

MaGiX: A Multi-Granular Adaptive Graph Intelligence Framework for Enhancing Cross-Lingual RAG

Nguyen Manh Hieu^{1*}, Vu Lam Anh^{1*}, Hung Pham Van^{2*}, Nam Le Hai¹,
Diep Thi-Ngoc Nguyen³, Linh Ngo Van^{1,†}, Thien Huu Nguyen⁴

¹Hanoi University of Science and Technology, ²Oraichain Labs Inc., US,

³VNU University of Engineering and Technology, ⁴University of Oregon,

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models by grounding their outputs in external knowledge. Recent advances in Graph-based RAG (GRAG) frameworks, such as GraphRAG, LightRAG, and HippoRAG2, integrate knowledge graphs into the retrieval process to improve multi-hop reasoning and semantic coherence. While effective in monolingual settings, these methods remain underexplored in cross-lingual scenarios and face limitations in semantic granularity and entity alignment. In this work, we propose **MaGiX**, the first GRAG framework tailored for English–Vietnamese cross-lingual question answering. MaGiX constructs a multi-granular cross-lingual knowledge graph using fine-grained attribute descriptions and cross-synonym edges, and incorporates a custom multilingual embedding model trained with contrastive learning for semantic alignment. During retrieval, MaGiX leverages graph-based reasoning and a semantic-aware reranking strategy to enhance cross-lingual relevance. Experiments across five benchmarks show that MaGiX substantially outperforms prior GRAG systems in both retrieval accuracy and generation quality, advancing structured retrieval for multilingual QA.

1 Introduction

In the rapidly evolving landscape of multilingual information access, the ability to retrieve and generate knowledge across languages is increasingly essential. While large language models (LLMs) have shown impressive capabilities in reasoning and generation, they often struggle to integrate new knowledge efficiently, particularly in multilingual or low-resource settings (Chirkova et al., 2024; Wu et al., 2024; Nguyen et al., 2025). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020)

frameworks offer a promising solution by allowing LLMs to ground their outputs in external documents. However, RAG systems face several challenges in continuous knowledge acquisition, including ensuring high-quality retrieval, effectively managing long contexts, and seamlessly integrating newly acquired knowledge (Jiang et al., 2024; Jin et al., 2024). Besides, traditional RAG systems depend heavily on dense embedding similarity and often treat documents as flat, unstructured units—limiting their performance on complex, multi-hop, or semantically rich queries (Gupta et al., 2024; Barnett et al., 2024; Tang and Yang, 2024; Gutiérrez et al., 2025).

To address these limitations, recent developments such as GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025) and HippoRAG 2 (Gutiérrez et al., 2025) have incorporated knowledge graph (KG) into the retrieval pipeline, enabling the modeling of semantic relationships between entities. Collectively, these graph-based RAG (GRAG) frameworks improve retrieval precision and reasoning performance by representing knowledge in a structured and interconnected form, which allows for more effective extraction of relevant information grounded in the contextual and relational structure of the underlying data. Overall, these GRAG methodologies generally operate in two main stages: *offline knowledge graph construction* (offline indexing) and *online graph-based inference retrieval* (online retrieval). In the offline stage, unstructured documents are processed (often via an LLM-based pipeline) to identify key entities and semantic relations, and these extracted elements are assembled into a knowledge graph; each entity node is further enriched with a concise description and each relation edge with a brief contextual annotation to capture their meaning. The indexing process also includes canonicalization, merging duplicate or co-referent entities into unified nodes to eliminate redundancy and maintain a clean graph

*Equal contribution

†Corresponding author: linhngv@soict.hust.edu.vn

index. In the online retrieval stage, the user’s query is encoded or analyzed to identify its key concepts, which are matched to corresponding seed nodes in the knowledge graph. The system then traverses outward from these seed nodes to explore neighboring nodes and retrieve a relevant subgraph. The candidate documents or chunks linked to this subgraph are aggregated and reranked by relevance, ensuring the most pertinent and contextually connected evidence for answer generation.

Although these approaches show promise of KG in enhancing RAG systems, they still exhibit several limitations. **(1) Underexploration in Cross-Lingual Settings:** Most existing GRAG implementations have seen limited application in cross-lingual question answering tasks. As a result, the potential of graph-based approaches in multilingual scenarios remains largely unexplored. **(2) Limitation in Knowledge Graph Construction** GraphRAG and LightRAG typically concatenate all descriptions of entities and relations extracted from multiple text chunks into a single representation, resulting in overly long and ambiguous summaries that hinder fine-grained query matching and accurate localization of relevant chunks. Furthermore, these frameworks often lack rich synonym edges and rely on exact string matching for entity deduplication, which is ineffective in cross-lingual contexts where semantically equivalent entities may appear in different surface forms. While HippoRAG 2 introduces synonym edges based on entity name embedding similarity, this method remains limited in multilingual scenarios, as it depends solely on short entity names and fails to leverage the richer semantic from contextual descriptions. **(3) Limitation in Graph-based Inference Retrieval** Prior GRAG frameworks typically rank or score candidate text chunks based on the connectivity of retrieved seed nodes within the graph. For instance, LightRAG prioritizes passages containing a high number of neighboring entities to the seed node, while HippoRAG 2 leverages Personalized PageRank (PPR), which inherently favors nodes with greater structural connectivity. However, such approaches may overemphasize co-occurrence patterns and underweight semantic relevance, potentially biasing retrieval toward text chunks that are densely linked rather than contextually appropriate. This limitation may be further exacerbated in cross-lingual scenarios, where semantically aligned entities across languages are often not directly connected or co-located in the same passage, leading

to the misranking for a text chunk and underestimating its relevance level.

In this paper, we present MaGiX —a Multi-Granular Adaptive Graph Intelligence framework that is the first to explore Graph-based RAG techniques for English-Vietnamese cross-lingual question answering. By addressing these challenges, MaGiX not only expands the applicability of GRAG to low-resource languages but also offers methodological innovations that improve retrieval accuracy and cross-lingual alignment (comparison with prior GRAG models shown in Table 5). In summary, our contributions are as follows:

- **Cross-Lingual KG Construction with Cross Semantic Canonicalization:** MaGiX constructs a cross-lingual knowledge graph by linking English and Vietnamese entities through *fine-grained attribute descriptions* extracted from diverse text chunks (cross-synonym edge). To further capture semantic equivalence across languages, we introduce a *custom embedding model* using contrastive learning and bilingual data, outperforming pretrained embeddings used in prior systems.
- **Granular Graph Reasoning for Retrieval:** Rather than relying on single unified representations (i.e. concatenated or summarized description), MaGiX considers *all contextual descriptions* linked to each entity and relation, improving text chunk localization. We further enhance retrieval by traversing across cross-synonym edges and provide a custom *multi-aspect, semantic-aware reranking strategy*, enabling robust semantic matching in cross-lingual scenarios.
- **Empirical Results on Cross-Lingual Benchmarks:** Our comprehensive experiments on five benchmarks show that MaGiX achieves *substantial improvements* over existing GRAG frameworks in both retrieval accuracy and generation quality, highlighting the effectiveness of our proposed methods across both the offline graph construction and online retrieval stages.

2 Related work

2.1 Graph-based RAG

Recent work has explored integrating knowledge graphs into Retrieval-Augmented Generation (RAG) to improve structured context retrieval, enhance LLM reasoning, and address challenges in

complex and multi-hop queries where naive RAG falls short. GraphRAG (Edge et al., 2024) constructs hierarchical knowledge graphs by extracting entities and relations from text and grouping them into communities, enabling multi-level retrieval and improving summarization tasks, though at high computational cost. To address this, LightRAG (Guo et al., 2025) introduces a more lightweight, update-friendly framework that embeds graph structure directly into retrieval. PathRAG (Chen et al., 2025) further improves coherence by pruning redundant information along relational paths. Inspired by human memory, HippoRAG (Gutiérrez et al., 2024) and HippoRAG 2 (Gutiérrez et al., 2025) applies Personalized PageRank over knowledge graphs to support continual knowledge integration, achieving strong results in multi-hop reasoning tasks.

2.2 Cross-lingual Information Retrieval

Cross-lingual Information Retrieval (CLIR) tackles the challenge of retrieving relevant documents written in a language different from that of the user’s query. Traditional CLIR techniques often rely on machine translation, bilingual dictionaries, or parallel corpora to bridge the language gap (Zbib et al., 2019; Litschko et al., 2021; Shi et al., 2021). However, these methods frequently suffer from translation errors and are particularly limited in coverage and quality for low-resource languages. In response, recent research has shifted toward neural retrieval models that employ cross-lingual embeddings to project text from different languages into a shared semantic space, allowing for retrieval without explicit translation (Pires et al., 2019; Conneau et al., 2019; Chen et al., 2024).

For example, Guo et al. (2024) enhance Chinese–English CLIR by guiding large language models using bilingual dictionaries and comparable corpora. Similarly, the WACSE framework (Miao et al., 2024) aligns word-level semantics across English and low-resource languages such as Tagalog, Vietnamese, and Swahili using word alignment models and translation-based learning objectives. Specifically in the English–Vietnamese setting, Nguyen et al. (2025) improve cross-lingual retrieval by constructing a bilingual corpus and training cross-lingual embedding models with auxiliary losses and synthetic data.

Despite these advancements, the use of graph-based retrieval-augmented generation (GRAG) remains largely unexplored in cross-lingual or mul-

tilingual scenarios, particularly for low-resource language pairs like English and Vietnamese. This presents a critical opportunity to investigate how structured knowledge representations can further enhance semantic alignment and retrieval effectiveness across languages.

3 Methodology

3.1 Overview

Following previous studies, our method comprises two main stages: offline KG construction and online graph-based inference retrieval. The overall of our methodology is depicted in Figure 1.

Offline KG construction (1) We employ an LLM to extract triples from raw text chunks, along with enclosed descriptions for each entity, following the approach of the LightRAG framework (Guo et al., 2025). (2) Each entity node may appear in multiple text chunks, yielding several context-specific attribute descriptions. We concatenate each with the entity name and encode them independently, producing N attribute embeddings for N occurrences. This fine-grained representation addresses limitations in prior work: LightRAG compresses all descriptions into one, losing detail, while HippoRAG uses only raw entity names without context. To detect synonyms, we apply a transformer-based encoder for semantic matching and introduce cross-synonym edges to link equivalent entities across languages (3.2). Cross-lingual alignment is further improved through a custom embedding model with multi-stage training strategy (3.3) that enhances semantic consistency across languages. (3) To maintain a balance between contextual and conceptual information, as emphasized in HippoRAG 2 (Gutiérrez et al., 2025), we preserve the linkage between graph nodes and edges and their corresponding raw text chunks, along with the attribute descriptions of entities and relation-specific descriptions extracted from those texts.

Online Graph-based Inference Retrieval: (1) Given an input query, we first encode it into an embedding vector, which is then used to perform semantic search over both attribute node descriptions and edge descriptions. Based on the highest cosine similarity scores, we identify the most relevant seed nodes and edges, and retrieve their associated text chunks as candidate contexts. (2) To enhance retrieval coverage, we expand the candidate set by

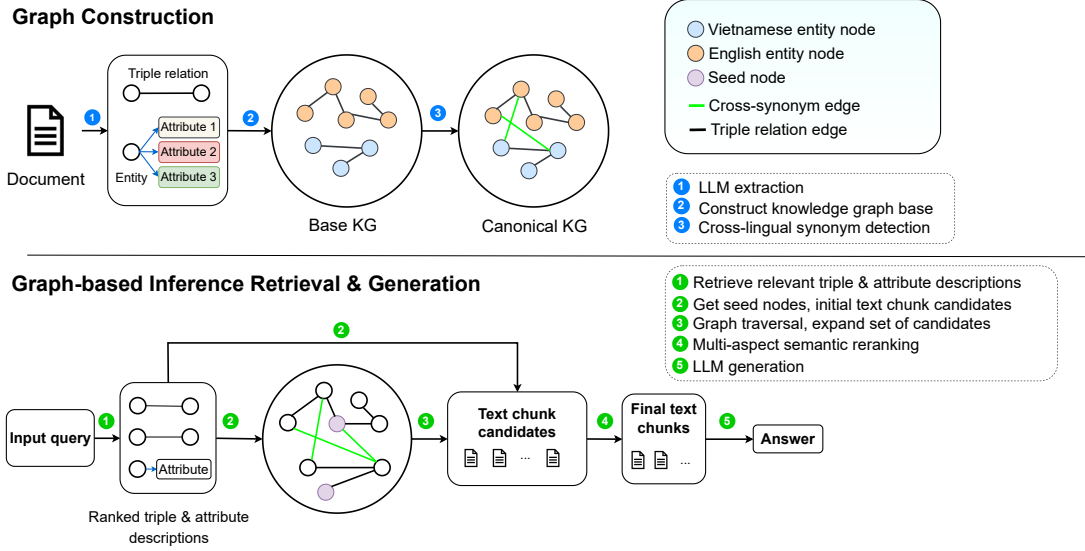


Figure 1: **MaGiX architecture overview.** The knowledge graph is built from extracted entities, relations, and attributes, with synonym edges ($\mathcal{CR}$) added via contextual similarity. During retrieval, queries match node attributes and triple descriptions to identify seed nodes and initial candidates. The candidate set is expanded via $\mathcal{CR}$ traversal. A semantic-aware reranking strategy selects the top- n text chunks for LLM generation.

traversing the graph from seed nodes, leveraging cross-synonym edge ($\mathcal{CR}$) to identify cousin nodes. For each cousin node, we select text chunks whose attribute embeddings are most similar to the query embedding, thereby increasing matching capacity by effectively utilizing the graph structure (3.4). (3) Text chunks are scored based on three score components that collectively capture multi-aspect semantic alignment (3.5). For robust cross-lingual matching, a graph-based score boosting mechanism is applied first, where, for nodes with cross-synonym relations, the attribute description score is updated to the maximum cosine similarity score of attribute descriptions from their cross-lingual cousin nodes to prioritize synonymous entities. The top- n chunks are then ranked by their composite scores and selected for LLM generation.

3.2 Cross-lingual Semantic Canonicalization

The construction of a knowledge graph from unstructured text involves extracting relational triplets of the form entity-relation-entity from a set of text chunks. A key challenge in this process is the *canonicalization problem*, where LLMs generate varied representations of the same entity due to their inherent randomness and linguistic diversity. For instance, the entity referring to the capital of Vietnam may appear as “Hanoi city,” “Hanoi,” “Capital Hanoi,” or historical names like “Thang Long” or “Ke Cho”. Our approach employs a fine-tuned multilingual embedding model tailored to

the target domain, enabling more effective semantic matching by leveraging contextual attributes extracted from text chunks to improve synonym detection.

Given a set of text chunks $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$, we utilize an LLM to extract a set of relational triplets $\mathcal{T} = \{(e_i, r_{ij}, e_j)\}$, where $e_i, e_j \in \mathcal{E}$ are entities, and the relation description $r_{ij} \in \mathcal{R}$ is the edge connecting them. Each text chunk $c_k \in \mathcal{C}$ may yield many triplets, and each entity e_i extracted from c_k is associated with a *contextual attribute* a_{ik} , which encapsulates the descriptive context of e_i within c_k . Formally, for an entity e_i , the set of contextual attributes is defined as:

$$A_i = \{a_{ik} \mid e_i \text{ is extracted from } c_k, c_k \in \mathcal{C}\}$$

To address the canonicalization problem, we encode each entity using its name and contextual attributes to create attribute embeddings. For an entity e_i with name n_i and contextual attribute a_{ik} , the attribute embedding is generated by concatenating the entity name and attribute, followed by encoding via a fine-tuned multilingual embedding model f :

$$\mathbf{v}_{ik} = f(n_i \parallel a_{ik})$$

where $\parallel$ denotes string concatenation, $\mathbf{v}_{ik} \in \mathbb{R}^d$ is the d -dimensional embedding vector and the set of attribute embeddings for e_i is:

$$V_i = \{\mathbf{v}_{ik} \mid a_{ik} \in A_i\}$$

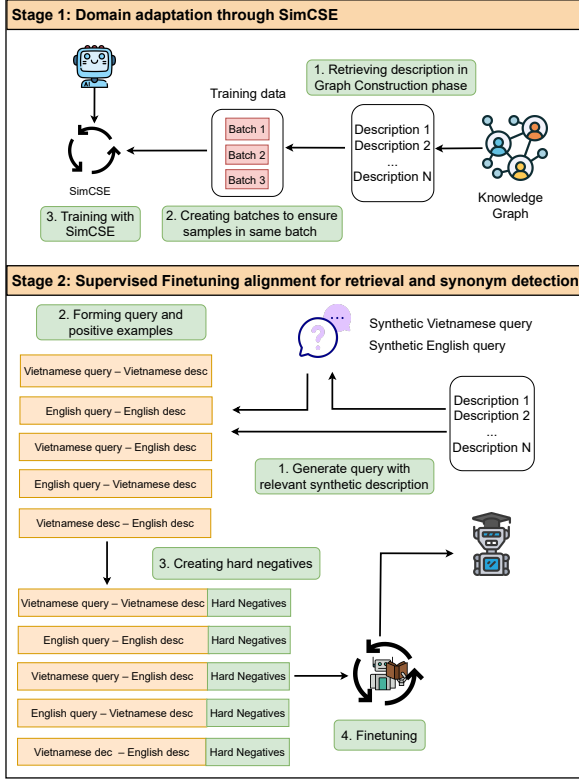


Figure 2: A multi-stage self-supervised pipeline for domain adaptation and cross-lingual alignment.

This approach preserves the semantic context of each text chunk in the embedding space, supporting robust entity comparison and improving online retrieval. To deduplicate entities and identify synonyms, we compute cosine similarity between all pairs of attribute embeddings from e_i and e_j .

$$s(\mathbf{v}_{ik}, \mathbf{v}_{jl}) = \frac{\mathbf{v}_{ik} \cdot \mathbf{v}_{jl}}{\|\mathbf{v}_{ik}\| \|\mathbf{v}_{jl}\|},$$

$$\forall \mathbf{v}_{ik} \in V_i, \mathbf{v}_{jl} \in V_j$$

If there exists any pair (k, l) such that $s(v_{ik}, v_{jl}) > \tau$, where $\tau \in (0, 1)$ is a predefined threshold, we establish a cousin relation/synonym edge between e_i and e_j . if the entities belong to different languages, this edge is designated as a cross-synonym connection. The set of cousin relations is:

$$\mathcal{CR} = \{(e_i, e_j) \mid \exists k, l \text{ s.t. } s(\mathbf{v}_{ik}, \mathbf{v}_{jl}) > \tau\}$$

This relation is represented as an edge in the graph, facilitating online retrieval of synonymous entities.

3.3 Domain Adaptation and Cross-Lingual Alignment

To develop a robust embedding model for multilingual and cross-lingual retrieval tasks, we propose

a multi-stage fine-tuning strategy that leverages synthetic data to enhance domain adaptation and cross-lingual alignment. The complete fine-tuning pipeline is depicted in the Figure 2. Further details on the training objectives for each stage are provided in Appendix B.

Stage 1: Self-supervised Fine-Tuning for Domain Adaptation To enhance domain adaptation, we adopt the SimCSE (Gao et al., 2021) contrastive learning framework for unsupervised training on synthetic English and Vietnamese attribute descriptions generated during graph construction. This step enables the model to capture domain-specific semantics in graph-based attributes, yielding embeddings optimized for canonicalization.

Stage 2: Supervised Fine-Tuning for Cross-Lingual Alignment The second stage performs supervised fine-tuning to align embeddings across English and Vietnamese for effective cross-lingual retrieval. The model is trained using in-batch contrastive loss with hard negatives, based on five types of synthetic bilingual query-description pairs, as illustrated in Figure 2. This setup encourages high similarity for correct pairs and low similarity for unrelated ones, enhancing the alignment of entity and relation embeddings across languages.

3.4 Granularity Retrieval

In the online retrieval phase, our method leverages the fine-grained attribute embeddings constructed during the offline phase to enable precise and contextually relevant retrieval of text chunks. This section details the granularity retrieval process, which enhances the explorability and accuracy of retrieved contexts by incorporating multi-level semantic matching and graph traversal.

Given a query q , we encode it using the fine-tuned multilingual model f to obtain the embedding $\mathbf{v}_q = f(q) \in \mathbb{R}^d$. A semantic search is performed over attribute embeddings $\mathbf{v}_{ik}$ of entity nodes and edge description embeddings $\mathbf{v}_{r_{ij}}$, selecting the top- k most similar items as the seed set $\mathcal{S}$. For each $\mathbf{v}_{ik} \in \mathcal{S}$, we retrieve the corresponding text chunk c_{ik} containing the attribute a_{ik} . Similarly, for each $\mathbf{v}_{r_{ij}} = f(d_{r_{ij}}) \in \mathcal{S}$, where $d_{r_{ij}}$ is the edge description, we retrieve the chunk c_{ij} containing $d_{r_{ij}}$. These form the initial text chunk candidates.

$$\mathcal{C}_{\text{init}} = \{c_{ik} \mid \mathbf{v}_{ik} \in \mathcal{S}\} \cup \{c_{ij} \mid \mathbf{v}_{r_{ij}} \in \mathcal{S}\}.$$

To extend explorability, we leverage the graph structure by traversing cousin relations defined in Section 3.2. For each attribute embedding $\mathbf{v}_{ik}$ in the top- k set, we identify the corresponding entity node e_i . We then collect all cousin nodes e_j such that $(e_i, \text{cousin}, e_j) \in \mathcal{CR}$. For each cousin node e_j , we compute the cosine similarity between $\mathbf{v}_q$ and each attribute embedding $\mathbf{v}_{jl} \in V_j$, selecting the attribute embedding with the highest similarity score:

$$\mathbf{v}_{jl^*} = \arg \max_{\mathbf{v}_{jl} \in V_j} s(\mathbf{v}_q, \mathbf{v}_{jl}),$$

$$\text{where } s(\mathbf{v}_q, \mathbf{v}_{jl}) = \frac{\mathbf{v}_q \cdot \mathbf{v}_{jl}}{\|\mathbf{v}_q\| \|\mathbf{v}_{jl}\|}$$

We include the text chunk c_{jl^*} containing the attribute description a_{jl^*} associated with $\mathbf{v}_{jl^*}$. The set of cousin text chunk candidates is:

$$\mathcal{C}_{\text{cousin}} = \{c_{jl^*} \mid (e_i, e_j) \in \mathcal{CR}, \mathbf{v}_{ik} \in \mathcal{S}\} \quad (1)$$

The final set of text chunk candidates is the union of the initial and cousin candidates:

$$\mathcal{C}_{\text{cand}} = \mathcal{C}_{\text{init}} \cup \mathcal{C}_{\text{cousin}} \quad (2)$$

3.5 Scoring and Reranking Mechanism

Given the candidate text chunks $\mathcal{C}_{\text{cand}}$, we rank them using a weighted average of three normalized components, enhanced by a graph-based boosting mechanism for cross-lingual alignment.

Scoring Components For each chunk $c \in \mathcal{C}_{\text{cand}}$, we consider three semantic aspects:

- **Chunk Similarity** (s_{chunk}): Cosine similarity between the query embedding $\mathbf{v}_q$ and chunk embedding $\mathbf{v}_c = f(c)$.
- **Attribute Similarity** (s_{attr}): Average similarity of top- m entity attribute embeddings $\mathbf{v}_{ik} \in \mathcal{S}$ in c to $\mathbf{v}_q$.
- **Triple Similarity** (s_{triple}): Average similarity of top- m triple embeddings $\mathbf{v}_{rij} \in \mathcal{S}$ in c to $\mathbf{v}_q$.

Cross-Lingual Score Boosting To boost s_{attr} , we incorporate similarities from cross-lingual cousin nodes via $\mathcal{CR}$:

$$s'_{\text{attr}}(e_i) = \max \left(\begin{aligned} &\{s(\mathbf{v}_q, \mathbf{v}_{ik}) \mid \mathbf{v}_{ik} \in \mathcal{S}\} \\ &\cup \{s(\mathbf{v}_q, \mathbf{v}_{jl}) \mid \mathbf{v}_{jl} \in \mathcal{S}, \\ &(e_i, e_j) \in \mathcal{CR}\} \end{aligned} \right). \quad (3)$$

The updated $s_{\text{attr}}(c)$ is the average of the top- m boosted $s'_{\text{attr}}(e_i)$.

Composite Score and Reranking Each score is min-max normalized:

$$s_x^{\text{norm}}(c) = \frac{s_x(c) - \min s_x}{\max s_x - \min s_x}, \quad (4)$$

$$x \in \{\text{chunk}, \text{attr}, \text{triple}\}.$$

Final score:

$$s_{\text{composite}}(c) = \sum_{x \in \{\text{chunk}, \text{attr}, \text{triple}\}} w_x s_x^{\text{norm}}(c) \quad (5)$$

where w_{chunk} , w_{attr} and w_{triple} are weighting coefficients used to balance the contribution of each score component. Top- n ranked chunks form $\mathcal{C}_{\text{final}}$ for LLM generation.

4 Experiments

4.1 Experimental Setup

More details on dataset statistics, implementation, and evaluation metrics are provided in Appendix C.

Datasets. To evaluate our approach on both simple and multi-hop QA, we use diverse datasets across domains. For simple QA, we include two Vietnamese benchmarks: ZaloWikipediaQA (Zalo AI Team, 2019) (general topics) and ZaloLegal2021 (Zalo AI Team, 2021) (legal domain), along with two English datasets from HippoRAG 2: 1,000 samples each from NaturalQuestions (NQ) (Wang et al., 2024) and PopQA (Mallen et al., 2022), using the December 2021 Wikipedia. For multi-hop QA, we sample 1,000 queries from MuSiQue (Trivedi et al., 2022), which requires reasoning over multiple passages. All datasets are translated between Vietnamese and English using Gemini 2.0 Flash (Google DeepMind, 2024) to create cross-lingual QA scenarios.

Baselines. We compare our approach against two categories of baselines: (1) a NaiveRAG setup using the multilingual dense retriever BGE-M3 (Chen et al., 2024), and (2) GRAG methods, including GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025), and HippoRAG 2 (Gutiérrez et al., 2025), all using Gemini 2.0 Flash (Google DeepMind, 2024) for both knowledge extraction and reasoning, with BGE-M3 as the retrieval encoder.

4.2 Main Results

In this section, we evaluate the principal experimental outcomes for English-Vietnamese cross-lingual

Retrieval	Simple QA								Multi-Hop QA	
	ZaloWikipediaQA		ZaloLegal2021		NQ		Popqa		MuSiQue	
	vi-en	en-vi	vi-en	en-vi	vi-en	en-vi	vi-en	en-vi	vi-en	en-vi
<i>Dense Retrieval</i>										
BGE-M3	76.29	<u>75.92</u>	64.17	61.95	81.62	81.63	45.15	46.22	46.58	44.86
<i>Graph-based Retrieval</i>										
GraphRAG - Local	18.24	32.80	18.76	16.06	22.99	19.47	17.12	18.14	17.25	15.44
LightRAG - Local	17.05	27.17	16.61	16.61	23.11	27.71	23.32	26.65	20.30	22.01
LightRAG - Global	40.79	48.01	15.57	15.57	15.54	18.02	14.27	17.52	13.81	10.05
HippoRAG 2	<u>79.27</u>	75.26	<u>66.19</u>	60.23	<u>84.56</u>	<u>84.14</u>	<u>49.45</u>	<u>48.49</u>	<u>55.53</u>	<u>50.98</u>
MaGiX	81.32	85.45	68.02	65.26	87.79	87.27	50.65	50.85	60.37	58.43

Table 1: Vietnamese-to-English (Vi-En) and English-to-Vietnamese (En-Vi) Passage Retrieval Performance (Recall@k=10) across four simple QA datasets and one multi-hop QA dataset. In the table, we highlight the **best** score in bold and the second-best score with an underline.

Method	NQ	PopQA	MuSiQue
LightRAG	23.11	23.32	20.30
+ Granular Retrieval	78.90	49.10	48.86
+ Cross-Synonym	80.73	49.20	51.21
+ Composite Score	88.58	50.08	56.96
+ Finetune (MaGiX)	87.79	50.65	60.37

Table 2: Ablation: Recall@10 for Vietnamese-to-English settings on two simple QA datasets (NQ and PopQA) and one multi-hop QA dataset (MuSiQue).

passage retrieval (Tables 1) and QA (Table 3) tasks, providing a comprehensive analysis of performance across diverse datasets.

Performance on Cross-lingual Retrieval: Table 3 presents retrieval results across five QA datasets, covering both simple and multi-hop queries under cross-lingual settings (vi-en and en-vi). The results compare our proposed method, MaGiX, against dense retriever and various GRAG baselines. MaGiX consistently achieves the best performance across all scenarios, clearly demonstrating its effectiveness in cross-lingual retrieval. For example, on the ZaloWikipediaQA dataset (en-vi), MaGiX achieves a score of 85.45%, outperforming both HippoRAG 2 and BGE-M3 by nearly 10%. On the multi-hop MuSiQue benchmark, MaGiX attains scores of 60.37% and 58.43% for vi-en and en-vi settings, respectively, both substantially outperforming the second-best baseline, HippoRAG 2.

Performance on Cross-lingual QA: Table 3 presents the win-rate evaluation results of MaGiX against four baselines in five evaluation dimensions: Accuracy, Comprehensiveness, Diversity, Logical-

ity, and Coherence. Judgments are conducted using Grok-3 and detailed in Appendix C. MaGiX consistently outperforms all baselines across nearly every metric and dataset, demonstrating strong generation quality in both simple and multi-hop cross-lingual QA. For instance, on the multi-hop MuSiQue dataset, MaGiX outperforms all baselines in every evaluation aspect, achieving win rates exceeding 65%. Similarly, on the simple QA benchmarks, MaGiX consistently outperforms all baselines, achieving high win rates across most evaluation dimensions. These results highlight MaGiX’s strengths in generating coherent, contextually rich, and logically consistent answers across languages enabled by its fine-grained attribute modeling, cross-lingual alignment strategies, and semantic-aware reranking.

4.3 Ablation Study

In this section, we present ablation experiments to evaluate the effectiveness of each proposed technique in MaGiX, as well as the contribution of individual components within the ranking strategy. Additional experiments and analysis are provided in Appendix D.

Effectiveness of Key Components in MaGiX:

Table 2 demonstrates that building upon LightRAG, the incremental integration of granular retrieval, cross-synonym linking, and composite scoring results in consistent performance gains. Adding granular retrieval—which encodes multiple fine-grained attribute descriptions per entity—results in a substantial improvement, particularly for NQ (from 23.11 to 78.90) and MuSiQue (from 20.30 to 48.86), demonstrating the importance of main-

	ZaloWikipedia		ZaloLegal2021		NQ		Popqa		MuSiQue	
	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX
Accuracy	32.60	67.40	38.12	61.88	38.83	61.17	38.12	69.07	32.53	67.47
Comprehensiveness	35.14	64.86	45.16	54.84	46.25	53.75	34.13	65.87	30.36	69.64
Diversity	38.53	61.47	49.84	50.16	50.35	49.65	35.79	64.21	33.47	66.53
Logicality	28.86	71.14	40.47	59.53	35.82	64.18	40.47	72.31	29.84	70.16
Coherence	29.16	70.84	40.00	60.00	34.63	65.37	28.27	71.73	26.49	73.51
Overall	33.70	66.30	41.25	58.75	42.50	57.50	33.33	66.67	31.66	68.34
	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX
Accuracy	31.70	68.30	23.44	76.56	26.00	74.00	22.57	77.43	28.01	71.99
Comprehensiveness	33.50	66.50	43.91	56.09	54.30	45.70	49.75	50.25	27.86	72.14
Diversity	41.80	58.20	59.94	40.06	66.90	33.10	57.01	42.99	33.87	66.13
Logicality	31.60	68.40	34.69	65.31	35.40	64.60	30.69	69.31	21.56	78.44
Coherence	31.10	68.90	41.88	58.13	39.22	60.78	34.37	65.53	16.83	83.17
Overall	32.60	67.40	34.84	65.16	40.70	59.30	36.44	63.56	23.55	76.45
	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX
Accuracy	17.45	82.55	21.25	78.75	25.15	74.85	20.56	79.44	24.10	75.90
Comprehensiveness	55.10	44.90	41.41	58.59	51.90	48.10	39.24	60.76	22.55	77.45
Diversity	67.07	32.93	60.31	39.69	64.10	35.90	47.09	52.91	29.59	70.41
Logicality	34.53	65.47	31.92	68.08	33.20	66.80	25.85	71.15	17.85	82.15
Coherence	41.90	58.10	37.52	62.48	39.24	60.76	26.75	73.25	13.33	86.67
Overall	17.45	82.55	31.09	68.91	40.70	59.30	27.93	72.07	20.14	79.86
	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX
Accuracy	28.96	71.04	40.69	59.31	32.49	67.51	25.30	74.70	21.23	78.77
Comprehensiveness	13.70	86.30	42.19	57.81	8.50	91.50	11.21	88.79	20.14	79.86
Diversity	18.00	82.00	49.22	50.78	12.00	88.00	14.31	85.69	27.61	72.39
Logicality	21.30	78.70	39.06	60.94	18.34	81.66	15.41	85.49	17.84	82.16
Coherence	11.00	89.00	37.34	62.66	6.70	93.30	7.11	92.89	12.32	87.68
Overall	17.20	82.80	40.00	60.00	13.80	86.20	12.71	87.29	17.23	82.77

Table 3: Performance on Vietnamese-to-English QA across five datasets and five evaluation dimensions, measured by win rates (%) based on judgments from the Grok-3. Additional results for English-to-Vietnamese QA are provided in Table 9 in the Appendix.

Composite Score	NQ	PopQA	MuSiQue
Ours	87.79	50.65	60.37
w/o s_{chunk}	73.91	47.55	49.91
w/o s_{attr}	87.37	49.70	58.34
w/o s_{triple}	85.69	48.50	57.12

Table 4: Ablation study on vi-en CLIR setting, Recall@10 on two simple QA datasets (NQ and PopQA) and one multi-hop QA dataset (MuSiQue), evaluating the effect of removing individual components from the Composite Score.

taining contextual semantics at the attribute level. Cross-synonym edges further strengthen cross-lingual alignment, while the composite scoring strategy improves performance across all datasets, highlighting the benefit of combining multiple semantic signals. The final MaGiX model, enhanced with fine-tuned embeddings, yields the best results on MuSiQue and PopQA, though it slightly drops on NQ, possibly due to the synthetic fine-tuning data introducing minor noise or fine-tuning favoring complex cross-lingual patterns.

Effectiveness of each Components in Scoring Mechanism: As described in Section 3.5, we introduce a semantic-aware scoring framework that considers three key aspects for ranking candidate

text chunks. Table 4 shows that removing any single component leads to a significant drop in retrieval performance, indicating that each semantic-aware aspect contributes meaningfully to the effectiveness of the proposed ranking strategy.

5 Conclusion

In this work, we introduced MaGiX, the first Graph-based RAG framework designed for English–Vietnamese cross-lingual question answering. By addressing key limitations in prior GRAG systems; such as limited multilingual adaptability, lack of fine-grained entity modeling, and reliance on shallow graph connectivity, we proposed a multi-granular architecture that integrates contextual attribute descriptions, cross-synonym reasoning, and multi-aspect semantic-aware reranking. Our framework leverages a domain-adapted multilingual embedding model and a structured retrieval strategy to achieve robust semantic alignment across languages. Extensive experiments on diverse cross-lingual benchmarks demonstrate that MaGiX significantly improves both retrieval and generation quality. We believe this work lays the foundation for future research in cross-lingual structured retrieval, especially in low-resource and complex multilingual settings.

Limitations

While MaGiX demonstrates strong performance in English–Vietnamese cross-lingual question answering, its current evaluation is limited to this language pair. Although Vietnamese presents meaningful challenges due to its syntactic differences and low-resource characteristics, the generalizability of our approach to other languages, especially those with different morphological structures or less available bilingual data, has not yet been tested in our scope. In addition, although synthetic data and bilingual embeddings offer promising results, their effectiveness may vary across domains or languages with weak translation models. We leave a broader multilingual evaluation and further refinement of domain adaptation strategies for future work.

Acknowledgements

This research has been supported by the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

References

- Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199.
- Boyue Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. Pathrag: Pruning graph-based retrieval augmented generation with relational paths. *arXiv preprint arXiv:2502.14902*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multi-lingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.
- Google DeepMind. 2024. Gemini 2.0 Flash: Multimodal Large Language Model. <https://developers.googleblog.com/2024/12/gemini-20-flash-experimental.html>. Accessed: May 16, 2025.
- Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–596.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. Lightrag: Simple and fast retrieval-augmented generation. *Preprint*, arXiv:2410.05779.
- Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.
- Ziyan Jiang, Xueguang Ma, and Wenhui Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

- Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I* 43, pages 342–358. Springer.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.
- Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. *arXiv preprint arXiv:2404.02490*.
- Toan Ngoc Nguyen, Nam Le Hai, Nguyen Doan Hieu, Dai An Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Sang Dinh. 2025. Improving vietnamese-english cross-lingual retrieval for legal and general domains. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 142–153.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.
- Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training with dense retrieval for document retrieval. *arXiv preprint arXiv:2109.01628*.
- Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*.
- Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *arXiv preprint arXiv:2410.21970*.
- Zalo AI Team. 2019. Zalo challenge dataset. Zalo AI Challenge 2019, <https://challenge.zalo.ai>. Accessed: 2025-02-08.
- Zalo AI Team. 2021. Zalo challenge dataset. Zalo AI Challenge 2021, <https://challenge.zalo.ai>. Accessed: 2025-02-08.
- Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, and 1 others. 2019. Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

Appendix

A Comparison between MaGiX and prior GRAG frameworks

	GraphRAG/LightRAG	HippoRAG 2	MaGiX
Node/edge description	Concatenated and Summarization	✗	Fine-grained
CS Edges	✗	✗	✓
Finetuning Embedding Model	✗	✗	✓
Node/Edge retrieval	Unified Description Retrieval	Triple Retrieval	Granular Description Retrieval
Extend searching with CS Edges	✗	✗	✓
Passage Ranking Strategy	Node Structural Connectivity	Personalize PageRank	Multi-Aspect Semantic-Aware

Table 5: **Comparison of Multi-Step Process between existing GRAG and MaGiX.** from Graph Construction to Graph Inference. In the table, the color **Green** denotes the Graph Construction stage, the color **Orange** denotes Graph Inference stage. Moreover, CS abbreviates for Cross-Synonym.

B Domain Adaptation and Cross-Lingual Alignment

To develop a robust embedding model for multilingual and cross-lingual retrieval tasks, we propose a multi-stage fine-tuning strategy that leverages synthetic data to enhance domain adaptation and cross-lingual alignment. The complete fine-tuning pipeline is depicted in the Figure 2.

Stage 1: Self-supervised Fine-Tuning for Domain Adaptation To better adapt to the domain, we employ the SimCSE (Gao et al., 2021) contrastive learning framework, training on synthetic English-Vietnamese attribute descriptions generated during graph construction. This stage helps the model capture domain-specific nuances in graph-derived attributes, producing embeddings well-suited for canonicalization.

In the training process, each description d_i in a batch of size B is encoded twice using different dropout masks, yielding two embeddings z_i^1 and z_i^2 . The SimCSE loss function maximizes the cosine similarity between these paired embeddings, while treating the embeddings of other descriptions in the batch as negative examples. The loss is defined as:

$$\mathcal{L}_{\text{SimCSE}} = -\frac{1}{2N} \sum_{i=1}^N \left[\log \frac{\exp(\text{sim}(z_i^1, z_i^2)/t)}{\sum_{j=1}^B \exp(\text{sim}(z_i^1, z_j^2)/t)} + \log \frac{\exp(\text{sim}(z_i^2, z_i^1)/t)}{\sum_{j=1}^B \exp(\text{sim}(z_i^2, z_j^1)/t)} \right] \quad (6)$$

where $\text{sim}(a, b) = \frac{a \cdot b}{\|a\| \|b\|}$ is the cosine similarity, t is the temperature parameter. This loss ensures that the model learns robust, domain-adapted embeddings for the synthetic descriptions, laying a strong foundation for cross-lingual tasks.

Stage 2: Supervised Fine-Tuning for Cross-Lingual Alignment The second stage performs supervised fine-tuning to align embeddings across English and Vietnamese, enabling effective cross-lingual retrieval. It focuses on matching queries to their corresponding descriptions, with emphasis on cross-lingual pairs (e.g., English queries and Vietnamese descriptions). To improve cosine similarity between aligned entity and relation embeddings, English–Vietnamese description pairs are also included. Synthetic training data is generated from Stage 1 descriptions by creating user-like queries in both languages, forming five types of training pairs, as illustrated in Figure 2. Each set includes positive and hard negative pairs to

enhance the model’s discriminative ability. The fine-tuning process trains the model on all five datasets simultaneously, using an in-batch contrastive loss that incorporates hard negatives to improve alignment accuracy. For each query q_i in a batch of size B , with its positive description d_i , and a set of hard negatives $\{h_{i,k}\}_{k=1}^K$, the loss is:

$$\mathcal{L}_i = -\log \left[\frac{\frac{\exp(\text{sim}(q_i, d_i)/t)}{\sum_{j=1}^B \exp(\text{sim}(q_i, d_j)/t)}}{1 + \sum_{k=1}^K \exp(\text{sim}(q_i, h_{i,k})/t)} \right] \quad (7)$$

where $\text{sim}(a, b)$ is the cosine similarity, t is the temperature parameter, and K is the number of hard negatives per query. The total loss is:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \mathcal{L}_i \quad (8)$$

This loss encourages the model to assign high similarity to correct query-description pairs, especially across languages, while effectively distinguishing them from both in-batch and hard negative examples.

C Additional Experimental Settings

Evaluation Metrics. Following prior work in cross-lingual document retrieval (Gutiérrez et al., 2025; Nguyen et al., 2025), we use Recall@10 as our evaluation metric for text chunk retrieval. To evaluate answer quality, inspired by previous approaches (Edge et al., 2024; Guo et al., 2025; Chen et al., 2025), we utilize LLMs (i.e. Grok-3) as evaluators to assess generated answers in relation to queries and ground truth documents. Our LLM evaluators examine five key dimensions: **Accuracy** (factual correctness), **Comprehensiveness** (topic coverage depth), **Diversity** (inclusion of multiple perspectives), **Logicity** (logical response to all parts of the question), and **Coherence** (internal consistency and smooth structure). For each query, we compare two answers across these five criteria, determining the superior response as the one that outperforms in more criteria. Win rates are calculated by averaging these results across all queries. The prompt used for LLM-based evaluation is provided below.

Implementation Details. In our experiments, to set up English-Vietnamese cross-lingual setting, we first translate all Vietnamese documents and queries to English version and vice versa. By utilizing Gemini 2.0 Flash (Google DeepMind, 2024) model we translate ZaloWikipediaQA (Zalo AI Team, 2019) and ZaloLegal2021 (Zalo AI Team, 2021) datasets from Vietnamese into English and reverse direction for the NQ (Wang et al., 2024), PopQA (Mallen et al., 2022), and MuSiQue datasets (Trivedi et al., 2022). We then construct bilingual knowledge graph based on the documents and their translated documents using a procedure adapted from LightRAG (Guo et al., 2025), except that we split entity and relation descriptions which tell specific aspect of entity and relations in a text chunk rather than concatenating them into a single node summary. To align embeddings across languages, we synthesize queries, keywords, entities, and relation descriptions from documents to fine-tune the BGE-M3 embedding model (Chen et al., 2024) with SimCSE (Gao et al., 2021) in two stages. In the unsupervised phase, we apply standard SimCSE on synthesized entities and relations documents for improving domain adaptation with the batch samples enforced low intra-batch semantic similarity (< 0.7) to maximize the quality of SimCSE training. In the supervised stage, we fine-tune on pairs of generated queries with their associated entity or relation descriptions—both monolingual and cross-lingual. We augment training with hard negatives selected for low semantic similarity (< 0.7) yet high lexical overlap (above the 75th percentile), where lexical overlap is measured via BM25 scores computed between each query and all candidate negative samples. In the knowledge graph construction stage, we define the cross-synonym edges between any pairs of entities/edges in different language on the graph whenever their embedding similarity exceeds threshold $\tau = 0.9$. Finally, in the last multi-aspect, semantic-aware passage reranking step, we selected $w_{chunk} = 1.0$, $w_{attr} = w_{triple} = 0.3$ after conducting parameter sensitivity analysis.

Dataset Statistics. Below is the detailed information showing the number of queries and documents in each dataset.

Datasets	ZaloWikipediaQA	ZaloLegal2021	NQ	PopQA	MuSiQue
Number of queries	4,399	640	1,000	1,000	1,000
Number of documents	15,957	61,062	9,365	8,676	11,656

Table 6: Details number of queries and documents of different datasets.

Details of hyperparameter. Below, we present the hyperparameter search space used in MaGiX. The selected values corresponding to the best performance are highlighted in **bold**.

- τ : {0.75, 0.8, 0.85, **0.9**, 0.95}
- $\text{top-}k$: {50, 60, 70, 80, 90, **100**, 110, 120, 130, 140, 150}
- w_{chunk} : {0.8, 0.85, 0.9, 0.95, **1.0**}
- w_{attr} : {0.1, 0.15, 0.2, 0.25, **0.3**, 0.35, 0.4, 0.45, 0.5}
- w_{triple} : {0.1, 0.15, 0.2, 0.25, **0.3**, 0.35, 0.4, 0.45, 0.5}

Prompt for LLM to judge

You will evaluate two answers to the same question based on five criteria: **Accuracy, Comprehensiveness, Diversity, Logicity, and Coherence**.

- **Accuracy**: How factually correct is the answer based on the provided context and ground truth? Does it avoid hallucinations or factual errors?

- **Comprehensiveness**: How much detail does the answer provide to cover all aspects and details of the question?

- **Diversity**: How varied and rich is the answer in providing different perspectives and insights on the question?

- **Logicity**: How logically does the answer respond to all parts of the question?

- **Coherence**: How well does the answer maintain internal logical connections between its parts, ensuring a smooth and consistent structure?

For each criterion, choose the better answer (either Answer 1 or Answer 2) and explain why.

When determining the overall winner, consider the following:

1. Each of the five criteria contributes to the quality of the answer.
2. Consider the relative strengths and weaknesses of each answer across all criteria.
3. The overall winner should generally be the answer that performs better across most criteria.
4. While accuracy is important, also consider the other dimensions that contribute to a comprehensive and useful response.

Here is the question:

{query_text}

Here is the context which supported to answer the question:

{context}

Here are the two answers:

Answer 1:

{answer1}

Answer 2:

{answer2}

Evaluate both answers using the five criteria listed above and provide detailed explanations for each criterion.

Output your evaluation in the following JSON format and NOTHING ELSE:

{json_format}

json_format in evaluation prompt

```
{
  "Accuracy": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide explanation here, specifically noting any
    ↪ factual errors or hallucinations in either answer]"
  },
  "Comprehensiveness": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide explanation here]"
  },
  "Diversity": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide explanation here]"
  },
  "Logicality": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide explanation here]"
  },
  "Coherence": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide explanation here]"
  },
  "Overall_Winner": {
    "Winner": "[Answer 1 or Answer 2]",
    "Explanation": "[Provide a balanced explanation considering all five
    ↪ criteria. Explain how the winning answer performed across the
    ↪ different dimensions and why its strengths in certain areas make
    ↪ it the overall better response.]"
  }
}
```

D Additional Experimental Results

This section supplements the main findings in Section 4 by providing extended evaluations, ablation studies, and further analysis to validate the effectiveness and generalizability of MaGiX.

Table 9 shows evaluation results for English-to-Vietnamese QA across five datasets and five dimensions, using Grok-3 win rates. Similar to the Vietnamese-to-English setting, MaGiX consistently outperforms all baselines across nearly all aspects, demonstrating strong cross-lingual generation quality. Additionally, Table 7 presents an ablation study for contribution of each proposed components in MaGiX on English-to-Vietnamese QA, mirroring the setup used for the Vietnamese-to-English setting. Starting from the LightRAG baseline, each added component—granular retrieval, cross-synonym linking, and composite scoring—yields significant performance gains across all datasets. Granular retrieval boosts NQ from 27.71 to 77.27, highlighting the importance of fine-grained semantic representation. Cross-synonym edges further enhance performance, especially on MuSiQue (46.82 to 49.17), by improving cross-lingual alignment. Composite scoring leads to the highest PopQA result (51.45), while the final MaGiX model, incorporating fine-tuned embeddings, achieves the best overall performance on NQ (87.27) and MuSiQue (58.43). Similar to the vi-to-en setting, these results confirm the complementary contributions of each module and the overall robustness of MaGiX in cross-lingual retrieval.

Table 8 presents a parameter sensitivity analysis of the attribute and triple similarity weights ($w_{\text{attr}} = w_{\text{triple}}$), while keeping $w_{\text{chunk}} = 1$ fixed. Results on the NQ and MuSiQue development sets show that

Method	NQ	PopQA	MuSiQue
LightRAG	27.71	26.65	22.01
+ Granular Retrieval	77.27	49.35	46.82
+ Cross-Synonym	77.86	50.29	49.17
+ Composite Score	86.02	51.45	54.22
+ Finetune (MaGiX)	87.27	50.85	58.43

Table 7: Ablation: Recall@10 for English-to-Vietnamese settings on two simple QA datasets (NQ and PopQA) and one multi-hop QA dataset (MuSiQue).

w_{attr} & w_{triple}	0.1	0.15	0.2	0.25	0.3
Avg En-Vi	69.26	69.97	69.86	69.99	69.68
Avg Vi-En	69.23	69.76	69.97	69.68	69.45

Table 8: Average passage Recall@10 in Vietnamese–English settings and English–Vietnamese settings on the five benchmarks consisting ZaloWikipedia, ZaloLegal2021, PopQA, NQ and MuSiQue, evaluating different weights for similarity scores of attributes and triples. We fix $w_{chunk} = 1$ and vary $w_{attr} = w_{triple}$ to reduce the search space.

	ZaloWikipedia		ZaloLegal2021		NQ		Popqa		MuSiQue	
	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX	NaiveRAG	MaGiX
Accuracy	38.15	<u>61.85</u>	31.87	<u>68.12</u>	37.21	<u>62.79</u>	29.64	<u>70.36</u>	30.94	<u>69.06</u>
Comprehensiveness	40.10	<u>59.90</u>	37.50	<u>62.50</u>	43.60	<u>56.40</u>	36.77	<u>63.23</u>	28.13	<u>71.87</u>
Diversity	41.70	<u>58.30</u>	43.73	<u>56.27</u>	47.15	<u>52.85</u>	35.86	<u>64.14</u>	32.59	<u>67.41</u>
Logicality	31.91	<u>68.09</u>	33.28	<u>66.72</u>	35.19	<u>64.81</u>	28.61	<u>71.39</u>	27.40	<u>72.60</u>
Coherence	30.48	<u>69.52</u>	32.81	<u>67.19</u>	33.06	<u>66.94</u>	29.43	<u>70.57</u>	23.01	<u>76.99</u>
Overall	39.60	<u>60.40</u>	33.12	<u>66.88</u>	41.80	<u>58.20</u>	34.07	<u>65.93</u>	30.13	<u>69.87</u>
	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX	GraphRAG	MaGiX
Accuracy	26.05	<u>73.95</u>	26.25	<u>73.75</u>	30.99	<u>60.01</u>	26.06	<u>73.94</u>	30.86	<u>69.14</u>
Comprehensiveness	36.00	<u>64.00</u>	<u>63.28</u>	<u>63.72</u>	<u>70.80</u>	<u>70.80</u>	44.14	<u>55.86</u>	<u>50.25</u>	<u>49.75</u>
Diversity	<u>71.30</u>	28.70	<u>88.91</u>	11.09	<u>81.10</u>	18.90	46.89	<u>53.11</u>	<u>59.06</u>	40.93
Logicality	35.37	<u>64.63</u>	49.06	<u>50.94</u>	<u>56.71</u>	43.29	28.41	<u>71.59</u>	31.16	<u>68.84</u>
Coherence	40.14	<u>58.86</u>	<u>72.03</u>	27.97	47.34	<u>52.66</u>	31.83	<u>68.17</u>	35.30	<u>64.70</u>
Overall	43.70	<u>56.30</u>	<u>56.25</u>	43.75	<u>51.70</u>	48.30	34.93	<u>65.07</u>	35.54	<u>64.46</u>
	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX	LightRAG	MaGiX
Accuracy	22.40	<u>77.60</u>	20.00	<u>80.00</u>	29.06	<u>70.94</u>	24.45	<u>75.55</u>	28.01	<u>71.99</u>
Comprehensiveness	<u>57.10</u>	42.90	<u>52.03</u>	47.97	<u>70.00</u>	30.00	43.34	<u>56.66</u>	45.95	<u>54.05</u>
Diversity	<u>68.27</u>	31.73	<u>84.22</u>	15.78	21.20	<u>78.80</u>	49.15	<u>50.85</u>	<u>54.25</u>	45.75
Logicality	31.76	<u>68.24</u>	38.75	<u>61.25</u>	37.70	<u>62.30</u>	29.55	<u>70.45</u>	29.66	<u>70.34</u>
Coherence	36.64	<u>63.36</u>	<u>62.66</u>	37.34	<u>62.30</u>	37.70	32.50	<u>67.50</u>	29.93	<u>70.07</u>
Overall	39.50	<u>60.50</u>	42.66	<u>57.34</u>	49.50	<u>50.50</u>	33.93	<u>66.07</u>	33.63	<u>66.37</u>
	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX	HippoRAG 2	MaGiX
Accuracy	34.33	<u>65.67</u>	40.69	<u>57.66</u>	23.80	<u>76.20</u>	18.13	<u>81.78</u>	18.79	<u>81.21</u>
Comprehensiveness	11.90	<u>88.10</u>	50.00	<u>50.00</u>	6.10	<u>93.90</u>	26.03	<u>73.97</u>	15.82	<u>84.18</u>
Diversity	17.40	<u>82.60</u>	<u>60.47</u>	39.53	8.10	<u>91.90</u>	41.44	<u>58.56</u>	22.75	<u>77.25</u>
Logicality	18.80	<u>81.20</u>	40.31	<u>59.69</u>	8.40	<u>91.60</u>	40.31	<u>59.69</u>	13.11	<u>86.89</u>
Coherence	6.90	<u>93.10</u>	43.28	<u>56.72</u>	2.90	<u>97.10</u>	23.22	<u>76.78</u>	7.91	<u>92.09</u>
Overall	15.80	<u>84.20</u>	43.75	<u>56.25</u>	8.00	<u>92.00</u>	20.52	<u>79.48</u>	13.71	<u>86.29</u>

Table 9: Performance on English-to-Vietnamese QA across five datasets and five evaluation dimensions, measured by win rates based on judgments from the Grok-3.

performance remains relatively stable across a moderate range of weight values. The best Recall@10 is observed when $w_{attr} = w_{triple} = 0.25$, achieving 69.99 for En–Vi and 69.68 for Vi–En. Performance slightly drops when the weights are set too low (e.g., 0.1) or too high (e.g., 0.3), indicating that a balanced contribution of attribute and triple similarity, relative to chunk similarity, is essential. These findings confirm the effectiveness of incorporating multi-aspect semantic signals in a well-calibrated scoring function for cross-lingual retrieval.

E Graph Statistics Analysis

In this appendix, we analyze the connectivity of the graph through several statistics. The analysis shows that our proposed MaGiX method demonstrates a consistent improvement in graph connectivity across all datasets compared to the baseline, as detailed in Table 10. By maintaining the same number of nodes while significantly increasing edge counts, MaGiX enhances the Edge-to-Node ratio. Notable examples include PopQA, where the ratio rises from 0.98 to 1.43 (+0.33), and ZaloLegal, with an increase from 1.85 to 2.25 (+0.40). These improvements, with Δ edge/node ratios ranging from +0.33 to +0.45, underscore a substantial boost in connectivity.

Dataset	Method	# Nodes	# Edges	Δ Edge	Edge/Node Ratio	Δ Ratio
ZalowWiki	Baseline	120,981	153,754	-	1.27	-
	MaGiX	120,981	192,974	+39,220	1.60	+0.33
ZaloLegal	Baseline	452,319	838,917	-	1.85	-
	MaGiX	452,319	1,016,622	+177,705	2.25	+0.40
NQ	Baseline	128,140	155,240	-	1.21	-
	MaGiX	128,140	204,482	+49,242	1.60	+0.39
Popqa	Baseline	153,554	151,008	-	0.98	-
	MaGiX	153,554	219,350	+68,342	1.43	+0.45
Musique	Baseline	155,901	171,335	-	1.10	-
	MaGiX	155,901	227,073	+55,738	1.46	+0.36

Table 10: Comparison of Graph statistics between baseline and MaGiX methods across datasets

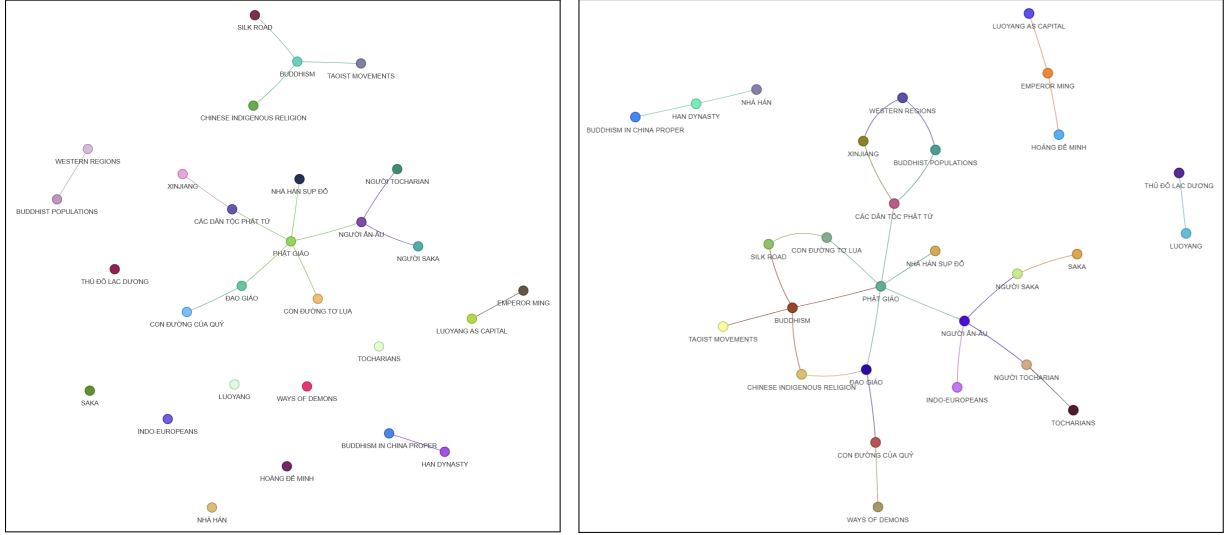


Figure 3: Structure of the graph built from small sample data before and after establishing cousin relations.

Additionally, graph visualization of a small sample set before and after creating cousin relations in figure 3 reveals MaGiX’s strength in forming more intra-entity and cross-lingual semantic connections, fostering cohesive substructures that enhance retrieval relevance.

F Graph Construction and Inference Time Analysis

To rigorously evaluate the efficiency and scalability of MaGiX, we present comprehensive data on inference duration, graph construction time, and index dimensions, leveraging the extensive NQ dataset. Concerning inference time, we applied our method alongside other baseline approaches to knowledge graphs built from an identical initial set of 19,266 raw text chunks. For graph construction time, we performed an experiment using 100 documents from the NQ dataset, where each document averages 656 words (approximately 190 tokens).

Method	MaGiX	GraphRAG	LightRAG	HippoRAG2
Time (min)	7:40	17:44	8:53	1:46

Table 11: Comparison of Graph Construction Time between MaGiX and other baselines

Inference Time. As shown in the table 12, MaGiX employs an advanced retrieval framework with 245,000 multi-attribute entity vectors and 19,000 chunk vectors, unlike LightRAG and GraphRAG, which use only 128,000 entity vectors without chunk-level retrieval. Despite this richer representation, MaGiX achieves a competitive inference time of 3.77 seconds, outperforming HippoRAG2 (7.38s) and closely matching LightRAG (2.19s) and GraphRAG (2.09s).

Method	Time (s)	Entity Vec.	Reltion Triplet Vec.	Chunk Vec.
LightRAG	2.19	128,140	155,240	-
GraphRAG	2.09	128,140	155,240	-
HippoRAG2	7.38	136,762	231,012	19,266
MaGiX (Our)	3.77	245,460	172,833	19,266

Table 12: **Inference Time and Size of the vector databases for retrieval on the NQ Dataset.** All methods construct their knowledge graphs from the same initial pool of 19,266 raw text chunks.

Graph Consutruction Time. For graph construction, the table 11 illustrates that MaGiX takes 7:40 minutes, exceeding HippoRAG2 but surpassing GraphRAG, and aligning with LightRAG, despite enhanced semantics, as details in . This one-time, parallelizable process optimizes efficiency.

Though incurring modest computational cost over GraphRAG and LightRAG, MaGiX delivers up to 4 times better in retrieval performance (Table 1) and outpaces HippoRAG2 in speed and Recall@10 by approximately 3%. These findings affirm that MaGiX achieves an optimal equilibrium between scalability and performance, providing highly competitive inference and construction times alongside exceptional retrieval accuracy.

G Qualitative Error Analysis and Cases Studies

To better understand MaGiX’s performance, we performed a qualitative error analysis by examining both successful and unsuccessful cases. We highlight two examples: one where MaGiX greatly outperforms HippoRAG 2, and another showing a failure case. These examples clarify the system’s strengths and existing limitations.

Vietnamese Query: Khi nào Elijah xuất hiện lần đầu trong Vampire Diaries?
English Query: When does Elijah first appear in Vampire Diaries?

Method	Top-3 Retrieved Documents	Answer (Vietnamese)	Answer (English)
HippoRAG 2	Doc 1: “Daniel Gillies Daniel Gillies (born 14 March 1976) is a Canadian-born New Zealand actor. He is best known for his role as Elijah Mikaelson on the television series “The Vampire Diaries” ...”. Doc 2: “hit The CW drama “The Vampire Diaries” as Elijah Mikaelson. In 2013, CW created a successful spin-off of The Vampire Diaries ...”. Doc 3: “ Klaus (The Vampire Diaries) “Klaus” is the 19th episode of the second season... Elijah (Daniel Gillies) wakes up in the Salvatore base-ment. ...”	Để trả lời câu hỏi này, chúng ta cần tìm thông tin về lần xuất hiện đầu tiên của Elijah trong series The Vampire Diaries ... Tuy nhiên, trong đoạn văn thứ 4, có đề cập đến một tập có thể là “Klaus” (tập 19 của mùa 2), nơi Elijah (do Daniel Gillies thủ vai) xuất hiện sau khi Elena rút con dao ra khỏi người anh ta. Điều này cho thấy Elijah đã xuất hiện trong mùa 2 của series, nhưng không nêu rõ tập chính xác anh ta lần đầu xuất hiện ... Trả lời: Mùa 2.	To answer this question, we need to find information about Elijah’s first appearance in the series The Vampire Diaries ... However, based on the available information, we can infer that Elijah first appeared in season 2 of The Vampire Diaries, although more specific information is needed to determine the exact episode. Answer: Season 2.
MaGiX	Doc 1: “... He is portrayed by New Zealand actor Daniel Gillies Elijah first appears in episode eight of season two, entitled “Rose”. Elijah is brought into the circle by two vampires named Rose”. Doc 2: “... Niklaus Mikaelson (more commonly called Klaus) is physically the most powerful immortal supernatural being in “The Vampire Diaries” and “The Originals”...”. Doc 3: “... He kidnaps Elena who tells him that Esther wants to kill them all...”	Giới thiệu về Elijah: Elijah là một nhân vật trong bộ phim “The Vampire Diaries” và “The Originals”. Ông là một trong những nhân vật quan trọng trong series này. Xuất hiện lần đầu của Elijah: Elijah xuất hiện lần đầu trong tập 8 của mùa 2, với tựa đề “Rose”. Ông được giới thiệu như một nhân vật quan trọng trong series và có mối quan hệ phức tạp với các nhân vật khác.	Introduction to Elijah: Elijah is a character in the TV series “The Vampire Diaries” and “The Originals”. He is one of the key characters in these series. Elijah’s First Appearance: Elijah first appears in episode 8 of season 2, titled “Rose.” He is introduced as an important character in the series and has complex relationships with other characters.

Table 13: Comparison of MaGiX and HippoRAG 2, including top-3 retrieved documents and their answers in Vietnamese and English.

Case Study 1: Our method is better than HippoRAG 2. In this use case, MaGiX demonstrates a clear advantage over HippoRAG 2 by retrieving the exact episode of Elijah’s first appearance in The Vampire Diaries ("Rose").

Method	Query	Top-3 Retrieved Documents	Groundtruth Documents
MaGiX	Vietnamese Query: “Các quốc gia châu phi bắt đầu bằng chữ a ?” English Version: “Which African countries that begin with the letter a ?”	Doc 1 : "19th century, European countries colonised almost all of Africa; most present states in Africa originated from a process of decolonisation in the 20th century. African nations cooperate through the establishment of the African Union ..." Doc 2 : "... In the French version, the letters were hand-printed in lower case only. Only 56 of the letters in the English version were listed – omitting the hooktop-z – and two further apostrophe-like letters (' and ') were included. ..." Doc 3 : "Alphabetical Africa Alphabetical Africa is a constrained writing experiment by Walter Abish. It is written in the form of a novel. the first chapter contains only words starting with the letter "a", the second chapter only words starting with "a" or "b", etc.;..."	"multi-media services: IRIN covers the following countries: AFRICA: Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad,..."

Table 14: Examples of MaGiX failure queries, top-3 documents retrieved by MaGiX, and ground truth documents.

The table 13 shows that unlike HippoRAG 2, which is swayed by high-connectivity nodes and general content like actor bios, MaGiX uses multi-granular knowledge representation and semantic-aware reranking to focus on context-specific mentions of "Elijah," prioritizing attribute similarity over broad graph connectivity to avoid popularity bias common in traditional GRAG systems.

Case Study 2: Failure Case of MaGiX. We have observed some notable cases that MaGiX fails to retrieve the correct document and other baseline models also fail.

In the table 14, MaGiX underperforms in this scenario, as entities like “Africa” and “letters” appear more prominently in unrelated documents, inflating their attribute similarity scores. Meanwhile, the correct document contains only sparse mentions, causing it to be ranked lower. As a result, This case illustrates a broader challenge: for queries referencing multiple dispersed entities, systems struggle when the key evidence is sparse and embedded within long documents, especially when such content is split across separate chunks during preprocessing.