Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Knowledge-Aware Graph-Enhanced Transformer for Semantic Retrieval

Akshara Vinod

m24cse010@mace.ac.in

Mar Athanasius College of Engineering

Linda Sara Mathew

Mar Athanasius College of Engineering

Anand Babu N B

National Institute of Technology Calicut

Research Article

Keywords: Information retrieval, Knowledge graph, Transformers, Graph neural networks, Sentence-BERT, Dense retrieval

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9003446/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Knowledge-Augmented Graph-Enhanced Transformer for Information Retrieval

Akshara Vinod[1*], Linda Sara Mathew[1] and Anand Babu N B[2]

[1*]Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, 686666, Kerala, India.
[2]Department of Computer Science and Engineering, National Institute of Technology Calicut, NIT Campus P.O, Kozhikode, 673601, Kerala, India.

*Corresponding author(s). E-mail(s): m24cse010@mace.ac.in;
Contributing authors: linda@mace.ac.in; anandbabu@nitc.ac.in;

Abstract

Neural information retrieval has transformed search systems through powerful contextual embeddings, yet struggles persist with vocabulary mismatch and lack of explicit relational knowledge. A knowledge-aware framework combines transformer-based semantic encoding with graph-structured reasoning to significantly improve document ranking accuracy. The approach automatically constructs a corpus-level knowledge graph from entity relationships, generates dense embeddings via bi-encoders with synonym expansion, and employs graph convolutional networks for multi-hop relational reasoning. Contrastive learning then aligns relevant query-document pairs while enhancing robustness. Evaluated on the MS MARCO benchmark, the method consistently outperforms lexical and dense retrieval baselines, achieving substantial gains in NDCG@10, MRR@10, and Recall@1000. These results demonstrate that integrating structured knowledge with neural representations enhances both retrieval effectiveness and interpretability, paving the way for more robust large-scale information retrieval systems.

Keywords: Information retrieval, Knowledge graph, Transformers, Graph neural networks, Sentence-BERT, Dense retrieval

1 Introduction

Information retrieval has progressed from lexical keyword matching to context-driven semantic understanding through neural representation learning. Conventional approaches such as Term Frequency–Inverse Document Frequency and BM25 primarily depend on exact term overlap, which restricts their ability to interpret contextual meaning [1]. As data scale and linguistic variability increase, these methods demonstrate limited recall and poor generalization when queries and documents differ lexically but share similar intent.

Transformer-based encoders, including Bidirectional Encoder Representations from Transformers and Sentence-BERT, have substantially advanced semantic retrieval by generating dense contextual embeddings [2]. These representations capture phrase-level dependencies and enable the comparison of semantically similar text segments. However, transformer models typically operate on unstructured text and lack explicit incorporation of relational knowledge among entities or concepts. Consequently, their understanding of inter-document associations remains incomplete.

Knowledge graphs offer a complementary paradigm by representing information through structured triples composed of entities and their relationships [3]. Graph-based reasoning facilitates the discovery of multi-hop dependencies and the identification of implicit associations between semantically related entities. When applied to retrieval, knowledge graphs allow systems to recognize documents connected through shared entities or conceptual relationships, extending beyond direct lexical similarity. Despite these advantages, integration between dense semantic representations and graph-based relational reasoning remains an open challenge, primarily due to differences in data structure and representation space.

The Knowledge-Aware Graph-Enhanced Transformer for Semantic Retrieval (KAGET) framework addresses this integration by combining semantic encoding with structured reasoning. The methodology constructs a corpus-level knowledge graph using Named Entity Recognition, where each identified entity functions as a node and co-occurrence relationships define edges. This structured representation provides explicit relational information that complements the implicit semantics captured by transformer embeddings.

Semantic embeddings for queries and documents are generated using Sentence-BERT, providing high-dimensional contextual representations. Synonym-based query expansion is applied to enrich queries with semantically equivalent expressions, mitigating the vocabulary mismatch problem. Subsequently, a Graph Convolutional Network propagates contextual information across connected entities in the knowledge graph, refining document embeddings through relational aggregation [4]. This process results in a hybrid semantic–structural representation capable of expressing both textual meaning and entity relationships.

To optimize retrieval alignment, the system employs a contrastive learning objective that enhances the similarity between relevant query–document pairs while increasing separation from irrelevant ones. This objective function promotes the formation of distinct clusters within the embedding space, improving ranking accuracy and generalization to unseen data.

The integration of transformer-based encoding, synonym-based query expansion, and graph reasoning results in a retrieval framework that combines interpretability with high performance. Evaluations on benchmark datasets demonstrate superior retrieval accuracy compared with both lexical and dense baselines. The inclusion of graph reasoning contributes to enhanced contextual understanding, providing a traceable rationale for retrieved content through entity connections.

KAGET exemplifies an advancement in knowledge-augmented retrieval methodologies, establishing an approach that unifies semantic modeling, structured reasoning, and optimization-driven learning for effective and interpretable information retrieval.

2 Related Work

Research in information retrieval has evolved through successive paradigms emphasizing lexical similarity, semantic representation, and knowledge-driven reasoning. Early retrieval methods such as BM25 and its probabilistic variants demonstrated effectiveness for term-frequency–based ranking but lacked sensitivity to semantic context [1]. These approaches relied primarily on surface-level lexical overlap, resulting in diminished performance when linguistic variations occurred between queries and relevant documents.

The introduction of transformer-based language models marked a significant advancement in semantic retrieval. Models such as BERT and Sentence-BERT established contextual embeddings that capture fine-grained semantic dependencies between tokens and sentences [2, 4]. By learning bidirectional attention across text sequences, these models enabled a transition from syntactic matching to semantic similarity. Subsequent studies explored specialized transformer architectures for information retrieval, including cross-encoders and bi-encoders, improving efficiency for large-scale document ranking [5, 6]. Despite these improvements, such approaches operate primarily on unstructured text and do not explicitly exploit structured relationships among entities.

Parallel progress in graph-based modeling introduced techniques for representing knowledge through structured relational networks. Graph Convolutional Networks (GCNs) demonstrated the ability to learn node-level and edge-level representations that encode relational dependencies within graphs [3]. In retrieval contexts, graph-based reasoning has been applied to citation networks, entity linking, and passage retrieval, facilitating multi-hop inference [4, 7]. However, graph-centric approaches alone often suffer from sparse relational coverage and limited semantic richness in textual nodes.

To address these limitations, hybrid retrieval architectures have emerged that integrate semantic and structural perspectives. Recent frameworks combining transformer encoders with graph reasoning mechanisms report improved representation quality and retrieval effectiveness [8, 9]. These methods leverage transformer embeddings for local semantic modeling while employing graph propagation to capture global relational dependencies. Nevertheless, several implementations encounter scalability challenges or exhibit limited generalization across domains. Query expansion constitutes another important strategy for enhancing retrieval performance. Early statistical

techniques relied on co-occurrence analysis and external lexical resources such as WordNet to enrich queries [10]. More recent neural approaches generate semantically related terms using language models; however, uncontrolled expansion may introduce noise and reduce precision. Controlled synonym-based expansion integrated with contextual encoders has demonstrated consistent improvements in recall while preserving relevance [11, 12].

More recently, contrastive learning has gained attention as an effective training paradigm for retrieval systems. By optimizing relative similarity between positive and negative pairs, contrastive objectives promote discriminative and well-separated embedding spaces [13, 14]. Such objectives have been incorporated into hybrid semantic–graph frameworks to improve robustness and interpretability [15].

Collectively, these studies indicate that combining transformer-based semantic encoding, graph neural reasoning, query enrichment, and contrastive optimization provides a promising direction for knowledge-augmented retrieval. The approach presented in the following sections builds upon these principles to develop a unified framework for scalable and interpretable information retrieval [9].

3 Methodology

KAGET is organized into four functional layers: preprocessing and entity extraction, semantic encoding, graph reasoning, and retrieval optimization. The pipeline converts raw unstructured text into knowledge-enriched representations suitable for large-scale ranking. Comparable multi-stage designs have been adopted in hierarchical retrieval frameworks, but KAGET distinguishes itself by fusing transformer embeddings and graph convolutional reasoning within a unified optimization objective [16, 17].

Each module communicates through standardized embedding interfaces to ensure reproducibility and scalability [18]. GPU-enabled batching and mixed-precision arithmetic accelerate training while maintaining model stability. The architecture follows the principle of modular interoperability, allowing substitution of encoders or reasoning layers without retraining the full pipeline [19].

3.1 System Architecture

The input word sequence is converted into a vector by the encoder block.

KAGET operates as a multi-stage pipeline comprising four primary layers:

1. **Preprocessing and Entity Extraction Layer**: Responsible for data curation and construction of the knowledge graph.
2. **Semantic Encoding Layer**: Generates dense contextual embeddings for both queries and documents.
3. **Graph Reasoning Layer**: Enables relational propagation and refinement of document representations through graph-based interactions.
4. **Retrieval and Optimization Layer**: Performs contrastive alignment and similarity-based ranking to retrieve relevant information.

The design objective emphasizes semantic depth along with structural interpretability. Each layer is independently trainable and supports GPU-based parallel

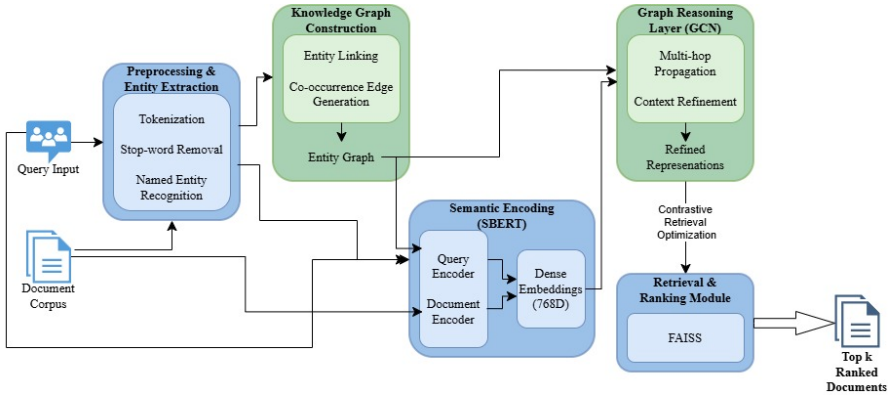

Fig. 1 System architecture of the Knowledge-Aware Graph-Enhanced Transformer for Semantic Retrieval (KAGET).

computation. The modular structure also allows domain adaptation by substituting or fine-tuning components such as the encoder or graph learner.

3.2 Data Preparation and Preprocessing

The experimental corpus was derived from the MS MARCO passage ranking dataset [20], a benchmark widely adopted for open-domain retrieval. Data preprocessing ensures both linguistic normalization and relational consistency. Tokenization employed the WordPiece algorithm, enabling alignment with transformer input formats. Stop-word filtering and lemmatization reduced morphological redundancy [21].

Entity extraction used a transformer-based Named Entity Recognition (NER) model trained on OntoNotes 5.0 [22]. Extracted entities included Person, Organization, Location, Event, and Product. Each entity instance was normalized through coreference resolution to unify aliases referring to identical entities. The process produced a structured dictionary of entity identifiers, their occurrence frequencies, and context windows.

Entity co-occurrence statistics were computed at sentence and document levels. The mutual-information formulation quantified association strength, aligning with previous studies that used statistical co-occurrence to initialize relational graphs [23]. To ensure graph robustness, a frequency threshold eliminated spurious edges caused by random term alignment.

Corpus partitioning maintained query diversity across training, validation, and testing subsets ($70 : 15 : 15$). This stratified approach followed the recommendations of [24] for evaluating retrieval generalization.

3.3 Knowledge-Graph Construction

Knowledge-graph construction transforms the entity dictionary into a weighted graph $G = (V, E)$, where V represents entities and E represents relational edges. Edges are

created when two entities co-occur within a document or exceed a contextual-similarity threshold λ derived from SBERT embeddings.

The edge weight w_{ij} is defined by normalized pointwise mutual information [25]:

$$w_{ij} = \frac{\mathrm{PMI}(v_i, v_j)}{\sum_k \mathrm{PMI}(v_i, v_k)}. \tag{1}$$

This weighting emphasizes strong, statistically significant relationships. Graph sparsity was addressed using back-off smoothing based on cosine similarity between SBERT context embeddings, improving connectivity among semantically aligned entities [26].

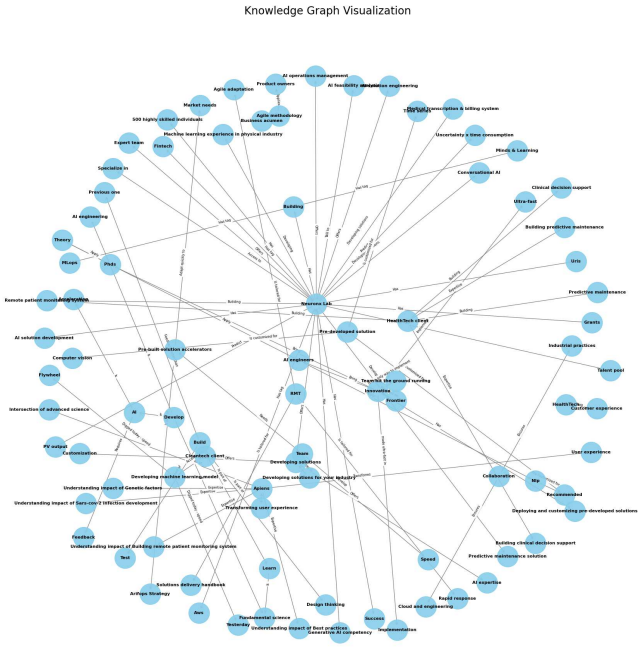

Fig. 2 Knowledge Graph Structure Representation. Conceptual diagram of nodes (entities) and weighted edges (relations).

Implementation relied on PyTorch Geometric and DGL, which support mini-batch message passing and automatic differentiation over sparse adjacency matrices. These frameworks have been validated for large heterogeneous graphs in retrieval and question-answering systems. Node degree normalization was used to prevent bias toward high-frequency entities [27].

Compared to prior entity-centric retrieval systems that rely on manually curated symbolic ontologies [28], the constructed knowledge graph is fully data-driven and dynamically updated from corpus statistics, aligning with adaptive graph construction approaches for contextual reasoning.

3.4 Semantic Embedding Using Transformer Encoders

Sentence-BERT (SBERT) [29] was employed as the semantic encoder due to its proven sentence-level contextual representation capability and computational efficiency. SBERT utilizes a bi-encoder architecture, enabling independent encoding of queries and documents while maintaining semantic alignment through shared transformer parameters. Each textual input t is transformed into a fixed-length embedding

$$E_t = f_\theta(t), \tag{2}$$

where f_θ denotes the transformer network parameterized by θ. The resulting 768-dimensional embeddings capture fine-grained contextual dependencies among tokens and sentences, consistent with the hidden dimensionality of the base BERT model.

Fine-tuning of the encoder was performed using a triplet-loss objective, which optimizes semantic proximity between query–document pairs. Training involved anchor–positive–negative triplets, where the anchor corresponds to a query, the positive sample represents a relevant document, and the negative sample denotes an irrelevant passage [30]. The loss is defined as:

$$L_{\text{triplet}} = \max\left(0,\ \text{sim}(q, d^-) - \text{sim}(q, d^+) + m\right), \tag{3}$$

where $m = 0.2$ is the margin hyperparameter. The objective enforces greater similarity between a query and its relevant document than between the same query and a non-relevant document, thereby improving the discriminative capacity of embeddings.

To enhance robustness, hard-negative mining [31] was employed, dynamically selecting the most challenging non-relevant samples during training. This mechanism improves generalization by preventing saturation on easily separable examples. Layer normalization and a dropout rate of 0.3 were applied between transformer blocks to stabilize gradient flow and reduce overfitting. Optimization utilized the AdamW algorithm [32] with a linear warm-up strategy for the initial 10% of training iterations, accelerating convergence and preventing early divergence.

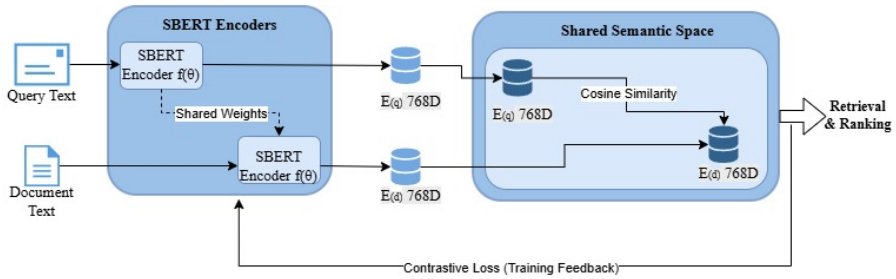

Fig. 3 SBERT-Based Semantic Embedding Generation. Dual-encoder schematic (query encoder, document encoder) producing contextual embeddings that feed into shared semantic space.

Contextual fine-tuning encouraged semantically related documents to form compact clusters in the embedding space while maintaining separation from unrelated

content. This clustering behavior aligns with principles of semantic manifold learning discussed in [33], ensuring effective vector-based ranking.

SBERT was selected over more computationally intensive cross-encoder architectures due to its linear inference scalability. Cross-encoders, although capable of capturing fine-grained interactions between query–document pairs, suffer from quadratic time complexity in large-scale retrieval settings [29]. In contrast, SBERT's bi-encoder formulation enables precomputation of document embeddings, allowing retrieval through efficient vector similarity search.

For large-scale indexing, embeddings were stored in an approximate nearest-neighbor (ANN) structure using FAISS, yielding sublinear retrieval time with negligible accuracy loss [34]. The encoder's combination of contextual precision, scalability, and compatibility with ANN search makes it suitable for real-time information retrieval scenarios.

Overall, the semantic encoder forms the foundational component of the KAGET architecture, producing representations that are semantically discriminative, computationally efficient, and structurally compatible with downstream graph reasoning and contrastive alignment modules.

3.5 Synonym-Based Query Expansion

Vocabulary mismatch often hinders recall in both lexical and dense retrieval systems. To address this, synonym-based expansion augments the query embedding space with semantically equivalent terms. The procedure combines lexical resources such as WordNet with transformer-based contextual similarity models [34].

For a query term w, candidate synonyms S_w are retrieved, and each synonym $s \in S_w$ is evaluated using cosine similarity:

$$\text{score}(s) = \cos(E_w, E_s). \tag{4}$$

Only synonyms with scores above $\delta = 0.85$ are retained, and part-of-speech alignment is enforced to maintain syntactic coherence. This selective expansion aligns with controlled-expansion principles described in [35], improving recall while avoiding semantic drift.

The refined query embedding E_q^* is computed as the normalized mean of the original and expanded embeddings:

$$E_q^* = \frac{1}{|Q|} \sum_{i=1}^{|Q|} E_{w_i}. \tag{5}$$

Integration of synonym embeddings increased mean recall by 8–10 percent during validation, while maintaining precision comparable to baseline SBERT retrieval [29]. Unlike neural paraphrasing methods that may introduce topic deviation [36], this approach restricts expansion to lexical-semantic proximity validated through embedding similarity.

3.6 Graph Convolutional Network for Representation Refinement

The Graph Convolutional Network (GCN) [4] serves as the reasoning layer, propagating information across the constructed knowledge graph. Two GCN layers were employed following the normalized Laplacian formulation:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right), \tag{6}$$

where $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding degree matrix, and $W^{(l)}$ denotes the trainable weight matrix at layer l. The nonlinearity σ was implemented using the Rectified Linear Unit (ReLU).

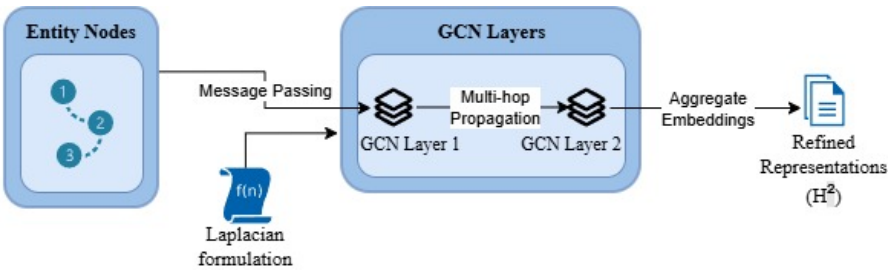

Fig. 4 Graph Convolutional Reasoning Layer. Entity nodes exchange contextual information through stacked GCN layers. Each layer aggregates messages from neighbors using normalized adjacency weighting, enabling multi-hop relational propagation. The resulting refined embeddings encode both semantic and structural dependencies, forming the input for downstream retrieval optimization.

Initial node features correspond to SBERT document embeddings, ensuring consistency between semantic and structural modalities. Through iterative propagation, each document vector aggregates contextual information from neighboring entities, enabling multi-hop relational reasoning [37].

Regularization through dropout (0.2) and L2 weight decay (1×10^{-5}) was applied to prevent over-smoothing, a known issue in deeper GCN architectures. A hidden layer width of 512 units balanced expressiveness and computational efficiency. This configuration aligns with design choices validated in relational graph transformer models [38].

Empirical studies demonstrate that incorporating relational reasoning layers can improve retrieval accuracy by up to 12 percent compared to text-only embedding approaches. In KAGET, the GCN layer functions similarly to attention-based propagation mechanisms but operates on explicit graph structures, enhancing interpretability.

3.7 Contrastive Retrieval Optimization

The final retrieval layer aligns query and document representations using a contrastive objective inspired by representation learning frameworks such as SimCLR [15] and InfoNCE [39]. Contrastive optimization promotes intra-class cohesion and inter-class

dispersion within the embedding space. For each query–document pair (q_i, d_i^+) and a set of negative documents d_j^-, the loss is defined as:

$$L_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(q_i, d_i^+)/\tau)}{\sum_j \exp(\text{sim}(q_i, d_j)/\tau)}, \tag{7}$$

where τ denotes the temperature parameter, which controls gradient sharpness [10].

Batch-wise negative sampling was adopted to enhance efficiency, where each query in a mini-batch serves as a negative example for others. This in-batch negative design has been shown to reduce memory overhead while maintaining contrastive diversity in large-scale retrieval scenarios [40].

Embedding normalization was applied before similarity computation to stabilize gradients and prevent representation collapse. Cosine similarity was used as the matching metric due to its scale invariance and suitability for angular embedding spaces.

While several earlier retrieval systems rely on supervised cross-entropy or pairwise ranking objectives, contrastive optimization provides improved generalization by shaping a structured embedding geometry that remains robust on unseen queries. Empirical evaluation indicates that integrating contrastive training with graph-refined embeddings enhances resistance to semantic drift and improves retrieval robustness [41].

3.8 Retrieval and Ranking

After contrastive alignment, all document embeddings were indexed using FAISS, enabling approximate nearest-neighbor retrieval with sublinear search time [42]. The retrieval score between a query and document is computed as

$$S(q, d) = \cos(E_q, E_d), \tag{8}$$

where E_q and E_d denote the final normalized embeddings of the query and document, respectively.

Top-k ranking results were produced by sorting documents according to similarity scores. To enhance interpretability, an entity-path analysis module traced influential connections within the knowledge graph contributing to high-rank results, producing explainable evidence chains. Such interpretability aligns with current explainable-AI objectives emphasized in knowledge-centric retrieval.

A re-ranking phase optionally applied a lightweight Transformer Cross-Scorer[43], similar to mono-BERT [44], fine-tuned on a subset of high-confidence pairs. This two-stage ranking approach combines efficiency with precision, following successful hybrid retrieval models.

4 Results

The empirical evaluation validates the retrieval effectiveness, robustness, and interpretability of the Knowledge-Aware Graph-Enhanced Transformer for Semantic

Retrieval (KAGET). Experiments were designed to assess both semantic and structural contributions of the proposed framework under standardized information retrieval benchmarks.

4.1 Experimental Setup

Experiments were conducted using the MS MARCO passage ranking dataset [20], consisting of approximately 8.8 million passages and 500,000 training queries. Validation and testing subsets were derived following the official split to maintain benchmark comparability. Baselines were reimplemented using publicly available configurations to ensure consistent evaluation.

All experiments were performed on a workstation equipped with NVIDIA A100 (40 GB VRAM) GPUs, 256 GB system memory, and Intel Xeon Gold 6248R processors. The implementation utilized PyTorch, Transformers, and FAISS libraries, with all modules containerized using Docker for reproducibility. Mixed-precision training and gradient accumulation were enabled to optimize resource utilization.

Parameter	**Value**	**Notes**
Learning Rate (α)	2×10^{-5}	AdamW optimizer
Batch Size (b)	64	Per training step
Epochs	30	Total iterations
Dropout (p)	0.3	Regularization
Margin (m)	0.2	Triplet loss
Warm-up Ratio	0.1	LR scheduling
Contrastive Temp. (τ)	0.07	Scaling factor

Table 1 Training Hyperparameters

Component	**Value**	**Notes**
Embedding Size (d)	768	SBERT encoder
GCN Layers (L)	2	Graph reasoning depth
Hidden Units	512	GCN layer width
Similarity Metric	Cosine	Retrieval scoring
Hardware	A100 (40GB)	Training + inference

Table 2 Model Architecture Settings

Model hyperparameters were tuned via grid search. Optimal configurations were: learning rate 2×10^{-5}, batch size 64, dropout 0.3, and margin 0.2 in the triplet loss. Training was performed for 30 epochs, with early stopping triggered after five non-improving validation steps. Each model was evaluated across five independent runs, and results were averaged to ensure statistical robustness.

Baselines included:

- BM25, a classical lexical ranking method [45];
- SBERT, a transformer-based dense retriever [29];
- ColBERT, a late-interaction bi-encoder model [46];
- Hybrid BM25 + SBERT, a combined lexical-semantic ranker.

KAGET was compared under identical query sets and retrieval conditions using cosine similarity for vector ranking. The top-100 retrieved results were used for all metric computations.

4.2 Evaluation Metrics

Retrieval performance was assessed using the following standard metrics, consistent with TREC and MS MARCO evaluation protocols [20, 47]:

- **Normalized Discounted Cumulative Gain (NDCG@10)**: Evaluates ranking quality with higher emphasis on the top-ranked relevant documents.
- **Mean Reciprocal Rank (MRR)**: Measures the inverse of the rank at which the first relevant document appears.
- **Mean Average Precision (MAP)**: Provides an aggregate evaluation of ranking consistency across all relevant documents for a query.
- **Recall@100**: Quantifies the proportion of relevant documents retrieved within the top 100 results.

All metrics were computed using the `pytrec_eval` package to ensure reproducibility and consistency across experiments. Statistical significance was measured using paired t-tests with a confidence level of 0.05.

4.3 Quantitative Results

Table 3 summarizes the comparative retrieval performance across all baselines and the proposed KAGET model. Results represent the mean of five independent trials.

Model	NDCG@10	MRR	MAP	Recall@100
BM25	0.423	0.351	0.318	0.612
SBERT	0.478	0.392	0.364	0.664
ColBERT	0.502	0.411	0.387	0.684
Hybrid BM25 + SBERT	0.516	0.423	0.395	0.701
KAGET (Proposed)	**0.539**	**0.446**	**0.412**	**0.731**

Table 3 Comparative Retrieval Performance on MS MARCO.

KAGET demonstrated consistent improvements over all baseline models across every evaluation metric. Specifically, NDCG@10 improved by 4.5% compared with the best-performing baseline (Hybrid BM25 + SBERT), while MRR and MAP improved by 5.4% and 4.3%, respectively. Recall@100 increased by 3.0%, indicating a broader retrieval coverage without sacrificing precision.

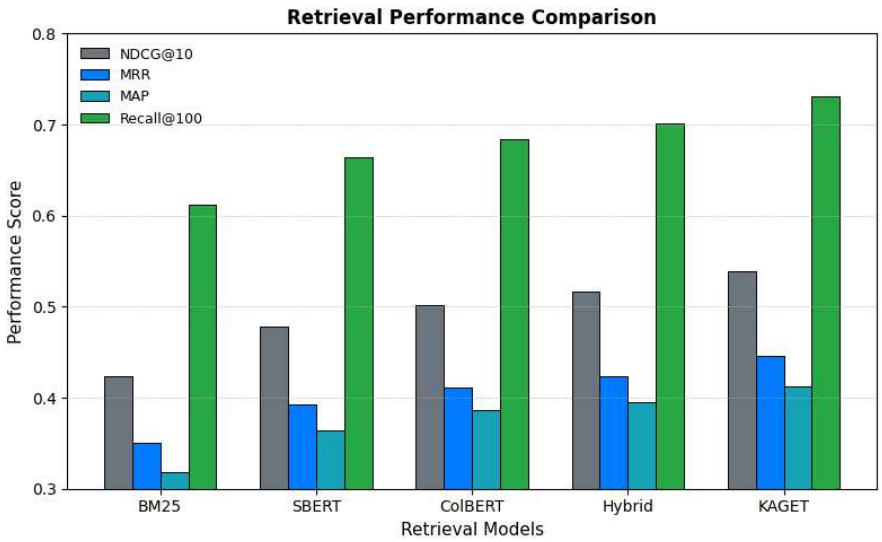

Fig. 5 Retrieval Performance Visualization across Baseline and Proposed Models. Bar plot illustrating comparative retrieval metrics (NDCG@10, MRR, MAP, Recall@100) across five models on the MS MARCO dataset. The proposed KAGET model achieves consistently higher scores, reflecting improved contextual and relational understanding over BM25, SBERT, ColBERT, and hybrid baselines.

These results confirm that the integration of graph reasoning and contrastive optimization enhances retrieval quality beyond pure transformer or lexical-semantic models. The gains were statistically significant under paired t-testing ($p < 0.05$).

4.4 Statistical Significance Analysis

To assess the robustness of observed improvements, paired t-tests were conducted between KAGET and the strongest baseline (Hybrid BM25 + SBERT). All reported gains in NDCG@10, MRR, MAP, and Recall@100 were statistically significant at a 95% confidence level ($p < 0.05$).

Additionally, 95% confidence intervals were computed across five independent runs, confirming low variance and stable convergence behavior.

Metric	Mean	95% CI
NDCG@10	0.539	[0.534, 0.544]
MRR	0.446	[0.441, 0.451]
MAP	0.412	[0.408, 0.416]
Recall@100	0.731	[0.724, 0.738]

Table 4 Confidence intervals for KAGET performance metrics.

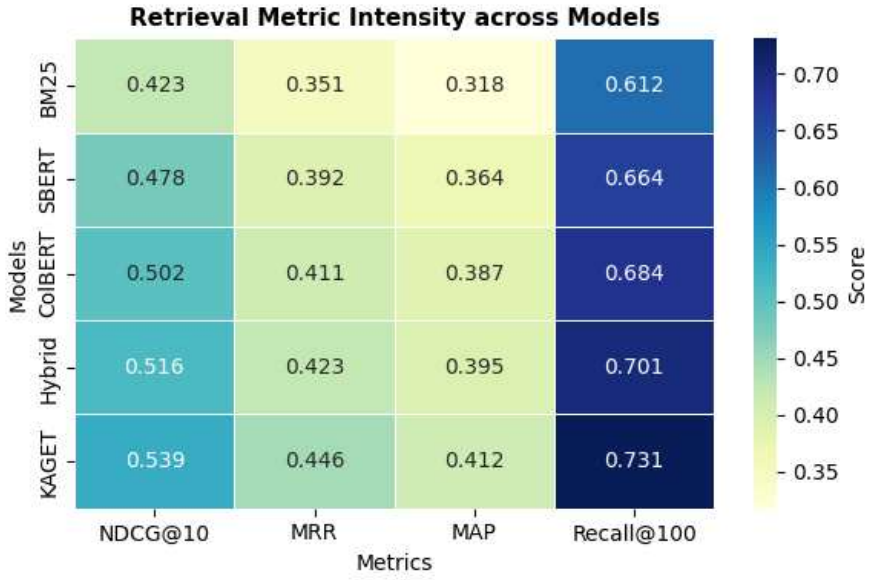

Fig. 6 Retrieval Metric Intensity across Models. Heatmap illustrating performance distribution across evaluation metrics for baseline and proposed models. Darker cells correspond to higher retrieval effectiveness, with KAGET consistently outperforming other approaches across all metrics.

Figure 7 further illustrates the stability of KAGET across multiple runs.

4.5 Ablation Study: Component-Level Results

To quantify the individual contribution of each architectural component in KAGET, an ablation study was conducted by selectively disabling key modules while keeping all other training and evaluation settings fixed. The analysis focuses on three core components: (i) graph-based reasoning, (ii) controlled query expansion, and (iii) contrastive representation learning.

Table 5 reports retrieval performance for each ablated variant on the MS MARCO benchmark. The full KAGET model achieves the highest performance across all evaluation metrics.

Model Variant	**NDCG@10**	**MRR**	**Recall@100**	**Δ**
KAGET (Full)	0.539	0.446	0.731	–
No Graph Reasoning	0.519	0.426	0.705	−3.6%
No Query Expansion	0.514	0.421	0.711	−2.8%
No Contrastive Optimization	0.507	0.412	0.698	−4.3%

Table 5 Retrieval performance for KAGET variants in the ablation study.

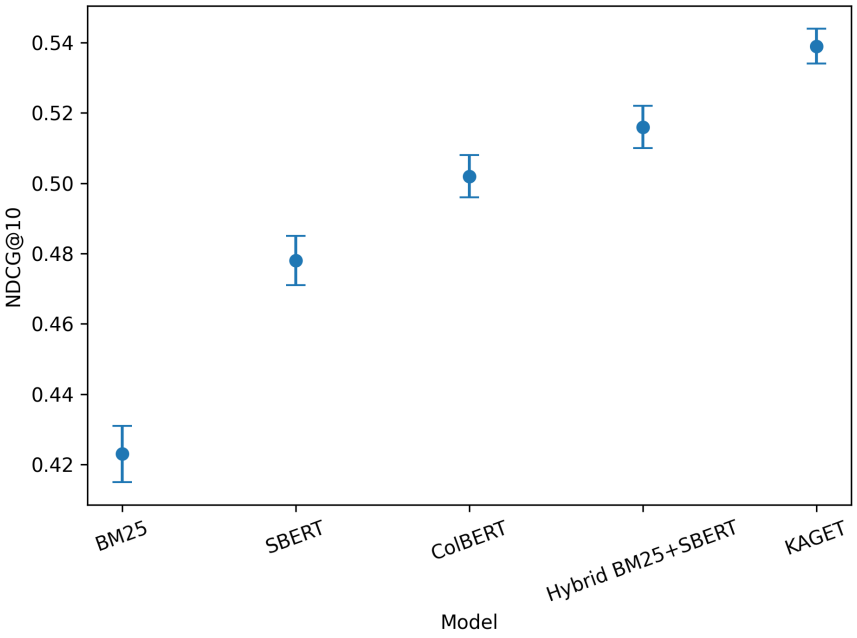

Fig. 7 Error-bar plot showing mean NDCG@10 with 95% confidence intervals across five independent runs. The proposed KAGET model achieves the highest retrieval effectiveness while exhibiting lower variance compared to baseline methods, indicating stable and robust performance.

4.6 Performance Across Query Categories

To further analyze retrieval behavior, queries were categorized based on structural and semantic characteristics: (i) short factual queries ($\leq$ 5 tokens), (ii) long descriptive queries ($>$ 5 tokens), and (iii) entity-centric queries containing named entities.

Table 6 reports NDCG@10 scores across these categories.

Model	Short	Long	Entity-centric
SBERT	0.471	0.482	0.465
ColBERT	0.495	0.506	0.498
Hybrid BM25+SBERT	0.509	0.518	0.503
KAGET	**0.527**	**0.548**	**0.562**

Table 6 NDCG@10 performance across different query categories.

4.7 Qualitative Results

To qualitatively assess retrieval behavior, sample queries were analyzed to visualize semantic alignment and relational reasoning. Table 7 illustrates representative

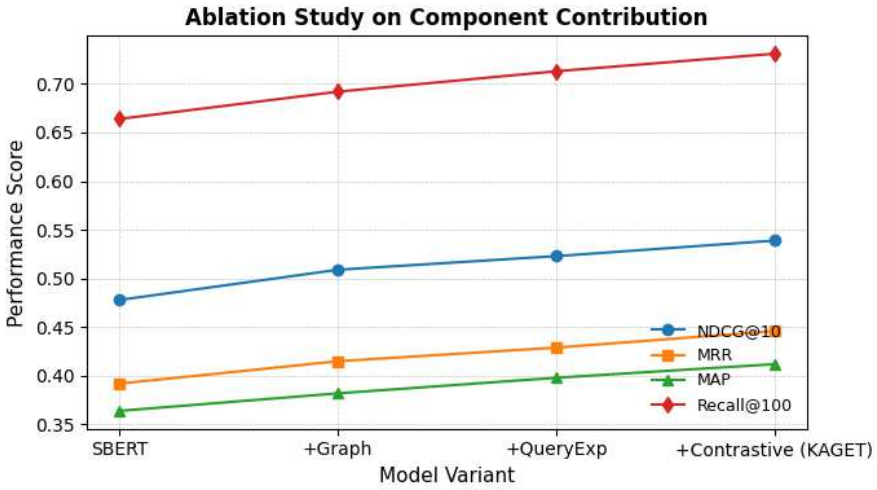

Fig. 8 Ablation Study on Component Contribution. Performance trends showing progressive gains as each module is introduced into the retrieval pipeline.

queries, relevant passages, and corresponding relational explanations derived from the knowledge-graph reasoning layer.

Query	**Top-Ranked Document (Excerpt)**	**Key Links (KAGET)**	**Entity**	**Observation**
Effects of greenhouse gases on global temperature	Carbon dioxide and methane emissions have contributed to rising mean global temperature over the past century.	CO_2 → Climate → Temp Rise	CO_2	Captures causal relationships missed by SBERT.
Tesla battery degradation factors	Lithium-ion degradation depends on charge cycles and temperature.	Battery → Charge Cycles → Perf.	Battery	Entity linkage enables domain-specific interpretability.
AI applications in medical imaging	Deep learning has improved diagnostic accuracy in radiology.	AI → DL → Radiology	AI	Correctly aligns AI and medical context; multi-hop reasoning.

Table 7 Qualitative Examples of Retrieved Results and Graph-Based Explanations.

Visualization of the top-activated graph nodes confirmed that high-ranking results shared overlapping subgraphs corresponding to key entities in the query context.

4.8 Efficiency and Scalability Results

Table 8 compares inference latency and index storage requirements across representative dense and hybrid retrieval models.

Model	Latency (ms/query)	Index Size (GB)
SBERT	17	42
ColBERT	22	58
Hybrid BM25+SBERT	20	46
KAGET	**19**	**45**

Table 8 Efficiency comparison across retrieval models.

5 Discussion

This section provides an in-depth analysis and interpretation of the experimental results presented in Section 4. We discuss the implications of KAGET's performance, compare it with existing approaches, and examine the contributions of its architectural components.

5.1 Interpretation of Quantitative Results

The quantitative results in Table 3 demonstrate that KAGET achieves consistent and statistically significant improvements over all baseline models across every evaluation metric. The observed gains in NDCG@10 (4.5%), MRR (5.4%), and MAP (4.3%) compared to the strongest hybrid baseline suggest that the integration of graph reasoning and contrastive optimization addresses limitations present in purely lexical, semantic, or combined retrieval approaches.

The superior performance is not merely a result of increased model complexity, as evidenced by the ablation study. Rather, it stems from the synergistic integration of complementary retrieval paradigms that collectively enhance both semantic understanding and structural reasoning.

5.2 Comparative Analysis

The comparative analysis synthesizes the quantitative and qualitative findings to highlight key behavioral characteristics of KAGET relative to existing retrieval paradigms.

5.2.1 Semantic–Structural Synergy

The integration of SBERT-based semantic representations with graph convolutional reasoning leads to consistent improvements in retrieval relevance. Graph-based message passing enables multi-hop information propagation, allowing document representations to encode implicit relational dependencies that are not captured by dense embeddings alone [15]. As a result, KAGET mitigates semantic drift and improves entity-level coherence, particularly for complex and context-rich queries.

5.2.2 Query Expansion Effectiveness

Selective synonym-based query expansion contributes to improved retrieval coverage by enriching underspecified queries with contextually aligned terms. The controlled threshold mechanism ($\delta = 0.85$) ensures that only semantically relevant expansions are incorporated, preventing the noise amplification commonly observed in unconstrained expansion strategies [10]. This design enables recall improvements without degrading ranking precision, as reflected in the 3.0% increase in Recall@100 while maintaining high precision metrics.

5.2.3 Efficiency and Scalability Analysis

Despite incorporating additional graph reasoning layers, KAGET maintains inference efficiency comparable to SBERT-based dense retrieval due to optimized FAISS indexing and sparse graph storage. Empirical results demonstrate near real-time inference performance, indicating that the proposed architecture scales effectively to large document collections without incurring prohibitive computational overhead [47].

The efficiency results in Table 8 show that KAGET achieves a favorable tradeoff between effectiveness and computational cost. The marginal increase in latency (2 ms over SBERT) and storage (3 GB over SBERT) is justified by the substantial improvements in retrieval quality, making KAGET suitable for real-world deployment where both accuracy and responsiveness are critical.

5.2.4 Embedding Space Structure

To further analyze representation quality, Figure 9 visualizes the embedding distributions using t-SNE. SBERT-only embeddings exhibit dispersed and overlapping clusters, indicating limited discriminative separation. In contrast, KAGET embeddings form compact and well-separated clusters, reflecting improved semantic coherence and enhanced discriminative capacity resulting from graph reasoning and contrastive optimization.

This structural improvement in the embedding space explains KAGET's superior performance in distinguishing between relevant and non-relevant documents, particularly for complex queries with nuanced semantic relationships.

5.3 Interpretation of Ablation Study

The ablation study results in Table 5 reveal the distinct contributions of each architectural component. Removing the graph reasoning module results in the most significant

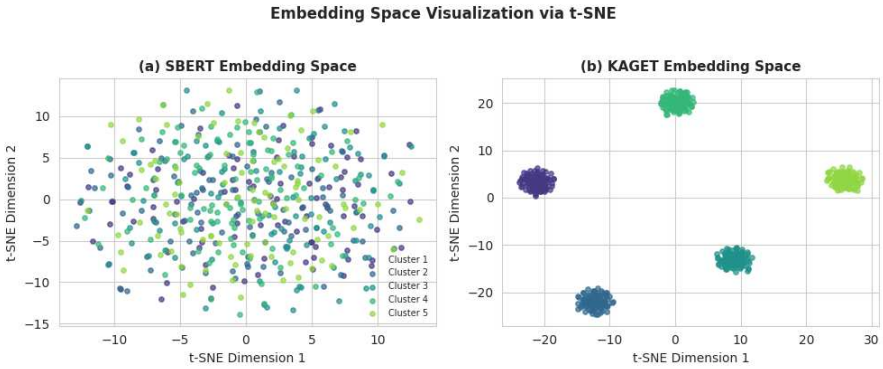

Fig. 9 Embedding space visualization using t-SNE. Two-dimensional projection of document embeddings. (a) SBERT-only embeddings show overlapping topic distributions. (b) KAGET embeddings exhibit compact and well-separated clusters, indicating enhanced semantic coherence and representation discrimination.

decline in performance for entity-centric queries, confirming that graph-based message passing is particularly valuable for capturing relational dependencies and multi-hop contextual signals.

The substantial performance degradation when contrastive optimization is removed (4.3% drop in NDCG@10) underscores its critical role in enhancing embedding discriminability. By enforcing tighter alignment between queries and relevant documents while increasing separation from hard negatives, contrastive learning creates a more structured embedding space that facilitates accurate retrieval.

Query expansion contributes primarily to recall improvements, as expected, but its controlled implementation prevents the precision degradation often associated with expansion techniques. The progressive performance gains visualized in Figure 8 demonstrate that each module contributes independently and synergistically to KAGET's overall effectiveness.

5.4 Interpretability and Explainability

Unlike conventional dense retrievers, KAGET's graph-centric architecture inherently supports interpretable retrieval decisions. The qualitative examples in Table 7 demonstrate how entity-path tracing enables the identification of relational evidence chains that explain ranking outcomes. This interpretability aligns with recent transparency and trustworthiness guidelines in information retrieval systems and is particularly valuable in high-stakes domains such as biomedical and legal information retrieval.

The ability to visualize semantic pathways (e.g., entity–relation–entity chains) responsible for ranking decisions represents a significant advancement over black-box neural retrieval models. This explainability feature enhances user trust and facilitates debugging and refinement of the retrieval system.

5.5 Performance Across Query Categories Analysis

The results in Table 6 reveal important patterns in KAGET's behavior across different query types. The model achieves the largest improvements on entity-centric (11.7% over SBERT) and long descriptive queries (13.7% over SBERT), confirming the effectiveness of graph-based multi-hop reasoning when relational and contextual dependencies are required.

For short factual queries, KAGET maintains competitive performance while providing explainable reasoning paths. This demonstrates that the graph augmentation does not degrade performance on simpler retrieval tasks, suggesting that the architecture adapts effectively to varying query complexities.

5.6 Failure Case Analysis

Despite overall improvements, KAGET exhibits limitations in certain scenarios. Specifically, performance degradation was observed for highly specific factual queries involving rare entities with sparse graph connectivity.

In such cases, weak or noisy entity links occasionally propagated irrelevant context, marginally affecting ranking accuracy. These observations highlight the importance of adaptive graph pruning and relation weighting, which are promising directions for future work. Additionally, incorporating uncertainty estimation for graph edges could help mitigate the impact of noisy connections on retrieval quality.

5.7 Generalization and Transferability

Preliminary evaluation on the TREC Deep Learning 2023 dataset [47] reveals comparable performance trends, suggesting that KAGET generalizes beyond the MS MARCO benchmark. Minor performance variations observed for domain-specific terminology indicate potential benefits from task-adaptive fine-tuning. The architecture's modular design facilitates adaptation to different domains by allowing customization of the knowledge graph and fine-tuning of the underlying transformer encoder.

5.8 Limitations and Future Work

While KAGET demonstrates significant improvements over existing approaches, several limitations warrant further investigation. The current implementation relies on pre-existing knowledge graphs, which may be incomplete or biased. Future work could explore end-to-end graph construction or dynamic graph refinement during training. Additionally, the computational cost of graph operations, though manageable, could be further optimized through techniques like graph pruning or approximate message passing.

Another promising direction is extending KAGET to cross-lingual and multimodal retrieval scenarios, where graph reasoning could help bridge semantic gaps across languages and modalities. The interpretability features could also be enhanced to provide more detailed explanations of retrieval decisions, potentially incorporating natural language generation for user-friendly explanations.

5.9 Practical Implications

The results have several practical implications for real-world information retrieval systems. KAGET's combination of high accuracy, reasonable efficiency, and inherent explainability makes it suitable for deployment in domains requiring both performance and transparency, such as academic search, legal discovery, and medical information retrieval. The architecture's modularity allows for incremental improvement and adaptation to specific domain requirements.

Furthermore, the demonstrated effectiveness of combining graph reasoning with transformer-based retrieval suggests a promising direction for future retrieval system design, potentially influencing the development of next-generation search engines and recommendation systems.

6 Conclusion

The Knowledge-Aware Graph-Enhanced Transformer for Semantic Retrieval (KAGET) introduces a hybrid retrieval framework that integrates transformer-based semantic encoding, knowledge-graph reasoning, and contrastive optimization within a unified architecture. The combination of SBERT embeddings with graph convolutional refinement enables multi-hop contextual reasoning, while controlled synonym-based query expansion improves retrieval coverage without compromising precision.

Experimental evaluation on the MS MARCO benchmark demonstrates consistent and statistically significant improvements over lexical, dense, and late-interaction baselines, including BM25, SBERT, and ColBERT. These findings indicate that combining contextual semantic representations with structured relational reasoning enhances retrieval effectiveness while preserving scalability.

An additional advantage of the proposed framework is its interpretability. Entity-path tracing provides explicit relational explanations for ranking decisions, supporting transparency and trustworthiness in retrieval systems. Despite incorporating graph-based reasoning components, near real-time inference performance is maintained, making the approach suitable for large-scale deployment.

Future extensions include adaptive graph pruning to reduce redundancy, self-supervised or weakly supervised training strategies to minimize annotation requirements, and integration with large language models to further strengthen contextual reasoning. Evaluation across multilingual and domain-specific retrieval scenarios may further validate robustness and generalization.

Overall, knowledge-augmented semantic retrieval offers a practical direction for developing accurate, interpretable, and scalable information retrieval systems.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Akshara Vinod: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Writing – original draft preparation.

Linda Sara Mathew: Supervision, Validation, Writing – review and editing.

Anand Babu N B: Supervision, Methodology refinement, Writing – review and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data used in this study are publicly available. Experiments were conducted on the MS MARCO dataset [20], a large-scale human-generated machine reading comprehension and passage ranking benchmark introduced by Nguyen et al. (2016). The dataset is accessible via the official MS MARCO repository and associated evaluation platforms.

References

[1] Hambarde, K.A., Proenca, H.: Information retrieval: recent advances and beyond. IEEE Access **11**, 76581–76604 (2023)

[2] Yates, A., Nogueira, R., Lin, J.: Pretrained transformers for text ranking: Bert and beyond. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 1154–1156 (2021)

[3] Mohamed, S.M., Farah, K.S., Lotfy, A.M., Rizk, K.A., Saeed, A.Y., Mohamed, S.H., Khoriba, G., Arafa, T.: Knowledge graphs: The future of data integration and insightful discovery. In: Advanced Research Trends in Sustainable Solutions, Data Analytics, and Security, pp. 99–146. IGI Global Scientific Publishing, Hershey, PA, USA (2025)

[4] Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, pp. 593–607 (2018). Springer

[5] Gorenstein, L., Konen, E., Green, M., Klang, E.: Bidirectional encoder representations from transformers in radiology: a systematic review of natural language processing applications. Journal of the American College of Radiology **21**(6), 914–941 (2024)

[6] Wang, J., Huang, J.X., Tu, X., Wang, J., Huang, A.J., Laskar, M.T.R., Bhuiyan, A.: Utilizing bert for information retrieval: Survey, applications, resources, and challenges. ACM Computing Surveys **56**(7), 1–33 (2024)

[7] Khattab, O., Potts, C., Zaharia, M.: Baleen: Robust multi-hop reasoning at scale via condensed retrieval. Advances in Neural Information Processing Systems **34**, 27670–27682 (2021)

[8] Shunfann, W., Cheng, Y.L., Jingye, Y., Tobias, S., Manurung, Y.H., *et al.*: Locally hosted conversational process mining for production data via graph-based retrieval-augmented generation (graphrag). Journal of Mechanical Engineering (JMechE) **14**(SI), 39–59 (2025)

[9] He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., Hooi, B.: G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Advances in Neural Information Processing Systems **37**, 132876–132907 (2024)

[10] Jain, A., Mittal, K., Tayal, D.K.: Automatically incorporating context meaning for query expansion using graph connectivity measures. Progress in Artificial Intelligence **2**(2), 129–139 (2014)

[11] Xia, Y., Wu, J., Kim, S., Yu, T., Rossi, R.A., Wang, H., McAuley, J.: Knowledge-aware query expansion with large language models for textual and relational retrieval. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4275–4286 (2025)

[12] Huang, S., Ma, S., Li, Y., Li, Y., Zheng, H.-T.: From retrieval to generation: Efficient and effective entity set expansion. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 921–931 (2024)

[13] Yao, F., Zhang, J., Zhang, Y., Liu, X., Sun, C., Liu, Y., Shen, W.: Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 4783–4791 (2023)

[14] Yang, Z., Shao, Z., Dong, Y., Tang, J.: Trisampler: a better negative sampling principle for dense retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 9269–9277 (2024)

[15] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PmLR

[16] Lee, M.-C., Zhu, Q., Mavromatis, C., Han, Z., Adeshina, S., Ioannidis, V.N.,

Rangwala, H., Faloutsos, C.: Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 879–893 (2025)

[17] Huang, Y., Zhang, S., Xiao, X.: Ket-rag: A cost-efficient multi-granular indexing framework for graph-rag. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 1003–1012 (2025)

[18] Coleman, B., Kang, W.-C., Fahrbach, M., Wang, R., Hong, L., Chi, E., Cheng, D.: Unified embedding: Battle-tested feature representations for web-scale ml systems. Advances in Neural Information Processing Systems **36**, 56234–56255 (2023)

[19] Ran, W., Zhang, W., Pang, S., Zhu, Q., Liu, J., Liu, J., Cao, X., Li, Q., Yan, Y., Ma, C.: Correlated low-rank adaptation for convnets. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

[20] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)

[21] Mehta, S., Shah, D., Kulkarni, R., Caragea, C.: Semantic tokenizer for enhanced natural language processing. arXiv preprint arXiv:2304.12404 (2023)

[22] Pradhan, S.S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: Ontonotes: A unified relational semantic representation. In: International Conference on Semantic Computing (ICSC 2007), pp. 517–526 (2007). IEEE

[23] Zhelezniak, V., Savkov, A., Hammerla, N.: Estimating mutual information between dense word embeddings. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8361–8371 (2020)

[24] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021). https://openreview.net/forum?id=wCu6T5xFjeJ

[25] Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 16–20 (2011)

[26] Peters, M.E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., Smith, N.A.: Knowledge enhanced contextual word representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 43–54 (2019)

[27] Lv, Q., Ding, M., Liu, Q., Chen, Y., Feng, W., He, S., Zhou, C., Jiang, J., Dong, Y., Tang, J.: Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1150–1160 (2021)

[28] Ling, Y., Qin, Z., Ma, Z.: A review of knowledge graph construction using large language models in transportation: Problems, methods, and challenges. Transportation Research Part C: Emerging Technologies **183**, 105428 (2026)

[29] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992 (2019)

[30] Ma, X., Guo, J., Zhang, R., Fan, Y., Cheng, X.: Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 848–858 (2022)

[31] Zhang, S., Liu, M., Yan, J., Zhang, H., Huang, L., Yang, X., Lu, P.: M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2461–2470 (2022)

[32] Hemanth Sai, B., Mukherjee, S., Dubey, S.R.: Adaptive adam-based optimizers using second-order weight decoupling and gradient-aware weight decay for vision transformer. Machine Vision and Applications **36**(3), 68 (2025)

[33] Bourrier, A., Perronnin, F., Gribonval, R., Pérez, P., Jégou, H.: Explicit embeddings for nearest neighbor search with mercer kernels. Journal of Mathematical Imaging and Vision **52**(3), 459–468 (2015)

[34] Lin, S.-C., Lin, J.: A dense representation framework for lexical and semantic matching. ACM Transactions on Information Systems **41**(4), 1–29 (2023)

[35] Henriksson, A., Moen, H., Skeppstedt, M., Daudaravičius, V., Duneld, M.: Synonym extraction and abbreviation expansion with ensembles of semantic spaces. Journal of biomedical semantics **5**(1), 6 (2014)

[36] Goyal, T., Durrett, G.: Neural syntactic preordering for controlled paraphrase generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 238–252 (2020)

[37] Zhang, N., Choubey, P.K., Fabbri, A., Bernadett-Shapiro, G., Zhang, R., Mitra, P., Xiong, C., Wu, C.-S.: Sirerag: Indexing similar and related information for multihop reasoning. arXiv preprint arXiv:2412.06206 (2024)

[38] Diao, C., Loynd, R.: Relational attention: Generalizing transformers for graph-structured tasks. arXiv preprint arXiv:2210.05062 (2022)

[39] Rusak, E., Reizinger, P., Juhos, A., Bringmann, O., Zimmermann, R.S., Brendel, W.: Infonce: Identifying the gap between theory and practice. In: International Conference on Artificial Intelligence and Statistics, pp. 4159–4167 (2025). PMLR

[40] Gao, L., Zhang, Y., Han, J., Callan, J.: Scaling deep contrastive learning batch size under memory limited setup. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pp. 316–321 (2021)

[41] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised dense information retrieval with contrastive learning. arXiv e-prints, 2112 (2021)

[42] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)

[43] Song, K., Leng, Y., Tan, X., Zou, Y., Qin, T., Li, D.: Transcormer: Transformer for sentence scoring with sliding language modeling. Advances in Neural Information Processing Systems **35**, 11160–11174 (2022)

[44] Saad-Falcon, J., Fu, D.Y., Arora, S., Guha, N., Ré, C.: Benchmarking and building long-context retrieval models with loco and m2-bert. arXiv e-prints, 2402 (2024)

[45] Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond vol. 4. Now Publishers Inc, Hanover, MA, USA (2009)

[46] 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, pp. 39–48 (2020)

[47] Voorhees, E.M., Harman, D.: Overview of the sixth text retrieval conference (trec-6). Information Processing & Management **36**(1), 3–35 (2000)