ComLQ: Benchmarking Complex Logical Queries in Information Retrieval

Ganlin Xu[1], Zhitao Yin[1], Linghao Zhang[1], Jiaqing Liang[1], Weijia Lu[2], Xiaodong Zhang[2], Zhifei Yang[2], Sihang Jiang[3], Deqing Yang[1]*,

[1]School of Data Science, Fudan University, Shanghai, China
[2]United Automotive Electronic Systems, Shanghai, China
[3]College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
glxu24@m.fudan.edu.cn, yangdeqing@fudan.edu.cn

Abstract

Information retrieval (IR) systems play a critical role in navigating information overload across various applications. Existing IR benchmarks primarily focus on simple queries that are semantically analogous to single- and multi-hop relations, overlooking *complex logical queries* involving first-order logic operations such as conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$). Thus, these benchmarks can not be used to sufficiently evaluate the performance of IR models on complex queries in real-world scenarios. To address this problem, we propose a novel method leveraging large language models (LLMs) to construct a new IR dataset **ComLQ** for **Com**plex **L**ogical **Q**ueries, which comprises 2,909 queries and 11,251 candidate passages. A key challenge in constructing the dataset lies in capturing the underlying logical structures within unstructured text. Therefore, by designing the subgraph-guided prompt with the subgraph indicator, an LLM (such as GPT-4o) is guided to generate queries with specific logical structures based on selected passages. All query-passage pairs in ComLQ are ensured *structure conformity* and *evidence distribution* through expert annotation. To better evaluate whether retrievers can handle queries with negation, we further propose a new evaluation metric, **Log-Scaled Negation Consistency** (**LSNC@**K). As a supplement to standard relevance-based metrics (such as nDCG and mAP), LSNC@K measures whether top-K retrieved passages violate negation conditions in queries. Our experimental results under zero-shot settings demonstrate existing retrieval models' limited performance on complex logical queries, especially on queries with negation, exposing their inferior capabilities of modeling exclusion. In summary, our ComLQ offers a comprehensive and fine-grained exploration, paving the way for future research on complex logical queries in IR.

Code and Datasets — https://anonymous.4open.science/r/ComLQR-main-6B8D/

Introduction

Information retrieval (IR) systems, as a cornerstone in addressing information overload, have been widely adopted in various AI applications, including recommendation systems (Dai et al. 2024), question answering (Karpukhin et al. 2020). A typical IR system retrieves relevant documents or passages from a designated corpus in response to user queries (Zhao et al. 2024). Traditional IR benchmarks such as MS-MARCO (Nguyen et al. 2016), TREC (Craswell et al. 2020a), and BEIR (Thakur et al. 2021a), predominantly focused on relatively simple queries that are semantically analogous to single- and multi-hop relations requiring limited compositional inference, which fail to satisfy complex retrieval needs of users in real-world applications. According to our analysis using GPT-4o to classify queries, over 93% of queries in existing IR benchmarks are such simple queries. Yet, real user queries are often more complex and involve compositional logical reasoning, highlighting the limitations of these IR benchmarks in capturing the diversity and complexity of user queries.

Complex logical queries can be represented with first-order logic (FOL) that involves logical operations such as

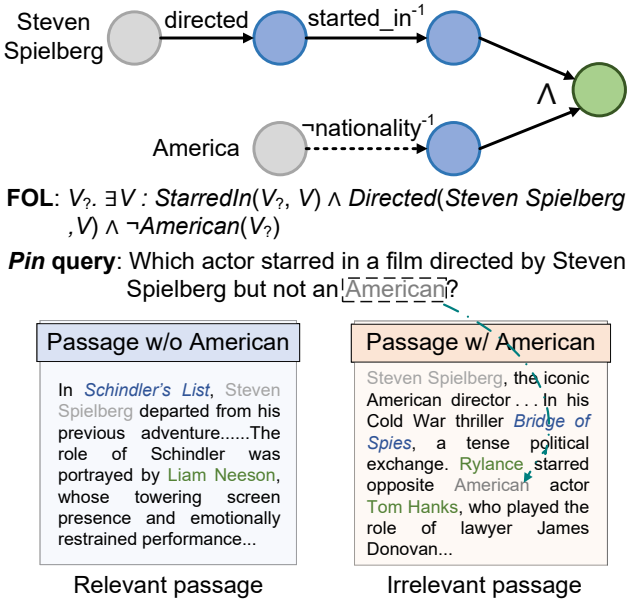

Figure 1: Given a complex logical query of *pin* type, the state-of-the-art retriever InteR tends to retrieve the passages containing the keyword 'American', which, however, is not relevant to the query.

*Corresponding author

conjunction ($\wedge$), disjunction ($\vee$), negation ($\neg$) and existential quantifier ($\exists$) (Ren and Leskovec 2020; Ren, Hu, and Leskovec 2020). For example, given a complex logical query of *pin* type in Figure 1 "*Which actor starred in a film directed by Steven Spielberg but not an American?*" where *p, i* and *n* represent *projection*, *intersection* and *negation*, respectively (Ren and Leskovec 2020; Ren, Hu, and Leskovec 2020), it can be formalized in FOL as:

$$V_?.\ \exists V : \textit{StarredIn}(V_?, V) \wedge \textit{Directed}(\text{Steven Spielberg}, V) \wedge \neg \textit{American}(V_?).$$

Although complex logical queries received considerable attention in the community of knowledge base question answering (KBQA) (Fang et al. 2024a; Ji et al. 2024), they remain underrepresented in existing IR benchmarks[1]. Compared to simple queries, handling complex logical queries requires the precise localization of information resources and thus is conducive to achieving logical reasoning. These queries involve intricate retrieval intents, which cannot be handled through simple word co-occurrence and thus pose significant challenges for current IR systems. Figure 1 shows that, for the given query, the state-of-the-art retriever InteR (Feng et al. 2024) relying on word co-occurrence favors the irrelevant passage containing the keyword 'American' rather than the relevant passage without the keyword, revealing that they fail to understand query intents correctly (Xu et al. 2025).

To overcome the underrepresentation of **Com**plex **L**ogical **Q**ueries in existing IR benchmarks, we introduce a novel IR dataset, namely **ComLQ** in this paper, to provide a comprehensive and fine-grained exploration for complex logical queries, which includes 2,909 queries and 11,251 candidate passages. A key challenge in constructing the dataset lies in capturing the underlying logical structures within unstructured text. To address the problem and ensure the quality of the dataset, we design a data synthesis process applicable across diverse domains. Specifically, we first select passages from the existing corpus. To automatically acquire queries with specific logical structures, we then design a subgraph-guided prompt to request an LLM to generate queries based on selected passages, where a key component *subgraph indicator* guides the LLM to learn subgraph patterns associated with different query types. Finally, experienced annotators review each generated query-passage pair concerning the following two criteria. ii) **Structure conformity** ensures that queries not strictly conforming to the intended query structure are filtered out. ii) **Evidence distribution** ensures that for queries generated from multiple passages, supporting evidence is indeed distributed across those passages. Furthermore, to better evaluate whether retrievers can handle queries with negation in ComLQ, we propose a new evaluation metric **Log-Scaled Negation Consistency** (**LSNC@**K), which measures the extent to which the retrieved top-K passages violate negation conditions in queries.

We conduct experiments on a wide range of retrieval models on our ComLQ under zero-shot settings, of which the significant findings include: ***I***) None of the experimented retrievers can consistently outperform other methods across all query types, highlighting the need for approaches tailored to different logical structures. ***II***) All retrievers exhibit consistent performance degradation as query complexity increases, revealing their limitations in capturing and reasoning over intricate logical structures. ***III***) The order of logical operations in query structures notably affects retrievers' performance, since their performance on projection-then-intersection queries (e.g., pi, pin and pni) is worse than that on intersection-then-projection queries (e.g, ip and inp). ***IV***) All experimented retrievers exhibit low LSNC@100 scores on queries with negation, revealing a critical gap in IR models' ability to model exclusion.

The main contributions of this paper include:

1. We introduce a new IR dataset **ComLQ** for benchmarking complex logical queries in IR. To the best of our knowledge, ComLQ is the first IR dataset offering a comprehensive and fine-grained exploration for complex logical queries.

2. We propose an effective method to automatically acquire queries with specific logical structures based on LLMs' generation, where a subgraph-guided prompt involving the *subgraph indicator* is specially designed to guide LLMs to learn subgraph patterns associated with different query types.

3. To evaluate whether retrievers handle queries with negation, we propose a novel metric, LSNC@K, to supplement existing relevance-based metrics (nDCG and mAP).

4. We conducted extensive experiments on a wide range of retrieval models, and found that existing retrievers exhibit significant limitations on complex logical queries, with consistently low scores on the proposed metric LSNC@100, revealing their inability to model exclusion. Our findings in this paper pave the way for future research on complex logical queries in IR.

Related Work

Complex Logical Queries

In recent years, reasoning over single-hop and multi-hop relational data (Yang et al. 2018; Lin et al. 2023) has made remarkable advances. In addition, subsequent research has explored more complex logical structures that involve unobserved edges, multiple entities, and variable interactions (Bai et al. 2023). In this paper, we focus on conjunctive logical queries (Hamilton et al. 2018), a subclass of first-order logic queries characterized by existential quantifier $\exists$ and conjunction $\wedge$. Conjunctive logical queries require a set of anchor entities, $\mathcal{V}$, a unique target entity $V_?$ representing the answer to the query, and a set of existential quantified variables $V_1, \cdots, V_m$, and are defined as the conjunction of literals $e_1, \cdots, e_n$:

$$q = V_?, \exists V_1, \cdots, V_m : e_1 \wedge e_2 \wedge \cdots \wedge e_n, \qquad (1)$$

where e_i is an edge involving variable nodes and anchor nodes, satisfying $e_i = r(v_j, V_k)$, $V_k \in \{V_?, V_1, \cdots, V_m\}$,

[1]Unlike KBQA's operations on structured triples with explicit entities and relations, IR systems execute set operations such as intersection, union, projection, and negation directly over unstructured text. Please refer to Section 2.1 for more details.

Model type	1p	2p	3p	2i	3i	pi	ip	2u	up	2in	3in	inp	pin	pni	total
Query number	176	227	189	207	267	168	240	258	193	278	271	173	140	122	2,909

Table 1: The statistics of all query types in ComLQ.

$v_j \in \mathcal{V}$, $r \in \mathcal{R}$, or $e_i = r(V_j, V_k)$, $V_j, V_k \in \{V_?, V_1, \cdots, V_m\}$, $j \neq k$, $r \in \mathcal{R}$. $\mathcal{R}$ is the set of relations defined in the knowledge base (KB).

Although prior work on complex logical queries has predominantly focused on knowledge base question answering (KBQA) (Fang et al. 2024b; Zhang et al. 2024), these approaches typically operate over structured triples with a predefined entity-relation schema, where reasoning is performed along explicit relation paths. In contrast, our retrieval setting shifts the focus from path-based reasoning on structured triplets to executing set-theoretic operations (such as projection, intersection, union and negation) directly over unstructured text. This requires retrievers to identify and combine relevant spans from natural language passages that jointly satisfy the query's logical intent. As a result, our task demands not only semantic understanding but also the recovery of implicit logical structures in open-domain text.

Information Retrieval Benchmarks

Traditionally, the evaluation of information retrieval (IR) systems has relied on standardized benchmarks and well-established evaluation metrics. Numerous datasets evaluated IR systems from Wikipedia (Lee, Chang, and Toutanova 2019), web queries (Bajaj et al. 2018) and biomedical questions (Voorhees and Tice 2000). Recently, several benchmarks combine multiple datasets and evaluate retrieval or embedding models across different domains and use cases, such as BEIR (Thakur et al. 2021b), and MTEB (Muennighoff et al. 2023). Besides, increasing efforts have focused on vertical domains, such as climate science (Schimanski et al. 2024), the legal domain (Su et al. 2024) and academic scholarship (Ajith et al. 2024). These domains are typically characterized by dense specialized terminology and a strong reliance on domain-specific knowledge, placing greater demands on the professional adaptability of IR systems.

Despite these advancements, current IR benchmarks still fall short in providing a comprehensive and fine-grained evaluation for complex logical queries involving projection, intersection, union and negation, as well as various combinations of these operations. NegConstraint (Xu et al. 2025) only focuses on negative-constraint queries, which are similar to the queries with negation in our work. Multi-hop benchmarks, such as HotpotQA (Yang et al. 2018), neglect other essential logical operations, including intersection, union and negation, which are explicitly represented in our ComLQ. Therefore, they represent only a subpart (or subset) of ComLQ's broader scope. The TREC Complex Answer Retrieval (CAR) track (Dietz et al. 2017) focuses on generating and restructuring long textual answers, whereas our ComLQ emphasizes the structural formulation of queries themselves. Wang et al. (2023) introduce complex scientific questions which, however, are long-form questions containing multiple simple sentences, and excluding complex logical structures.

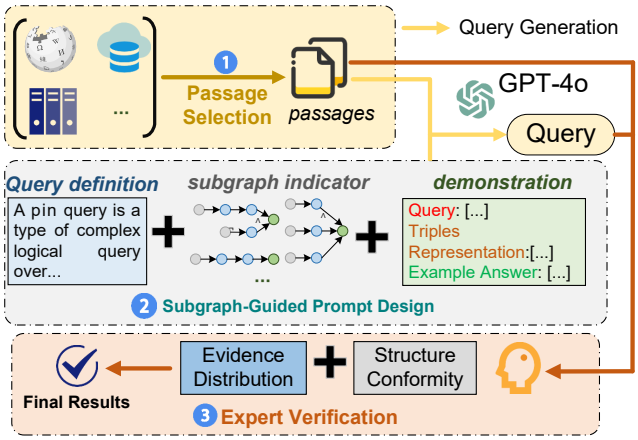

Figure 2: The data synthesis process of ComLQ.

Dataset Construction

Dataset Overview

In this paper, we adopt the standard definition of complex logical queries from (Ren and Leskovec 2020; Ren, Hu, and Leskovec 2020), consisting of 9 query types without negation (denoted as *1p/2p/3p/2i/3i/pi/ip/2u/up*) and 5 query types with negation (denoted as *2in/3in/inp/pin/pni*), where *p, i, u* and *n* represent *projection*, *intersection*, *union* and *negation* in query structure, respectively. All data in our ComLQ are organized in a standard format (corpus, queries, qrels[2]) akin to the BEIR benchmark (Thakur et al. 2021b). The average length of a query in ComLQ is 19.95 words, and the average length of a passage is 112.74 words. Table 1 lists the query number of each type, where queries with negation account for 33.8%, ensuring balanced distributions to evaluate retrieval performance on modeling exclusion. Query examples and structures are introduced in Appendix A. We request the annotators to score each query-passage pair using a 3-point grading scale (0-2):

- **Level-0**. The passage is irrelevant (fully mismatched) to the query.
- **Level-1**. The passage is partially relevant to the query and partly satisfies the query's information needs. That is, if supporting evidence for the query is distributed across two or three passages, each passage is labeled as Level-1.
- **Level-2**. The passage content is customized to satisfy the information needs of the query and precisely contains the

[2]Qrels (query relevance judgments) are ground-truth annotations indicating which passages are relevant to specific queries.

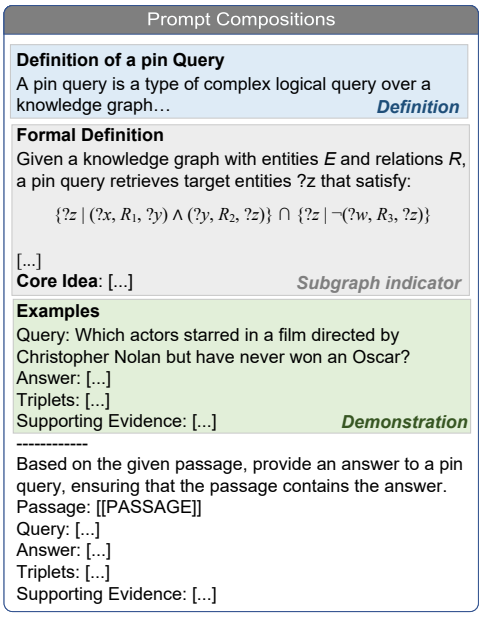

Figure 3: Prompt compositions for generating *pni* query sample.

Positive Query	**Query**: Which actor starred in a film directed by Steven Spielberg but not an American?	✓
	Triplets: **Projection**: (StevenSpielberg, directed, ?y) (?y, starred by, ?z) **Negation**: ¬(?z, nationality, America) **Intersection**: (StevenSpielberg, directed, ?y) ∧ (?y, starred by, ?z) ∧ ¬(?z, nationality, America)	
Negative Query	**Query**: Which microprocessors were considered for the original IBM PC but were not used in its final design?	×
	Triples: **Projection**: (IBM, considered_processors, ?y) **Negation**: ¬(IBM PC, used_processor, ?y) **Intersection**: (IBM, considered_processors, ?y) ∧ ¬(IBM PC, used_processor, ?y)	

Table 2: Illustrations of structure conformity.

answer to the query. If supporting evidence for the query is entirely contained within a single passage, that passage is labeled as Level-2.

Data Synthesis

Figure 2 shows the data synthesis process of ComLQ, which is applicable across diverse domains. We first select passages from the existing corpus, for which we use the standard Wikipedia dump as the knowledge source in this paper. Then, we design a *subgraph-guided prompt* to request LLMs to generate queries conforming to specific query structures based on the selected passages. Although queries are generated by LLMs, the process is structure-constrained and passage-grounded, ensuring that each query is anchored in real content and follows well-defined logical structures. Finally, in the *expert verification*, three experienced annotators carefully examine each query-passage pair by two criteria: *structure conformity* and *evidence distribution*.

Passage Selection According to (Xu et al. 2025), we use the Wikipedia dump as the knowledge source in this paper, from which 20 million passages are obtained by segmenting 5 million titled articles. We select one or multiple passages from the same article each time to generate corresponding queries. When selecting multiple passages for query generation, we ensure that all passages are topically relevant to the query. That is, queries require reasoning across passages sharing the same title, which aims at evaluating the retrievers' ability to aggregate relevant information from multiple passages.

Subgraph-Guided Prompt Design To automatically acquire queries with specific logical structures, we design a subgraph-guided prompt that enables LLMs to generate queries based on selected passages. Specifically, LLMs often struggle to internalize and reproduce complex reasoning patterns (such as projection, intersection, and negation) by relying solely on natural language descriptions. To address this, we incorporate symbolic subgraph patterns into the prompt, allowing the LLM to align natural language queries with corresponding logical structures. As shown in Figure 3, a full prompt consists of three components: *query definition*, *subgraph indicator*, and *demonstration*, where the *subgraph indicator* guides the LLM to learn subgraph patterns associated with different query types, enabling consistent query generation. This design combines symbolic logic for structural control with LLMs' strengths in natural language generation, enabling LLMs to generate queries with complex logical structures. For example, given a *pni* query "Which actors starred in a film directed by Christopher Nolan but have never won an Oscar?", the corresponding *subgraph indicator* is formulated as:

$$\{?z \mid (?x, R1, ?y) \wedge (?y, R2, ?z)\} \cap \{?z \mid \neg(?w, R3, ?z)\},$$

where $?x$ and $?y$ denote the starting constant entities *Christopher* and *Oscar*, respectively. In addition, $?w$ is an intermediate variable representing *film* entities, and $?z$ refers to the target variable entities (*actors*). $R1$, $R2$, and $R3$ denote the respective relations between entities. The prompt defines a projection chain and a negation constraint, which are combined via a set intersection to identify the set of actors. All prompt examples are provided in Appendix J.

Expert Verification To ensure the quality of generated queries, we apply expert verification to all query-passage pairs. Although LLMs can generate fluent outputs, they often exhibit structural hallucinations, i.e., produce queries deviating from the intended structure, or rely on evidence not properly distributed across the supporting passages. To address this problem, we also provide auxiliary query triplets

Models	1p	2p	3p	2i	3i	pi	ip	2u	up	2in	3in	inp	pin	pni	total
HyDE	65.8	56.7	52.1	58.4	57.1	44.7	48.3	49.7	45.2	36.5	31.8	38.3	31.7	34.0	45.6
BGE	66.3	56.6	53.9	60.1	58.2	45.7	48.5	49.5	43.4	37.9	31.3	38.9	33.3	34.8	47.4
PromptReps	64.6	63.4	58.0	61.4	57.4	47.7	47.8	53.5	46.7	35.7	30.2	37.6	35.7	29.6	48.6
BM25	66.1	60.2	57.4	63.5	60.3	46.2	51.6	52.7	39.5	39.3	32.2	36.6	32.4	31.7	50.5
LameR	69.6	58.8	56.9	**65.9**	58.0	50.6	52.1	55.7	45.7	37.0	33.0	38.1	**35.8**	33.5	52.8
Contriever	70.2	65.3	**61.7**	60.7	61.2	52.0	54.8	57.4	48.8	36.9	33.2	38.3	32.1	35.5	53.4
AGR	**74.3**	61.2	60.3	62.3	**62.7**	48.4	53.0	**59.1**	**52.4**	35.5	34.5	**42.3**	35.5	33.8	54.3
InteR	71.8	**64.4**	58.7	63.6	62.6	**52.3**	**55.8**	58.5	50.3	**39.3**	**34.7**	40.3	34.6	**37.5**	**55.7**

Table 3: All retrieval models' nDCG@10 (%) scores on ComLQ queries across all query types.

and supporting evidence to assist three experienced annotators in reviewing each generated query-passage pair, concerning *structure conformity* and *evidence distribution* as follows.

- **Structure Conformity** As shown in Table 2, we provide auxiliary query triplets to assist filtering low-quality samples. Three annotators manually review all queries and their corresponding triple forms to validate strict adherence to the target query structure. For example, the positive query "*Which actor starred in a film directed by Steven Spielberg but not an American?*" conforms to pin structure, while the negative query "*Which microprocessors were considered for the original IBM PC but were not used in its final design?*" is not a strict pin query. Furthermore, we use majority voting to achieve a consensus for ambiguous queries. This two-step process helps minimize false positives (well-formed queries are incorrectly rejected) and false negatives (ill-formed queries are incorrectly accepted).
- **Evidence Distribution** In addition, we assess the evidence distribution of generated samples. In other words, for queries generated from multiple passages, the annotators verify whether necessary supporting evidence is indeed distributed across those passages. If the evidence is not properly distributed, we discard the corresponding query-passage pair. This annotation stage follows the same protocol as structure conformity, including majority voting and a random sample check.

Furthermore, to assess the model's ability to disregard irrelevant information, we augment the corpus with distractor passages whose titles are entirely disjoint from those of the sampled passages, ensuring they are completely unrelated to the generated queries.

Experiments

Retrieval Models

We first consider several zero-shot retrieval models not involving query-document relevance labels in our experiments, including sparse retriever BM25 (Robertson and Zaragoza 2009), dense retriever BGE (Xiao et al. 2024), and Contriever (Izacard et al. 2021). Besides, we consider some LLM-based retrieval models, including HyDE (Gao et al. 2023), InteR (Feng et al. 2024), LameR (Shen et al. 2024), AGR (Chen et al. 2024), and PromptReps (Zhuang et al. 2024).

Implementation Details

To ensure a fair and consistent comparison, we reproduce results on HyDE, LameR, AGR, and InteR, using `bge-small-en-v1.5` as the embedding model and `GPT-4o` with a temperature of 0.5 as the underlying LLM. For PromptReps, we adopt `LLaMA3-70B-Instruct` to generate hybrid representations. All experiments are conducted on three `NVIDIA A800 80GB` GPUs.

Evaluation Metrics

In our experiment results, we report the nDCG@10 scores of all retrieval models on ComLQ queries, given nDCG's robustness in capturing the effectiveness of IR models across tasks with binary and graded relevance judgments. Especially, to evaluate the ability of retrievers to handle queries with negation in ComLQ, we propose a novel evaluation metric, **Log-Scaled Negation Consistency (LSNC@K)**, which measures whether the Top-K retrieved passages violate negation conditions specified in queries. Unlike standard relevance-based metrics (e.g., nDCG and mAP) measuring overall relevance, LSNC targets negation consistency evaluation, which is a credible metric for evaluating IR performance on queries with negation. Formally, let $\mathcal{D}_k$ denote the set of top-K retrieved passages, its LSNC@K score is computed as

$$\text{LSNC@}K = -\frac{\log\left(\big(\sum_{d\in\mathcal{D}_k} V(d)+1\big)/(K+1)\right)}{\log(K+1)}, \quad (2)$$

where $V(d) \in \{0,1\}$ is an indicator function, and equals 1 if the retrieved passage d violates negation conditions, otherwise 0. A higher LSNC@K score indicates there are fewer violations in the top-K passages.

Experiment Results

Overall Results Table 3 presents the nDCG@10 scores of all retrieval models across 14 query types, along with the overall performance in *total* column, where the best scores are highlighted in bold and second-best scores are underlined. The results show that none of the retrieval models can consistently achieve the best performance across all query types, highlighting the importance of developing specialized retrieval approaches tailored to the unique reasoning demands of various logical structures. In addition, there are some significant findings from the results as follows.

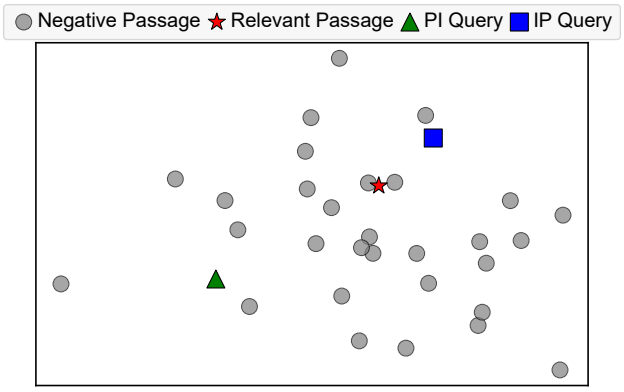

Figure 4: A visualized illustration of query and passage embeddings from ComLQ.

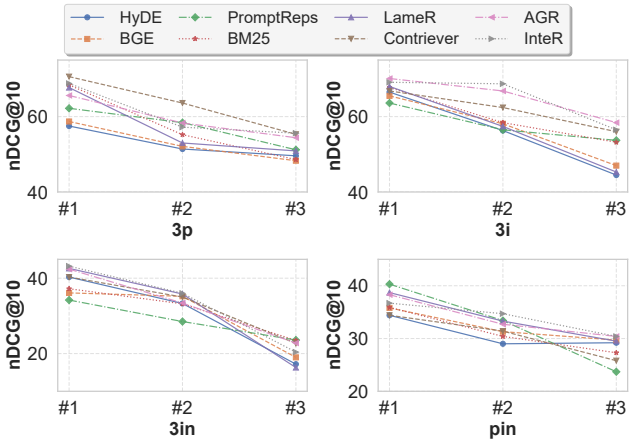

Figure 5: All retrievers' performance with varying numbers of supporting passages on queries of *3p*, *3i*, *3in* and *pin*, respectively.

Models	2in	3in	inp	pin	pni
HyDE	27.8	26.0	23.1	24.4	24.9
BGE	30.2	31.8	26.6	27.2	25.7
PromptReps	31.3	33.4	27.4	26.8	29.7
BM25	32.2	29.0	30.5	29.3	27.4
LameR	35.6	34.7	29.9	27.2	26.4
Contriever	34.4	37.4	28.1	30.4	31.2
AGR	36.4	37.3	31.2	**32.8**	**32.2**
InteR	**38.4**	**38.3**	**33.2**	31.7	31.7

Table 4: LSNC@100 results (%) on ComLQ across queries with negation.

1. **All retrievers' performance degrades as query complexity increases**. The trend of performance degradation is observed in the results across *1p*, *2p*, and *3p* queries, demonstrating the inherent difficulty of multi-hop queries. A similar trend is also observed in the results of other query types (*2i* vs. *3i* and *2in* vs. *3in*). Besides, the retrievers perform better on *2u* queries compared to *up* queries, as the latter introduces an additional projection operation that complicates reasoning. The performance gaps reveal the limited capabilities of the retrievers in capturing and reasoning over increasingly intricate logical structures.
2. **All retrievers perform poorly on queries with negation**. Their performance on queries with negation (*2in*, *3in*, *inp*, *pin*, and *pni*) is consistently lower than that on queries without negation. This reveals that current retrieval models struggle to handle the reasoning scenarios requiring the exclusion of specific entities or relations. The performance gap aligns with prior research (Xu et al. 2025), suggesting that relying on word co-occurrence makes retrievers particularly ill-suited for queries with negation.
3. **Sparse retrievers remain competitive with dense retrievers**. BM25 consistently outperforms both BGE and HyDE across all query types. These results violate the common assumption that dense retrievers universally outperform sparse ones, but conform to the observations from the BEIR benchmark (Thakur et al. 2021a) where traditional sparse methods maintain robust performance across diverse datasets.

Impact of Logical Operation Order From the experiment results, we also observe that the retrievers exhibit inferior retrieval performance on projection-then-intersection queries than intersection-then-projection queries (*pi* vs. *ip*, *pin* and *pni* vs. *inp*). To illustrate it, we compare a *pi* query example ("*Find people who acted in a movie directed by Christopher Nolan and who also won an Oscar*") with an *ip* query example ("*Which universities have produced individuals who are both Nobel Prize winners and Fields Medalists?*") from ComLQ. The embeddings of the two queries and their corresponding passages are visualized using t-SNE in Figure 4. Although two queries have the same relevant passage, the results show that the embedding of *ip* query is closer to that of the passage than the *pi* query. We argue that projection-then-intersection queries generally involve more complex semantic compositions, which tend to confuse retrievers and lead queries to less alignment with relevant passages.

Impact of Evidence Distribution To investigate the impact of the number of supporting passages on retrievers' performance, in Figure 5, we depict all retrievers' performance lines when varying evidence distribution levels (we only display the results on the representative query types *3p*, *3i*, *3in* and *pin* due to space limitation). Specifically, #1, #2 and #3 denote the cases where the correct answer must be inferred from one, two or three supporting passages, respectively. The results show that queries requiring evidence from more passages are more challenging for retrievers.

Performance on Negation Queries To evaluate whether retrievers can correctly handle queries with negation in ComLQ, we evaluate retrieval models with our proposed metric LSNC@K. As shown in Table 4, all retrievers exhibit very low LSNC@100 scores on the five query types with negation, i.e., *2in*, *3in*, *inp*, *pin*, and *pni*, and are weaker on queries involving multiple operations (*inp*, *pin* and *pni*). This suggests that increased logical complexity further im-

Pin query: Which actor starred in a film directed by Steven Spielberg but not an American?	
Models	**Rewriting queries**
HyDE	Steven Spielberg, one of the most influential American directors... One notable non-American actor who starred in a Spielberg-directed film is Daniel Day-Lewis... Although *Lincoln* is a deeply American story, Day-Lewis himself is not an American citizen, making him a prominent example of a non-American actor starring in a Spielberg film.
InteR	Which actor starred in a film directed by Steven Spielberg but not an American?...One notable non-American actor who starred in a Spielberg-directed film is Ralph Fiennes...while not being an American, making him a correct answer to the question.
LameR	Which actor starred in a film directed by Steven Spielberg but not an American? Steven Spielberg has worked with many actors from around the world, including several who are not American...in an American-directed film. His British nationality makes him a correct answer to the question of which actor starred in a Spielberg film but is not American.
AGR	Ben Kingsley, a British actor, starred in Steven Spielberg's *Schindler's List* (1993) as Itzhak Stern. Kingsley is not American, making him a correct example of a non-American actor who starred in a Spielberg-directed film.

Table 5: Query examples rewritten by LLM-based query rewriting models HyDE, InteR, LameR and AGR, respectively.

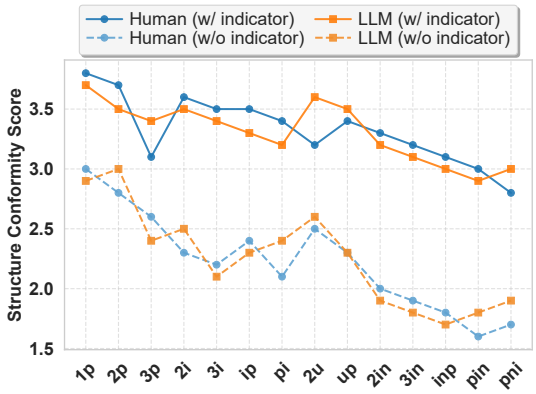

Figure 6: Structure conformity scores of human and LLM assessment across 14 query types with and without the subgraph indicator, respectively.

pairs the models' ability to handle negation conditions, highlighting the need for retrievers capable of modeling exclusion. The alignment between low LSNC and nDCG scores on queries with negation further reveals that retrievers tend to prioritize passages containing negation conditions, thus resulting in poor overall relevance.

Effect of Subgraph Indicator

To evaluate the effect of the subgraph indicator, we conduct an ablation study by removing the corresponding component in the prompt of data generation. Then, the structure conformity of the generated queries is assessed by the human and LLM (GPT-4o), respectively. The assessment score scale is from 1 to 4, where higher scores indicate better generation qualities. Figure 6 reports the average structure conformity scores of human and LLM assessments across 14 query types with and without the subgraph indicator, respectively. Two assessments show that removing the subgraph indicator leads to a drop in structure conformity scores across all query types. Furthermore, the more complex query types (in the right part of Figure 6) get lower scores, verifying the challenges LLMs encounter when generating more complex queries. These ablation study results highlight the significance of the subgraph indicator in guiding LLMs to generate structurally consistent queries.

Case Study

It has been proven that LLM-based query rewriting models such as HyDE, InteR, LameR, and AGR achieve strong performance on some benchmarks, including BEIR (Thakur et al. 2021a), TREC DL'19 (Craswell et al. 2020b) and DL'20 (Craswell et al. 2020a). However, our empirical studies show that these models exhibit catastrophic failures on queries with negation. To illustrate this finding, in Table 5 we display queries rewritten respectively by the four models for the *pin* query. It shows that all rewriting queries explicitly retain the negation condition 'American'. Such incorrectly rewriting queries cause these word co-occurrence-based retrievers to favor the irrelevant documents containing the term 'American', demonstrating their misunderstanding of the original query intent (Xu et al. 2025).

Conclusion

In this paper, we introduce **ComLQ**, a novel dataset to evaluate IR systems on complex logical queries, which are overlooked by existing IR benchmarks. We propose an LLM-based data synthesis method to construct ComLQ, which consists of passage selection, data generation by the subgraph-guided prompt with a subgraph indicator, expert verification for structure conformity and evidence distribution. Our experiment results under zero-shot settings demonstrate that existing retrieval models exhibit significant limitations on complex logical queries. Our findings also emphasize the need for retrieval models capable of reasoning over complex structures. To further measure how well retrievers handle negation conditions, we further propose a new evaluation metric, LSNC, of which the scores reveal consistently low negation consistency across all retrieval models on queries with negation.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and publication of this article: This research was supported by the Chinese NSF Major Research Plan (No.92270121), General Program (No.62572129) and the AI Laboratory of United Automotive Electronic Systems (UAES) Co. (Grant no. 2025-3944).

References

Ajith, A.; Xia, M.; Chevalier, A.; Goyal, T.; Chen, D.; and Gao, T. 2024. LitSearch: A Retrieval Benchmark for Scientific Literature Search. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15068–15083. Miami, Florida, USA: Association for Computational Linguistics.

Bai, J.; Liu, X.; Wang, W.; Luo, C.; and Song, Y. 2023. Complex Query Answering on Eventuality Knowledge Graph with Implicit Logical Constraints. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 30534–30553. Curran Associates, Inc.

Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; Rosenberg, M.; Song, X.; Stoica, A.; Tiwary, S.; and Wang, T. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268.

Chen, X.; Chen, X.; He, B.; Wen, T.; and Sun, L. 2024. Analyze, Generate and Refine: Query Expansion with LLMs for Zero-Shot Open-Domain QA. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11908–11922. Bangkok, Thailand: Association for Computational Linguistics.

Craswell, N.; Mitra, B.; Yilmaz, E.; and Campos, D. 2020a. Overview of the TREC 2020 Deep Learning Track. In Voorhees, E. M.; and Ellis, A., eds., *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Craswell, N.; Mitra, B.; Yilmaz, E.; Campos, D.; and Voorhees, E. M. 2020b. Overview of the TREC 2019 deep learning track. arXiv:2003.07820.

Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; and Xu, J. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 6437–6447. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704901.

Dietz, L.; Verma, M.; Radlinski, F.; and Craswell, N. 2017. TREC Complex Answer Retrieval Overview. In *TREC*.

Fang, T.; Chen, Z.; Song, Y.; and Bosselut, A. 2024a. Complex Reasoning over Logical Queries on Commonsense Knowledge Graphs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11365–11384. Bangkok, Thailand: Association for Computational Linguistics.

Fang, T.; Chen, Z.; Song, Y.; and Bosselut, A. 2024b. Complex Reasoning over Logical Queries on Commonsense Knowledge Graphs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11365–11384. Bangkok, Thailand: Association for Computational Linguistics.

Feng, J.; Tao, C.; Geng, X.; Shen, T.; Xu, C.; Long, G.; Zhao, D.; and Jiang, D. 2024. Synergistic Interplay between Search and Large Language Models for Information Retrieval. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9571–9583. Bangkok, Thailand: Association for Computational Linguistics.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1762–1777. Toronto, Canada: Association for Computational Linguistics.

Hamilton, W.; Bajaj, P.; Zitnik, M.; Jurafsky, D.; and Leskovec, J. 2018. Embedding Logical Queries on Knowledge Graphs. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv e-prints*, arXiv:2112.09118.

Ji, Y.; Wu, K.; Li, J.; Chen, W.; Zhong, M.; Jia, X.; and Zhang, M. 2024. Retrieval and Reasoning on KGs: Integrate Knowledge Graphs into Large Language Models for Complex Question Answering. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7598–7610. Miami, Florida, USA: Association for Computational Linguistics.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6086–6096. Florence, Italy: Association for Computational Linguistics.

Lin, Q.; Mao, R.; Liu, J.; Xu, F.; and Cambria, E. 2023. Fusing topology contexts and logical rules in language models for knowledge graph completion. *Information Fusion*, 90: 253–264.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037. Dubrovnik, Croatia: Association for Computational Linguistics.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A Hu-

man Generated MAchine Reading COmprehension Dataset. In Besold, T. R.; Bordes, A.; d'Avila Garcez, A. S.; and Wayne, G., eds., *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ren, H.; Hu, W.; and Leskovec, J. 2020. Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ren, H.; and Leskovec, J. 2020. Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 19716–19726. Curran Associates, Inc.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Schimanski, T.; Ni, J.; Martín, R. S.; Ranger, N.; and Leippold, M. 2024. ClimRetrieve: A Benchmarking Dataset for Information Retrieval from Corporate Climate Disclosures. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17509–17524. Miami, Florida, USA: Association for Computational Linguistics.

Shen, T.; Long, G.; Geng, X.; Tao, C.; Lei, Y.; Zhou, T.; Blumenstein, M.; and Jiang, D. 2024. Retrieval-Augmented Retrieval: Large Language Models are Strong Zero-Shot Retriever. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 15933–15946. Bangkok, Thailand: Association for Computational Linguistics.

Su, W.; Hu, Y.; Xie, A.; Ai, Q.; Bing, Q.; Zheng, N.; Liu, Y.; Shen, W.; and Liu, Y. 2024. STARD: A Chinese Statute Retrieval Dataset Derived from Real-life Queries by Non-professionals. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10658–10671. Miami, Florida, USA: Association for Computational Linguistics.

Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021a. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021b. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Voorhees, E. M.; and Tice, D. M. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, 200–207. New York, NY, USA: Association for Computing Machinery. ISBN 1581132263.

Wang, J. A.; Wang, K.; Wang, X.; Naidu, P.; Bergen, L.; and Paturi, R. 2023. Scientific Document Retrieval using Multi-level Aspect-based Queries. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 38404–38419. Curran Associates, Inc.

Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; and Nie, J.-Y. 2024. C-Pack: Packed Resources For General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 641–649. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.

Xu, G.; Zhang, Z.; Mei, W.; Liang, J.; Lu, W.; Zhang, X.; Yang, Z.; Ma, X.; Xiao, Y.; and Yang, D. 2025. Logical Consistency is Vital: Neural-Symbolic Information Retrieval for Negative-Constraint Queries. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 1828–1847. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Zhang, C.; Peng, Z.; Zheng, J.; and Ma, Q. 2024. Conditional Logical Message Passing Transformer for Complex Query Answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 4119–4130. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704901.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2024. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.*, 42(4).

Zhuang, S.; Ma, X.; Koopman, B.; Lin, J.; and Zuccon, G. 2024. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4375–4391. Miami, Florida, USA: Association for Computational Linguistics.
fi