the proposed framework. The process begins with the Daily Drilling Reports corpus, which constitutes the primary unstructured textual source. In the first stage, domain knowledge is extracted through Named Entity Recognition (NER) and rule-based relation extraction, producing a semi-structured representation that links operational descriptions with domain entities and metadata. These elements are then integrated into a **heterogeneous knowledge graph** composed of three node types: textual reports, extracted entities, and metadata attributes.

In the third stage, all graph nodes are initialized using contextual embeddings produced by the domain-adapted language model PetroBERT, allowing heterogeneous node types to be projected into a shared semantic space. The resulting node features are used to train a **Heterogeneous Graph Attention Network** under a self-supervised learning objective that reconstructs graph connectivity. This step captures higher-order structural dependencies between reports, entities, and operational metadata, producing topology-aware embeddings for textual nodes.

In the fifth stage, the learned graph representations are used to derive **structure-aware**

hard negatives. The textual node embeddings are clustered using density-based methods, grouping reports that share similar structural contexts. For each anchor document, semantically similar documents from the same cluster are sampled as hard negatives, forcing the downstream retriever to distinguish between highly similar operational events.

Finally, in the last stage, a large language model generates **synthetic queries** aligned with the anchor documents. These queries are paired with the anchor, positive, and negative documents to form contrastive training tuples used to train the final dense retriever. By combining graph-derived structural supervision with linguistically coherent query signals, the proposed pipeline enables effective representation learning without requiring manually annotated relevance labels.

3.3.1 Problem Formulation

To formally capture the dependencies between unstructured text and structured domain knowledge, we model the dataset as a Heterogeneous Graph $G = (V, E, \mathcal{T}_V, \mathcal{T}_E)$. Here, V denotes the set of nodes and E the set of edges, with node type mapping function $\phi : V \rightarrow \mathcal{T}_V$ and edge type mapping function $\psi : E \rightarrow \mathcal{T}_E$, satisfying the heterogeneity constraint $|\mathcal{T}_V| + |\mathcal{T}_E| > 2$.

The semantic interpretation of the nodes and relations used in this formulation follows the ontology-driven Knowledge Graph construction process described in Section 3.2. In particular, entity and relation types correspond to concepts and predicates defined in the RIPPO ontology, which provides the domain-specific schema for representing operational events extracted from the Daily Drilling Reports. In this work, the set of node types $\mathcal{T}_V$ consists of three disjoint sets:

- **Textual Nodes (V_T):** Representing the unstructured description of the daily operations.
- **Entity Nodes (V_E):** Representing domain-specific entities (e.g., equipment, formations) identified via Named Entity Recognition (NER) and linked to the RIPPO ontology.
- **Metadata Nodes (V_M):** Representing categorical attributes such as well IDs or operational phases.

These node categories correspond to abstractions derived from the RIPPO ontology described in Section 3.2. Textual nodes represent the primary operational records, while entity and metadata nodes correspond to ontology concepts and contextual attributes extracted from the reports.

The objective is to learn an embedding function

$$f_\theta : V \rightarrow \mathbb{R}^d$$

parameterized by θ , which maps nodes in the heterogeneous graph to a low-dimensional representation space. In particular, the embeddings of textual nodes V_T are used to represent drilling reports in the downstream dense retrieval task.

In addition to the node and edge sets, each node $v \in V$ is associated with an initial feature vector $\mathbf{x}_v \in \mathbb{R}^{d_0}$. Collectively, these features form the node feature matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d_0}$. The relational structure of the graph is represented by a set of adjacency matrices $\{\mathbf{A}_r\}_{r \in \mathcal{T}_E}$, where each matrix corresponds to a specific relation type r .

In the proposed knowledge graph, edges capture semantic and contextual relationships between nodes. Specifically, we define three primary relation categories:

- **Text–Entity edges (E_{TE}):** connecting drilling reports to extracted domain entities.
- **Text–Metadata edges (E_{TM}):** linking reports to operational metadata such as well identifiers or drilling phases.
- **Entity–Entity edges (E_{EE}):** capturing semantic relations between domain entities derived from ontology links or co-occurrence patterns.

For a node v , the neighborhood under relation type r is defined as:

$$\mathcal{N}_v^r = \{u \in V \mid (v, u) \in E \text{ and } \psi((v, u)) = r\} \quad (17)$$

This relational structure enables the graph neural network to distinguish interactions between different semantic components of the drilling reports.

3.3.2 Graph-Based Representation Learning

To directly address the limitations identified in Chapter 2, namely the absence of structural signals in contrastive learning for dense retrieval, we reformulate the training process to incorporate graph-based structural similarity as an explicit supervisory signal.

3.3.2.1 Feature Initialization

To ensure a cohesive semantic representation across the heterogeneous graph, we employ a unified initialization strategy using *PetroBERT* (RODRIGUES et al., 2022), a domain-specific language model pre-trained on domain. We leverage the transformer’s capability to process all node inputs as textual sequences. This decision is motivated by the specific characteristics of the domain ontology, which exhibits a relatively flat hierarchy. By projecting Textual, Entity, and Metadata nodes through the same encoder, they naturally occupy a shared vector space $\mathbb{R}^d$, thereby substantially reducing cross-modal alignment complexity and the need for complex linear projection layers to harmonize varying feature dimensions.

The initialization process treats every node as a sequence of tokens. For Textual Nodes $\mathbf{h}_T^{(0)}$, the input corresponds to the full unstructured description of the operation. For Entity Nodes $\mathbf{h}_E^{(0)}$ and Metadata Nodes $\mathbf{h}_M^{(0)}$, we utilize their canonical labels or attribute values as the input strings. This allows the model to leverage the rich semantic knowledge embedded in the model, capturing the linguistic relationships between technical terms and the operational context more effectively than structural embeddings could in a sparse-hierarchy environment. To generate the fixed-size feature vectors from the variable-length token sequences, we deviate from the standard practice of utilizing the [CLS] token, which is often under-optimized in domain-adapted models that lack next sentence prediction fine-tuning. Instead, we perform mean pooling over the token embeddings of the last hidden layer. Formally, given a node input sequence of N tokens with final layer representations $\mathbf{e}_1, \dots, \mathbf{e}_N$, the initial node feature $\mathbf{h}^{(0)}$ is computed as $\frac{1}{N} \sum_{i=1}^N \mathbf{e}_i$. This results in a dense vector that serves as the foundation for the subsequent message-passing phases.

3.3.2.2 Heterogeneous Graph Encoder

To capture the complex interactions between these diverse node types, we employ a Heterogeneous Graph Attention Network (HeteroGAT) as the backbone encoder. Graph Attention Networks were selected due to their ability to assign adaptive importance weights to heterogeneous neighbors, a critical requirement in operational knowledge graphs where different entity types and relations contribute unequally to semantic similarity. Unlike standard GATs which treat all neighbors isotropically, HeteroGAT employs type-specific attention mechanisms. For a target textual node $t \in V_T$, the aggregation process at layer l distinguishes neighbors by type $\Phi \in \{E, M\}$:

$$\mathbf{h}_t^{(l+1)} = \sigma \left(\sum_{\Phi \in \mathcal{T}_V} \sum_{j \in \mathcal{N}_t^\Phi} \alpha_{tj}^\Phi \cdot \mathbf{W}_\Phi^{(l)} \mathbf{h}_j^{(l)} \right) \quad (18)$$

Where $\mathbf{W}_\Phi^{(l)}$ is a learnable weight matrix specific to node type Φ , allowing the model to learn distinct projection spaces for Entities and Metadata. The attention coefficient α_{tj}^Φ determines the importance of neighbor j to node t , computed via a Softmax over the type-specific neighborhood. This mechanism enables textual nodes to selectively attend to the specific entities or metadata values that are most discriminative for their representation.

The attention coefficients α_{tj}^Φ quantify the relative importance of neighbor j when updating the representation of node t . They are computed using a type-specific attention mechanism:

$$e_{tj}^\Phi = \text{LeakyReLU} \left(\mathbf{a}_\Phi^\top [\mathbf{W}_\Phi^{(l)} \mathbf{h}_t^{(l)} \parallel \mathbf{W}_\Phi^{(l)} \mathbf{h}_j^{(l)}] \right) \quad (19)$$

where $\mathbf{a}_\Phi$ is a learnable attention vector associated with node type Φ , and $\parallel$ denotes vector concatenation.

The coefficients are normalized using a softmax function over the type-specific neighborhood:

$$\alpha_{tj}^\Phi = \frac{\exp(e_{tj}^\Phi)}{\sum_{k \in \mathcal{N}_t^\Phi} \exp(e_{tk}^\Phi)} \quad (20)$$

This mechanism allows the model to selectively emphasize the most informative neighboring entities and metadata attributes when constructing textual node representations.

3.3.3 Unsupervised Optimization Strategy

Since the dataset lacks explicit relevance labels for retrieval training, we employ a self-supervised learning objective to pre-train the graph encoder. This strategy focuses on learning node representations by maximizing the reconstruction capability of the graph structure.

To achieve this, we utilize a Heterogeneous Graph Autoencoder (HeteroGAE) framework, where the HeteroGAT described previously serves as the encoder $\mathbf{Z} = \text{Enc}(\mathbf{X}, \mathbf{A})$. The objective is link prediction reconstructing the graph structure. We employ an inner-product decoder to model the probability of an edge between a textual node u and a context node v :

$$p(\mathbf{A}_{uv} = 1 | \mathbf{z}_u, \mathbf{z}_v) = \sigma(\mathbf{z}_u^\top \mathbf{z}_v) \quad (21)$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function. Let E^+ denote the set of observed edges in the heterogeneous graph and E^- denote a set of sampled non-existent edges used as negative examples. The model is trained to reconstruct the observed graph structure by minimizing a binary cross-entropy objective:

$$\mathcal{L}_{\text{GAE}} = - \sum_{(u,v) \in E^+} \log \sigma(\mathbf{z}_u^\top \mathbf{z}_v) - \sum_{(u,v) \in E^-} \log (1 - \sigma(\mathbf{z}_u^\top \mathbf{z}_v)) \quad (22)$$

In practice, the number of possible non-existent edges in a graph grows quadratically with the number of nodes. Therefore, rather than considering all negative pairs, we employ negative sampling, where a fixed number of node pairs not connected by an edge are randomly sampled during training.

This optimization process forces the learned embeddings to preserve local connectivity and geometric proximity, ensuring that reports are spatially aligned with their associated entities and metadata in the vector space.

3.3.4 Structure-Aware Negative Sampling

The culmination of the graph learning phase is the generation of a robust training dataset for the final Dense Retrieval model. While previous works utilize retrieval-based signals to find negatives (which can be noisy or insufficiently challenging), our framework utilizes the learned graph embeddings $\mathbf{Z}_T$ to perform Structure-Aware Negative Sampling. Post-training, we apply dimensionality reduction (UMAP) and density-based clustering (HDBSCAN) to the textual embeddings $\mathbf{Z}_T$. This process groups reports that share latent structural and semantic dependencies into high-density clusters. These clusters serve as a rich source of hard negatives. For a given query document, we sample negative documents from the same semantic cluster. These documents are semantically and structurally similar to the query (sharing entities and metadata) but are distinct instances. By training the final bi-encoder to distinguish between these highly similar intra-cluster documents, the training objective encourages the encoder to distinguish between highly similar instances, significantly boosting retrieval performance in the target domain.

A methodological consideration in the proposed clustering-based negative sampling strategy concerns whether dimensionality reduction is strictly required prior to density-based clustering. In principle, HDBSCAN can be applied directly to high-dimensional embedding spaces. However, transformer-based sentence embeddings are known to exhibit distance concentration and strong anisotropy, which degrade the reliability of density estimation in high dimensions. Under these conditions, density-based clustering may produce unstable or semantically incoherent clusters, particularly in low-supervision settings where no corrective signal is available. Dimensionality reduction is therefore employed not for visualization purposes, but as a metric regularization step. By projecting embeddings into a lower-dimensional manifold that preserves local neighborhood structure, the clustering process becomes more sensitive to semantically meaningful variation while suppressing noise-dominated directions. This improves the consistency of cluster assignments and, by extension, the quality of structure-aware hard negatives derived from them. While direct clustering remains theoretically possible, the adopted approach provides a more stable and reproducible foundation for density-based negative sampling in specialized technical domains.

The choice of UMAP for dimensionality reduction is motivated by its ability to preserve local neighborhood structure and relative density in high-dimensional embedding spaces. Unlike linear techniques such as PCA, which prioritize global variance, UMAP is designed to approximate the underlying semantic manifold by maintaining meaningful local proximities between samples. This property is particularly important in the context of structure-aware negative sampling, where density-based clustering is used to identify semantically related but operationally distinct documents. Alternative methods such as t-SNE were not considered suitable due to their known distortion of density and lack of scalability, which would undermine the assumptions of density-based clustering algorithms such as HDBSCAN. UMAP therefore provides a principled

compromise between geometric fidelity and computational tractability, serving as a metric regularization step rather than a visualization tool.

Embeddings were projected from their original dimensionality into a 124-dimensional latent space prior to clustering. This dimensionality was selected to balance the preservation of local semantic neighborhoods against the need to mitigate distance concentration effects inherent to high-dimensional spaces. Empirically, dimensions in this range have been shown to retain sufficient representational capacity for technical language while enabling stable density estimation for clustering. Reducing to a substantially lower dimensionality risks collapsing semantically distinct operational states, whereas higher-dimensional projections reintroduce the geometric instability that motivates dimensionality reduction in the first place.

3.3.5 Synthetic Query Generation

For each training instance, structure-aware negative sampling produces a document triplet (d_a, d^+, d^-) , where d_a denotes the anchor document, d^+ a structurally similar document, and d^- a structurally dissimilar hard negative.

A well-known characteristic of information retrieval systems is the asymmetry between queries and documents. User queries are typically short, incomplete expressions of information needs, whereas documents contain detailed and structured descriptions of events or topics. Training dense retrievers using only document–document pairs may therefore lead to representations that are poorly aligned with the linguistic form of real search queries.

To mitigate this asymmetry, we employ a large language model to generate synthetic queries that approximate the form and length of typical user search requests. Given the document triplet (d_a, d^+, d^-) , the language model is prompted to produce a query q that captures the primary informational intent of the anchor document while remaining consistent with the positive document and excluding information characteristic of the negative document.

Formally, query generation can be expressed as:

$$q = g_{\text{LLM}}(d_a, d^+, d^-) \quad (23)$$

where $g_{\text{LLM}}(\cdot)$ denotes the prompting procedure used to guide the language model.

The prompt provides the anchor, positive, and negative documents as contextual inputs and instructs the model to generate a concise search query that retrieves documents related to the operational context of d_a and d^+ while avoiding elements specific to d^- . This constraint encourages the generated query to capture the shared semantic content between the anchor and positive documents while reinforcing the distinction from the negative instance.

This approach follows recent work in synthetic query generation for dense retrieval, where language models are used to bridge the gap between document semantics and user

query formulations. After query generation, the training dataset is expanded by constructing an additional query-based triplet:

$$(q, d^+, d^-) \quad (24)$$

Thus, the final training dataset contains two complementary forms of supervision:

$$(d_a, d^+, d^-) \quad \text{and} \quad (q, d^+, d^-) \quad (25)$$

The document-based triplets preserve the structural supervision derived from the knowledge graph, while the query-based triplets introduce linguistic signals that approximate real user information needs. This combination allows the dense retriever to learn representations that capture both structural relationships between documents and the asymmetric nature of query–document interactions typical of information retrieval tasks.

3.3.6 Dense Retriever Training Objective

The final stage of the proposed framework consists of training a dense retrieval model capable of mapping queries and documents into a shared semantic embedding space. This model is responsible for computing relevance scores between user queries and drilling reports during retrieval.

3.3.6.1 Model Architecture

We adopt a bi-encoder architecture implemented using the *Sentence Transformers* library (REIMERS; GUREVYCH, 2019). In this formulation, queries and documents are encoded independently by the same transformer-based encoder, producing fixed-dimensional vector representations.

The encoder is initialized from the *DistilBERT* transformer model (SANH et al., 2019), a compressed variant of BERT that retains strong semantic representation capabilities while offering improved computational efficiency. DistilBERT was selected due to its favorable trade-off between model capacity and inference latency, which is particularly important for large-scale retrieval systems where thousands of candidate documents may need to be encoded or compared during search.

Formally, given a query q and a document d , the encoder $f_\theta(\cdot)$ maps each input sequence into a dense vector representation:

$$\mathbf{q} = f_\theta(q), \quad \mathbf{d} = f_\theta(d) \quad (26)$$

where f_θ denotes the shared transformer encoder parameterized by θ . Relevance between queries and documents is computed using cosine similarity between the resulting embeddings:

$$\text{sim}(q, d) = \frac{\mathbf{q}^\top \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|} \quad (27)$$

This formulation allows efficient retrieval by enabling approximate nearest neighbor search in the embedding space.

3.3.6.2 Contrastive Learning Objective

Training is performed using a contrastive learning objective that encourages relevant query–document pairs to have higher similarity than non-relevant pairs. Specifically, we employ the *Multiple Negatives Ranking Loss* provided by the Sentence Transformers framework, which corresponds to an implementation of the InfoNCE objective.

Given a training batch containing N query–document pairs $\{(q_i, d_i^+)\}_{i=1}^N$, the loss for a single query q_i is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(q_i, d_i^+)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(q_i, d_j^+)/\tau)} \quad (28)$$

where τ is a temperature hyperparameter controlling the concentration of the similarity distribution. In this formulation, the positive document d_i^+ is treated as the relevant document for query q_i , while the remaining documents within the batch implicitly serve as negative examples. This in-batch negative sampling strategy significantly increases training efficiency by leveraging the diversity of documents present in each mini-batch.

The structure-aware negative sampling process described previously further enhances the difficulty of the contrastive task by ensuring that many of the in-batch negatives correspond to semantically similar but operationally distinct drilling reports.

3.3.6.3 Training Configuration

The dense retriever was trained using the training dataset constructed from both document-based and synthetic query-based triplets described in the previous section. Optimization was performed using the AdamW optimizer with a final learning rate of 2×10^{-5} .

Training was conducted for a maximum of 12 epochs, with early stopping applied based on validation performance to prevent overfitting. A warmup schedule was used during optimization, with the learning rate linearly increased during the first 10% of the total training steps before transitioning to standard decay.

This training configuration balances stability and convergence speed, allowing the model to effectively learn domain-specific semantic representations while maintaining generalization across operational contexts.

3.4 Experimental Design and Baseline Configuration

This section describes the experimental design adopted in this dissertation, with particular emphasis on the controlled comparison between the proposed method and existing synthetic query-based dense retrieval training strategies. In addition to the proposed knowledge-enhanced framework, three established approaches were implemented as baselines: UDAPDR, SWIM-IR, and ReadQG. These methods were previously introduced in Chapter 2 and are summarized here only to the extent necessary to clarify their role in the experimental setup. Although these approaches differ substantially in how synthetic queries are generated and how training signals are constructed, they share a common assumption: the absence of manually annotated query-document relevance data in specialized domains. Consequently, all methods rely on large language models to generate synthetic queries that serve as supervision for contrastive training of dense retrievers.

3.4.1 Baseline Methods

This study compares the proposed approach against three representative synthetic query-based dense retrieval training frameworks: UDAPDR, SWIM-IR, and ReadQG. Detailed conceptual descriptions of these methods are provided in Chapter 2. Here, we summarize only their experimental instantiation as baselines.

UDAPDR (SAAD-FALCON et al., 2023) is implemented as a multi-stage synthetic query training pipeline. Synthetic queries are generated using Qwen3-4B and paired with their source documents to form training instances. Following the original formulation, supervision is derived via distillation from a domain-adapted reranker trained on the synthetic data. The distilled signals are used to train a single bi-encoder retriever, which is employed for all evaluations.

SWIM-IR (THAKUR et al., 2024) is instantiated using the Summarize-then-Ask Prompting strategy for synthetic query generation. For each training document, Qwen3-4B first produces an extractive summary, which is then used to generate a single synthetic query. These query-document pairs are directly used to train a bi-encoder retriever without additional reranking or distillation stages.

ReadQG (JU et al., 2024) is implemented by generating multiple synthetic queries per document, each conditioned on a target relevance score. Soft prompt tuning is used to control the relevance signal during generation, following the calibrated likelihood objective described in

the original work. The resulting query–document pairs, associated with graded relevance labels, are converted into contrastive training instances for bi-encoder optimization.

Across all baselines, the same encoder architecture, optimization settings, and training schedule are used. Synthetic queries are generated exclusively for training, while evaluation is conducted on held-out expert-curated queries, ensuring that differences in performance reflect the nature of the training supervision rather than architectural or evaluation artifacts.

3.4.2 Synthetic Query Generation Control

To ensure experimental comparability, all synthetic queries across all methods, including the proposed approach, were generated using the same large language model: **Qwen3-4B**. This choice reduces confounding effects arising from differences in language model capacity, pretraining data, or linguistic bias. Method-specific prompting strategies were preserved in accordance with the original formulations of UDAPDR, SWIM-IR, and ReadQG, with adaptations restricted to domain terminology and formatting constraints required for Daily Drilling Reports. Synthetic queries were used exclusively during training. Evaluation was conducted using expert-curated queries from the test collection described in Section 3.1, ensuring that retrieval performance reflects generalization to realistic information needs rather than memorization of generated queries.

3.4.3 Architectural and Training Controls

All methods were instantiated using the same bi-encoder architecture based on *DistilBERT*, with identical embedding dimensionality, optimization parameters, batch size, and training schedule. This uniformity was maintained to isolate the effect of training signal construction from architectural or optimization differences. While UDAPDR, SWIM-IR, ReadQG, and the proposed method differ in how they generate queries and structure training supervision, all approaches ultimately produce contrastive training instances used to optimize a dense retriever. Differences in performance should therefore be interpreted as arising from the nature of the supervisory signal rather than from model capacity or evaluation artifacts.

From a computational perspective, all compared methods share identical bi-encoder inference costs, ensuring that observed differences do not arise from runtime advantages. The primary additional overhead introduced by the proposed approach occurs during the offline graph construction and representation learning phases, which are executed once prior to retriever training. In contrast to late-interaction models, which incur quadratic token-level costs at query time, the proposed framework preserves constant-time dense retrieval during inference, making it more suitable for large-scale industrial deployment.

3.5 Evaluation Metrics

Evaluating the performance of information retrieval systems is essential for understanding how effectively they retrieve relevant documents in response to user queries. Various evaluation metrics have been developed to quantify the effectiveness of retrieval models, especially those used for dense retrieval. These metrics assess not only the ability to retrieve relevant documents but also the quality of their ranking. This section outlines the most commonly used evaluation metrics in IR: Recall, Precision, Mean Average Precision and Normalized Discounted Cumulative Gain.

Recall measures the proportion of relevant documents retrieved from the total number of relevant documents in a dataset. It evaluates the system's ability to find as many relevant documents as possible. A high recall indicates that the system is effective at identifying relevant documents within the top- k results. Recall at a specific cutoff k (Recall@ k) is defined as:

$$\text{Recall@}k = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\#retr_{q,k}}{\#rel_q} \quad (29)$$

Here, the numerator is the number of relevant documents for a query q found within the top k results, and the denominator is the total number of documents relevant to q in the entire dataset. The final score is averaged over all queries in the set Q .

Precision assess the proportion of retrieved documents that are relevant to a given query. It emphasizes the quality of the results provided by the system, especially regarding the top-ranked documents. A high precision indicates that most of the retrieved documents are relevant. Precision at a specific cutoff k (Precision@ k) is defined as:

$$\text{Precision@}k = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\#retr_{q,k}}{k} \quad (30)$$

Precision and recall are often in tension: improving recall by retrieving more documents can sometimes decrease precision if the additional documents are irrelevant.

Mean Average Precision is a metric that provides a single-figure measure of quality across recall levels. MAP is particularly sensitive to the rank of each relevant document, penalizing systems that return relevant documents late in the result list. It is calculated by first finding the Average Precision for each query and then averaging these scores across all queries. The Average Precision for a single query, AP_q , averages the precision scores calculated at each position where a relevant document is found.

$$AP_q = \frac{1}{\#rel_q} \sum_{k=1}^N P@k \times \text{rel}(k) \quad (31)$$

Here, N is the total number of retrieved documents, $\#rel_q$ is the total number of known relevant documents for query q , $P@k$ is the precision at cutoff k , and $rel(k)$ is an indicator function which is 1 if the document at rank k is relevant and 0 otherwise. The MAP is then the mean of the AP scores over the set of all queries Q :

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP_q, \quad (32)$$

Since MAP considers both precision and the rank of relevant items, it provides a more comprehensive assessment than Precision@k alone.

The Normalized Discounted Cumulative Gain is a metric used to assess the quality of ranked results by considering both the relevance of documents and their positions. A key feature of NDCG is its ability to handle multiple levels of relevance (e.g., scores of 0 for irrelevant, 1 for relevant, 2 for highly relevant). The core idea is that retrieving highly relevant documents earlier in the ranked list is more valuable than retrieving them later. The metric starts with Cumulative Gain (CG), which simply sums the relevance scores of the top k documents. To account for position, Discounted Cumulative Gain (DCG) applies a logarithmic penalty that increases with rank:

$$DCG_q@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}, \quad (33)$$

Here, rel_i is the graded relevance score of the document at position i . To make this score comparable across different queries, it is normalized by the Ideal Discounted Cumulative Gain (IDCG). The IDCG is the DCG score of a perfect ranking, where documents are sorted by their true relevance in descending order.

$$NDCG@k = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{DCG_q@k}{IDCG_q@k}, \quad (34)$$

The final NDCG@k for a set of queries Q is the average of the individual NDCG scores. A score of 1.0 indicates a perfect ranking, while 0 indicates the worst possible ranking.

4 Results

This chapter presents the empirical results of this work. It begins by presenting a comparative analysis of the fine-tuned model against a suite of baseline models using standard retrieval metrics. We then conduct a detailed evaluation of the model, incorporating domain-specific analyses and qualitative assessments to provide a comprehensive understanding of its performance and practical utility. Throughout this chapter, results are reported descriptively and comparatively, with emphasis on observable performance patterns across models and evaluation settings. Interpretative analysis and broader implications are discussed in Chapter 5 to maintain a clear separation between empirical findings and their theoretical or practical interpretation.

4.1 Comparative Performance Analysis

To assess the efficacy of our proposed methodology, we compare it against a diverse set of baselines representing distinct retrieval paradigms. These include traditional probabilistic models (*TF-IDF*, *BM25*) and a domain specific Word2Vec model trained on Oil and Gas corpora *petrovec-o&g* (GOMES, 2021). For neural retrieval, we evaluate sparse methods *splade-v3* (LASSANCE et al., 2024) and a broad range of dense embedding models: the commercial state-of-the-art *text-embedding-3-large*, standard dual-encoders (*mcontriever*, *e5-large*), and recent LLM-based architectures including *multilingual-e5-large-instruct* (WANG et al., 2024), *qwen3-embedding-0.6b* (ZHANG et al., 2025), *embedding-gemma-300m* (VERA et al., 2025), and the high-performance *bge-multilingual-gemma2*. We also evaluate the late interaction paradigm using both the standard *colbertv2.0* (SANTHANAM et al., 2022) and the multilingual-optimized *jina-colbert-v2* (JHA et al., 2024). This selection combines top-ranked open-weight models from the MTEB leaderboard with widely adopted industry standards. Table 4 provides a technical summary of the selected models.

To provide a comprehensive evaluation of retrieval effectiveness, we report complementary ranking metrics that capture distinct aspects of model behavior. *nDCG@k* emphasizes ranking quality with graded relevance, *MAP* reflects consistency across the ranked list, and *Recall@k* measures coverage of relevant documents as the retrieval depth increases. Reporting these metrics jointly allows for a nuanced comparison of lexical, dense, and interaction-based models under both limited and exhaustive annotation regimes.

4.1.1 Analysis of Zero-Shot Performance

The experimental results presented in Tables 7 through 5 offer a comprehensive view of the capabilities and limitations of current state-of-the-art retrieval architectures when applied

Model	Params	Description
TF-IDF / BM25	-	Standard probabilistic baselines; unsupervised lexical matching via term frequency and inverse document frequency.
petrovec-o&g	-	Domain-specific Word2Vec embeddings (300d) trained on internal Oil & Gas corpora; represents legacy semantic search.
splade-v3	110M	BERT-based sparse expansion; addresses vocabulary mismatch via explicit term re-weighting.
text-embedding-3-large	Closed	OpenAI's SOTA proprietary model; high-dimensional (3072d) multilingual embeddings.
mcontriever-base	110M	BERT-based dense retriever trained with unsupervised contrastive learning; designed for zero-shot generalization.
e5-large	335M	Weakly-supervised dense retriever trained on large-scale text pairs; serves as a standard dense baseline.
multilingual-e5-large-instruct	560M	Instruction-tuned XLM-RoBERTa; trained with contrastive loss for zero-shot retrieval.
qwen3-embedding-0.6b	600M	Optimized for multilingual retrieval with extended context (32k tokens); high MTEB performance.
embedding-gemma-300m	300M	Distilled efficiency model based on Google's Gemma architecture; optimized for low-latency.
bge-multilingual-gemma2	9B	Large-scale LLM-based embedding model; uses the Gemma-2 backbone for state-of-the-art semantic understanding.
colbertv2.0	110M	A late-interaction model that performs scalable and fine-grained text matching by interacting (MaxSim operation) with the query and document embeddings at the token level
jina-colbert-v2	560M	Multilingual late-interaction model based on XLM-RoBERTa; features Matryoshka embeddings and extended context window (8k).

Table 4 – Overview of retrieval baselines used for evaluation, spanning probabilistic lexical models, domain-specific static embeddings, sparse neural retrievers, dense bi-encoder architectures, and late-interaction models. The selection reflects both widely adopted industry standards and top-performing open-weight models, enabling a comprehensive comparison across retrieval paradigms.

Model Name	Gold Standard				Silver Standard			
	@1	@5	@10	@50	@1	@5	@10	@50
tf-idf	0.6454	0.4889	0.4426	0.3859	0.6596	0.5322	0.4951	0.4692
bm25	0.7660	0.6629	0.6404	0.4762	0.7660	0.6629	0.6404	0.5460
petrovec-o&g	0.3830	0.3079	0.2875	0.2236	0.3901	0.3170	0.3051	0.2831
splade-v3	0.5177	0.4500	0.4164	0.3608	0.6064	0.5243	0.5026	0.4856
text-embedding-3-large	0.4468	0.3877	0.3857	0.3003	0.4468	0.3877	0.3871	0.3447
mcontriever-base-msmarco	0.4894	0.4724	0.4804	0.3456	0.4894	0.4724	0.4804	0.4098
e5-large	0.4752	0.3777	0.3413	0.2996	0.5142	0.4255	0.3987	0.3820
multilingual-e5-large-instruct	0.5390	0.4203	0.3999	0.3096	0.5887	0.4713	0.4526	0.3879
bge-multilingual-gemma2	0.8227	0.6955	0.6740	0.5174	0.8227	0.6955	0.6740	0.5867
qwen3-embedding-0.6b	0.5532	0.4509	0.4023	0.3240	0.5745	0.4837	0.4472	0.4039
embedding-gemma-300m	0.3262	0.2671	0.2350	0.1974	0.3333	0.2795	0.2595	0.2448
colbert-v2	0.6667	0.6461	0.6347	0.4727	0.6667	0.6461	0.6352	0.5358
jina-colbert-v2	0.8511	0.7732	0.7393	0.5554	0.8511	0.7732	0.7426	0.6124
distilbert-udapdr	0.4681	0.4064	0.3678	0.2223	0.4894	0.4448	0.4161	0.2480
distilbert-swim	0.5248	0.5033	0.4695	0.3450	0.6135	0.6103	0.6043	0.4119
distilbert-readqg	0.6170	0.6115	0.6427	0.4719	0.6986	0.6944	0.7352	0.4947
distilbert-our	0.6028	0.6530	0.6509	0.4904	0.6844	0.7252	0.7394	0.5045

Table 5 – Normalized Discounted Cumulative Gain (nDCG@k) for all retrieval models evaluated on the Gold and Silver Standard datasets. Results are reported at multiple cutoff values to assess ranking quality under varying retrieval depths. The table enables comparison across lexical, dense, and late-interaction paradigms under different annotation granularities.

Model Name	Gold Standard				Silver Standard			
	@1	@5	@10	@50	@1	@5	@10	@50
tf-idf	0.0193	0.0606	0.0903	0.1735	0.0005	0.0019	0.0030	0.0105
bm25	0.0241	0.0856	0.1490	0.2453	0.0009	0.0026	0.0044	0.0138
petrovec-o&g	0.0113	0.0352	0.0545	0.0971	0.0002	0.0008	0.0014	0.0040
splade-v3	0.0166	0.0560	0.0864	0.1593	0.0007	0.0026	0.0044	0.0129
text-embedding-3-large	0.0152	0.0436	0.0745	0.1367	0.0004	0.0011	0.0020	0.0080
mcontriever-base-msmarco	0.0143	0.0558	0.0996	0.1632	0.0003	0.0013	0.0025	0.0087
e5-large	0.0140	0.0394	0.0574	0.1170	0.0006	0.0015	0.0026	0.0087
multilingual-e5-large-instruct	0.0163	0.0446	0.0686	0.1237	0.0004	0.0018	0.0030	0.0102
bge-multilingual-gemma2	0.0237	0.0891	0.1557	0.2718	0.0009	0.0028	0.0048	0.0162
qwen3-embedding-0.6b	0.0145	0.0440	0.0693	0.1392	0.0003	0.0014	0.0026	0.0088
embedding-gemma-300m	0.0084	0.0231	0.0320	0.0676	0.0003	0.0009	0.0015	0.0049
colbert-v2	0.0195	0.0853	0.1563	0.2521	0.0007	0.0030	0.0053	0.0140
jina-colbert-v2	0.0262	0.1029	0.1756	0.2874	0.0009	0.0031	0.0053	0.0163
distilbert-udapdr	0.0139	0.0445	0.0650	0.0886	0.0005	0.0019	0.0030	0.0047
distilbert-swim	0.0210	0.0838	0.1305	0.2012	0.0009	0.0042	0.0082	0.0161
distilbert-readqg	0.0221	0.0958	0.1784	0.2741	0.0009	0.0045	0.0088	0.0165
distilbert-our	0.0222	0.1019	0.1796	0.2883	0.0009	0.0045	0.0090	0.0170

Table 6 – Mean Average Precision (MAP@k) results for the evaluated retrieval models on the Gold and Silver Standard datasets. MAP emphasizes precision across the full ranked list, providing insight into the consistency of relevance ordering across different retrieval approaches and supervision regimes.

Model Name	Gold Standard				Silver Standard			
	@1	@5	@10	@50	@1	@5	@10	@50
tf-idf	0.0193	0.0675	0.1167	0.3124	0.0005	0.0022	0.0039	0.0172
bm25	0.0241	0.0934	0.1731	0.3721	0.0009	0.0029	0.0053	0.0201
petrovec-o&g	0.0113	0.0417	0.0790	0.1851	0.0002	0.0009	0.0021	0.0074
splade-v3	0.0166	0.0668	0.1150	0.2860	0.0007	0.0029	0.0051	0.0182
text-embedding-3-large	0.0152	0.0509	0.0984	0.2334	0.0004	0.0013	0.0028	0.0134
mcontriever-base-msmarco	0.0143	0.0633	0.1267	0.2523	0.0003	0.0015	0.0035	0.0137
e5-large	0.0140	0.0471	0.0800	0.2333	0.0006	0.0017	0.0035	0.0137
multilingual-e5-large-instruct	0.0163	0.0542	0.0932	0.2193	0.0004	0.0024	0.0039	0.0155
bge-multilingual-gemma2	0.0237	0.0921	0.1722	0.3877	0.0009	0.0029	0.0055	0.0225
qwen3-embedding-0.6b	0.0145	0.0535	0.0915	0.2441	0.0003	0.0019	0.0036	0.0139
embedding-gemma-300m	0.0084	0.0294	0.0439	0.1411	0.0003	0.0012	0.0018	0.0076
colbert-v2	0.0195	0.0938	0.1835	0.3519	0.0007	0.0032	0.0063	0.0187
jina-colbert-v2	0.0262	0.1097	0.1956	0.3959	0.0009	0.0032	0.0057	0.0217
distilbert-udapdr	0.0139	0.0529	0.0871	0.1328	0.0005	0.0022	0.0035	0.0061
distilbert-swim	0.0210	0.0924	0.1570	0.2697	0.0009	0.0043	0.0086	0.0173
distilbert-readqg	0.0221	0.1037	0.2015	0.3246	0.0009	0.0046	0.0091	0.0174
distilbert-our	0.0222	0.1103	0.2007	0.3408	0.0009	0.0046	0.0091	0.0171

Table 7 – Recall@k scores for all retrieval models evaluated on the Gold and Silver Standard datasets. Recall at increasing cutoff values reflects the ability of each method to retrieve a broader set of relevant documents, complementing precision-oriented metrics and highlighting trade-offs across retrieval paradigms.

zero-shot to the specialized domain of offshore drilling. Several key patterns emerge regarding the efficacy of interaction-based models and the impact of the Silver Standard on evaluation metrics.

4.1.1.1 The Dense Retrieval Gap

Figure 8 presents the comparative performance across models measured using NDCG@10. A critical observation from the Gold Standard evaluation is the underperformance of many general-purpose dense retrieval models compared to the lexical baseline. The *BM25* model demonstrates remarkable robustness, achieving an NDCG@10 of 0.6404 and Precision@10 of 0.6894. Surprisingly, it outperforms widely used dense models such as *text-embedding-3-small* (NDCG@10: 0.3724) and *e5-large* (NDCG@10: 0.3413).

This pattern suggests the presence of a dense retrieval gap in specialized industrial domains, where general-purpose semantic models trained on web-scale data may struggle to capture the precise technical nomenclature and constraints present in operational reports. Without domain adaptation, these models often prioritize broad semantic relatedness over the specific keyword constraints required by engineering queries (e.g., distinguishing between specific tool assemblies like “BHA” and “LWD”). The notable exceptions to this trend are *bge-multilingual-gemma2* and *jina-colbert-v2*. The BGE model significantly outperforms BM25 (NDCG@10: 0.6740), suggesting that large-scale, instruction-tuned embeddings can bridge this semantic gap where smaller or older dense models fail.

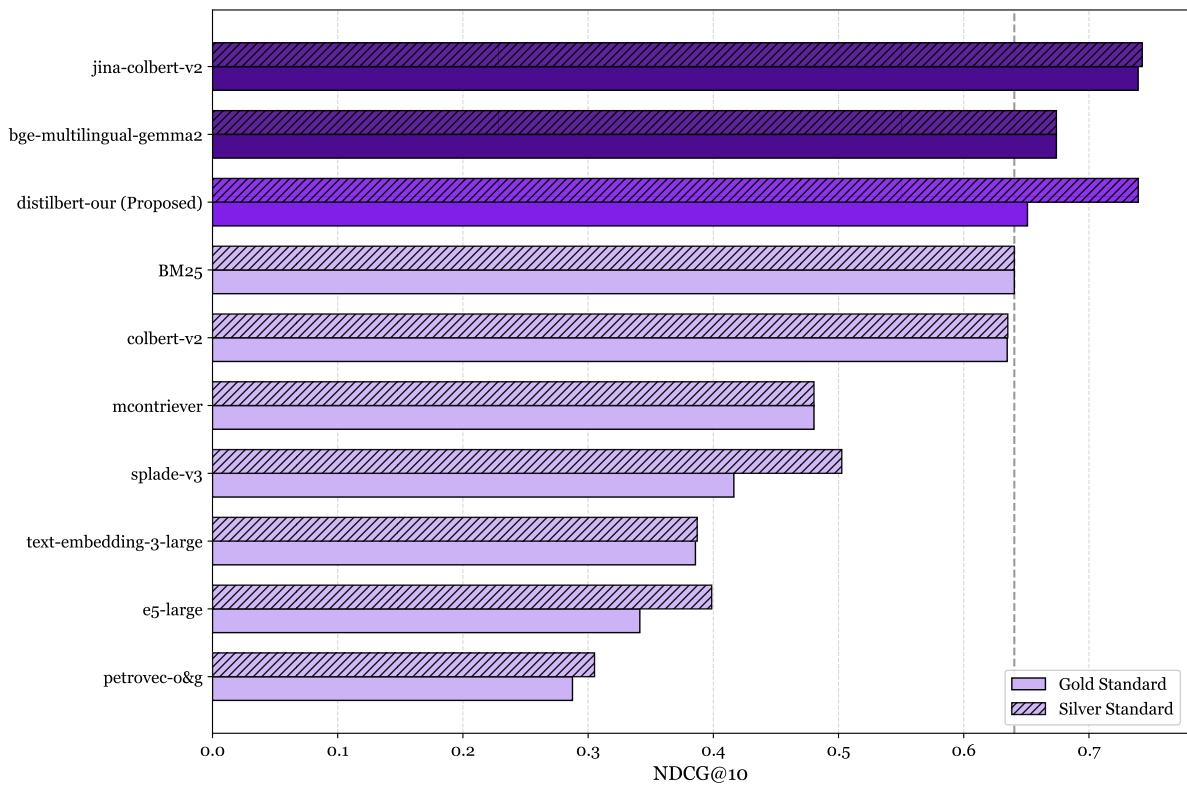


Figure 8 – Comparative retrieval performance across models under gold-standard and silver-standard datasets NDCG@10.

4.1.1.2 Superiority of Late Interaction

The results indicate that late-interaction architectures are particularly effective in this domain, consistently achieving strong performance across multiple ranking metrics. *jina-colbert-v2* emerges as the top-performing model across most metrics, achieving the highest Gold Standard NDCG@10 (0.7393) and Precision@10 (0.7787). *colbert-v2* also performs competitively, closely matching or outperforming the strong baselines. This superiority suggests that the fine-grained, token-level interaction mechanism inherent in ColBERT architectures is better suited for matching the complex constraints of drilling operations than the single-vector compression of standard bi-encoders. By preserving token-level features, these models prevent the loss of critical technical details during the encoding phase.

4.1.1.3 Impact of the Silver Standard: Recall Drop and Precision Gain

Comparing the metrics between the Gold and Silver Standards reveals a distinct structural divergence that validates the methodological approach of this dissertation.

Recall and MAP Drop: In the Silver Standard, metrics sensitive to the total number of relevant documents, specifically Recall and MAP drop significantly (e.g., BM25 Recall@50 drops from 0.3721 to 0.0201). This behavior should be interpreted in light of the differing annotation regimes rather than as an absolute degradation in model performance. The Gold

Standard calculation relies on a limited pool (where the total number of relevant documents is small), effectively measuring “relative recall.” In contrast, the Silver Standard annotated the entire corpus of 108,000 documents, identifying a much larger set of relevant documents per query. Consequently, retrieving only 50 documents covers a naturally smaller fraction of this total relevant set, resulting in lower absolute scores.

Precision Stability and Gap filling: Conversely, Precision metrics remain stable or even improve for several models in the Silver Standard. For instance, *splade-v3* sees its Precision@10 rise from 0.4447 (Gold) to 0.6936 (Silver), and *multilingual-e5-large-instruct* improves from 0.4234 to 0.5723. This is a relevant finding that suggests that the Gold Standard suffered from the pooling artifact problem, where relevant documents retrieved by models not included in the original pool (or ranked lower than the pooling depth) were incorrectly classified as non-relevant. The Silver Standard, generated by the global classifier, identified these documents as relevant. This demonstrates that the Silver Standard provides a fairer and more comprehensive evaluation, acknowledging the utility of models that retrieve relevant content outside the narrow consensus of the initial manual pool. However, it is important to mention that, ideally, these documents should be posteriorly re-evaluated by a human specialist in order to be validated and added to the Golden Standard.

4.1.2 Analysis of Fine-Tuned Models

Having established the limitations of zero-shot baselines, particularly the dense retrieval gap where standard bi-encoders fall short of lexical matching, we now turn to the evaluation of our domain-adapted models. The performance of the fine-tuned DistilBERT models reveals the significant impact of the training data construction strategy on retrieval effectiveness. By isolating the encoder architecture (using *distilbert-base-multilingual-cased* across all fine-tuned variants), the performance differences observed in Tables 5–7 can be primarily associated with differences in the construction and semantic constraints of the synthetic training triples, given that the encoder architecture and optimization procedure are held constant. Figure 9 presents the comparative performance of these methods across NDCG@10, MAP@10, and Recall@10. The results indicate that the proposed framework consistently achieves the strongest overall performance across metrics, suggesting that the combination of graph-based negative sampling and synthetic query generation provides a strong training signal for dense retrievers operating in specialized technical domains.

4.1.2.1 Synthetic Generation vs. Unsupervised Adaptation

A clear performance hierarchy emerges between methods that rely on standard unsupervised signals versus those that employ advanced synthetic query generation. The *distilbert-udapdr* model, despite its sophisticated distillation pipeline, struggles in this highly technical domain, achieving an NDCG@10 of only 0.3678 in the Gold Standard, marginally outperforming

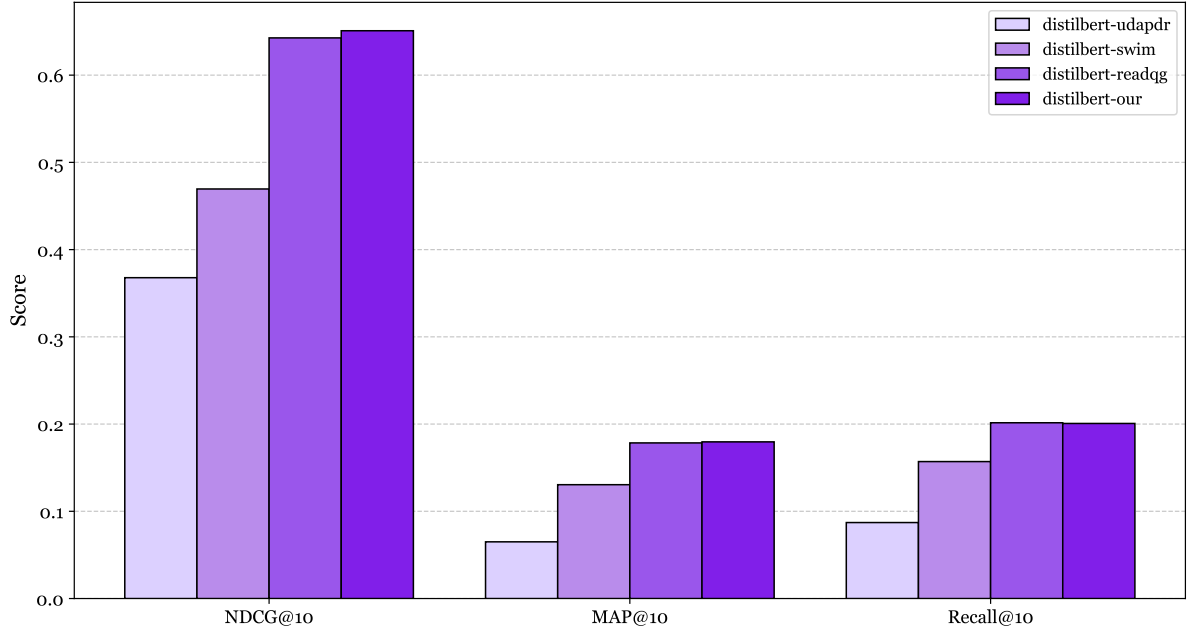


Figure 9 – Performance comparison of fine-tuning strategies across NDCG@10, MAP@10, and Recall@10.

the static *petrovec* embeddings. This suggests that without a domain-specific query generation mechanism that deeply understands engineering context, general-purpose adaptation frameworks may fail to capture the necessary semantic granularity. In contrast, methods that enforce structured generation logic show immediate gains. *distilbert-swim*, which employs a “Summarize-then-Ask” prompting strategy, reaches an NDCG@10 of 0.4695. By grounding queries in extractive summaries, SWIM-IR ensures better alignment with the core operational facts of the drilling reports, leading to a noticeable improvement over the baseline UDAPDR approach.

4.1.2.2 Efficacy of Relevance-Aware and Graph-Enhanced Strategies

The most significant improvement is observed with *distilbert-readqg* and our proposed method (*distilbert-our*). *distilbert-readqg*, which uses soft prompt-tuning to generate queries with continuous relevance scores, demonstrates exceptional precision at the top of the ranking (NDCG@1: 0.6170). This indicates that training on a diverse spectrum of relevance signals helps the model distinguish between highly relevant and marginally relevant documents effectively. Our proposed model, which leverages graph-based structural sampling to mine hard negatives, demonstrates superior robustness at deeper ranking positions. While performing comparably to ReadQG at the top rank (NDCG@1: 0.6028), it surpasses all other fine-tuned models in broader retrieval metrics, achieving the highest NDCG@5 (0.6530) and NDCG@10 (0.6509) among the DistilBERT variants.

4.1.2.3 Surpassing the Lexical Baseline

Crucially, at the NDCG@10 cutoff, our proposed model achieves a score of 0.6509, successfully outperforming the strong BM25 baseline (0.6404). This result represents an important empirical outcome for this dissertation. It confirms that our fine-tuning strategy (which injects structural domain knowledge via graph-based negative sampling) successfully instills domain-specific ranking capabilities that exceed traditional term-matching, even when utilizing a significantly smaller and more efficient encoder architecture (110M parameters) compared to large-scale state-of-the-art baselines like *bge-multilingual-gemma2* (9B parameters).

4.1.2.4 Generalization in the Silver Standard

The robustness of the proposed method is further validated by the Silver Standard evaluation. Unlike the zero-shot baselines which saw precipitously dropping recall, the fine-tuned models maintain strong performance. *distilbert-our* achieves a Silver Standard NDCG@10 of 0.6509, virtually identical to its Gold Standard performance. This stability implies that the model has not merely overfitted to the small pooling set of the Gold Standard but has genuinely learned a generalized relevance function applicable to the entire corpus. Furthermore, in the Silver Standard, our model's Recall@10 (0.0091) matches that of ReadQG, indicating that both advanced generation strategies are equally effective at recovering hidden relevant documents that were missed by the initial pooling phase.

4.2 In-Depth Analysis

While the comparative analysis establishes the superior ranking effectiveness of the proposed *distilbert-our* model, it is necessary to investigate the underlying mechanisms driving this performance. This section moves beyond aggregate metrics to analyze the geometric properties of the learned latent space, the model's robustness across query difficulty levels, and the qualitative nature of the semantic relationships captured.

4.2.1 Latent Space Geometry: Alignment and Uniformity

To quantify the quality of the learned representations, we adopt the *Alignment and Uniformity* framework proposed by Wang e Isola (2020). High-quality dense retrieval representations should satisfy two properties: Alignment, where positive pairs are mapped close together, and Uniformity, where feature vectors are uniformly distributed on the hypersphere, preserving maximal information. We calculated the Alignment loss ($\mathcal{L}_{align}$) and Uniformity loss ($\mathcal{L}_{uniform}$) for the baseline models and our fine-tuned *distilbert-our* model. Table 8 presents the calculated metrics for the baseline and fine-tuned models on the Silver Standard dataset.

Model	Alignment (↓)	Uniformity (↓)
mcontriever-base-msmarco	1.1809	-2.4989
embedding-gemma-300m	1.0249	-2.1998
e5-large	0.3326	-0.8107
multilingual-e5-large-instruct	0.2748	-0.5984
distilbert-udapdr	1.3705	-3.3039
distilbert-swim	1.3308	-3.1564
distilbert-readqg	1.2453	-3.1924
distilbert-our	1.2311	-3.1813

Table 8 – Alignment and Uniformity metrics calculated on highly relevant pairs (Score=3) from the Silver Standard. Lower values indicate better performance for both metrics.

Trade-off Between Alignment and Uniformity: A distinct dichotomy emerges between the massive supervised baselines and the domain-adapted models. The *e5-large* and *multilingual-e5-large-instruct* models achieve exceptionally low Alignment scores (0.3326 and 0.2748, respectively). This is a direct result of their large-scale supervised contrastive pre-training on millions of positive pairs, which aggressively pulls positive samples together. However, this comes at the cost of Uniformity: their scores (-0.8107 and -0.5984) are significantly higher (worse) than the adapted models. This indicates a degree of representation collapse or anisotropy, where embeddings are clustered in narrow cones of the vector space, potentially limiting their ability to distinguish between fine-grained domain nuances. In contrast, the domain-adapted models (*udapdr*, *swim*, *readqg*, and *our*) exhibit significantly better Uniformity (ranging from -3.15 to -3.30). This suggests that the adaptation process, driven by synthetic data and hard negatives, forces the models to utilize the semantic space more efficiently, spreading documents out to resolve ambiguities.

Superiority of Graph-Enhanced Alignment: Among the fine-tuned models, our proposed method (*distilbert-our*) demonstrates the most effective balance. While *udapdr* achieves the lowest Uniformity (-3.3039), it suffers from the worst Alignment (1.3705), indicating that its embeddings are well-spread but fail to bring relevant documents sufficiently close to their queries. Our model achieves the best Alignment (1.2311) among all fine-tuned variants, surpassing *readqg* (1.2453), *swim* (1.3308), and *udapdr* (1.3705). This quantitative result is consistent with the hypothesis that structure-aware negative sampling provides a more informative learning signal than purely text-based generation strategies. By utilizing graph-based hard negatives with documents that share structural context but are distinct, the model is forced to pull the positive document closer to the query to satisfy the contrastive margin, resulting in tighter semantic coupling without sacrificing the desirable uniformity properties of the latent space.

4.2.2 Query-Level Analysis and Robustness

Aggregate retrieval metrics such as Mean Average Precision provide a summary measure of system performance but do not capture variability across individual queries. This limitation is particularly relevant in datasets characterized by substantial variation in query difficulty. To evaluate the robustness of the proposed method under such conditions, a stratified query-level analysis was conducted. The fine-tuned *distilbert-our* model was compared against a lexical baseline (*BM25*) and two neural baselines (*bge-multilingual-gemma2* and *jina-colbert-v2*).

4.2.2.1 Stratified Performance by Difficulty

Test queries were partitioned into three difficulty categories based on the NDCG@10 performance of the lexical baseline: *Hard* ($\text{NDCG} \leq 0.2$), *Medium* ($0.2 < \text{NDCG} \leq 0.6$), and *Easy* ($\text{NDCG} > 0.6$). Table 9 reports the corresponding performance of each model within these strata.

Difficulty	BM25	bge-multilingual-gemma2	jina-colbert-v2	distilbert-our
Hard	0.1176	0.4551	0.4584	0.4849
Medium	0.4116	0.4760	0.6215	0.6098
Easy	0.8481	0.8121	0.8484	0.7010

Table 9 – NDCG@10 performance stratified by lexical query difficulty. The proposed model demonstrates superior performance on "Hard" queries, surpassing even the largest baselines.

While Table 9 reports the exact performance values, Figure 10 provides a visual comparison of model behavior across difficulty levels, highlighting how the relative ranking of models changes as query complexity increases.

On Hard queries, the proposed model achieves the highest NDCG@10 score (0.4849), exceeding the performance of BM25 (0.1176), *bge-multilingual-gemma2* (0.4551), and *jina-colbert-v2* (0.4584). These queries typically involve vocabulary mismatch, implicit relationships, or domain-specific terminology, conditions under which lexical matching is less effective. The observed performance is consistent with the hypothesis that incorporating structured domain information through graph-based negative sampling improves semantic discrimination in such cases. For Easy queries, the proposed model underperforms relative to BM25 and *jina-colbert-v2*. This behavior is consistent with prior observations in dense retrieval, where bi-encoder architectures may retrieve semantically related documents that do not satisfy strict lexical constraints required for precise factual queries.

4.2.2.2 Win/Loss Analysis

To further contextualize model behavior at the query level, pairwise Win Rates were computed against each baseline. A Win is defined as a query for which *distilbert-our* achieves

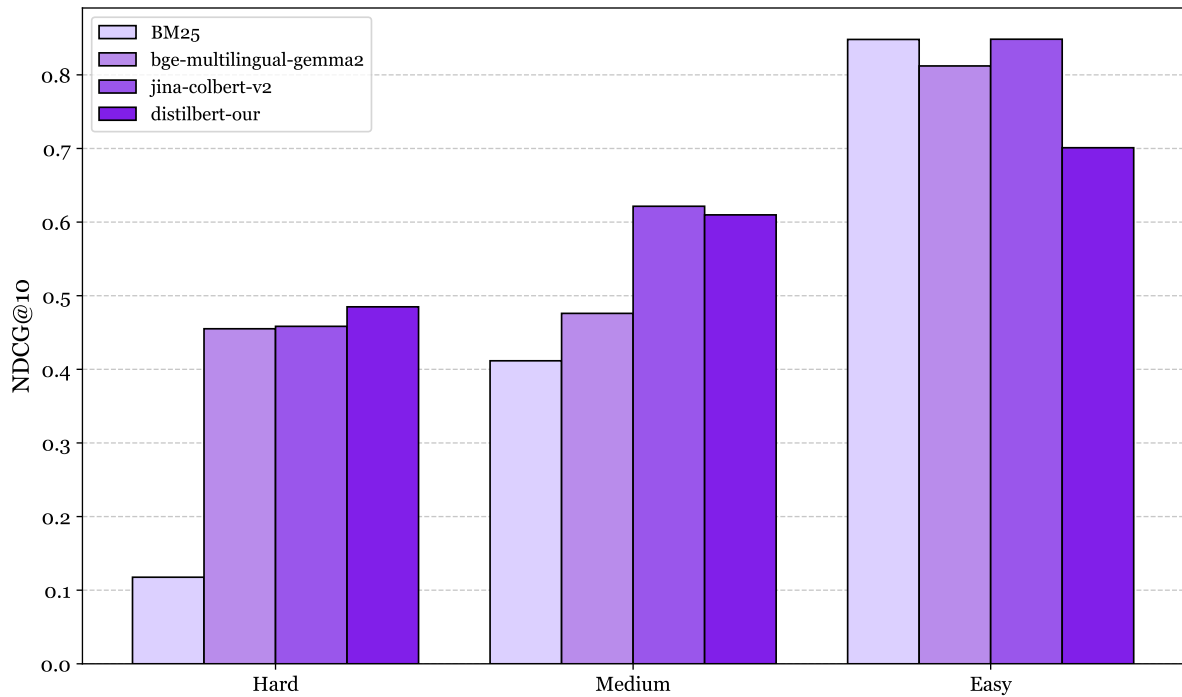


Figure 10 – NDCG@10 performance stratified by query difficulty.

a higher nDCG@10 score than the comparator.

Against BM25, the proposed model achieves a Win Rate of 51.1%, indicating comparable performance overall, with relative gains on semantically oriented queries and lower performance on exact-match queries. When compared to *bge-multilingual-gemma2*, the model achieves a Win Rate of 44.7%, despite having substantially fewer parameters. This result indicates that domain-specific fine-tuning can partially offset differences in model scale under the evaluated conditions. Against *jina-colbert-v2*, the Win Rate is 40.4%, reflecting competitive performance relative to a late-interaction architecture, while maintaining lower inference complexity.

4.2.2.3 Qualitative Inspection

While aggregate metrics quantify overall retrieval effectiveness, they do not reveal the specific semantic behaviors that enable a model to succeed on difficult queries. To better understand the mechanisms through which the proposed model outperforms the strongest zero-shot baseline (*jina-colbert-v2*), we conducted a qualitative inspection of retrieval outputs for a subset of hard queries characterized by vocabulary mismatch, implicit causality, and domain-specific measurements. Table 10 presents three representative cases.

These examples illustrate that the performance gains observed on hard queries are not driven by superficial lexical matching, but by the model’s ability to recover latent operational relationships. In QID 31, the model distinguishes between a system state and an operational event, prioritizing documents that reflect causal interruption rather than static conditions. In QID 14, it links a qualitative phenomenon (drag) to its quantitative manifestation, expressed in

Query (QID 31) <i>Blackout occurred</i>	
jina-colbert-v2	Retrieves a generic description: “ <i>Partial blackout affecting the electrical system...</i> ”. The model matches the keyword <i>blackout</i> but retrieves a document describing a system condition rather than the occurrence of an operational event.
distillbert-our	Retrieves a causally grounded description: “ <i>Operation interrupted due to partial blackout...</i> ”. The model correctly interprets the query as an event occurrence and prioritizes a document linking the blackout to an operational interruption of the rig.
Query (QID 14) <i>Drag observed during casing run</i>	
jina-colbert-v2	Retrieves a procedural note: “ <i>Drag identified during the first attempt...</i> ”, which acknowledges the phenomenon but lacks operational specificity.
distillbert-our	Retrieves a quantitative report: “ <i>Drag between 50 klbf and 100 klbf.</i> ”. The model associates the abstract concept of drag with its domain-specific numerical measurement, reflecting an understanding of how the phenomenon is operationally quantified.
Query (QID 22) <i>High current observed</i>	
jina-colbert-v2	Retrieves: “ <i>Operational difficulty due to high current conditions...</i> ”, largely matching the surface phrasing of the query.
distillbert-our	Retrieves: “ <i>Strong current caused reduced running speed...</i> ”. The model resolves synonymy between <i>high</i> and <i>strong</i> current and prioritizes the document describing the operational consequence of the environmental condition.

Table 10 – Qualitative comparison of retrieval outputs for hard queries. The proposed model demonstrates superior handling of implicit causality, synonymy, and domain-specific quantitative associations compared to a strong late-interaction baseline.

kilopounds-force, a relationship that is rarely explicit in general-language corpora. In QID 22, the model resolves synonymy while simultaneously ranking documents that describe downstream operational impact.

Crucially, these behaviors align with the training strategy employed in this work. The Graph-Enhanced Negative Sampling mechanism exposes the model to documents that are structurally adjacent (sharing entities, metadata, or operational context) while differing in their underlying event semantics. As a result, the encoder is forced to learn distinctions that go beyond surface-level similarity, enabling it to generalize to queries where relevant information is expressed indirectly or through domain-specific shorthand. This qualitative evidence complements the quantitative results, providing explanatory support for the observed improvements on high-difficulty queries.

4.3 Generalization and Domain Specificity

A central question in domain adaptation research concerns the extent to which optimizing a model for a highly specialized target distribution compromises its ability to perform general-domain information retrieval. In this work, fine-tuning was explicitly oriented toward maximizing

retrieval effectiveness within the Oil & Gas drilling domain, without introducing constraints aimed at preserving open-domain semantic representations. To examine the consequences of this design choice, we evaluated the resulting model on the Quati benchmark (BUENO et al., 2024), a large-scale dataset designed to assess information retrieval performance in Brazilian Portuguese.

On the Quati benchmark, the domain-adapted model achieved an nDCG@10 score of 0.2225, substantially underperforming both lexical and dense general-purpose baselines. In comparison, the unsupervised BM25 retriever reported an nDCG@10 of 0.3991, while state-of-the-art dense retrievers ranged from approximately 0.49 to 0.65, with the strongest configuration combining BM25 with an mT5-XL re-ranker. Given that both the benchmark and the evaluation setup operate within the same language and retrieval paradigm, this performance gap cannot be attributed to linguistic mismatch or evaluation artifacts, but rather reflects the representational effects induced by domain-specific fine-tuning.

In an attempt to mitigate this representational drift, preliminary experiments were conducted using partial parameter freezing during the fine-tuning stage. Specifically, configurations were evaluated in which the lower layers of the encoder were kept frozen while only the higher transformer layers were updated, following the intuition that earlier layers encode more general linguistic representations while later layers capture task-specific semantics. However, these configurations exhibited unstable optimization behavior and failed to achieve satisfactory convergence. This difficulty is likely a consequence of the strong distributional shift between the general-domain pretraining corpus and the highly specialized drilling reports used for adaptation. When a large portion of the model parameters is constrained, the remaining trainable layers appear insufficient to fully accommodate the domain-specific terminology and relational patterns required for effective retrieval.

More recent parameter-efficient fine-tuning techniques, such as adapter modules and Low-Rank Adaptation (LoRA), may offer a more principled mechanism to address this trade-off. By introducing additional trainable components while preserving the original pretrained weights. Although such approaches were not explored within the scope of the present work, they represent a promising direction for future research aimed at maintaining broader semantic generalization while still benefiting from domain specialization.

The observed degradation in general-domain retrieval effectiveness is consistent with phenomena commonly described in the literature as catastrophic forgetting or representation drift. During the fine-tuning process described in Chapter 3, model parameters were optimized exclusively to minimize a contrastive loss over drilling reports, reinforcing sensitivity to highly technical terminology and operational patterns while progressively weakening the alignment with broader semantic relationships required for open-domain retrieval. As a result, the embedding space becomes increasingly structured around domain-specific concepts, reducing its capacity to represent queries whose relevance depends on general world knowledge.

This interpretation is further supported by the contrast between the Quati results and the in-domain evaluation presented earlier, where the same model demonstrated superior performance relative to both BM25 and the *bge-multilingual-gemma-2* baseline. Queries grounded in drilling operations and engineering phenomena are effectively embedded and ranked within the specialized corpus, whereas queries requiring general semantic coverage, such as those related to tourism or everyday topics are no longer mapped to meaningful regions of the vector space. This divergence illustrates the plasticity stability trade-off inherent to unconstrained domain adaptation, whereby gains in target-domain performance are accompanied by a corresponding loss of generalization.

Taken together, these findings indicate that the fine-tuned encoder has transitioned from a generalist multilingual representation model to a domain-specialized retriever tailored to the information needs of the O&G drilling context. While this specialization renders the model unsuitable for open-domain retrieval without further adaptation, it directly enables the efficiency and effectiveness observed within the target domain. Accordingly, the reduction in general-domain performance should be understood not as a limitation of the modeling approach, but as an expected and theoretically grounded consequence of a domain-focused optimization strategy.

4.3.1 Computational Efficiency and Environmental Impact

While retrieval effectiveness is the primary objective of this study, practical deployment in industrial Oil and Gas environments requires careful consideration of computational efficiency. These settings are often characterized by constrained hardware resources, strict latency requirements, and the need to index and query millions of legacy documents. In addition, in alignment with the principles of Green AI (SCHWARTZ et al., 2019), model efficiency is treated not as a secondary concern, but as a first-class evaluation criterion alongside retrieval quality.

4.3.1.1 Computational Efficiency Analysis

To assess deployability, we benchmarked the proposed model against representative baseline architectures using a hardware environment equipped with a single NVIDIA A40 GPU. Three complementary dimensions were evaluated: Indexing Throughput (documents encoded per second), Inference Latency (time to encode a single query), and Peak GPU Memory Utilization. The results are summarized in Table 11.

The results highlight a trade-off between model scale and operational agility. The state-of-the-art LLM-based retriever, *bge-multilingual-gemma2*, while offering strong zero-shot performance, incurs a prohibitive computational cost. With over 9 billion parameters, it achieves a throughput of only ≈ 17 documents per second. Indexing a typical industrial corpus of 1 million documents with this model would require approximately 16 hours on a high-end A40 GPU. In stark contrast, our domain-adapted model demonstrates exceptional efficiency. With

Model	VRAM (GB)*	Throughput (Docs/s)	Latency (ms)	P99 Latency (ms)
bge-multilingual-gemma2	39.82	16.98	35.28	43.72
multilingual-e5-large-instruct	4.42	431.23	7.59	8.50
e5-large	2.33	388.19	7.41	8.31
mcontriever-base	0.99	1,117.52	4.15	6.96
Our Model (DistilBERT)	0.37	1,903.09	2.50	3.49

Table 11 – Computational efficiency benchmark on NVIDIA A40. *Note: VRAM usage for the smaller models reflects the peak system allocation during the batched benchmarking process rather than the minimum memory requirement for deployment, which is strictly correlated with parameter count.

only 66 million parameters, approximately 0.7% of the size of the gemma baseline, it achieves an indexing throughput of 1,903 documents per second. This represents a $112\times$ speedup over the LLM-based baseline and a $4.4\times$ speedup over the *multilingual-e5-large* model. Crucially, the inference latency is reduced to 2.50 ms per query, enabling real-time search applications with negligible lag.

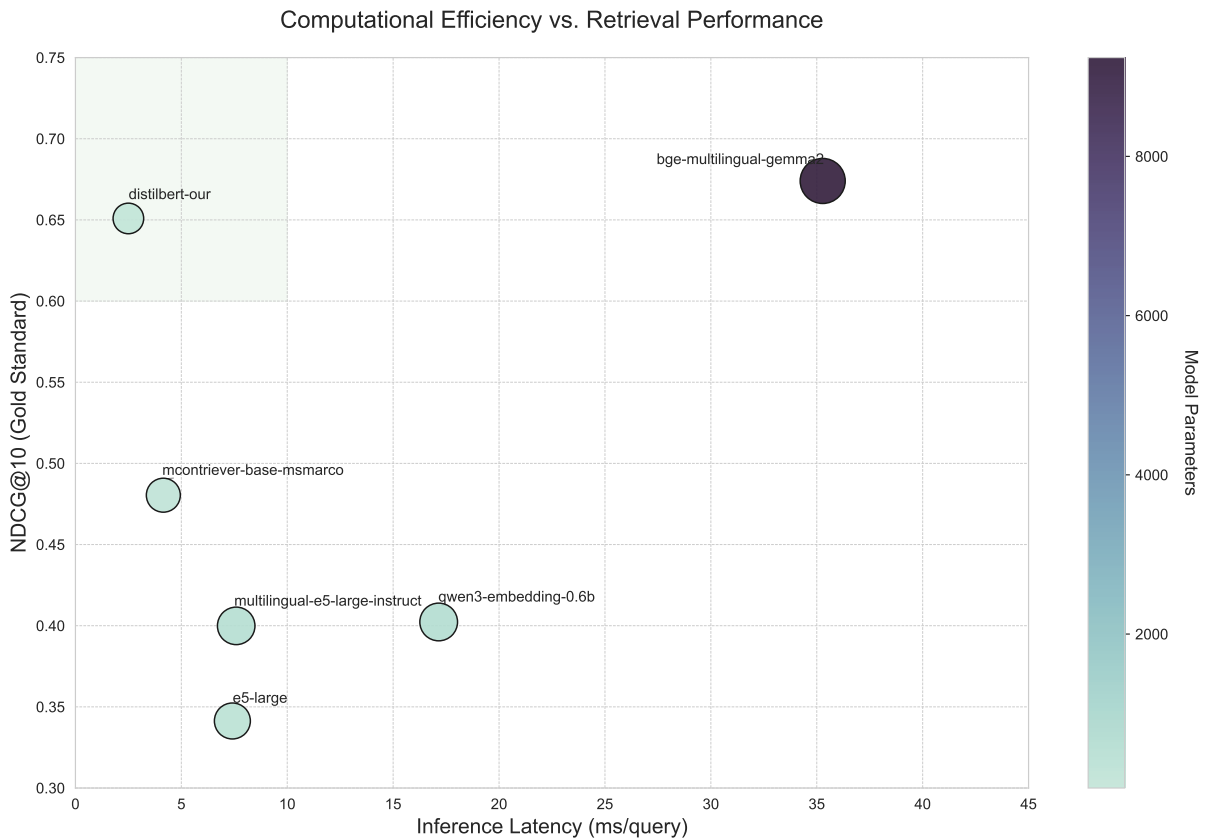


Figure 11 – Pareto Frontier of Retrieval Effectiveness vs. Inference Latency.

Figure 11 presents a Pareto frontier relating retrieval effectiveness to inference latency. The proposed model occupies the optimal region of this trade-off, achieving performance comparable to the large-scale LLM baseline while being an order of magnitude faster and several orders of magnitude smaller. This result reinforces the central thesis of this dissertation:

domain adaptation and structurally informed training signals can outperform brute-force parameter scaling in efficiency-critical industrial settings.

4.3.1.2 Environmental Impact Analysis

Beyond operational efficiency, we estimated the carbon footprint associated with the fine-tuning of our models. Our methodology adopts the framework proposed by Lacoste et al. (2019), updated with the specific emission factors for the Brazilian National Interconnected System (SIN) for the year 2025. The carbon footprint is calculated based on the total energy consumption (E) and the carbon intensity of the local grid (I):

$$\text{CO}_2\text{e} = E \times I = (P_{avg} \times t \times PUE) \times I \quad (35)$$

Where P_{avg} is the average power draw of the hardware (conservatively estimated at 300W for the NVIDIA A40 during fine-tuning), t is the training duration, and PUE (Power Usage Effectiveness) is set to 1.58, a standard average for data centers. Unlike previous years where the average emission factor was lower, the updated 2025 factors from the Ministry of Science, Technology, and Innovation (MCTI) reflect a period of higher carbon intensity, likely due to hydrological variations requiring increased thermal generation. The factors ranged from 0.2146 kg/kWh in March to a peak of 0.5780 kg/kWh in November.

We conducted our experiments primarily during November, corresponding to an emission factor of 0.5780 kg/kWh. With a total cumulative training time of approximately 20 hours for the DistilBERT model. Total emissions are estimated to be 3.47 kgCO₂eq of which 0 percents were directly offset. Estimations were conducted using the MachineLearning Impact calculator. This is the equivalent of burning 1.73 kg of coal or driving an average ICE car for 14 km. To contextualize, training a large-scale model would require hundreds of GPU-hours, resulting in a carbon footprint significantly larger. Our approach’s low footprint, demonstrates that high-performance domain adaptation can be achieved sustainably. By utilizing efficient architectures and targeted fine-tuning strategies, we minimize the environmental cost while maximizing retrieval effectiveness for the Oil and Gas domain.

4.3.2 Summary of Empirical Findings

Across all evaluation settings, three consistent empirical patterns emerge. First, zero-shot dense retrievers trained on general-purpose corpora exhibit limitations when applied to highly technical industrial data, often underperforming strong lexical baselines. Second, domain adaptation through synthetic query generation improves retrieval effectiveness, particularly when generation strategies enforce semantic grounding. Third, incorporating graph-derived structural constraints during training yields more robust performance at deeper ranking positions and under exhaustive evaluation, without sacrificing computational efficiency. These findings

motivate the discussion in the following chapter, where their broader implications, limitations, and generalizability are examined.

5 Discussion

This dissertation set out to investigate whether incorporating explicit structural information, derived from a domain ontology, can improve document retrieval performance in low-supervision industrial settings. The results presented in Chapter 4 indicate that structure-aware training strategies consistently outperform text-only baselines across multiple evaluation metrics and annotation regimes, particularly at deeper ranking positions. Rather than demonstrating a universal superiority of any single model architecture, these findings highlight the importance of training signal design when operating under domain and supervision constraints.

The central thesis of this research is that, for information retrieval in specialized technical domains, incorporating intrinsic structural knowledge via heterogeneous information networks can be more effective for representation learning than increasing model scale or relying exclusively on general-purpose semantic representations. The empirical results suggest that a domain-adapted bi-encoder, refined through structure-aware negative sampling, achieves a more favorable balance between alignment and uniformity in the latent space when compared to larger, non-specialized architectures. This effect is most evident on complex queries characterized by vocabulary mismatch and technical shorthand, which have historically posed challenges for dense retrieval models.

The initial findings regarding the superiority of the fine tuned DistilBERT model over the BM25 baseline indicate a successful bridging of the lexical semantic gap. While lexical models like BM25 are robust in handling technical jargon when exact matches exist, they fail to capture the nuanced operational relationships inherent in Daily Drilling Reports. The fact that a relatively small transformer model achieved an NDCG@10 of 0.6509, surpassing the lexical baseline, indicates that the learned representations extend beyond surface-level lexical matching. These results are consistent with an adaptation to the specialized syntactic and contextual patterns of the Oil and Gas domain, enabling the model to distinguish between operational concepts based on contextual usage rather than term frequency alone.

This adaptation is fundamentally driven by the use of the Heterogeneous Information Network as a source of supervision. A potential counterargument to this approach might suggest that synthetic data generation via Large Language Models, as seen in frameworks like InPars or UDAPDR, would provide a more diverse training signal. However, the results suggest that leveraging the intrinsic topology of the data provides a more constrained and semantically grounded signal for negative mining than purely text-based synthetic generation strategies. By utilizing graph based hard negatives, the model was forced to resolve semantic ambiguities between documents that share similar metadata hubs but describe distinct operational states. The geometric analysis of the latent space is consistent with this interpretation, indicating that

the model achieves a favorable balance between alignment and uniformity without exhibiting representation collapse. This indicates that the structure aware sampling successfully clustered relevant documents without causing representation collapse, which is a common failure mode in unsupervised dense retrieval.

Furthermore, the query level analysis reinforces the significance of domain specific structural knowledge. The observation that the proposed model outperformed the 9B parameter Gemma based model on hard queries, specifically those involving high vocabulary mismatch, suggests that, within this domain and task setting, increased model scale alone does not substitute for explicit domain context. Large scale models, while linguistically capable, often rely on general semantic proximities that do not hold in high stakes technical environments. In contrast, our model utilized the structural relationships defined by the RIPPO ontology to identify synonymy and quantitative associations that are invisible to general purpose encoders. For instance, the ability to link the term drag to its corresponding unit of measure is consistent with the model internalizing relational constraints encoded in the knowledge graph during training.

The practical implications of these findings are particularly relevant for large-scale industrial retrieval systems in the energy sector. Given that the proposed architecture is significantly faster at indexing than LLM based alternatives, it offers a scalable solution for processing millions of legacy technical reports. This efficiency is achieved without a corresponding degradation in retrieval effectiveness under the evaluated conditions. While it is acknowledged that late interaction models currently maintain a higher zero shot performance due to their token level interaction mechanisms, the computational overhead of such models often precludes their use in large scale industrial applications.

Despite these advancements, certain limitations must be addressed to contextualize the scope and generalizability of this work. A primary constraint lies in the reliance on the Silver Standard relevance set. While this automated labeling approach allowed for the evaluation of retrieval performance at a scale previously inaccessible through manual annotation, it introduces a potential margin of error. The leniency observed in the LLM based labeling phase may result in the inclusion of false positive relevance signals, which can introduce uncertainty in absolute performance estimates and partially obscure fine-grained differences between retrieval models. Furthermore, the current scope is strictly limited to the drilling sub field of the Oil and Gas industry. Extending this methodology to other domains, such as geophysics, reservoir engineering, or offshore production, would require a significant effort in ontological alignment to ensure that the Heterogeneous Information Network correctly captures the unique relational schemas of those disciplines.

An additional limitation concerns the degree of domain specialization induced by the proposed training strategy. As discussed in Section 4.3, fine-tuning the encoder exclusively on drilling-related data without constraints aimed at preserving general-domain semantics resulted

in a substantial degradation of open-domain retrieval performance. While this behavior is theoretically consistent with the plasticity–stability trade-off observed in domain adaptation and continual learning, it nonetheless restricts the applicability of the resulting model to the target domain. In its current form, the retriever should therefore be regarded as a domain-specialized system rather than a general-purpose multilingual encoder. Applications requiring both broad semantic coverage and domain-specific precision would necessitate additional mechanisms, such as multi-domain training, parameter isolation, or regularization strategies designed to mitigate representation drift.

From an architectural perspective, the choice of a bi encoder involves an inherent trade off regarding the information bottleneck. By compressing a document into a single dense vector, the model loses the ability to perform fine grained, token level cross attention during the retrieval phase. This limitation helps explain why late-interaction models such as ColBERT often exhibit stronger zero-shot performance in certain complex retrieval scenarios. Although our structure aware negative sampling partially compensates for this by forcing the encoder to learn higher order semantic boundaries, the model remains bounded by its inability to model explicit query document token interactions at inference time. Furthermore, the evaluation was conducted in an offline setting using static benchmarks. Real world performance is likely to be influenced by user specific search behaviors, varying levels of expertise among engineers, and the dynamic nature of operational logs that are updated in real time.

Future research should focus on mitigating these limitations through several promising avenues. Integrating a cross encoder re ranker on top of the initial bi encoder retrieval results could provide the necessary token level precision while maintaining scalable indexing speeds. Furthermore, the exploration of Deep Graph Infomax (DGI) tailored for heterogeneous graphs represents a significant theoretical extension. By maximizing the Mutual Information between local node patches and a global graph summary, the model could learn to capture high level community structures and global semantic information that transcends local adjacency. This direction may be particularly valuable for capturing long-range operational patterns that are not readily observable through local graph neighborhoods alone.

Additionally, the integration of this retriever into Retrieval Augmented Generation systems could facilitate the development of technical chatbots for the Oil and Gas sector. Such systems would benefit from the enhanced grounding provided by the RIPPO ontology, ensuring that generated answers are not only linguistically coherent but also operationally accurate. Future iterations of this work should also explore the use of distinct embedding spaces for different modalities to optimize information density. Finally, conducting longitudinal A/B testing with domain experts in a live industrial environment would be essential to quantify the actual productivity gains and decision support utility provided by knowledge enhanced dense retrieval.

6 Conclusion

The rapid integration of artificial intelligence into industrial sectors has indicated significant potential in how organizations manage and retrieve knowledge. In the context of the Oil and Gas industry, where decision-making is critical and data volume is overwhelming, the ability to instantly access precise technical information is not merely a convenience but an operational necessity. However, this dissertation has demonstrated that the transition from theoretical AI capabilities to practical, high-stakes application is characterized by complexities that general-purpose models cannot resolve in isolation. While modern Deep Learning architectures, particularly those based on Transformers, have achieved remarkable feats in general language understanding, this research highlights a dense retrieval gap when such models are applied to highly specialized technical domains, abbreviated syntax, and rigorous semantic constraints of drilling operations. The idealized concept of a “one-size-fits-all” model fails when faced with the reality of Daily Drilling Reports, where the distinction between a “loss of circulation” and a “fluid influx” is not merely semantic but a matter of operational safety. Consequently, the central argument of this research has been that domain adaptation is not an optional refinement but a fundamental requirement for the viability of neural search in technical industries.

The investigation began by exposing the inadequacy of traditional evaluation frameworks for this specific domain. A significant finding of this work was that standard pooling methods, relied upon by decades of Information Retrieval research, are insufficient when expert annotators are scarce and the relevant documents are hidden within massive, unstructured corpora. The construction of the test collection revealed that without rigorous, human-in-the-loop methodologies, the performance of novel retrieval systems is often underestimated. The development of the Silver Standard dataset, generated through a classifier trained on expert judgements, represents a methodological contribution that extends beyond the specific results of this thesis. It challenges the conventional reliance on shallow pooling depths and demonstrates that combining human expertise with automated classification can uncover hidden relevance, providing a fairer and more robust benchmark for future domain-specific models. This rigorous evaluation framework was essential to proving that the limitations of standard pre-trained models were not artifacts of the data, but genuine limitations in their semantic understanding.

Addressing these limitations required a fundamental shift in how training data is conceptualized. This dissertation proposed that the solution to the scarcity of labeled training data lies not in massive generic corpora or purely synthetic generation, but in the intrinsic topological structure of the data itself. By constructing a domain-specific Knowledge Graph, rooted in the RIPPO ontology, we were able to transform unstructured text into a structured Heterogeneous Information Network. This transformation was pivotal; it allowed us to formulate

the retrieval problem not just as text matching, but as a graph learning task. The introduction of the Heterogeneous Graph Attention Network (HeteroGAT) to mine “structure-aware” hard negative samples constitutes the technical core of this work. This mechanism forced the retrieval model to learn fine-grained distinctions, distinguishing between reports that are superficially similar but structurally distinct, thereby injecting domain expertise directly into the embedding space without the high cost of manual annotation. The success of this approach provides empirical support for the hypothesis that structural knowledge can serve as an effective supervision signal, effectively guiding the model to understand the causal and hierarchical relationships that define the engineering domain.

The empirical results of this dissertation provide strong evidence for the efficacy of the proposed graph-enhanced approach within the evaluated domain and experimental setting. The fine-tuned bi-encoder model did not merely improve upon baselines; it significantly improved retrieval performance for hard queries, particularly those characterized by technical jargon and ambiguity where traditional lexical models fail. The analysis revealed that the model achieved a superior balance of alignment and uniformity in the vector space, pulling semantically related operational contexts together while maintaining the discriminative power necessary to separate distinct events. Crucially, this performance was achieved not by scaling up the model parameters to billions, but by refining the quality of the learning signal. This finding has profound implications for the deployment of AI in resource-constrained industrial environments. We demonstrated that a compact, well-adapted model (66 million parameters) could outperform large-scale state-of-the-art Large Language Models (9 billion parameters) while being more than $100\times$ faster and significantly more energy-efficient. This aligns with the principles of green AI, proving that high performance does not strictly require high computational cost, provided the training methodology is sufficiently sophisticated.

However, the development of a fully autonomous semantic search system is ongoing. While this study successfully addressed the retrieval effectiveness in an offline setting, the complexities of real-time user interaction and the dynamic nature of drilling data present new frontiers. The limitation of the bi-encoder architecture in capturing token-level interactions suggests that future research could explore hybrid systems that combine the speed of our proposed model with the granular precision of late-interaction architectures or cross-encoders. Furthermore, the methodology established here opens the door for integration into broader generative workflows. The retrieval system developed in this work is ideally positioned to serve as the retrieval module in Retrieval-Augmented Generation systems, potentially enabling chatbots that can not only find documents but synthesize technical answers with high fidelity.

In conclusion, this dissertation has moved beyond the initial expectations of general artificial intelligence to address the complex practicalities of industrial application. It has established that while the promise of AI is vast, its value is realized only through deep, structural adaptation to the domain. By bridging the gap between unstructured narrative

text and structured engineering knowledge, this research proposes a scalable and efficient framework grounded in rigorous empirical evaluation. The methodologies developed herein offer a framework for how technical domains can harness the power of deep learning, ensuring that the vast archives of historical data become active assets in driving future innovation and operational excellence.

Bibliography

AGOSTI, M.; MARCHESIN, S.; SILVELLO, G. Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval. *ACM Transactions on Information Systems (TOIS)*, v. 38, n. 4, p. 1–48, 2020.

AMATI, G.; RIJSBERGEN, C. J. V. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, v. 20, n. 4, p. 357–389, 2002. ISSN 1046-8188. Disponível em: <<https://doi.org/10.1145/582415.582416>>.

ASIF, N. A.; SARKER, Y.; CHAKRABORTTY, R. K.; RYAN, M. J.; AHAMED, M. H.; SAHA, D. K.; BADAL, F. R.; DAS, S. K.; ALI, M. F.; MOYEEN, S. I.; ISLAM, M. R.; TASNEEM, Z. Graph neural network: A comprehensive review on non-euclidean space. *IEEE Access*, v. 9, p. 60588–60606, 2021.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. 2011. Disponível em: <<https://www2.dcc.ufmg.br/livros/irbook/>>.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, v. 284, n. 5, p. 34–43, 2001.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, v. 3, n. null, p. 993–1022, mar. 2003. ISSN 1532-4435.

BODENREIDER, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, v. 32, n. Database-Issue, p. 267–270, 2004. Disponível em: <<http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04>>.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. *Enriching Word Vectors with Subword Information*. arXiv, 2017. ArXiv:1607.04606. Disponível em: <<http://arxiv.org/abs/1607.04606>>.

BONIFACIO, L.; ABONIZIO, H.; FADAE, M.; NOGUEIRA, R. Inpars: Unsupervised dataset generation for information retrieval. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2022. (SIGIR '22), p. 2387–2392. ISBN 9781450387323. Disponível em: <<https://doi.org/10.1145/3477495.3531863>>.

BORDES, A.; USUNIER, N.; GARCIA-DURAN, A.; WESTON, J.; YAKHNENKO, O. Translating embeddings for modeling multi-relational data. In: BURGESS, C.; BOTTOU, L.; WELLING, M.; GHAHRAMANI, Z.; WEINBERGER, K. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013. v. 26. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf>.

BOTEVA, V.; GHOLIPOUR, D.; SOKOLOV, A.; RIEZLER, S. A full-text learning to rank dataset for medical information retrieval. In: . [s.n.], 2016. Disponível em: <<http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf>>.

BRONSTEIN, M. M.; BRUNA, J.; LECUN, Y.; SZLAM, A.; VANDERGHEYNST, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, v. 34, n. 4, p. 18–42, 2017.

BUENO, M.; OLIVEIRA, E. S. de; NOGUEIRA, R.; LOTUFO, R. A.; PEREIRA, J. A. *Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers*. arXiv, 2024. ArXiv:2404.06976 [cs]. Disponível em: <<http://arxiv.org/abs/2404.06976>>.

BURGES, C.; SHAKED, T.; RENSHAW, E.; LAZIER, A.; DEEDS, M.; HAMILTON, N.; HULLENDER, G. Learning to rank using gradient descent. In: *Proceedings of the 22nd international conference on Machine learning*. New York, NY, USA: Association for Computing Machinery, 2005. (ICML '05), p. 89–96. ISBN 978-1-59593-180-1. Disponível em: <<https://doi.org/10.1145/1102351.1102363>>.

CAO, G.; NIE, J.-Y.; GAO, J.; ROBERTSON, S. Selecting good expansion terms for pseudo-relevance feedback. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 2008. (SIGIR '08), p. 243–250. ISBN 978-1-60558-164-4. Disponível em: <<https://doi.org/10.1145/1390334.1390377>>.

CHANG, W.-C.; YU, F. X.; CHANG, Y.-W.; YANG, Y.; KUMAR, S. *Pre-training Tasks for Embedding-based Large-scale Retrieval*. arXiv, 2020. ArXiv:2002.03932. Disponível em: <<http://arxiv.org/abs/2002.03932>>.

CINELLI, L. P.; de Oliveira, J. F.; de Pinho, V. M.; PASSOS, W. L.; PADILLA, R.; BRAZ, P. F.; GALVES, B.; DALVI, D. P.; LEWENFUS, G.; FERREIRA, J. O.; JI, A. Y.; de Oliveira, F. L.; GONÇALVES, C. J.; NETTO, S. L.; da Silva, E. A.; de Campos, M. L. Automatic event identification and extraction from daily drilling reports using an expert system and artificial intelligence. *Journal of Petroleum Science and Engineering*, v. 205, p. 108939, 2021. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0920410521005982>>.

CLINCHANT, S.; PERRONNIN, F. Aggregating Continuous Word Embeddings for Information Retrieval. In: ALLAUZEN, A.; LAROCHELLE, H.; MANNING, C.; SOCHER, R. (Ed.). *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 100–109. Disponível em: <<https://aclanthology.org/W13-3212>>.

CORDEIRO, F. C.; SILVA, P. F. d.; GOMES, D. d. S. M.; SOUZA, R. R.; COELHO, F. C.; ELL, B. SSRN Scholarly Paper, *Petro Kgraph: A Methodology for Extracting Knowledge Graph from Technical Documents - an Application in the Oil and Gas Industry*. Rochester, NY: [s.n.], 2024. Disponível em: <<https://papers.ssrn.com/abstract=4776804>>.

DAI, Z.; CALLAN, J. Context-Aware Term Weighting For First Stage Passage Retrieval. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR '20), p. 1533–1536. ISBN 978-1-4503-8016-4. Disponível em: <<https://dl.acm.org/doi/10.1145/3397271.3401204>>.

DAI, Z.; XIONG, C.; CALLAN, J.; LIU, Z. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2018. (WSDM '18), p. 126–134. ISBN 978-1-4503-5581-0. Disponível em: <<https://dl.acm.org/doi/10.1145/3159652.3159659>>.

DAI, Z.; ZHAO, V. Y.; MA, J.; LUAN, Y.; NI, J.; LU, J.; BAKALOV, A.; GUU, K.; HALL, K.; CHANG, M.-W. Promptagator: Few-shot dense retrieval from 8 examples. In: *The Eleventh*

International Conference on Learning Representations. [s.n.], 2023. Disponível em: <<https://openreview.net/forum?id=gml46Ympu2J>>.

DAVIS, R.; SHROBE, H.; SZOLOVITS, P. What Is a Knowledge Representation? *AI Magazine*, v. 14, n. 1, p. 17–17, mar. 1993. ISSN 2371-9621. Number: 1. Disponível em: <<https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1029>>.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990. ISSN 1097-4571. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>>.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: <<http://arxiv.org/abs/1810.04805>>.

DIGGELMANN, T.; BOYD-GRABER, J. L.; BULIAN, J.; CIARAMITA, M.; LEIPPOLD, M. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614, 2020. Disponível em: <<https://arxiv.org/abs/2012.00614>>.

EFRON, M.; ORGANISCIK, P.; FENLON, K. Improving retrieval of short texts through document expansion. In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. [s.n.], 2012. Disponível em: <<https://api.semanticscholar.org/CorpusID:11138893>>.

FAN, M.; ZHOU, Q.; CHANG, E.; ZHENG, T. F. Transition-based knowledge graph embedding with relational mapping properties. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Phuket, Thailand: Department of Linguistics, Chulalongkorn University, 2014. p. 328–337. Disponível em: <<https://aclanthology.org/Y14-1039/>>.

FEI, H.; REN, Y.; ZHANG, Y.; JI, D.; LIANG, X. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, v. 22, n. 3, p. bbaa110, maio 2021. ISSN 1467-5463, 1477-4054. Disponível em: <<https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa110/5854405>>.

FELDMAN, Y.; EL-YANIV, R. Multi-Hop Paragraph Retrieval for Open-Domain Question Answering. In: KORHONEN, A.; TRAUM, D.; MÀRQUEZ, L. (Ed.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 2296–2309. Disponível em: <<https://aclanthology.org/P19-1222>>.

FENG, J.; HUANG, M.; WANG, M.; ZHOU, M.; HAO, Y.; ZHU, X. Knowledge graph embedding by flexible translation. In: *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*. [S.l.]: AAAI Press, 2016. (KR'16), p. 557–560.

FORMAL, T.; LASSANCE, C.; PIWOWARSKI, B.; CLINCHANT, S. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. *arXiv*, 2021. ArXiv:2109.10086. Disponível em: <<http://arxiv.org/abs/2109.10086>>.

FORMAL, T.; PIWOWARSKI, B.; CLINCHANT, S. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*. arXiv, 2021. ArXiv:2107.05720. Disponível em: <<http://arxiv.org/abs/2107.05720>>.

FREJ, J.; MULHEM, P.; SCHWAB, D.; CHEVALLET, J.-P. Learning Term Discrimination. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR '20), p. 1993–1996. ISBN 978-1-4503-8016-4. Disponível em: <<https://doi.org/10.1145/3397271.3401211>>.

GAO, L.; DAI, Z.; CALLAN, J. *COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List*. arXiv, 2021. ArXiv:2104.07186. Disponível em: <<http://arxiv.org/abs/2104.07186>>.

GLOROT, X.; BORDES, A.; WESTON, J.; BENGIO, Y. *A Semantic Matching Energy Function for Learning with Multi-relational Data*. arXiv, 2013. ArXiv:1301.3485. Disponível em: <<http://arxiv.org/abs/1301.3485>>.

GOMES, D. d. S. M. *PetroVec: desenvolvimento e avaliação de modelos vetoriais de palavras em português para o domínio de óleo e gás*. Tese (Tese (Doutorado em Engenharia Civil)) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2021. COPPE. Disponível em: <<https://pantheon.ufrj.br/handle/11422/27739>>.

GUO, J.; CAI, Y.; FAN, Y.; SUN, F.; ZHANG, R.; CHENG, X. Semantic Models for the First-stage Retrieval: A Comprehensive Review. *ACM Transactions on Information Systems*, v. 40, n. 4, p. 1–42, out. 2022. ISSN 1046-8188, 1558-2868. ArXiv:2103.04831 [cs]. Disponível em: <<http://arxiv.org/abs/2103.04831>>.

GUO, J.; FAN, Y.; AI, Q.; CROFT, W. B. A Deep Relevance Matching Model for Ad-hoc Retrieval. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2016. (CIKM '16), p. 55–64. ISBN 978-1-4503-4073-1. Disponível em: <<https://doi.org/10.1145/2983323.2983769>>.

GUO, K.; HUANG, R. A Chinese Knowledge Graph Q&A System Based on Dense Relationship Retrieval. In: *2023 IEEE 3rd International Conference on Software Engineering and Artificial Intelligence (SEAI)*. [s.n.], 2023. p. 148–153. Disponível em: <<https://ieeexplore.ieee.org/document/10217649>>.

GYSEL, C. V.; RIJKE, M. d.; KANOULAS, E. *Neural Vector Spaces for Unsupervised Information Retrieval*. arXiv, 2018. ArXiv:1708.02702. Disponível em: <<http://arxiv.org/abs/1708.02702>>.

HAMBARDE, K. A.; PROENCA, H. *Information Retrieval: Recent Advances and Beyond*. arXiv, 2023. ArXiv:2301.08801 [cs]. Disponível em: <<http://arxiv.org/abs/2301.08801>>.

HAMILTON, W. L.; YING, R.; LESKOVEC, J. Inductive representation learning on large graphs. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 1025–1035. ISBN 9781510860964.

HE, X.; TIAN, Y.; SUN, Y.; CHAWLA, N. V.; LAURENT, T.; LECUN, Y.; BRESSON, X.; HOOL, B. *G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and*

Question Answering. arXiv, 2024. ArXiv:2402.07630. Disponível em: <<http://arxiv.org/abs/2402.07630>>.

HOOGEVEEN, D.; VERSPOOR, K. M.; BALDWIN, T. Cqadupstack: A benchmark data set for community question-answering research. In: *Proceedings of the 20th Australasian Document Computing Symposium (ADCS)*. New York, NY, USA: ACM, 2015. (ADCS '15), p. 3:1–3:8. ISBN 978-1-4503-4040-3. Disponível em: <<http://doi.acm.org/10.1145/2838931.2838934>>.

HUANG, P.-S.; HE, X.; GAO, J.; DENG, L.; ACERO, A.; HECK, L. Learning deep structured semantic models for web search using clickthrough data. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2013. (CIKM '13), p. 2333–2338. ISBN 978-1-4503-2263-8. Disponível em: <<https://doi.org/10.1145/2505515.2505665>>.

IZACARD, G.; CARON, M.; HOSSEINI, L.; RIEDEL, S.; BOJANOWSKI, P.; JOULIN, A.; GRAVE, E. *Unsupervised Dense Information Retrieval with Contrastive Learning*. 2022. Disponível em: <<https://arxiv.org/abs/2112.09118>>.

JANG, K.-R.; KANG, J.; HONG, G.; MYAENG, S.-H.; PARK, J.; YOON, T.; SEO, H. Ultra-High Dimensional Sparse Representations with Binarization for Efficient Text Retrieval. In: MOENS, M.-F.; HUANG, X.; SPECIA, L.; YIH, S. W.-t. (Ed.). *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 1016–1029. Disponível em: <<https://aclanthology.org/2021.emnlp-main.78>>.

JHA, R.; WANG, B.; GÜNTHER, M.; MASTRAPAS, G.; STURUA, S.; MOHR, I.; KOUKOUNAS, A.; WANG, M. K.; WANG, N.; XIAO, H. Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever. In: SÄLEVÄ, J.; OWODUNNI, A. (Ed.). *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*. Miami, Florida, USA: Association for Computational Linguistics, 2024. p. 159–166. Disponível em: <<https://aclanthology.org/2024.mrl-1.11/>>.

JI, G.; HE, S.; XU, L.; LIU, K.; ZHAO, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In: ZONG, C.; STRUBE, M. (Ed.). *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015. p. 687–696. Disponível em: <<https://aclanthology.org/P15-1067>>.

JI, G.; LIU, K.; HE, S.; ZHAO, J. Knowledge Graph Completion with Adaptive Sparse Transfer Matrix. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 30, n. 1, fev. 2016. ISSN 2374-3468. Number: 1. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/10089>>.

JI, S.; SHAO, J.; YANG, T. Efficient Interaction-based Neural Ranking with Locality Sensitive Hashing. In: *The World Wide Web Conference*. New York, NY, USA: Association for Computing Machinery, 2019. (WWW '19), p. 2858–2864. ISBN 978-1-4503-6674-8. Disponível em: <<https://dl.acm.org/doi/10.1145/3308558.3313576>>.

JU, J.-H.; YANG, H. C.-H.; FU, S.-W.; TSAI, M.-F.; WANG, C.-J. Relevance-aware diverse query generation for out-of-domain text ranking. In: ZHAO, C.; MOSBACH, M.; ATANASOVA, P.; GOLDFARB-TARRENT, S.; HASE, P.; HOSSEINI, A.; ELBAYAD, M.; PEZZELLE, S.; MOZES, M. (Ed.). *Proceedings of the 9th Workshop on Representation Learning for NLP*

(*RepL4NLP-2024*). Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 26–36. Disponível em: <<https://aclanthology.org/2024.repl4nlp-1.3/>>.

KARPUKHIN, V.; OĞUZ, B.; MIN, S.; LEWIS, P.; WU, L.; EDUNOV, S.; CHEN, D.; YIH, W.-t. *Dense Passage Retrieval for Open-Domain Question Answering*. arXiv, 2020. ArXiv:2004.04906. Disponível em: <<http://arxiv.org/abs/2004.04906>>.

KENTER, T.; RIJKE, M. de. Short Text Similarity with Word Embeddings. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2015. (CIKM '15), p. 1411–1420. ISBN 978-1-4503-3794-6. Disponível em: <<https://doi.org/10.1145/2806416.2806475>>.

KHATTAB, O.; ZAHARIA, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR '20), p. 39–48. ISBN 978-1-4503-8016-4. Disponível em: <<https://dl.acm.org/doi/10.1145/3397271.3401075>>.

KHOT, T.; CLARK, P.; GUERQUIN, M.; JANSEN, P.; SABHARWAL, A. QASC: A dataset for question answering via sentence composition. *CoRR*, abs/1910.11473, 2019. Disponível em: <<http://arxiv.org/abs/1910.11473>>.

KIPF, T. N.; WELLING, M. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. Disponível em: <<http://arxiv.org/abs/1609.02907>>.

KURLAND, O.; LEE, L. Corpus structure, language models, and ad hoc information retrieval. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 2004. (SIGIR '04), p. 194–201. ISBN 978-1-58113-881-8. Disponível em: <<https://doi.org/10.1145/1008992.1009027>>.

KWIATKOWSKI, T.; PALOMAKI, J.; REDFIELD, O.; COLLINS, M.; PARIKH, A.; ALBERTI, C.; EPSTEIN, D.; POLOSUKHIN, I.; DEVLIN, J.; LEE, K.; TOUTANOVA, K.; JONES, L.; KELCEY, M.; CHANG, M.-W.; DAI, A. M.; USZKOREIT, J.; LE, Q.; PETROV, S. 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. Disponível em: <<https://aclanthology.org/Q19-1026>>.

LACOSTE, A.; LUCCIONI, A.; SCHMIDT, V.; DANDRES, T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

LASSANCE, C.; DÉJEAN, H.; FORMAL, T.; CLINCHANT, S. *SPLADE-v3: New baselines for SPLADE*. arXiv, 2024. ArXiv:2403.06789. Disponível em: <<http://arxiv.org/abs/2403.06789>>.

LE, Q. V.; MIKOLOV, T. *Distributed Representations of Sentences and Documents*. arXiv, 2014. ArXiv:1405.4053. Disponível em: <<http://arxiv.org/abs/1405.4053>>.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P.; HELLMANN, S.; MORSEY, M.; KLEEF, P. van; AUER, S.; BIZER, C. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. In: . [S.l.]: Semantic Web Journal 6, 2015. p. 167–195.

LESK, M. E. Word-word associations in document retrieval systems. *American Documentation*, v. 20, n. 1, p. 27–38, 1969. ISSN 1936-6108. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630200106>. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630200106>.

LI, J.; ZENG, H.; PENG, L.; ZHU, J.; LIU, Z. Learning to rank method combining multi-head self-attention with conditional generative adversarial nets. *Array*, v. 15, p. 100205, set. 2022. ISSN 2590-0056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2590005622000522>.

LIN, J.; MA, X. *A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques*. arXiv, 2021. ArXiv:2106.14807. Disponível em: <http://arxiv.org/abs/2106.14807>.

LIN, Y.; LIU, Z.; SUN, M.; LIU, Y.; ZHU, X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 29, n. 1, fev. 2015. ISSN 2374-3468. Number: 1. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/9491>.

LIU, J.; LIU, J.; YANG, Y.; WANG, J.; WU, W.; ZHAO, D.; YAN, R. *GNN-encoder: Learning a Dual-encoder Architecture via Graph Neural Networks for Dense Passage Retrieval*. arXiv, 2022. ArXiv:2204.08241 [cs]. Disponível em: <http://arxiv.org/abs/2204.08241>.

LIU, J.; YAN, J.; ZHU, C.; LIU, X.; QING, L.; RAO, Y. Neural topic modeling via contextual and graph information fusion. In: CHRISTODOULOPOULOS, C.; CHAKRABORTY, T.; ROSE, C.; PENG, V. (Ed.). *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025. p. 13247–13263. ISBN 979-8-89176-332-6. Disponível em: <https://aclanthology.org/2025.emnlp-main.670/>.

LIU, X.; CROFT, W. B. Cluster-based retrieval using language models. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 2004. (SIGIR '04), p. 186–193. ISBN 978-1-58113-881-8. Disponível em: <https://doi.org/10.1145/1008992.1009026>.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. Disponível em: <http://arxiv.org/abs/1907.11692>.

LU, Z.; DU, P.; NIE, J.-Y. *VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification*. arXiv, 2020. ArXiv:2004.05707 [cs, stat]. Disponível em: <http://arxiv.org/abs/2004.05707>.

MACAVANEY, S.; YATES, A.; COHAN, A.; GOHARIAN, N. CEDR: Contextualized Embeddings for Document Ranking. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2019. (SIGIR'19), p. 1101–1104. ISBN 978-1-4503-6172-9. Disponível em: <https://doi.org/10.1145/3331184.3331317>.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. 2008. ISBN: 9780511809071 Publisher: Cambridge University Press. Disponível em: <https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C>.

MITRA, B.; DIAZ, F.; CRASWELL, N. *Learning to Match Using Local and Distributed Representations of Text for Web Search*. arXiv, 2016. ArXiv:1610.08136. Disponível em: <<http://arxiv.org/abs/1610.08136>>.

MITRA, B.; NALISNICK, E.; CRASWELL, N.; CARUANA, R. *A Dual Embedding Space Model for Document Ranking*. arXiv, 2016. ArXiv:1602.01137. Disponível em: <<http://arxiv.org/abs/1602.01137>>.

MITRA, M.; BUCKLEY, C.; SINGHAL, A.; CARDIE, C. An analysis of statistical and syntactic phrases. In: *Computer-Assisted Information Searching on Internet*. Paris, FRA: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1997. (RIA0 '97), p. 200–214.

NANGI, S. R.; YASUNAGA, M.; REN, H.; HUANG, Q.; LIANG, P.; LESKOVEC, J. Dense Retrieval of Knowledge Graphs for Question Answering. 2023.

NGUYEN, T.; ROSENBERG, M.; SONG, X.; GAO, J.; TIWARY, S.; MAJUMDER, R.; DENG, L. Ms marco: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016. Disponível em: <<http://dblp.uni-trier.de/db/journals/corr/corr1611.html#NguyenRSGTMD16>>.

NICKEL, M.; ROSASCO, L.; POGGIO, T. Holographic embeddings of knowledge graphs. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, Arizona: AAAI Press, 2016. (AAAI'16), p. 1955–1961.

NICKEL, M.; TRESP, V.; KRIEGER, H.-P. A three-way model for collective learning on multi-relational data. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Madison, WI, USA: Omnipress, 2011. (ICML'11), p. 809–816. ISBN 978-1-4503-0619-5.

NOGUEIRA, R.; LIN, J.; EPISTEMIC, A. *From doc2query to doctttttquery*. 2019. Online preprint.

OLIVEIRA, L. L. D.; ROMEU, R. K.; MOREIRA, V. P. REGIS: A Test Collection for Geoscientific Documents in Portuguese. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual Event Canada: ACM, 2021. p. 2363–2368. ISBN 978-1-4503-8037-9. Disponível em: <<https://dl.acm.org/doi/10.1145/3404835.3463256>>.

OLSEN, M. et al. Information retrieval from daily drilling reports using natural language processing. In: *ECMOR XVII – European Conference on the Mathematics of Oil Recovery*. European Association of Geoscientists & Engineers, 2025. Disponível em: <<https://www.earthdoc.org/content/papers/10.3997/2214-4609.202510629>>.

PONTE, J. M.; CROFT, W. B. A Language Modeling Approach to Information Retrieval. *SIGIR Forum*, v. 51, n. 2, p. 202–208, 2017. ISSN 0163-5840. Disponível em: <<https://doi.org/10.1145/3130348.3130368>>.

QIN, Y.; LIN, Y.; TAKANOBU, R.; LIU, Z.; LI, P.; JI, H.; HUANG, M.; SUN, M.; ZHOU, J. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. *CoRR*, abs/2012.15022, 2020. Disponível em: <<https://arxiv.org/abs/2012.15022>>.

An Ontology of Well Engineering Entities to Extract and Structure Text Data from Daily Reports, SPE/IADC International Drilling Conference and Exhibition de SPE/IADC Drilling Conference and Exhibition, (SPE/IADC Drilling Conference and Exhibition, SPE/IADC International Drilling Conference and Exhibition). D022S009R001 p. Disponível em: <<https://doi.org/10.2118/223801-MS>>.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. Language models are unsupervised multitask learners. 2019.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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. 3982–3992. Disponível em: <<https://aclanthology.org/D19-1410/>>.

REN, R.; LV, S.; QU, Y.; LIU, J.; ZHAO, W. X.; SHE, Q.; WU, H.; WANG, H.; WEN, J. PAIR: leveraging passage-centric similarity relation for improving dense passage retrieval. *CoRR*, abs/2108.06027, 2021. Disponível em: <<https://arxiv.org/abs/2108.06027>>.

ROBERTSON, S.; WALKER, S.; JONES, S.; HANCOCK-BEAULIEU, M.; GATFORD, M. Okapi at trec-3. In: *Proceedings of The Third Text REtrieval Conference*. [S.l.]: National Institute of Standards and Technology (NIST), 1994. (NIST Special Publication, v. 500-225), p. 109–126.

ROBERTSON, S.; ZARAGOZA, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, v. 3, n. 4, p. 333–389, 2009. ISSN 1554-0669. Disponível em: <<https://doi.org/10.1561/1500000019>>.

ROBERTSON, S. E.; JONES, K. S. Relevance weighting of search terms. In: *Document retrieval systems*. GBR: Taylor Graham Publishing, 1988. p. 143–160. ISBN 978-0-947568-21-4.

ROCCHIO, J. J. Relevance feedback in information retrieval. In: SALTON, G. (Ed.). *The Smart retrieval system - experiments in automatic document processing*. [S.l.]: Englewood Cliffs, NJ: Prentice-Hall, 1971. p. 313–323.

RODRIGUES, R. B. M.; PRIVATTO, P. I. M.; SOUSA, G. J. de; MURARI, R. P.; AFONSO, L. C. S.; PAPA, J. P.; PEDRONETTE, D. C. G.; GUILHERME, I. R.; PERROUT, S. R.; RIENTE, A. F. Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In: PINHEIRO, V.; GAMALLO, P.; AMARO, R.; SCARTON, C.; BATISTA, F.; SILVA, D.; MAGRO, C.; PINTO, H. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2022. p. 101–109. ISBN 978-3-030-98305-5.

RUGGERI, F.; MESGAR, M.; GUREVYCH, I. *ArgSciChat: A Dataset for Argumentative Dialogues on Scientific Papers*. 2022.

SAAD-FALCON, J.; KHATTAB, O.; SANTHANAM, K.; FLORIAN, R.; FRANZ, M.; ROUKOS, S.; SIL, A.; SULTAN, M.; POTTS, C. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. 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. 11265–11279. Disponível em: <<https://aclanthology.org/2023.emnlp-main.693/>>.

SALAKHUTDINOV, R.; HINTON, G. Semantic hashing. *International Journal of Approximate Reasoning*, v. 50, n. 7, p. 969–978, jul. 2009. ISSN 0888-613X. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0888613X08001813>>.

SALTON, G. Developments in Automatic Text Retrieval. *Science*, v. 253, n. 5023, p. 974–980, ago. 1991. Publisher: American Association for the Advancement of Science. Disponível em: <<https://www.science.org/doi/10.1126/science.253.5023.974>>.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Commun. ACM*, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782. Disponível em: <<https://dl.acm.org/doi/10.1145/361219.361220>>.

SANDERSON, M. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, v. 4, n. 4, p. 247–375, 2010. ISSN 1554-0669, 1554-0677. Disponível em: <<http://www.nowpublishers.com/article/Details/INR-009>>.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. Disponível em: <<http://arxiv.org/abs/1910.01108>>.

SANTHANAM, K.; KHATTAB, O.; SAAD-FALCON, J.; POTTS, C.; ZAHARIA, M. *ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction*. 2022. Disponível em: <<https://arxiv.org/abs/2112.01488>>.

SANTOS, S. M. *Relações Semânticas*. [S.l.]: Universidade Federal de Minas Gerais, 2018.

SAP, M.; BRAS, R. L.; ALLAWAY, E.; BHAGAVATULA, C.; LOURIE, N.; RASHKIN, H.; ROOF, B.; SMITH, N. A.; CHOI, Y. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146, 2018. Disponível em: <<http://arxiv.org/abs/1811.00146>>.

SCHLICHTKRULL, M.; KIPF, T. N.; BLOEM, P.; BERG, R. van den; TITOV, I.; WELLING, M. Modeling Relational Data with Graph Convolutional Networks. In: GANGEMI, A.; NAVIGLI, R.; VIDAL, M.-E.; HITZLER, P.; TRONCY, R.; HOLLINK, L.; TORDAI, A.; ALAM, M. (Ed.). *The Semantic Web*. Cham: Springer International Publishing, 2018. p. 593–607. ISBN 978-3-319-93417-4.

SCHWARTZ, R.; DODGE, J.; SMITH, N. A.; ETZIONI, O. Green AI. *CoRR*, abs/1907.10597, 2019. Disponível em: <<http://arxiv.org/abs/1907.10597>>.

SELVALAKSHMI, B.; SUBRAMANIAM, M. Intelligent ontology based semantic information retrieval using feature selection and classification. *Cluster Computing*, v. 22, n. 5, p. 12871–12881, set. 2019. ISSN 1573-7543. Disponível em: <<https://doi.org/10.1007/s10586-018-1789-8>>.

SEO, M.; KWIATKOWSKI, T.; PARIKH, A.; FARHADI, A.; HAJISHIRZI, H. Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension. In: RILOFF, E.; CHIANG, D.; HOCKENMAIER, J.; TSUJII, J. (Ed.). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 559–564. Disponível em: <<https://aclanthology.org/D18-1052>>.

SEO, M.; LEE, J.; KWIATKOWSKI, T.; PARIKH, A.; FARHADI, A.; HAJISHIRZI, H. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index. In: KORHONEN, A.; TRAUM, D.; MÀRQUEZ, L. (Ed.). *Proceedings of the 57th Annual Meeting of the Association*

for *Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 4430–4441. Disponível em: <<https://aclanthology.org/P19-1436>>.

SHERMAN, G.; EFRON, M. Document Expansion Using External Collections. 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. 1045–1048. ISBN 978-1-4503-5022-8. Disponível em: <<https://doi.org/10.1145/3077136.3080716>>.

SINGHAL, A.; PEREIRA, F. Document expansion for speech retrieval. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 1999. (SIGIR '99), p. 34–41. ISBN 978-1-58113-096-6. Disponível em: <<https://dl.acm.org/doi/10.1145/312624.312645>>.

SPARCK-JONES, K.; RIJSHERGEN, C. J. V. Report on the need for and provision of an ideal 'ideal' information retrieval test collection. *Computer Laboratory*, 1975.

SPEER, R.; CHIN, J.; HAVASI, C. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016. Disponível em: <<http://arxiv.org/abs/1612.03975>>.

TAMINE, L.; GOEURLOT, L. Semantic Information Retrieval On Medical Texts: Research Challenges, Survey and Open Issues. *ACM Comput. Surv.*, v. 1, n. 1, 2020.

TAY, Y.; TRAN, V. Q.; DEHGHANI, M.; NI, J.; BAHRI, D.; MEHTA, H.; QIN, Z.; HUI, K.; ZHAO, Z.; GUPTA, J.; SCHUSTER, T.; COHEN, W. W.; METZLER, D. *Transformer Memory as a Differentiable Search Index*. 2022.

THAKUR, N.; NI, J.; ABREGO, G. H.; WIETING, J.; LIN, J.; CER, D. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. 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. 7699–7724. Disponível em: <<https://aclanthology.org/2024.naacl-long.426/>>.

THAKUR, N.; REIMERS, N.; RÜCKLÉ, A.; SRIVASTAVA, A.; GUREVYCH, I. *BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models*. 2021. Disponível em: <<https://arxiv.org/abs/2104.08663>>.

TSATSARONIS, G.; BALIKAS, G.; MALAKASIOTIS, P.; PARTALAS, I.; ZSCHUNKE, M.; ALVERS, M. R.; WEISSENBORN, D.; KRITHARA, A.; PETRIDIS, S.; POLYCHRONOPOULOS, D.; ALMIRANTIS, Y.; PAVLOPOULOS, J.; BASKIOTIS, N.; GALLINARI, P.; ARTIERES, T.; NGONGA, A.; HEINO, N.; GAUSSIER, E.; BARRIO-ALVERS, L.; SCHROEDER, M.; ANDROUTSOPOULOS, I.; PALIOURAS, G. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, v. 16, p. 138, 2015. Disponível em: <<http://www.biomedcentral.com/content/pdf/s12859-015-0564-6.pdf>>.

VASHISHTH, S.; SANYAL, S.; NITIN, V.; TALUKDAR, P. *Composition-based Multi-Relational Graph Convolutional Networks*. arXiv, 2020. ArXiv:1911.03082. Disponível em: <<http://arxiv.org/abs/1911.03082>>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VELIČKOVIĆ, P.; CUCURULL, G.; CASANOVA, A.; ROMERO, A.; LIÒ, P.; BENGIO, Y. Graph attention networks. In: *International Conference on Learning Representations*. [s.n.], 2018. Disponível em: <<https://openreview.net/forum?id=rJXMpikCZ>>.

VERA, H. S.; DUA, S.; ZHANG, B.; SALZ, D.; MULLINS, R.; PANYAM, S. R.; SMOOT, S.; NAIM, I.; ZOU, J.; CHEN, F.; CER, D.; LISAK, A.; CHOI, M.; GONZALEZ, L.; SANSEVIERO, O.; CAMERON, G.; BALLANTYNE, I.; BLACK, K.; CHEN, K.; WANG, W.; LI, Z.; MARTINS, G.; LEE, J.; SHERWOOD, M.; JI, J.; WU, R.; ZHENG, J.; SINGH, J.; SHARMA, A.; SREEPAT, D.; JAIN, A.; ELARABAWY, A.; CO, A.; DOUMANOGLOU, A.; SAMARI, B.; HORA, B.; POTETZ, B.; KIM, D.; ALFONSECA, E.; MOISEEV, F.; HAN, F.; GOMEZ, F. P.; ÁBREGO, G. H.; ZHANG, H.; HUI, H.; HAN, J.; GILL, K.; CHEN, K.; CHEN, K.; SHANBHOGUE, M.; BORATKO, M.; SUGANTHAN, P.; DUDDU, S. M. K.; MARISERLA, S.; ARIAFAR, S.; ZHANG, S.; ZHANG, S.; BAUMGARTNER, S.; GOENKA, S.; QIU, S.; DABRAL, T.; WALKER, T.; RAO, V.; KHAWAJA, W.; ZHOU, W.; REN, X.; XIA, Y.; CHEN, Y.; CHEN, Y.-T.; DONG, Z.; DING, Z.; VISIN, F.; LIU, G.; ZHANG, J.; KENEALY, K.; CASBON, M.; KUMAR, R.; MESNARD, T.; GLEICHER, Z.; BRICK, C.; LACOMBE, O.; ROBERTS, A.; SUNG, Y.; HOFFMANN, R.; WARKENTIN, T.; JOULIN, A.; DUERIG, T.; SEYEDHOSSEINI, M. Embeddinggemma: Powerful and lightweight text representations. Google DeepMind, 2025. Disponível em: <<https://arxiv.org/abs/2509.20354>>.

VOORHEES, E. M. Query expansion using lexical-semantic relations. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Berlin, Heidelberg: Springer-Verlag, 1994. (SIGIR '94), p. 61–69. ISBN 978-0-387-19889-7.

VOORHEES, E. M. The trec robust retrieval track. *SIGIR Forum*, Association for Computing Machinery, New York, NY, USA, v. 39, n. 1, p. 11–20, jun 2005. ISSN 0163-5840. Disponível em: <<https://doi.org/10.1145/1067268.1067272>>.

WANG, L.; YANG, N.; HUANG, X.; YANG, L.; MAJUMDER, R.; WEI, F. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

WANG, Q.; MAO, Z.; WANG, B.; GUO, L. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, v. 29, n. 12, p. 2724–2743, dez. 2017. ISSN 1558-2191. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

WANG, R.; TANG, D.; DUAN, N.; WEI, Z.; HUANG, X.; JI, J.; CAO, G.; JIANG, D.; ZHOU, M. K-adapter: Infusing knowledge into pre-trained models with adapters. *CoRR*, abs/2002.01808, 2020. Disponível em: <<https://arxiv.org/abs/2002.01808>>.

WANG, T.; ISOLA, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *CoRR*, abs/2005.10242, 2020. Disponível em: <<https://arxiv.org/abs/2005.10242>>.

WANG, X.; YANG, Q.; QIU, Y.; LIANG, J.; HE, Q.; GU, Z.; XIAO, Y.; WANG, W. KnowledGPT: Enhancing Large Language Models with Retrieval and Storage Access on Knowledge Bases. *arXiv*, 2023. ArXiv:2308.11761. Disponível em: <<http://arxiv.org/abs/2308.11761>>.

WANG, Z.; CHEN, X. *DERA: Dense Entity Retrieval for Entity Alignment in Knowledge Graphs*. 2024. Disponível em: <<https://arxiv.org/abs/2408.01154v1>>.

WANG, Z.; ZHANG, J.; FENG, J.; CHEN, Z. Knowledge Graph Embedding by Translating on Hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 28, n. 1, jun. 2014. ISSN 2374-3468. Number: 1. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/8870>>.

WEI, X.; WANG, S.; ZHANG, D.; BHATIA, P.; ARNOLD, A. *Knowledge Enhanced Pretrained Language Models: A Comprehensive Survey*. arXiv, 2021. ArXiv:2110.08455 [cs]. Disponível em: <<http://arxiv.org/abs/2110.08455>>.

WRZALIK, M.; KRECHEL, D. Cort: Complementary rankings from transformers. *CoRR*, abs/2010.10252, 2020. Disponível em: <<https://arxiv.org/abs/2010.10252>>.

XIAO, H.; HUANG, M.; HAO, Y.; ZHU, X. *TransA: An Adaptive Approach for Knowledge Graph Embedding*. 2015. Disponível em: <<https://arxiv.org/abs/1509.05490>>.

XIAO, H.; HUANG, M.; ZHU, X. From one point to a manifold: knowledge graph embedding for precise link prediction. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2016. (IJCAI'16), p. 1315–1321. ISBN 9781577357704.

XIONG, C.; DAI, Z.; CALLAN, J.; LIU, Z.; POWER, R. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. 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. 55–64. ISBN 978-1-4503-5022-8. Disponível em: <<https://dl.acm.org/doi/10.1145/3077136.3080809>>.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 1996. (SIGIR '96), p. 4–11. ISBN 978-0-89791-792-6. Disponível em: <<https://dl.acm.org/doi/10.1145/243199.243202>>.

YANG, B.; YIH, W.-t.; HE, X.; GAO, J.; DENG, L. *Embedding Entities and Relations for Learning and Inference in Knowledge Bases*. 2014. Disponível em: <<https://arxiv.org/abs/1412.6575v4>>.

YANG, J.; HU, X.; XIAO, G.; SHEN, Y. *A Survey of Knowledge Enhanced Pre-trained Models*. arXiv, 2023. ArXiv:2110.00269. Disponível em: <<http://arxiv.org/abs/2110.00269>>.

YANG, M.; YANG, C.; ZHU, J.; LI, J.; ZHANG, J.; LI, Y.; LI, Y. SLiNT: Structure-aware language model with injection and contrastive training for knowledge graph completion. In: CHRISTODOULOPOULOS, C.; CHAKRABORTY, T.; ROSE, C.; PENG, V. (Ed.). *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, 2025. p. 13658–13671. ISBN 979-8-89176-335-7. Disponível em: <<https://aclanthology.org/2025.findings-emnlp.736/>>.

YASUNAGA, M.; BOSSELUT, A.; REN, H.; ZHANG, X.; MANNING, C. D.; LIANG, P.; LESKOVEC, J. *Deep Bidirectional Language-Knowledge Graph Pretraining*. arXiv, 2022. ArXiv:2210.09338 [cs]. Disponível em: <<http://arxiv.org/abs/2210.09338>>.

YE, Z.; KUMAR, Y. J.; SING, G. O.; SONG, F.; WANG, J. A Comprehensive Survey of Graph Neural Networks for Knowledge Graphs. *IEEE Access*, v. 10, p. 75729–75741, 2022. ISSN 2169-3536. Conference Name: IEEE Access.

ZAMANI, H.; DEGHANI, M.; CROFT, W. B.; LEARNED-MILLER, E.; KAMPS, J. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2018. (CIKM '18), p. 497–506. ISBN 978-1-4503-6014-2. Disponível em: <<https://doi.org/10.1145/3269206.3271800>>.

ZHAN, J.; MAO, J.; LIU, Y.; ZHANG, M.; MA, S. Repbert: Contextualized text embeddings for first-stage retrieval. *CoRR*, abs/2006.15498, 2020. Disponível em: <<https://arxiv.org/abs/2006.15498>>.

ZHANG, D.; YUAN, Z.; LIU, Y.; FU, Z.; ZHUANG, F.; WANG, P.; CHEN, H.; XIONG, H. E-BERT: A phrase and product knowledge enhanced language model for e-commerce. *CoRR*, abs/2009.02835, 2020. Disponível em: <<https://arxiv.org/abs/2009.02835>>.

ZHANG, S.; YAO, L.; SUN, A.; TAY, Y. Deep Learning based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys*, v. 52, n. 1, p. 1–38, jan. 2020. ISSN 0360-0300, 1557-7341. ArXiv:1707.07435 [cs]. Disponível em: <<http://arxiv.org/abs/1707.07435>>.

ZHANG, Y.; LI, M.; LONG, D.; ZHANG, X.; LIN, H.; YANG, B.; XIE, P.; YANG, A.; LIU, D.; LIN, J.; HUANG, F.; ZHOU, J. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

ZHANG, Y.; NIE, P.; GENG, X.; RAMAMURTHY, A.; SONG, L.; JIANG, D. DC-BERT: decoupling question and document for efficient contextual encoding. *CoRR*, abs/2002.12591, 2020. Disponível em: <<https://arxiv.org/abs/2002.12591>>.

ZHAO, W. X.; LIU, J.; REN, R.; WEN, J.-R. *Dense Text Retrieval based on Pretrained Language Models: A Survey*. arXiv, 2022. ArXiv:2211.14876 [cs]. Disponível em: <<http://arxiv.org/abs/2211.14876>>.

ZHENG, G.; CALLAN, J. Learning to Reweight Terms with Distributed Representations. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2015. (SIGIR '15), p. 575–584. ISBN 978-1-4503-3621-5. Disponível em: <<https://doi.org/10.1145/2766462.2767700>>.

ZHOU, J.; AGICHTEIN, E. RLIRank: Learning to Rank with Reinforcement Learning for Dynamic Search. In: *Proceedings of The Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 2842–2848. ISBN 978-1-4503-7023-3. Disponível em: <<https://doi.org/10.1145/3366423.3380047>>.

ZHOU, J.; DING, G.; GUO, Y. Latent semantic sparse hashing for cross-modal similarity search. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. New York, NY, USA: Association for Computing Machinery, 2014. (SIGIR '14), p. 415–424. ISBN 978-1-4503-2257-7. Disponível em: <<https://doi.org/10.1145/2600428.2609610>>.

ZHU, Y.; YUAN, H.; WANG, S.; LIU, J.; LIU, W.; DENG, C.; CHEN, H.; DOU, Z.; WEN, J.-R. *Large Language Models for Information Retrieval: A Survey*. arXiv, 2024. ArXiv:2308.07107 [cs]. Disponível em: <<http://arxiv.org/abs/2308.07107>>.

ZHUKOVA, A.; LÜHRS, J.; LOBMÜLLER, C. E.; GIPP, B. Contrastive learning using graph embeddings for domain adaptation of language models in the process industry. In: POTDAR, S.; ROJAS-BARAHONA, L.; MONTELLA, S. (Ed.). *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Suzhou (China): Association for Computational Linguistics, 2025. p. 1472–1484. ISBN 979-8-89176-333-3. Disponível em: <<https://aclanthology.org/2025.emnlp-industry.103/>>.

ZUCCON, G.; KOOPMAN, B.; BRUZA, P.; AZZOPARDI, L. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In: *Proceedings of the 20th Australasian Document Computing Symposium*. New York, NY, USA: Association for Computing Machinery, 2015. (ADCS '15), p. 1–8. ISBN 978-1-4503-4040-3. Disponível em: <<https://doi.org/10.1145/2838931.2838936>>.