Q2EI: Query-to-Entity Inference for Semantic Condensation in Domain-Specific Retrieval

Yixuan Sun[1,2,3,*], Zhenqin Xu[1,3,*], Hanfeng Zhai[3,4], Zishu Yu[1,3,†], Xiaohui Peng[1,3]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]College of Computer Science and Software Engineering, Hohai University

Correspondence: yuzishu@ict.ac.cn,

Abstract

Retrieval-Augmented Generation (RAG) remains unreliable in specialized domains due to semantic and lexical mismatch between lay queries and professional terminology, and existing generative expansion often introduces redundancy or hallucinations that cause semantic drift. We propose Generative Query Condensation (GQC), a query rewriting strategy that reframes rewriting as semantic condensation rather than expansion. To operationalize GQC, we introduce Query-to-Entity Inference (Q2EI), an entity-centric rewriting method that realizes semantic condensation through explicit inference of the underlying target entity. By moving semantic alignment from retrieval-time vector matching to the rewriting stage, Q2EI produces information-dense query representations. Experimental results on medical and legal benchmarks show that Q2EI consistently outperforms strong baselines across retrievers, improving retrieval effectiveness while substantially reducing rewriting token consumption compared to generative expansion methods. Further analysis confirms that these gains primarily arise from accurate entity inference, and that Q2EI's semantic condensation design limits error amplification when inference is imperfect, leading to more stable and interpretable retrieval behavior [1].

1 Introduction

Retrieval-Augmented Generation (RAG) alleviates knowledge hallucination and temporal obsolescence in Large Language Models (LLMs) by grounding generation in external non-parametric knowledge, and has achieved strong performance on general tasks such as open-domain question answering (Lewis et al., 2020; Guu et al., 2020). However, retrieval still faces significant bottlenecks in specialized domains such as medicine, law, and IT operations. A key challenge is that user queries often reside in a non-professional semantic space, whereas documents are written in a professional one. This semantic misalignment is difficult to resolve by vector retrieval alone: retrievers primarily capture surface-level semantic similarity, but cannot reliably infer the underlying domain concept implied by a lay description. We refer to such non-expert user queries as *lay queries*. This representation mismatch leads to lexical mismatch and semantic gaps, thereby compromising retrieval effectiveness (see Figure 1 (a)).

Specifically, due to a lack of domain knowledge, user queries are often phenomenon-level descriptions. However, the documents use precise, domain-specific-level terminology. For instance, a patient might describe symptoms as *"Ate home-canned food, now blurred vision, droopy eyelids, hard to swallow—could this be deadly food poisoning?"* whereas the relevant document is indexed under the term *"Botulism."* On the one hand, this representation misalignment makes it difficult for sparse retrievers (e.g., TF-IDF (Salton and Buckley, 1988) and BM25 (Robertson and Zaragoza, 2009)) to retrieve target documents due to their reliance on lexical overlap. On the other hand, dense retrievers (such as Multilingual E5 (mE5) (Wang et al., 2024), BGE-M3 (Chen et al., 2024), or Contriever (Izacard et al., 2022)) also struggle to establish a semantic mapping from phenomenon-level descriptions to professional terminology.

To alleviate these issues, existing works predominantly adopt Generative Query Expansion (GQE). Generation-Augmented Retrieval (GAR) (Mao et al., 2021) and its successors, such as HyDE (Gao et al., 2023) and Query2Doc (Q2D) (Wang et al., 2023), expand query semantics by generating pseudo-documents to improve

[*]Equal contribution.
[†]Corresponding authors.
[1]We release code and data in an anonymized GitHub repository at https://github.com/xu2023-ICT/Q2EI

Findings of the Association for Computational Linguistics: ACL 2026, pages 35943–35958
July 2-7, 2026

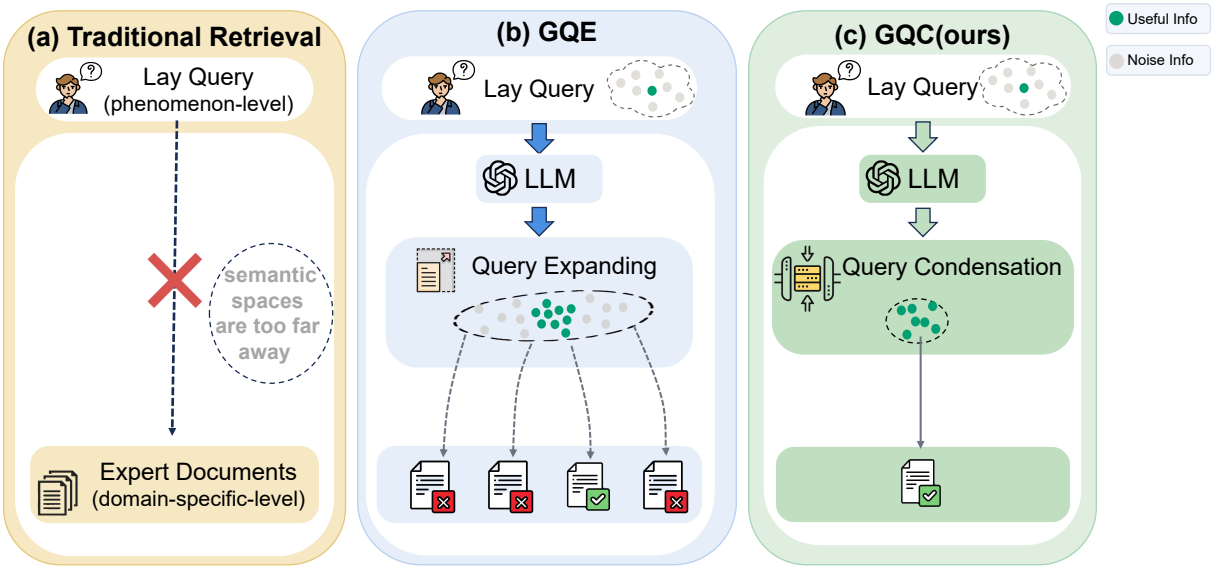


Figure 1: Schematic comparison of different retrieval strategies. (a) **Traditional Retrieval** suffers from significant lexical mismatch between lay queries and expert documents. (b) **GQE** bridges the gap but often introduces redundancy and noise, which may mislead the retriever toward irrelevant documents. (c) The proposed **GQC** (ours) employs an inference engine to condense semantics into high-density representations.

semantic coverage and lexical overlap (see Figure 1 (b)). However, our motivation experiments in section 5.1 indicate that in specialized domains, such verbose generated content tends to dilute information density and induce semantic drift. The result is also consistent with recent studies (Weller et al., 2024). Another category of research employs Knowledge Graph (KG) methods, utilizing explicit relational constraints to align semantics (Luo et al., 2024; Zhu et al., 2025; Mavromatis and Karypis, 2025). However, their high graph construction costs and dependence on graph coverage (Pan et al., 2024) limit their application.

In this paper, we propose the Generative Query Condensation (GQC) strategy (see Figure 1 (c)). Unlike conventional dense retrieval methods that rely on implicit vector matching during retrieval, this strategy leverages the parametric knowledge of LLMs to perform explicit semantic extraction and compression at the query rewriting stage. GQC compresses the semantic information within the query into a high-density representation, thereby effectively improving the signal-to-noise ratio. We introduce Query-to-Entity Inference (Q2EI) as a concrete implementation of GQC. Unlike GQE methods, Q2EI utilizes LLMs to condense ambiguous user descriptions into a core domain-specific entity and reconstructs an entity-centric query representation. This effectively mitigates both lexical and semantic mismatch.

Experimental results on two specialized-domain benchmarks (MedQuAD and COLIEE) show that Q2EI yields substantial and consistent improvements across retrievers. On COLIEE, mE5 achieves an nDCG@10 improvement of 3.35 over the strongest baseline. On MedQuAD, BM25 yields a Recall@10 improvement of 13.97 compared with the best-performing baseline, while maintaining significantly lower computational overhead.

The main contributions of this paper are summarized as follows:

- We propose GQC as a query rewriting strategy for specialized domain retrieval. By prioritizing semantic condensation over text expansion, it alleviates lexical mismatch and semantic gaps. We further analyze the critical role of high-information-density representations in specialized domain retrieval.
- We introduce Q2EI, an entity inference method that requires no fine-tuning and does not rely on manually constructed knowledge bases. By guiding LLMs with specific inference instructions to infer potential entities from lay queries, we generate high signal-to-noise ratio entity-centric queries.
- Extensive experiments on multiple special-

ized domain datasets and various retriever architectures demonstrate that Q2EI significantly outperforms strong baseline methods. Further analysis attributes the gains to accurate potential entity inference, providing interpretability, while incurring substantially lower computational overhead than GQE methods.

2 Related Work

2.1 Sparse and Dense Retrieval

In specialized domains such as medicine and law, a substantial semantic gap often exists between user queries and professional literature (Chalkidis et al., 2020). Traditional sparse retrieval models (e.g., BM25 (Robertson and Zaragoza, 2009)) rely heavily on lexical overlap and therefore struggle with the lexical mismatch between colloquial expressions and domain-specific terminology. Dense retrieval methods (e.g., DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2021)) embed texts into a low-dimensional vector space, relaxing the strict requirement of exact lexical matching. However, geometric proximity in the embedding space is often insufficient to capture deeper logical associations (Marcus, 2018; Onoe and Durrett, 2020).

2.2 Query Augmentation via Generative and Pseudo-Relevance Feedback

Traditional Pseudo-Relevance Feedback (PRF) methods (e.g., RM3 (Lavrenko and Croft, 2003)) rely solely on initial retrieval results, making it difficult to introduce external knowledge. With the development of LLMs, Generative Query Expansion (GQE) has become a mainstream strategy. HyDE (Gao et al., 2023) and Query2Doc (Q2D) (Wang et al., 2023) utilize LLMs to expand short queries into pseudo-documents to enhance semantic matching. This "additive strategy" is effective in open-domain question answering. However, in specialized domain retrieval tasks, long texts tend to introduce hallucinations and irrelevant noise, leading to semantic drift (Weller et al., 2024). Unlike GQE, this paper explores a "subtractive strategy", utilizing inference to condense redundant query descriptions, thereby improving the signal-to-noise ratio of the rewritten query and reducing the risk of semantic drift during retrieval.

2.3 Knowledge-Enhanced and Entity-Centric Retrieval

Although utilizing structured knowledge can improve precision (Pan et al., 2024), such methods suffer from a significant scalability bottleneck. The construction of explicit Knowledge Graphs (KGs) relies highly on expert human intervention, which incurs high maintenance costs (Ding et al., 2024). In contrast, our method leverages the parametric memory of LLMs as a "soft knowledge base," achieving logical inference without the need to explicitly construct a graph (Petroni et al., 2019).

2.4 Reasoning for Information Retrieval

Recent research has begun exploring the use of LLM reasoning capabilities (e.g., Chain-of-Thought (CoT) (Wei et al., 2022)) to assist Information Retrieval (IR) tasks. However, most existing works (Sun et al., 2023; Pradeep et al., 2023; Qin et al., 2024) apply reasoning during the reranking stage or the offline data augmentation stage (Jeronymo et al., 2023). The former sits at the end of the retrieval pipeline; the latter is primarily applied in an offline phase. We shift reasoning to the pre-retrieval stage by proposing GQC, which uses inference to generate retrieval anchors. This shift enables the model to bridge the semantic gap through logical reasoning rather than relying on text similarity matching. Consequently, our approach is orthogonal to existing work and complements prior reasoning-based IR methods.

3 Methodology

3.1 Problem Formulation

We formulate the specialized domain retrieval problem as a matching problem across semantic spaces. The query space $\mathcal{Q}_{\text{lay}}$ comprises phenomenon-level queries q_{lay} from lay users, while the document space $\mathcal{D}_{\text{expert}}$ consists of documents d with professional terminologies and normalized entities.

Due to significant distributional differences between the two spaces, $\text{Sim}(q_{\text{lay}}, d_{\text{target}})(d_{\text{target}} \in \mathcal{D}_{\text{expert}})$ struggles to effectively characterize the relevance between the q_{lay} and its target document d_{target}, thereby degrading retrieval performance.

Consequently, the objective of retrieval optimization is to construct a mapping $f : \mathcal{Q}_{\text{lay}} \rightarrow \mathcal{Q}$ that transforms a lay query q_{lay} into a rewritten

query $q = f(q_{\text{lay}})$ closer to the semantic distribution of the target document, thereby maximizing the similarity between the rewritten query and the target document d_{target}:

$$\max_{f} \ \mathrm{Sim}\big(\mathrm{Enc}(f(q_{\text{lay}})), \mathrm{Enc}(d_{\text{target}})\big), \quad d_{\text{target}} \in \mathcal{D}_{\text{expert}}. \tag{1}$$

where $\mathrm{Enc}(\cdot)$ denotes the retriever encoder, and $\mathrm{Sim}(\cdot,\cdot)$ represents the similarity function.

3.2 Strategy Comparison: GQE vs. GQC

To implement the mapping f, existing works predominantly adopt Generative Query Expansion (GQE), whereas this paper employs the Generative Query Condensation (GQC) strategy.

GQE (Query Expansion). GQE introduce an intermediate variable c_{gen} to approximate the semantic distribution of the target document by supplementing context. The mapping process can be expressed as:

$$f_{\text{GQE}}(q_{\text{lay}}) = \mathcal{T}(q_{\text{lay}}, c_{\text{gen}}), \quad c_{\text{gen}} \sim P_{\text{LLM}}\big(\cdot \mid q_{\text{lay}}; \mathcal{E}\big). \tag{2}$$

where $\mathcal{T}$ denotes a text concatenation or fusion operation, and $\mathcal{E}$ represents the set of expansion instructions used to guide the LLM in generating supplementary context c_{gen} to improve the semantic coverage of the query.

GQC (Semantic Condensation). GQC does not approximate the document distribution by supplementing redundant context; instead, it leverages the parametric knowledge of LLMs to perform semantic condensation on the query, outputting a rewritten query with high information density:

$$f_{\text{GQC}}(q_{\text{lay}}) = q_{\text{GQC}}, \quad q_{\text{GQC}} \sim P_{\text{LLM}}\big(\cdot \mid q_{\text{lay}}; \mathcal{C}\big). \tag{3}$$

where $\mathcal{C}$ is the instruction set for semantic condensation, used to constrain the output to focus on retrievable core semantics. This process establishes a semantic bridge between ambiguous phenomenon-level descriptions and precise professional knowledge. Compared to GQE, GQC significantly reduces redundant information and improves the signal-to-noise ratio of the query.

This distinction is illustrated by the example in Appendix A (Table 2). It can be observed that while GQE methods cover key semantics, the introduced redundancy and hallucinations can lead the retriever to incorrect documents. In contrast, the proposed condensation strategy (implemented as Q2EI, detailed in section 3.3) retains only core entity semantics, thereby avoiding such retrieval errors caused by noise.

3.3 Q2EI: Entity-Centric Query Rewriting

Based on the GQC strategy, we propose Query-to-Entity Inference (Q2EI), an entity-centric query rewriting method. In this work, an entity refers to a highly condensed semantic abstraction that captures the core conceptual meaning underlying a query, serving as an anchor for domain-specific knowledge.

As shown in Figure 2. This method comprises three steps:

Step 1: Instruction Construction and Constraint Setting. We construct instructions $\mathcal{I}_{\text{prompt}}$ to guide the model's inference and rewriting process. The instructions (see Figure 9 in Appendix B for details) include three key elements:"

(1) *Persona-based Prompting:* Guides the LLM to act as a domain expert via specific role-play instructions (Xu et al., 2023; Kong et al., 2024), thereby enhancing the model's ability to understand and parse domain semantics;

(2) *Entity-first inference with reformulation constraint:* Requires the model to first infer the normalized entity e_{core} and generate an entity-centric query q_{ent} based on this entity. This ensures that the rewritten query focuses on core entity semantics and aligns with domain-specific expression conventions;

(3) *Adaptive Demonstration:* Supports both zero-shot and few-shot settings, relying on the model's internal knowledge and a small number of in-context mapping examples (Brown et al., 2020), respectively.

Step 2: Model Inference and Entity-Centric Rewriting. Under the instruction constraints constructed in Step 1, given q_{lay} and $\mathcal{I}_{\text{prompt}}$, the LLM infers the core entity within the query and normalizes it into an entity-centric query q_{ent}. Through entity rewriting, this process ensures the query closely adheres to the core semantics of q_{lay}.

Step 3: Entity-Centric Retrieval. We utilize q_{ent} for retrieval, transforming the match from "Phenomenon Description → Professional Document" to "Entity-Centric Query → Professional Document." This effectively alleviates the seman-

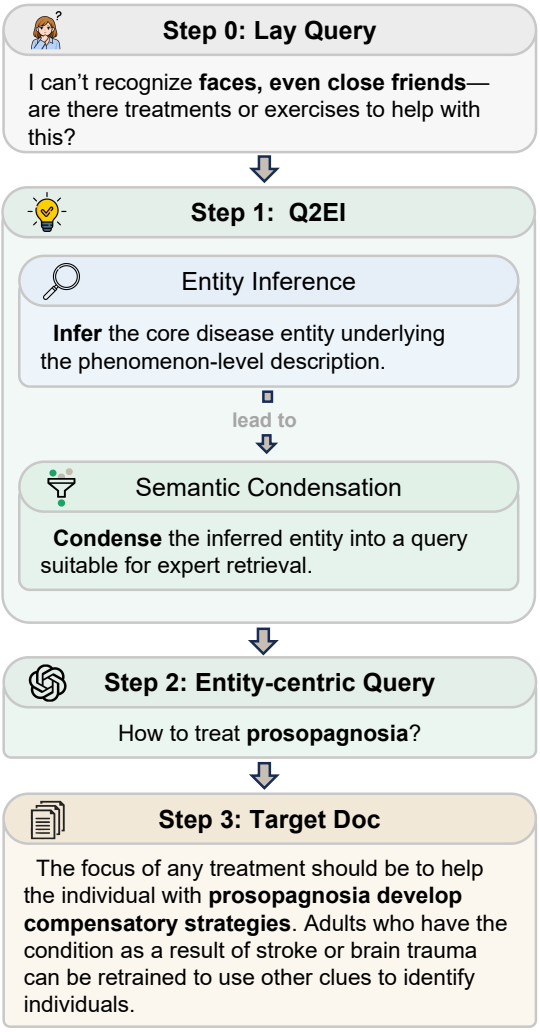


Figure 2: Overview of the Q2EI method.

tic gap and improves the signal-to-noise ratio of the query.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate Q2EI in two specialized domains: medicine and law, utilizing the MedQuAD (Abacha and Demner-Fushman, 2019) and COLIEE (Kim et al., 2022) datasets, respectively. Existing benchmarks typically use expert-written normalized questions as queries. To simulate the questioning style of lay users in specialized domains, we constructed lay queries: we utilized LLMs to rewrite original professional questions into lay descriptions of surface-level descriptions (for MedQuAD) or real-life dilemmas (for COLIEE). All rewritten queries underwent manual review by multiple reviewers to ensure the original semantic intent was maintained and professional terminology was removed. The specific prompt templates are provided in Appendix B.

Baselines. We compare Q2EI with the four methods: (1) **Lay Query:** Directly uses the lay query for retrieval. (2) **Keyword Extraction:** Utilizes an LLM to extract keywords from the query as retrieval input. This baseline is included to demonstrate that the performance gains of Q2EI stem from deep entity inference rather than shallow lexical extraction. (3) **HyDE** (Gao et al., 2023): A generative retrieval method that expands queries by generating hypothetical documents. (4) **Query2Doc (Q2D)** (Wang et al., 2023): Generates pseudo-documents using few-shot prompting to introduce contextual information. The specific prompts employed for all the LLM-based baselines mentioned above are provided in Appendix B. We provide additional analysis comparing Q2EI with a Chain-of-Thought based variant, Query Expansion(CoT) (Jagerman et al., 2023), in Appendix D. Query Expansion(CoT) leverages CoT prompting to generate step-by-step reasoning, producing query expansions with many related terms, serving as a strong expansion-based baseline.

Implementation Details. We primarily employ GPT-5 (OpenAI, 2026) as the generative language model for the experiments reported in this section. To further verify the robustness of our method across different models, we conduct supplementary experiments using Claude Sonnet 4.5 (Anthropic, 2025) and Qwen2.5-72B-Instruct (Yang et al., 2024); these results are discussed in Appendix C. We evaluate performance across three retrievers, including the sparse retriever BM25 (Robertson and Zaragoza, 2009) and the dense retrievers Multilingual E5 (mE5) (Wang et al., 2024) and BGE-M3 (Chen et al., 2024). Q2EI is tested under both zero-shot and few-shot settings. In the few-shot setting, we use the same examples as Q2D to ensure a fair comparison. All methods are evaluated using Recall@1, Recall@10, and nDCG@10 metrics. All experiments are implemented by the open-source LlamaIndex framework (Liu, 2022). Importantly, we do not perform any filtering or selection of samples based on retrieval performance. Moreover, all generated outputs are used as-is without any manual post-editing, ensuring the originality of the data and the fairness of the evaluation.

Method	BM25			BGE-M3			mE5		
	R@1	R@10	N@10	R@1	R@10	N@10	R@1	R@10	N@10
	MedQuAD								
Lay Query	7.35	21.88	13.94	7.17	31.25	17.94	5.70	24.82	14.71
HyDE	–	–	–	16.91	46.88	30.93	8.46	30.33	18.29
Q2D	13.97	38.05	24.77	18.01	49.08	32.35	7.90	31.25	18.70
Keyword Extraction	12.32	41.18	25.90	12.13	35.29	23.27	9.38	34.19	21.04
Q2EI (zero-shot)	**14.34**	49.63	30.90	22.06	55.51	38.39	24.82	60.85	42.38
Q2EI (few-shot)	13.79	**52.02**	**31.31**	**23.16**	**57.17**	**39.71**	**25.92**	**61.40**	**43.13**
	COLIEE								
Lay Query	2.59	9.43	6.00	7.02	16.08	11.12	6.47	14.23	10.23
HyDE	–	–	–	6.28	22.55	15.62	6.65	15.71	11.26
Q2D	12.94	27.91	20.53	11.28	27.54	19.80	8.32	22.00	15.77
Keyword Extraction	5.55	13.68	9.51	5.91	14.60	9.73	6.84	15.16	11.11
Q2EI (zero-shot)	8.13	22.74	15.44	10.91	23.66	17.58	9.24	19.59	14.36
Q2EI (few-shot)	**13.49**	**28.10**	**21.01**	**13.49**	**28.10**	**20.67**	**12.20**	**25.69**	**19.12**

Table 1: Retrieval performance (%) comparison of Q2EI and baseline methods on medical (MedQuAD) and legal (COLIEE) domain datasets. Note: R@k stands for Recall@k, N@k stands for nDCG@k.

4.2 Main Results

4.2.1 Performance on MedQuAD

Table 1 compares the retrieval accuracy of Q2EI and various baselines on the medical (MedQuAD) and legal (COLIEE) datasets. Overall, Q2EI demonstrates clear and consistent gains across domains and retriever architectures, with the largest improvements on dense retrievers-suggesting that entity-level semantic condensation is particularly effective for alignment in embedding-based retrieval spaces.

Taking mE5 on MedQuAD as an example, the nDCG@10 scores for HyDE and Q2D are 18.29 and 18.70, respectively, whereas Q2EI (zero-shot) improves this to 42.38, and Q2EI (few-shot) further achieves 43.13. These results indicate that, compared to query expansion methods relying on long text generation, entity-centric query rewriting aligns better with the semantic alignment requirements in dense representation spaces. Compared with Keyword Extraction, Q2EI achieves consistently higher performance across retrievers. For example, on BGE-M3, Q2EI (few-shot) reaches an nDCG@10 of 39.71, substantially higher than the 23.27 achieved by Keyword Extraction, indicating that the gains cannot be explained by shallow lexical cues alone but instead stem from explicit entity inference.

4.2.2 Performance on COLIEE

In the legal domain, the few-shot setting proves particularly critical for performance improvement. On the COLIEE dataset, after introducing the same number of examples, Q2EI (few-shot) achieves the best results across all retrievers. For instance, on mE5, its nDCG@10 improves from 15.77 (Q2D) to 19.12. This indicates that, under equal utilization of in-context learning, semantic condensation-based query rewriting exhibits more stable retrieval advantages in cross-domain scenarios.

5 Analysis

5.1 Information Density Hypothesis: Redundancy Induces Semantic Drift

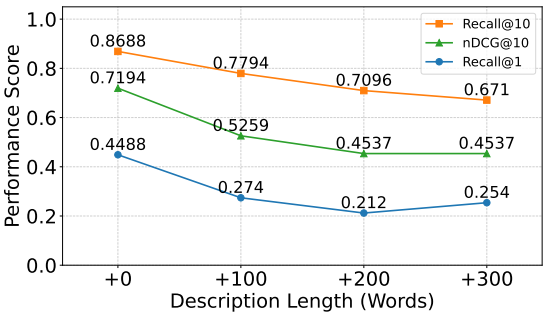


Figure 3: Validation of Information Density Hypothesis: Trend analysis of the impact of appending the generative content description length on retrieval performance.

To verify the information density hypothesis, we designed a controlled experiment. Using 500 original queries from the MedQuAD test set as a

baseline, we employed an LLM to generate query-relevant pathological descriptions of target lengths ($k \in \{100, 200, 300\}$ words) (approximately k words each). These descriptions are appended to the original queries (see Appendix E for prompts). All experiments are evaluated using the Multilingual E5 (mE5) retriever.

As shown in Figure 3, retrieval performance exhibits a decline as the length of the appended description increases. Compared to the original query, appending just 100 words causes Recall@10 to drop from 86.88 to 77.94; when the length increases to 300 words, Recall@10 further degrades to 67.10. These results indicate that retrieval effectiveness can weaken even if the generated content is semantically relevant. Appending the generated content induces a shift in the embedding centroid. This shift ultimately leads to the degradation of retrieval performance.

5.2 Attribution Analysis: Impact of Entity Inference Accuracy

To analyze the relationship between retrieval performance gains and the accuracy of entity inference, we divided the evaluation queries in the MedQuAD test set into two mutually exclusive subsets: *Entity-Aligned Group*, where the inferred entity holds an equivalence (same, hyponym, or hypernym) relationship with the target entity of the original question; and *Entity-Misaligned Group*, where this relationship is not met. In the zero-shot setting, *Entity-Aligned Group* accounts for 83.4 of queries, while in the few-shot setting, it accounts for 84.0. Notably, failed inference cases constitute only a small portion of the evaluation set; for completeness, we report the overall retrieval performance (Recall@1 and Recall@10 over all queries) in Appendix G.

We evaluated the retrieval performance of both groups on BM25, BGE-M3, and mE5. As shown in Figure 4 (a), the *Entity-Aligned Group* consistently outperforms the *Entity-Misaligned Group* by large absolute margins across retriever architectures. Taking mE5 in the zero-shot setting as an illustrative example, the nDCG@10 of the *Entity-Aligned Group* reaches 52.12, whereas the *Entity-Misaligned Group* achieves 2.72, yielding an absolute gap of 49.40 points. Comparable absolute gaps are observed on BM25 (38.42 vs. 5.60) and BGE-M3 (48.24 vs. 9.16). These results demonstrate that Q2EI's retrieval effectiveness is strongly correlated with entity inference accuracy. When entity inference is misaligned with the target concept, retrieval performance drops sharply, indicating that accurate entity inference constitutes a critical prerequisite for effective retrieval within the Q2EI framework.

5.3 Misalignment Analysis: Entity Misalignment Reflects Query-Intrinsic Hardness

To characterize retrieval behavior under entity misalignment, we compare all evaluated retrieval methods (Lay Query, HyDE, and Q2D) using the same group partition. We focus on nDCG@10 in this section; additional metrics and full experimental results are reported in Appendix F.

As shown in Figure 4 (b), all methods yield consistently low nDCG@10 across BM25, BGE-M3, and mE5. Specifically, in the Entity-Misaligned Group, nDCG@10 ranges from 2.72 to 9.10 across all methods and retrievers, whereas in the Entity-Aligned Group it spans 15.96 to 38.92. This uniform degradation across methods suggests that performance bottlenecks under entity misalignment are largely method-agnostic and instead stem from query-intrinsic issues: the phenomenon-level descriptions are often vague, incomplete, or even incorrect, making it difficult-and in some cases highly unreliable-to infer the correct underlying condition from the described manifestations alone.

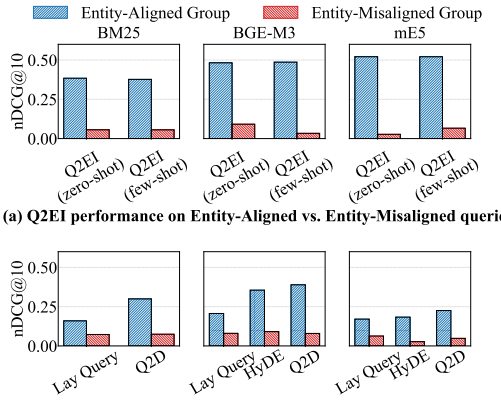


Figure 4: Retrieval performance (nDCG@10) under an entity-alignment split. (a) Performance comparison of Q2EI (zero-shot and few-shot) in Entity-Aligned vs. Entity-Misaligned Groups. (b) Performance of baseline methods (Lay Query, HyDE, Q2D) under the same grouping.

In the Entity-Aligned Group, Q2EI consistently achieves the best (or tied-best) retrieval perfor-

mance across retriever architectures. For example, on mE5, Q2EI (few-shot) reaches an nDCG@10 of 52.11, substantially outperforming HyDE (18.34) and Q2D (22.53) under the same retriever. Comparable margins are observed on BM25 and BGE-M3. This pattern suggests that in the Entity-Aligned Group, Q2EI infers an entity aligned with the target and condenses the query into its core domain semantics, producing a high-information-density representation. By contrast, HyDE and Q2D generate substantially longer expansions; while these expansions may include relevant content, they often contain redundant information that lowers the effective signal-to-noise ratio for retrieval.

Notably, Q2EI remains competitive in the Entity-Misaligned Group, which represents a particularly challenging regime where the correct condition is often hard to disambiguate from the query alone. In this setting, performance differences mainly reflect how much a rewriting method drifts when its hypothesis is off-target. Even when Q2EI infers an incorrect entity, the predicted condition may still share overlapping symptom clusters or manifestations with the true target, leading to a comparatively mild semantic mismatch. In contrast, HyDE and Q2D tend to generate verbose expansions under misalignment, which can introduce additional off-target details and increase the risk of semantic drift. As a result, Q2EI does not exhibit disproportionate degradation and is often among the best-performing methods in this group; for instance, on mE5, Q2EI (few-shot) achieves an nDCG@10 of 6.61, compared to 4.82 for Q2D. Overall, these findings suggest that semantic condensation is relatively failure-tolerant: when inference is imperfect, compact, high-density rewrites help limit error amplification and yield more stable retrieval behavior.

5.4 Efficiency and Deployment Cost: Token Consumption Comparison

To compare the computational overhead of different methods during the query rewriting stage, we measure the average token consumption per query on the MedQuAD dataset. This metric accounts for the entire rewriting process, including the input prompt, intermediate reasoning tokens generated by the LLM, and the final rewritten output.

The results reveal substantial differences in generation cost across methods. HyDE incurs the highest overhead, consuming 13,408.55 tokens per query on average due to the need to generate multiple hypothetical documents. Q2D reduces this cost to 2,767.93 tokens per query, but still requires generating relatively long pseudo-contexts. In contrast, Q2EI dramatically lowers generation cost: the zero-shot variant consumes only 905.21 tokens per query, corresponding to 6.8% of HyDE's token usage, while the few-shot variant consumes 1,718.34 tokens per query, or 12.8% of HyDE's cost.

These reductions directly reflect the design principle of the GQC strategy. Rather than producing verbose textual expansions, Q2EI infers a compact, entity-centric representation that preserves core semantics while minimizing redundant generation. As a result, Q2EI substantially reduces deployment cost at the query rewriting stage, offering significantly improved efficiency without sacrificing retrieval effectiveness.

6 Conclusion

To address lexical mismatch and semantic gaps in specialized-domain retrieval, this paper proposes the GQC strategy and introduces Q2EI as its concrete instantiation. Unlike the additive paradigm of GQE, GQC adopts a subtractive approach via semantic condensation. Q2EI operationalizes GQC through explicit entity inference. It maps users' phenomenon-level descriptions to a core domain entity and rewrites them into entity-centric queries. This reasoning-driven condensation constructs a semantic bridge between user intent and domain knowledge, improving the query's signal-to-noise ratio.

Experimental results demonstrate that Q2EI significantly outperforms existing baselines across multiple datasets and diverse retriever architectures. Our in-depth analysis confirms four key observations: (1) redundant generated content tends to induce semantic drift; (2) the performance gains of Q2EI are attributed to accurate entity inference; (3) Q2EI exhibits failure-tolerant behavior-when entity inference is imperfect, its semantic condensation design limits error amplification and leads to more stable retrieval behavior compared to generative expansion methods; and (4) Q2EI achieves these improvements with significantly reduced computational overhead.

In summary, Q2EI offers an efficient and cost-effective query rewriting strategy for specialized domain retrieval. This approach demonstrates ex-

ceptional robustness and exhibits significant potential practical value in professional scenarios such as medicine and law.

Limitations

While Q2EI demonstrates superior performance in specialized domain retrieval tasks, we acknowledge the following limitations, which also illuminate directions for future research.

Inference Latency & Cost: Although the computational overhead of Q2EI is significantly lower than that of GQE, on-the-fly inference inevitably introduces higher latency compared to traditional sparse or dense retrieval methods. For industrial-grade deployment scenarios with strict real-time requirements, the current inference cost remains a primary bottleneck. Fortunately, this bottleneck is likely to ease over time, as hardware cost-efficiency continues to improve and model compression/acceleration techniques increasingly enable faster and cheaper LLM inference in production.

Dependency on LLM Capabilities: The performance of Q2EI is constrained by the parametric domain knowledge of the underlying LLM. When handling extremely ambiguous queries or those involving highly long-tail knowledge, incorrect entity inference can trigger a cascading effect, directly propagating errors to the retrieval stage and degrading retrieval effectiveness. Conversely, Q2EI is expected to benefit directly from ongoing advances in foundation models: as LLMs become more capable and better grounded in domain knowledge, the quality of entity inferenceand thus retrieval performanceshould improve accordingly.

Domain & Task Applicability: The core value of Q2EI lies in realizing the inference from "phenomenon-level descriptions" to "core entity." In simple matching tasks that do not require complex reasoning or in non-entity-centric retrieval scenarios, the marginal utility of this method may diminish.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. U23B2004 and U24B6012.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Anthropic. 2025. Claude sonnet 4.5 system card. API version: claude-sonnet-4-5-20250929.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 2318–2335. Association for Computational Linguistics.

Linyi Ding, Sizhe Zhou, Jinfeng Xiao, and Jiawei Han. 2024. Automated construction of theme-specific knowledge graphs. *CoRR*, abs/2404.19146.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1762–1777. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 3929–3938. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Vitor Jeronymo, Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, Roberto A. Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *CoRR*, abs/2301.01820.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. COLIEE 2022 summary: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12-17, 2022, Revised Selected Papers*, Lecture Notes in Computer Science, pages 51–67. Springer.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4099–4113. Association for Computational Linguistics.

Victor Lavrenko and W Bruce Croft. 2003. Relevance models in information retrieval. In *Language modeling for information retrieval*, pages 11–56. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jerry Liu. 2022. Llamaindex. GitHub repository for LLM data framework.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4089–4100. Association for Computational Linguistics.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Costas Mavromatis and George Karypis. 2025. GNN-RAG: graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 16682–16699. Association for Computational Linguistics.

Yasumasa Onoe and Greg Durrett. 2020. Interpretable entity representations through large-scale typing. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 612–624. Association for Computational Linguistics.

OpenAI. 2026. Openai GPT-5 system card. *CoRR*, abs/2601.03267.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *CoRR*, abs/2312.02724.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Findings of ACL, pages 1504–1518. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Orion Weller, Kyle Lo, David Wadden, Dawn J. Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. 2024. When do generative query and document expansions fail? A comprehensive study across methods, retrievers, and datasets. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, Findings of ACL, pages 1987–2003. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *CoRR*, abs/2305.14688.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 8912–8924. Association for Computational Linguistics.

A Case Study

We conduct a qualitative case study to analyze the behavior of baseline methods and Q2EI.

Table 2 presents an example from the MedQuAD dataset. The table lists the lay query and the rewritten queries produced by Query2Doc (Q2D) and Q2EI. We exclude HyDE rewrites, as HyDE concatenates multiple hypothetical documents with the query during retrieval, lacking direct interpretability. The table also shows an irrelevant passage retrieved by Q2D.

In this example, baselines fail to retrieve target passages due to lexical mismatches or irrelevant information in expanded queries. This redundancy increases the risk of semantic drift. It can also introduce unsupported details. In contrast, Q2EI correctly infers the target entity, Crimean-Congo hemorrhagic fever (CCHF). It rewrites the query into a concise, entity-centric form. This condensation improves the signal-to-noise ratio. As a result, Q2EI retrieves the target passage with high confidence.

B Prompt Design

We detail the specific prompts used in our experiments.

Lay Query Generation. Figure 5 shows the prompt used to rewrite professional questions into lay queries. The rewritten queries simulate the questioning style of non-expert users in specialized domains. This process removes domain-specific terminology while preserving the original intent.

Keyword Extraction. Figure 6 presents the prompt used for the Keyword Extraction baseline. The prompt instructs the model to extract salient keywords from the query. This baseline demonstrates that the performance gains of Q2EI stem from deep entity inference rather than shallow lexical extraction.

HyDE. Figure 7 shows the prompt used by the HyDE method. For cost efficiency, we generate four hypothetical documents per query. We average their embeddings with the lay query embedding for retrieval.

Target Passage	Laboratory tests that are used to diagnose CCHF include antigen-capture enzyme-linked immunosorbent assay (ELISA), real time polymerase chain reaction (RT-PCR), virus isolation attempts... Laboratory diagnosis... by using the combination of detection of the viral antigen... Later in the course... antibodies can be found in the blood...
Query	After forest hike, tick bite, now high fever, bad headache, body aches, vomiting, nosebleeds. Should I worry?
Q2D	Tick-borne infections can cause abrupt high fever... including Rocky Mountain spotted fever (RMSF), ehrlichiosis/anaplasmosis, Colorado tick fever... Nosebleeds (epistaxis), easy bruising... suggest thrombocytopenia... seen in severe RMSF... Doxycycline is first-line...
Q2EI	What is Crimean-Congo hemorrhagic fever (CCHF) in a patient with a recent tick bite, sudden high fever, severe myalgias, and epistaxis/easy bruising?
Wrong passage	The first symptoms of Rocky Mountain spotted fever (RMSF)... spotted (petechial) rash of RMSF is usually not seen... Rickettsia rickettsii infects the endothelial cells... can be performed on a skin biopsy... Recommended Dosage Doxycycline is the first line treatment... such as ehrlichiosis and anaplasmosis...

Table 2: Case study from the MedQuAD dataset. Comparisons between Q2D and our Q2EI. Yellow indicates parts similar to the target passage, Pink marks "distractors" that induced retrieval errors; Blue represents redundant context.

	BM25			BGE-M3			mE5		
Method	R@1	R@10	N@10	R@1	R@10	N@10	R@1	R@10	N@10
	Claude Sonnet 4.5								
Lay query	7.35	21.88	13.94	7.17	31.25	17.94	5.70	24.82	14.71
Q2D	12.50	47.92	29.72	25.00	52.08	37.26	12.50	37.50	25.06
Q2EI (zero-shot)	**14.58**	**62.50**	**35.89**	**45.83**	**68.70**	**57.82**	**39.58**	**72.92**	**55.93**
Q2EI (few-shot)	12.50	56.25	31.75	29.17	64.58	45.04	29.17	68.75	48.39
	Qwen2.5-72B-Instruct								
Lay query	7.35	21.88	13.94	7.17	31.25	17.94	5.70	24.82	14.71
Q2D	**12.68**	**39.52**	**24.79**	15.07	42.28	27.98	9.93	33.27	20.93
HyDE	—	—	—	12.87	43.38	27.27	6.62	26.29	16.10
Q2EI (zero-shot)	10.66	33.82	21.64	**18.57**	**44.49**	**30.88**	**19.30**	**46.32**	**32.73**

Table 3: Retrieval performance (%) of Q2EI and baselines under different query rewriting models.

Q2D. Figure 8 shows the prompt used for Q2D. Following the original implementation, we use few-shot prompting to generate pseudo-passages. For each query, we randomly select three pairs of lay queries and target passages from the dataset as in-context demonstrations.

Q2EI. Figure 9 presents the instruction prompt template used in Q2EI. The prompt include three key elements (i) a persona to place the model in a domain-expert role, (ii) an entity-first constraint that requires inferring and normalizing the most likely core entity from the lay query, and (iii) a format constraint that enforces a standard professional-question form and asks the model to output only the rewritten question. Optionally, a small set of in-context examples can be appended for the few-shot setting.

C Robustness to the Generative Backbone

We examine the robustness of Q2EI to the choice of the generative backbone. We repeat the query rewriting step using Claude Sonnet 4.5 (API: claude-sonnet-4-5-20250929) and Qwen2.5-72B-Instruct.

Setup. We compare Q2EI with baselines. For cost efficiency, in the experiment with Claude Sonnet 4.5, we evaluate a random subset of 50 queries from MedQuAD. We exclude HyDE, which requires multiple hypothetical documents. This substantially increases token usage (Section 5.4). In the experiment with Qwen2.5-72B-Instruct, we exclude Q2EI (few-shot) and conduct the evaluation on 500 queries from the MedQuAD dataset. We treat this experiment as a small-scale robustness

Lay Query Generation Prompt

```
You are a simulation of an anxious patient with NO medical background. Your task is to generate a CASUAL, LAYMAN search query based **ONLY** on the medical condition mentioned in the provided Professional Question.

### INSTRUCTIONS:
### 1. IDENTIFY THE DISEASE: Look at the Professional Question and identify the specific disease or condition.
### 2. USE INTERNAL KNOWLEDGE: Use your own internal knowledge to imagine how a regular person would describe the symptoms, causes, or transmission **WITHOUT knowing the medical name**.
### 3. SIMULATE THE SCENARIO: Write a query as if you are experiencing the issue, but don't know what it is.
### 4. STRICTLY NO ENTITIES: Do NOT use the disease name. Use vague descriptions (e.g., 'weird virus', 'stomach bug').
### 5. LENGTH: Keep it under 20 words.

Query:
### Professional Question: {query}
### Rewritten Question :
```

Figure 5: Prompt for rewriting professional questions into lay queries.

Keyword Extraction Prompt

```
You are a keyword extractor. Extract the most important keywords from the query.

Query:
### Query: {query}
### Rewritten Question :
```

Figure 6: Prompt used for Keyword Extraction from a lay query.

HyDE Prompt

```
Please write a passage to answer the question.

Query:
### Query: {query}
### Rewritten Question :
```

Figure 7: Prompt used by HyDE to generate a hypothetical passage for retrieval.

Q2D Prompt

```
Write a passage that answers the given query.

Examples (optional):
### Query : {Example Query 1}
### Output : {Example Output 1}
### Query : {Example Query 2}
### Output : {Example Output 2}
### Query : {Example Query 3}
### Output : {Example Output 3}
### Query : {Example Query 4}
### Output : {Example Output 4}

Query:
### Query: {query}
### Rewritten Question :
```

Figure 8: Few-shot prompt used by Q2D to generate a pseudo-document.

Q2EI Prompt

```
You are a Medical Search Specialist optimizing queries for a Medical Database.
### 1: INFER THE SPECIFIC ENTITY
You MUST infer the most likely Specific Disease Name or Parasite Name based on the transmission method or unique symptoms described.
### 2: STANDARDIZE FORMAT
Use standard question formats. Such as 'What is [Disease Name]?' or 'How to diagnose [Disease Name]?' or 'How to prevent [Disease Name]?'.
### Note:
Output ONLY the rewritten professional question.
Examples (optional):
### Query : {Example Query 1}
### Output : {Example Output 1}
### Query : {Example Query 2}
### Output : {Example Output 2}
### Query : {Example Query 3}
### Output : {Example Output 3}
### Query : {Example Query 4}
### Output : {Example Output 4}

Query:
### Query : {query}
### Rewritten Question :
```

Figure 9: Prompt used by Q2EI to infer the core entity and rewrite a lay query into a professional, entity-centric query (MedQuAD example).

check.

Results. Results are reported in Table 3. In the experiment with Claude Sonnet 4.5, The overall trend is consistent with results obtained using GPT-based backbones. Q2EI consistently outperforms Lay Query and Q2D across retrievers. For instance, on the Multilingual E5 (mE5) retriever, Q2EI (zero-shot) achieves a Recall@10 of 72.92, compared to 37.50 for Q2D and 24.82 for the lay query. Interestingly, in this setting, the zero-shot performance of Q2EI occasionally surpasses the few-shot setting. One possible explanation is the strong inference capability of this backbone. This may reduce the reliance on in-context examples. This observation aligns with the objective of semantic condensation.

At the same time, the results with Qwen2.5-72B-Instruct show an overall decline in performance across query rewriting methods as the reasoning capability of the rewriting model weakens. However, Q2EI still outperforms the baselines on dense retrievers such as BGE-M3 and mE5. On BM25, it underperforms compared to Q2D. This is because Q2EI is sensitive to the quality of entity inference. When the generative backbone is weaker, the inferred entity can be less reliable, and the condensation process may remove surface lexical cues that are still useful for sparse matching. In contrast, Q2D expands query semantics by generating pseudo-documents to improve semantic coverage and lexical overlap, which is particularly beneficial for lexical matching in BM25. This suggests that, under weaker query rewriting models, expansion methods with stronger lexical coverage (e.g., Q2D) can be more robust for sparse retrievers.

D Additional Comparison with Query Expansion(CoT)

We further compare Q2EI with Query Expansion(CoT), a strong baseline based on Chain-of-Thought prompting. For cost efficiency, we evaluate a random subset of 50 queries from the MedQuAD dataset, and use GPT-5 as the rewriting model. As shown in Table 4, Q2EI consistently outperforms Query Expansion(CoT) across all evaluation metrics. Specifically, Q2EI improves Recall@1 from 26.00 to 40.00 and nDCG@10 from 37.56 to 58.30. This results further validate the effectiveness of our entity-centric semantic condensation strategy.

E Controlled Redundancy Prompts

Short Description

```
Identify the main medical entity in the
following query and write a 100-word
definition of it. Do not answer the query
itself.

Query:
### Query: {query}
### Rewritten Question :
```

Figure 10: Prompt for generating a short description to control redundancy.

Medium Description

```
Identify the main medical entity in the
following query and write a 100-word
definition of it. Do not answer the query
itself.

Query:
### Query: {query}
### Rewritten Question :
```

Figure 11: Prompt for generating a medium-length description to control redundancy.

Long Description

```
Identify the main medical entity in the
following query and write a 100-word
definition of it. Do not answer the query
itself.

Query:
### Query: {query}
### Rewritten Question :
```

Figure 12: Prompt for generating a long description to control redundancy.

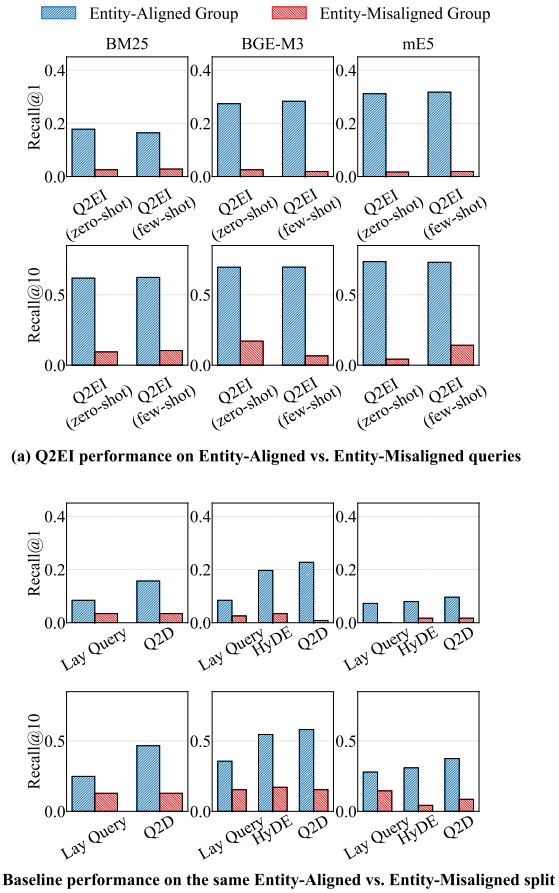
Figures 10, 11, and 12 show the prompts used in our motivation analysis (Section 5.1). We control the description length to vary the amount of redundant information.

F Additional Results for Attribution Analysis

We provide additional metrics to support the attribution analysis. Figure 13 reports Recall@1 and

Method	BM25			BGE-M3			mE5		
	R@1	R@10	N@10	R@1	R@10	N@10	R@1	R@10	N@10
Lay query	2.00	22.00	10.83	4.00	34.00	17.25	8.00	32.00	19.55
Q2D	10.00	50.00	27.51	18.00	50.00	34.40	14.00	46.00	29.22
HyDE	—	—	—	24.00	56.00	39.31	10.00	36.00	21.90
Query Expansion(CoT)	12.00	52.00	30.76	26.00	50.00	37.56	14.00	44.00	28.07
Q2EI (zero-shot)	**20.00**	**72.00**	**43.38**	**40.00**	**70.00**	**58.30**	**50.00**	**72.00**	**60.55**

Table 4: Retrieval performance (%) comparison between Q2EI and Query Expansion(CoT) on the MedQuAD dataset.



(a) Q2EI performance on Entity-Aligned vs. Entity-Misaligned queries

(b) Baseline performance on the same Entity-Aligned vs. Entity-Misaligned split

Figure 13: Retrieval performance (Recall@1 and Recall@10) under an entity-alignment split. (a) Performance comparison of Q2EI (zero-shot and few-shot) in Entity-Aligned vs. Entity-Misaligned Groups. (b) Performance of baseline methods (Lay Query, HyDE, Q2D) under the same grouping.

Recall@10 under different retrievers and rewriting methods. Results are shown separately for the Entity-Aligned Group and the Entity-Misaligned Group. The observed trends are consistent with the nDCG@10 results reported in Section 5.2 and Section 5.3. Specifically, Q2EI achieves substantially higher recall when entity inference is correct. When inference fails, all methods exhibit uniformly low performance. These results further support the conclusion that retrieval gains mainly stem from accurate entity inference. They also indicate that Q2EI does not amplify failure cases.

G Entity Candidate Ablation

Two Entities Prompt

```
You are a Medical Search Specialist optimizing
queries for a Medical Database.
### 1: INFER THE SPECIFIC ENTITY
You MUST infer the most likely **Specific
Disease Name** or **Parasite Name** based on
the transmission method or unique symptoms
described. If you can't make sure, you can
guess up to two possible diseases and use 'OR'
to connect them.
### 2: STANDARDIZE FORMAT
Use standard question formats. Such as 'What is
[Disease Name]?' or 'How to diagnose [Disease
Name]?' or 'How to prevent [Disease Name]?'.
### Note:
Output ONLY the rewritten professional
question.
Query:
### Query : {query}
### Rewritten Question :
```

Figure 14: Prompt for Q2EI with two potential entities candidates on MedQuAD.

We analyze the impact of inferring multiple potential entities. Figures 14, 15, and 16 illustrate the prompts used to infer varying numbers of entities. Table 5 reports the inference accuracy. Specifi-

Three Entities Prompt

```
You are a Medical Search Specialist optimizing
queries for a Medical Database.
### 1: INFER THE SPECIFIC ENTITY
You MUST infer the most likely **Specific
Disease Name** or **Parasite Name** based on
the transmission method or unique symptoms
described. If you can't make sure, you can
guess up to three possible diseases and use
'OR' to connect them.
### 2: STANDARDIZE FORMAT
Use standard question formats. Such as 'What is
[Disease Name]?' or 'How to diagnose [Disease
Name]?' or 'How to prevent [Disease Name]?'.
### Note:
Output ONLY the rewritten professional
question.
Query:
### Query : {query}
### Rewritten Question :
```

Figure 15: Prompt for Q2EI with three potential entities candidates on MedQuAD.

Four Entities Prompt

```
You are a Medical Search Specialist optimizing
queries for a Medical Database.
### 1: INFER THE SPECIFIC ENTITY
You MUST infer the most likely **Specific
Disease Name** or **Parasite Name** based on
the transmission method or unique symptoms
described. if you can't make sure, you can
guess up to four possible diseases and use
'OR' to connect them.
### 2: STANDARDIZE FORMAT
Use standard question formats. Such as 'What is
[Disease Name]?' or 'How to diagnose [Disease
Name]?' or 'How to prevent [Disease Name]?'.
### Note:
Output ONLY the rewritten professional
question.
Query:
### Query : {query}
### Rewritten Question :
```

Figure 16: Prompt for Q2EI with four potential entities candidates on MedQuAD.

Number of Candidates (N)	**Accuracy (%)**
$N = 1$	83.4
$N = 2$	87.4
$N = 3$	89.0
$N = 4$	90.2

Table 5: Entity inference accuracy (%) vs. number of candidates N for Q2EI (zero-shot).

cally, it shows the probability that the set of inferred entities includes one that holds an equivalence (same, hyponym, or hypernym) relationship with the target entity. The results show that increasing the number of inferred entities improves coverage. However, this also introduces noise. Future work will explore confidence scores to balance coverage gains against noise.