Bridging Underspecified Queries and Multimodal Retrieval: A Two-Stage Query Rewriting Approach

Szu-Ting Liu, Wen-Yu Cho, Hsin-Wei Wang, Berlin Chen
National Taiwan Normal University
{szutingliu, wenyu, hsinweiwang, berlin}@ntnu.edu.tw

Abstract

Retrieval-Augmented Generation (RAG) has proven effective for text-only question answering, yet expanding it to visually rich documents remains a challenge. Existing multimodal benchmarks, often derived from visual question answering (VQA) datasets, or large vision-language model (LVLM)-generated query-image pairs, which often contain underspecified questions that assume direct image access. To mitigate this issue, we propose a two-stage query rewriting framework that first generates OCR-based image descriptions and then reformulates queries into precise, retrieval-friendly forms under explicit constraints. Experiments show consistent improvements across dense, hybrid and multimodal retrieval paradigms, with the most pronounced gains in visual document retrieval—Hits@1 rises from 21.0% to 56.6% with VDocRetriever and further to 79.3% when OCR-based descriptions are incorporated. These results indicate that query rewriting, particularly when combined with multimodal fusion, provides a reliable and scalable solution to bridge underspecified queries and improve retrieval over visually rich documents.

Keywords: RAG, Query Rewriting, Visually Rich Documents, LVLMs, Information Retrieval, Multimodal Retrieval, Optical Character Recognition (OCR)

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm for building knowledge-intensive QA systems (Gao et al., 2023; Cheng et al., 2025), where large language models (LLM) are paired with retrieval modules to ensure a factual foundation and a broader domain coverage. In text-only settings, such as Wikipedia, news archives, or enterprise databases, RAG systems have been extensively studied, with dense, sparse, and hybrid retrievers achieving strong performance on well-established benchmarks (Lewis et al., 2020; Pan et al., 2022; Abdallah et al., 2025; Sawarkar et al., 2024).

Many real-world enterprise documents are visually rich, containing tables, charts, diagrams, and layout-dependent structures such as those found in product manuals, engineering drawings, or quality control reports. In these cases, relying solely on OCR text extraction leads to partial information loss, as complex visual semantics cannot be fully captured, limiting the effectiveness of traditional text-based RAG pipelines (Appalaraju et al., 2021; Xu et al., 2020). Developing effective RAG systems for such documents requires appropriate datasets and evaluation settings. However, most existing multimodal benchmarks originate from VQA tasks (Tanaka et al., 2025; Wang et al., 2025) are automatically constructed by prompting LVLMs to generate multiple queries for each image, which are then aggregated to form large-scale query-image datasets. These datasets often contain underspecified queries (e.g. 'What does this figure show') that presuppose direct image access; In retrieval settings where only textualized or embedding-based representations are available, such queries fail to identify the correct document reliably, leading to poor retrieval performance.

This limitation motivates our study. We propose a two-stage query rewriting framework that leverages OCR-informed context to reformulate underspecified queries in visually rich RAG settings. Our approach enriches query semantics and produces retrieval-friendly reformulations that better align with multimodal document representations. Extensive experiments demonstrate consistent gains across retrieval paradigms, particularly in multimodal

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 63–70
November 20-22, 2025 ©2025 Association for Computational Linguistics

settings. Our main contributions are as follows:

- We formalize the problem of underspecified queries in multimodal RAG systems.
- We propose a two-stage query rewriting framework that uses OCR-informed image descriptions and prompt constraints to produce retrieval-friendly queries.
- Our evaluations show that query rewriting significantly improves retrieval performance on visually rich documents.

2 Related Work

2.1 Optical Character Recognition

PaddleOCR PP-OCRv5 [1] (Cui et al., 2025) is an open-source multilingual OCR system supporting Simplified Chinese, Traditional Chinese, Chinese Pinyin, English, Japanese, and over 80 additional languages. It follows a three-stage pipeline of text detection, direction classification, and text recognition. Compared with PP-OCRv4, PP-OCRv5 reports a 13-percentage-point improvement in end-to-end benchmark accuracy and includes enhancements for challenging cases such as handwritten text, vertical text, and complex document layouts. Other widely used open-source OCR systems include docTR [2] and EasyOCR [3]. PP-OCRv5's open-source availability and comprehensive documentation make it a practical choice for research and production use.

2.2 Multimodal Document Retrieval

Retrieval-augmented generation (RAG) retrieves external knowledge to enhance large language models (Lewis et al., 2020), but most prior work assumes text-only corpora. Recent visual RAG studies leverage LVLMs to encode document images directly (Tanaka et al., 2025), enabling retrieval over visually rich documents. However, existing datasets such as ViDoRe (Wang et al., 2025) cover limited document types and often contain questions that do not truly require retrieval, and previous approaches typically lack dedicated training to adapt LVLMs for retrieval tasks.

VDocRAG (Tanaka et al., 2025) addresses these gaps with a dual-encoder retriever, where query tokens and document image features (processed by image encoder + projector) are fed into the same LVLM block to produce embeddings for similarity search. Its generator then uses the top-k retrieved images to produce answers. The model is built on Phi-3-Vision-128K-Instruct (4.2B parameters, image encoder + connector + projector + Phi-3 Mini LLM, 128K context length) and pre-trained with retrieval- and generation-oriented objectives (RCR, RCG) to align visual and textual features. OpenDocVQA [4], the accompanying dataset, provides open-domain and multi-hop questions, forming a comprehensive benchmark for visually rich document understanding.

2.3 Query Rewriting in Information Retrieval

Query rewriting is a common technique in information retrieval for reformulating user queries into semantically richer or more precise forms to improve retrieval performance. Existing approaches include rule-based methods, neural sequence-to-sequence models (Yu et al., 2020; Ma et al., 2023), and reinforcement learning strategies that optimize retrieval metrics (Ma et al., 2023). Recent work such as the Rewrite–Retrieve–Read framework demonstrates that rewriting can substantially improve dense retrievers by bridging the semantic gap between user queries and relevant documents (Ye et al., 2023; Kostric and Balog, 2024; Mo et al., 2023). However, most prior research focuses on text-only corpora, leaving open challenges for visually rich documents where key information may be embedded in layouts, figures, and tables.

3 Materials and Methods

3.1 Problem Definition

Let D denote a collection of image-centric documents (e.g., charts, tables, engineering drawings). We represent it as:

$$D = \{(Q_i, I_i)\}_{i=1}^{N}, \quad Q_i = \{q_{i1}, q_{i2}, \ldots, q_{ik}\},$$

where each image I_i may correspond to multiple associated queries, which are often ambiguous and underspecified. Our objective is

[1] https://www.paddleocr.ai/latest/en/version3.x/algorithm/PP-OCRv5/PP-OCRv5.html

[2] https://github.com/mindee/doctr

[3] https://github.com/JaidedAI/EasyOCR

[4] https://huggingface.co/datasets/NTT-hil-insight/OpenDocVQA-Corpus

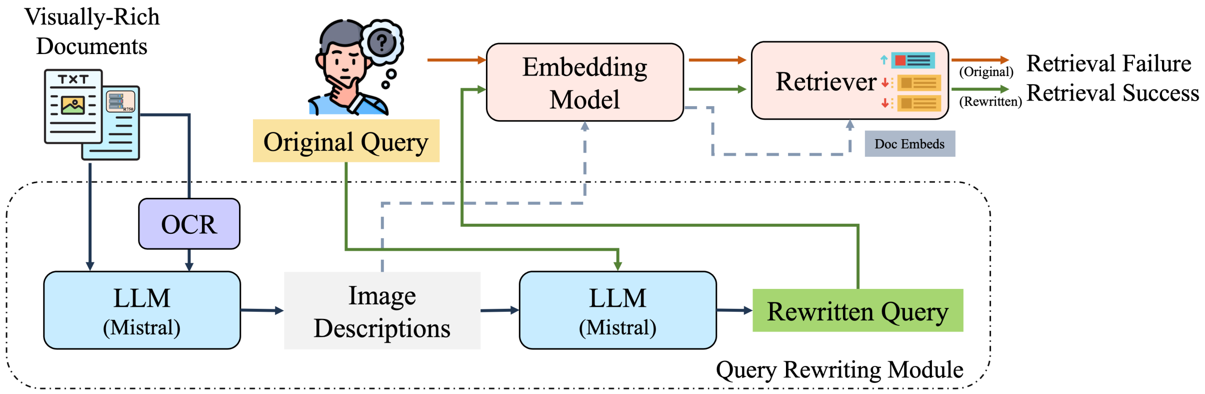

Figure 1: *Overview of the two-stage query rewriting framework.* Given an ambiguous query and a visually rich document, OCR text is extracted and summarized by a language model (Mistral) into an image description. The description, together with the original query, is used to generate a rewritten query that clarifies entities and avoids answer leakage. Both the original and rewritten queries are compared in the retriever to evaluate improvements in retrieval success.

to rewrite each query q_{ij} into a semantically complete and retrieval-friendly form $\tilde{q}_{ij}$.

$$\tilde{q}_{ij} = f_{\text{rewrite}}(q_{ij}) \qquad (1)$$

The rewritten query $\tilde{q}_{ij}$, generated through the two-stage process (Section 3.2), is constrained by a predefined system prompt (Section 3.3.1). Its objective is to retrieve the corresponding document more accurately.

3.2 Two-Stage Query Rewriting

The proposed two-stage framework (Figure 1) comprises image description generation and constrained query reformulation, detailed in the following subsections.

3.2.1 Image Description Generation

Each document image I_i is first processed by an OCR engine (PP-OCRv5; (Cui et al., 2025)) to extract the raw textual content $t_i = \text{OCR}(I_i)$. Since OCR outputs are often fragmented or incomplete (e.g. isolated labels or numbers), we employ Mistral-Small 3.2 (24B)[5] as the description generator f_{desc} to produce a context rich description d_i conditioned on both t_i and I_i:

$$d_i = f_{\text{desc}}(t_i, I_i) \qquad (2)$$

The generated description supplements missing or implicit OCR details, providing essential context for the subsequent rewriting stage.

The choice of Mistral-Small 3.2 (24B) was validated through comparisons with lighter multimodal models (LLaVA-7B and Qwen2.5-VL-7B) in the dense retrieval configuration. Although the smaller models achieved moderate accuracy (68 Hits@1) with shorter and less coherent descriptions, the 24B variant generated richer and layout-aware outputs, yielding +8-9 higher Hits@1 and a favorable cost-performance balance.

3.2.2 Controlled Query Rewriting

In the second stage, the original query q_{ij} is rewritten into $\tilde{q}_{ij}$ with the help of the image description d_i. The concatenated pair (q_{ij}, d_i) s fed into an LLM-based rewriting model f_{rewrite} (Mistral-Small 3.2 (24B)), guided by a structured prompt P and few-shot exemplars ξ (Section 2.3):

$$\tilde{q}_{ij} = f_{\text{rewrite}}(q_{ij}, d_i \mid P, \xi) \qquad (3)$$

This design allows the model to contextualize visual information via d_i and generate retrieval-friendly reformulations.

3.3 Prompt and Constraint Design

3.3.1 System Prompt

We design the rewriting prompt with explicit instructions that serve as hard constraints to ensure retrieval-oriented outputs. Specifically, the prompt requires that the rewritten query adhere to the following rules:

1. **Preserve interrogative form**: retain the question structure (e.g., "what," "how

[5] https://ollama.com/library/mistral-small3.2:24b

	# Queries	# Docs	Representative Visual Elements
Sales	135	25	Workflow and configuration diagrams; market analysis charts; wiring schematics; product dimension and application illustrations
Manufacturing	35	6	Process flow diagrams; Gantt charts; dimensional drawings; production statistics plots
Quality Control	52	9	Pareto and pie charts; Gantt charts; statistical performance plots
Technical	95	19	System layouts; architecture diagrams; measurement charts; circuit schematics
Others	40	7	Organizational and process flow diagrams
Total	357	66	—

Table 1: Domain-level statistics of the proprietary dataset containing 66 visually rich document images and 357 queries across five enterprise domains, each characterized by distinct visual elements common to industrial documentation.

many,” “why”) when the original query is interrogative.

2. **Avoid answer leakage**: exclude factual answers or numeric values appearing in the image text.

3. **Disambiguate references**: replace vague terms (e.g., “this chart,” “the server”) with concrete entities from d_i.

4. **Maintain source language**: keep the rewritten query in the same language as the input.

By enumerating these constraints in the system prompt, the model adheres to the intended query style and retrieval objectives.

3.3.2 Few-Shot Exemplars

To further guide model behavior, the prompt includes a few demonstration pairs of original and rewritten queries. Positive exemplars show effective reformulations where ambiguous queries are clarified with explicit entities or technical terms without leaking answers, while negative exemplars illustrate undesirable cases such as declarative rewrites, answer exposure, or language alteration. Together with the system prompt constraints (Section 3.3.1), these exemplars provide complementary supervision that steers f_{rewrite} toward generating well-formed, retrieval-oriented queries.

3.4 Post-hoc Validation

After rewriting, a lightweight validation step verifies compliance with the constraints in Section 3.3.1. This step ensures that each query preserves interrogative form, retains the original language, and avoids revealing factual answers or numeric values. Queries failing validation are replaced with the original input and logged with a status code, serving as a safeguard for overall quality and consistency.

Original Query	Rewritten Query
What kind of coating is applied to the machine surface?	What color of heat-resistant paint is applied on the surface of the ZX-200 industrial machine?
What is the efficiency improvement shown in the chart?	In the Q3 operations report, what is the percentage of efficiency improvement related to production output and cost reduction?
What is the memory specification of this server?	What is the memory configuration of the NovaEdge R720 server used in enterprise datacenter deployments?

Table 2: Query rewriting examples illustrating how ambiguous user questions are refined into precise, retrieval-oriented formulations.

3.5 Dataset

We evaluate the proposed method on a proprietary dataset provided by an industry partner, comprising 66 visually rich document images across five enterprise domains—sales, manufacturing, quality control, technical, and others. Each document contains layout-dependent visual structures such as charts and diagrams. A LVLM (Qwen3-VL-235B) was prompted to generate multiple natural-language queries per image, yielding 357 query-image pairs that simulate realistic but often underspecified informa-

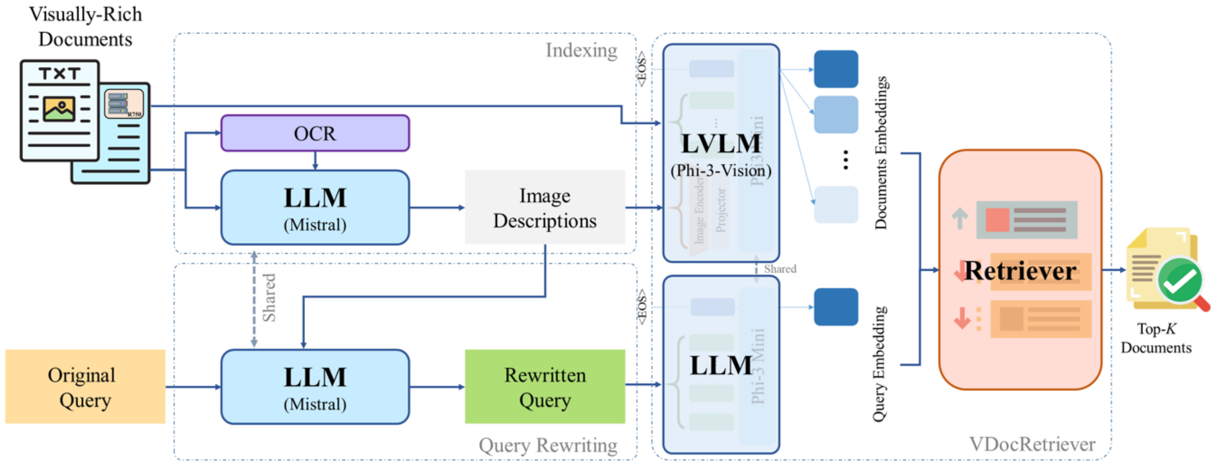

Figure 2: *Integration of the query rewriting module with VDocRetriever.* Each document is processed with OCR and a language model (Mistral) to generate enriched descriptions, which are combined with original queries for rewriting. Document and query embeddings are encoded by a vision—language model (Phi-3-Vision) to support multimodal retrieval. The red arrow marks VDocRetriever†, a variant that augments the document encoder with OCR-based image descriptions as additional textual context.

tion needs (Table 2). Each query q_{ij} is paired with its originating image I_i as the sole relevant item for retrieval.

To quantify query ambiguity, we manually classified all queries into three levels—clear (33.3%), partially underspecified (35.3%), and severely underspecified (31.4%)—based on the contextual information required for accurate retrieval. Although the dataset cannot be released due to confidentiality, detailed statistics (Table 1) and experimental results (Table 3) illustrate its diversity, the prevalence of ambiguous queries, and the effectiveness of the proposed framework.

4 Experiment Setup

4.1 Evaluation of Query Rewriting Across Retrieval Methods

We evaluate the proposed framework across three representative retrieval paradigms—dense, hybrid, and visual document retrieval (Figure 2)—covering neural, neural-lexical, and multimodal approaches. In each setting, rewritten queries replace the originals under identical conditions to isolate the effect of rewriting. Details of each retrieval configuration are provided in the following sections.

4.1.1 Dense Retrieval

For dense retrieval, we adopt BGE-M3 (Chen et al., 2024), a multilingual embedding model trained with contrastive objectives for retrieval tasks. Both queries and OCR-derived document descriptions are encoded into the same semantic space, and cosine similarity is used to rank document candidates. This text-only setup provides a strong baseline for evaluating whether query rewriting enhances semantic alignment between queries and OCR-based document representations.

4.1.2 Hybrid Retrieval

To leverage both semantic and lexical signals, we adopt a hybrid retrieval strategy combining BGE-M3 (Chen et al., 2024) and BM25 (Robertson and Zaragoza, 2009). Each rewritten query is simultaneously encoded by BGE-M3 for dense similarity matching and submitted to a BM25 index built from OCR-derived document text. BM25 first retrieves the top-k candidates; then both BM25 and cosine similarity scores are normalized to [0,1] and linearly combined (0.6 × BM25 + 0.4 × BGE-M3), as tuned on validation data. This design prioritizes exact lexical matches while allowing semantic reranking, enabling controlled analysis of how rewriting affects both retrieval signals.

4.1.3 Visual Document Retrieval

We further evaluate VDocRetriever (Tanaka et al., 2025), a state-of-the-art system for visually rich document retrieval. Unlike dense or hybrid retrievers that rely solely on textual representations, VDocRetriever jointly encodes multimodal signals (layout, visual appearance, and OCR text) making it an ideal baseline for testing the robustness of query rewriting under multimodal retrieval.

Two configurations are considered: the original VDocRetriever, which jointly encodes queries and document images, and VDocRetriever†, which augments document embeddings with OCR-based descriptions as additional textual context. The latter allows us to examine whether explicit textual anchors further enhance cross-modal alignment when combined with query rewriting.

4.2 Evaluation Metric

Retrieval effectiveness is measured using the Hits@k metric, reported at $k = 1, 5, 10$. A query is counted as successful if its relevant document appears within the top-k retrieved results. Formally, for a set of queries $\{q_i\}_{i=1}^{N}$, Hits@k is defined as:

$$\text{Hits@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{rank}(d_i^*|q_i) \leq k] \quad (4)$$

where d_i^* denotes the ground-truth document for query q_i, and $\text{rank}(d_i^*|q_i)$ denotes the rank position of d_i^* returned by the retrieval system. $\mathbf{1}[\text{rank}(d_i^*|q_i) \leq k]$ is an indicator function that equals 1 if the condition is true (i.e., if the relevant document d_i^* for query q_i is ranked within the top-k results) and 0 otherwise.

Since each query in our dataset has a single relevant document, Hits@k directly reflects the ability of each retrieval configuration to surface the correct document near the top of the ranked list. Higher values (especially for small k) indicate better retrieval effectiveness.

5 Results & Discussion

Rewriting Models	Hits@1	Hits@5	Hits@10
Qwen 3 (4B)	56.3	75.9	79.6
Qwen 3 (14B)	57.4	76.6	79.6
Llama 3 (8B)	56.0	75.4	78.1
Mistral-Nemo (12B)	74.5	82.9	84.0
Mistral-Small 3.2 (24B)	**76.8**	**82.9**	**84.6**

Table 4: Performance comparison of different query rewriting models evaluated under the dense retrieval configuration (BGE-M3).

Across all retrieval configurations, query rewriting consistently improves retrieval effectiveness (Table 3). For the dense retriever (BGE-M3), Hits@1 increases from 57.4% to 76.8% (+33.8%), showing stronger semantic alignment between rewritten queries and OCR-based document embeddings. The hybrid retriever (BGE-M3 + BM25) exhibits a similar pattern (Hits@1 + 37.8 %), suggesting that rewriting introduces lexical cues that complement dense representations.

The most pronounced gains occur in multimodal retrieval. The baseline Hits@1 of VDocRetriever (21.0%) is considerably lower than that of dense or hybrid retrievers, reflecting the difficulty of aligning vague queries with image-based embeddings. Rewritten queries introduce explicit anchors—such as entity names, field labels, and technical terms—that facilitate cross-modal alignment, raising Hits@1 to 56.6% (+169.5%). With additional OCR-based image descriptions (VDocRetriever†), performance further improves to 79.3% Hits@1 and 97.8% Hits@10, approaching near-perfect retrieval. These results highlight the value of multimodal fusion, where textual anchors extracted from images mitigate ambiguity in visual representations and strengthen query—document alignment.

Beyond retrieval paradigms, we also analyzed the influence of the rewriting backbone (Table 4). Model capacity correlates with rewriting precision: smaller models such as Qwen 3 (4B/14B) and Llama 3 (8B) produced syntactically correct but semantically shallow rewrites, while Mistral-Nemo (12B) and Mistral-Small 3.2 (24B) generated more contextually grounded reformulations, achieving 74.5 and 76.8 Hits@1, respectively. The 24B model slightly outperformed the 12B variant while maintaining acceptable inference latency, making Mistral-Nemo (12B) a practical choice for cost-sensitive deployments, whereas Mistral-Small 3.2 (24B) remains preferable for high-precision retrieval.

Taken together, these findings reveal several key insights. First, query rewriting benefits both dense and hybrid retrieval, but has the greatest impact in multimodal settings. Second, the disproportionate gains observed for VDocRetriever highlight that query rewriting is most critical when retrieval relies heavily on visual or layout-based representations. Finally, the strong performance of VDocRe-

Target	Retrieval	Original Queries			Rewritten Queries		
Document	Method	Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10
d_i	BGE-M3	57.4	76.8	79.6	76.8	82.9	84.6
d_i	BGE-M3+BM25	55.5	77.3	80.1	76.5	83.2	87.1
I_i	VDocRetriever	21.0	51.5	68.6	56.6	88.5	94.1
$I_i + d_i$	VDocRetriever[†]	29.4	52.4	64.2	79.3	93.8	97.8

Table 3: Retrieval performance with and without ablation study on the effect of query rewriting across different retrieval methods. Retrieval effectiveness is reported using Hits@k (%). [†] indicates the variant of VDocRetriever that incorporates the image description as additional context to improve retrieval precision. The column ”Target Document” specifies the representation used as the retrieval target, such as document image embeddings I_i or OCR-based descriptions d_i.

triever[†] shows that combining rewriting with textualized visual context offers a powerful strategy for visually rich document retrieval. Overall, the results position query rewriting as a robust and versatile technique, capable of enhancing retrieval effectiveness across both text-centric and multimodal paradigms.

6 Conclusion

This paper presents a two-stage query rewriting framework for addressing underspecified queries in RAG systems over visually rich documents. By leveraging OCR-informed image descriptions and applying constrained reformulation, the framework produces retrieval-friendly queries that reduce ambiguity and improve alignment with document content. Experimental results demonstrate that query rewriting consistently enhances retrieval effectiveness across dense, hybrid, and visual document paradigms, with particularly strong benefits in visual document settings. Overall, the findings establish query rewriting as a robust and general strategy for RAG over visually rich documents, with promising potential for scaling to larger datasets and integration into end-to-end question answering pipelines.

References

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, Mohammed Ali, and Adam Jatowt. 2025. From retrieval to generation: Comparing different approaches. *arXiv preprint arXiv:2502.20245*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. *Proceedings of the IEEE International Conference on Computer Vision*, pages 973–983.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2318–2335.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. 2025. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, pages 166–169.

Ivica Kostric and Krisztian Balog. 2024. A surprisingly simple yet effective multi-query rewriting method for conversational passage retrieval. *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2271–2275.

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. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 5303–5315.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:4998–5012.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *arXiv preprint arXiv:2203.16714*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework. *Foundations and Trends in Information Retrieval*, 3:333–389.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *Proceedings of the International Conference on Multimedia Information Processing and Retrieval, MIPR*, pages 155–161.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. Vdocrag: Retrieval-augmented generation over visually-rich documents. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24827–24837.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 20:1192–1200.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1:1933–1936.