

KAP: MLLM-assisted OCR Text Enhancement for Hybrid Retrieval in Chinese Non-Narrative Documents

Hsin-Ling Hsu
National Chengchi University
Taipei, Taiwan
112306092@nccu.edu.tw

Jing-Di Lin
National Chengchi University
Taipei, Taiwan
111301029@nccu.edu.tw

Ping-Sheng Lin
National Chengchi University
Taipei, Taiwan
111307050@nccu.edu.tw

Jengnan Tzeng*
National Chengchi University
Taipei, Taiwan
glophy@g.nccu.edu.tw

Abstract

Hybrid Retrieval systems, combining Sparse and Dense Retrieval methods, struggle with Traditional Chinese non-narrative documents due to their complex formatting, rich vocabulary, and the insufficient understanding of Chinese synonyms by common embedding models. Previous approaches inadequately address the dual needs of these systems, focusing mainly on general text quality improvement rather than optimizing for retrieval. We propose Knowledge-Aware Preprocessing (KAP), a novel framework that transforms noisy OCR outputs into retrieval-optimized text. KAP adopts a two-stage approach: it first extracts text using OCR, then employs Multimodal Large Language Models to refine the output by integrating visual information from the original documents. This design reduces OCR noise, reconstructs structural elements, and formats the text to satisfy the distinct requirements of sparse and dense retrieval. Empirical results demonstrate that KAP consistently and significantly outperforms conventional preprocessing approaches. Our code is available at <https://github.com/JustinHsu1019/KAP>.

Keywords

Post-OCR, MLLM, RAG, Hybrid Retrieval, BM25, Dense Retrieval, Structural Reconstruction, Text Preprocessing, Prompt Engineering

1 Introduction

Retrieval-Augmented Generation (RAG) [5] has emerged as a key technology for knowledge-driven applications, enabling systems to retrieve relevant information from large-scale knowledge bases to enhance tasks such as question answering and decision support. However, when dealing with non-narrative documents (e.g., financial statements, contractual clauses, and tables embedded in PDFs), traditional RAG-based retrieval methods face significant challenges due to poor OCR quality, loss of document structure, and suboptimal text chunking.

A major issue arises from OCR processing errors, particularly in Traditional Chinese financial documents, where character misrecognition, formatting loss, and disrupted table structures degrade retrieval accuracy. Common OCR techniques struggle to preserve tabular relationships, causing critical numerical and textual data to be misaligned or fragmented. Additionally, existing chunking methods are not optimized for non-narrative text, often splitting semantically related content incorrectly, further reducing retrieval performance.

One of the most widely adopted retrieval methods today is Hybrid Retrieval [3, 7], which combines Sparse Retrieval (e.g., BM25) and Dense Retrieval to leverage both exact keyword matching and semantic search. Sparse Retrieval methods like BM25 [10] rank documents based on term frequency-inverse document frequency (TF-IDF) and document length normalization, making them efficient for queries where keyword overlap strongly indicates relevance. Meanwhile, Dense Retrieval [4] encodes queries and documents into a shared embedding space for semantic similarity search, capturing contextual relationships beyond exact matches. Nevertheless, the effectiveness of both methods depends heavily on the quality of the retrieved text. Dense Retrieval models perform poorly on noisy OCR output and struggle to capture meaning from fragmented text. Meanwhile, Sparse Retrieval (BM25) relies on exact term matching, making it sensitive to synonym variations and phrasing differences, particularly in Traditional Chinese, where word order plays a crucial role in retrieval performance, and the language’s vast vocabulary with abundant synonyms makes semantic matching even more challenging for retrieval systems.

While existing post-OCR processing methods attempt to address character recognition errors and text correction [8], they are fundamentally designed for general text quality improvement rather than optimizing for retrieval systems. These methods typically focus on linguistic correctness and readability, overlooking the specific requirements of retrieval mechanisms where term frequency, keyword distribution, and semantic coherence are critical for performance. Current correction approaches improve textual accuracy but fail to consider how the processed text will ultimately be consumed by retrieval systems. This disconnect creates a significant gap—the ultimate purpose of knowledge preprocessing is to facilitate retrieval, yet existing methods do not adapt their strategies to the specific requirements of different retrieval approaches.

To address these challenges, we propose Knowledge-Aware Preprocessing (KAP), a two-stage preprocessing framework that optimizes textual representations specifically for Hybrid Retrieval in non-narrative documents. Rather than applying generic text improvement techniques, KAP recognizes that knowledge preprocessing must be tailored to the downstream retrieval task. Our framework enhances input data quality by leveraging Multimodal Large Language Models (MLLMs) with LLM-driven post-OCR processing that is explicitly designed to improve retrieval performance.

KAP refines extracted text, corrects OCR errors, restores lost structures, and strategically restructures content to improve compatibility with both Sparse and Dense Retrieval components of Hybrid systems.

The main contributions of this study are as follows:

- **Proposing the KAP framework:** A two-stage MLLM-based preprocessing approach that corrects OCR errors, reconstructs table structures, optimizes text representation in non-narrative documents, and enhances compatibility with both Sparse and Dense Retrieval.
- **Empirical validation of KAP’s effectiveness:** Tests on the E.SUN Bank dataset demonstrate that our framework significantly improves retrieval accuracy and structure preservation in Traditional Chinese financial documents.
- **Designing specialized validation methodology:** Creating an LLM-driven approach for augmenting validation datasets specifically optimized for Hybrid Retrieval evaluation, generating diverse query variations that comprehensively test both sparse and dense retrieval components.

2 Related Work

Optical Character Recognition (OCR) technology has been widely applied to text extraction from scanned documents and PDF files, but still faces significant challenges when processing non-narrative documents such as financial statements, contracts, and tables. These challenges include character recognition errors, format loss, and structural misalignment, which hinder accurate retrieval of content. Existing post-OCR processing methods generally fall into three categories: manual correction (accurate but labor-intensive), isolated-word methods (dictionary-based approaches that struggle with context), and context-dependent methods (from feature-based techniques to sequence-to-sequence models) [8]. Recent advancements have introduced sequence-to-sequence [11] and Transformer-based models [6] that improve text correction, while Large Language Models (LLMs) further enhance OCR correction by incorporating textual and structural understanding [13], but they are not optimized for the preferences of retrieval systems where term frequency and semantic coherence are critical.

In Retrieval-Augmented Generation (RAG) and Hybrid Retrieval systems, text quality significantly impacts retrieval outcomes. Sparse Retrieval methods (e.g., BM25) are sensitive to spelling errors and formatting disruptions, while Dense Retrieval methods perform poorly with fragmented text lacking context—particularly evident when handling table data. Previous studies have proposed improvements such as Contextual Retrieval [1] and Structure-Aware Transformers for table verification [16], but these primarily focus on semantic modeling rather than text preprocessing optimization.

The KAP framework proposed in this study builds on MLLM-based post-OCR techniques and is specifically optimized for Hybrid Retrieval requirements. Unlike traditional OCR correction methods, KAP incorporates a prompt-based approach that ensures generated text formats are suitable for retrieval systems, leverages LLMs for contextual understanding and structural restoration of tables, and precisely controls text structure and keyword distribution to enhance both Sparse and Dense Retrieval performance.

3 Methods

This section presents the proposed KAP framework, a two-stage preprocessing pipeline designed to enhance the effectiveness of Hybrid Retrieval in non-narrative documents. The framework addresses challenges associated with processing non-narrative documents, particularly financial statements and contractual clauses, which often contain complex tabular structures and require precise semantic preservation.

KAP consists of two main stages: (1) OCR Processing, which extracts text from PDFs, and (2) MLLM Post-OCR Processing, which refines the extracted text for better compatibility with Sparse and Dense Retrieval. By leveraging Multimodal Large Language Models (MLLMs) and prompt engineering, KAP corrects OCR errors, restores table structures, and optimizes text format to enhance retrieval accuracy. Figure 1 illustrates the overall architecture of KAP.

3.1 OCR Processing

The first stage of KAP involves extracting textual content from PDFs using Optical Character Recognition (OCR).

While OCR facilitates text extraction, it also introduces common errors, particularly when handling scanned financial documents. Character misrecognition often occurs, with errors such as "0" being recognized as "O" or "1" as "l." Additionally, the extraction process results in the loss of table structures, causing the extracted text to lack the original tabular layout and disrupting data relationships.

To mitigate these issues, we introduce an LLM-driven post-processing stage that refines OCR output and ensures compatibility with retrieval systems.

3.2 MLLM Post-OCR Processing

The second stage of KAP employs Multimodal Large Language Models (MLLMs) to enhance OCR-extracted text quality. Through prompt engineering with a single prompt, KAP effectively corrects OCR errors, reconstructs table structures, and optimizes text formatting to improve retrieval accuracy.

As illustrated in Figure 2, the prompt template takes as input the OCR-extracted text along with the corresponding original image, guiding the MLLM to perform targeted improvements. The prompt was originally written in Traditional Chinese; its English translation is provided at the beginning of each subsection in this section.

This process consists of three key components: error correction, layout-aware format reconstruction, and retrieval-aware rewriting. This streamlined approach simplifies the entire workflow described in this section while maintaining high performance across all processing tasks.

3.2.1 Error Correction. As shown in Figure 2, the first component (red box, lines 5-6) of our prompt template:

1. OCR Error Correction
- Correct errors in OCR conversion (e.g., typos, omissions, inverted sentence order, etc.), making the text more fluent and grammatically correct.

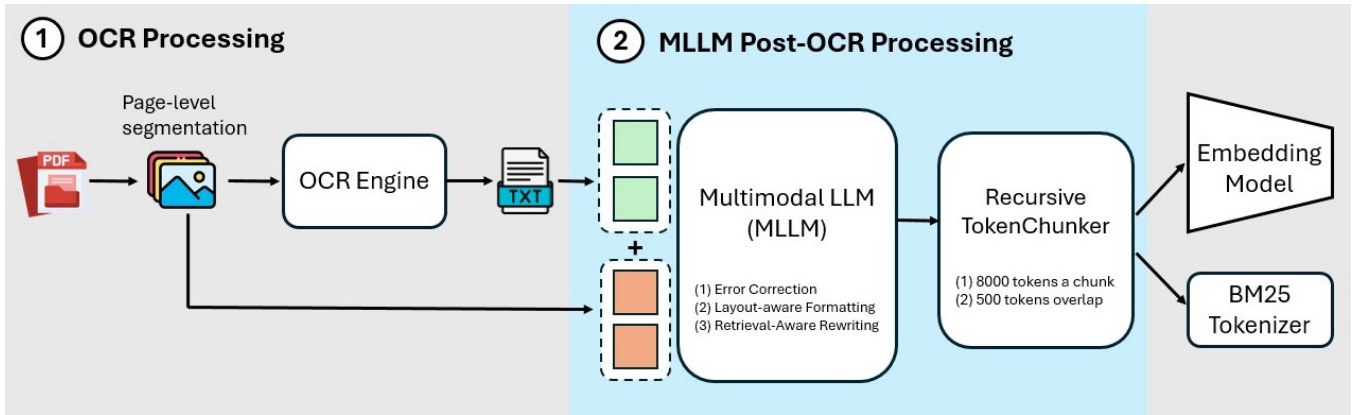


Figure 1: Overall architecture of the proposed KAP framework.

OCR outputs often contain spelling mistakes, incorrect numerical values, and misplaced punctuation. To address this, we apply an LLM-based correction mechanism that fixes common OCR recognition errors in Traditional Chinese and ensures grammatical correctness and readability while preserving semantic meaning.

3.2.2 *Layout-Aware Format Reconstruction.* As shown in Figure 2, the component highlighted in the blue box (lines 38) of our prompt template:

You can refer to the attached image to help you understand how this text is presented in the original PDF file (e.g., tables, narrative sentences), and what each text and number represents and where they appear in the original text.

Financial documents often rely on structured layouts, such as tables, section headers, and alignment, which standard OCR processing does not fully capture. To enhance text representation, KAP leverages MLLM’s multimodal capabilities within LLM-driven post-OCR processing. Specifically, MLLM’s vision capability is used to interpret the original PDF layout and refine extracted text accordingly, ensuring tables and formatting are preserved.

3.2.3 *Retrieval-Aware Rewriting.* KAP optimizes extracted text to improve compatibility with Hybrid Retrieval methods, ensuring better alignment with both Sparse and Dense retrieval strategies:

(1) **Dense Retrieval Optimization:** As shown in Figure 2, the component highlighted in the green box (lines 8-20) of our prompt template:

2. Retrieval-Friendly Rewriting (Suitable for Dense Retrieval)
 - For "tabular/non-pure narrative" content, convert into semantic retrieval-friendly narrative sentences.
 - For example, if the original text contains invoice, financial statement, or other tabular information, please rewrite into coherent descriptive sentences, ensuring the data and background information are complete:
 - Original tabular content:
 “

Date: 2022/03/03
 Company: XX Company
 Purchase item: XXX
 Amount: YY dollars
 “
 - After rewriting:
 "On March 3, 2022, XX Company purchased XXX, spending a total of YY dollars."
 - If the text content is messy, please optimize the paragraph structure to make it more suitable for semantic understanding and retrieval.

Reformulates structured tabular content into natural language descriptions to improve semantic embedding quality, ensuring better performance in embedding models:

• Original table:

日期 (Date)	公司 (Company)	金額 (Amount)
2025/03/03	XX	\$10,000

• Reformulated as:

2025年3月3日，XX 公司記錄了一筆 \$10,000 的交易。
 (On March 3, 2025, XX Corp recorded a transaction of \$10,000.)

(2) **Sparse Retrieval (BM25) Optimization:** As shown in Figure 2, the component highlighted in the orange box (lines 22-29) of our prompt template:

3. Retrieval-Friendly Rewriting (Suitable for BM25)
 - After "Retrieval-Friendly Rewriting (Suitable for Dense Retrieval)," rewrite synonyms in a retrieval-friendly way, naturally incorporating synonyms and near-synonyms while maintaining the original text’s meaning.
 - For example:

- Original text:
 > The system can analyze data, enhancing the enterprise’s decision-making ability.
 - After rewriting (ensuring the original keywords are preserved and expanding synonyms commonly used by the general public when asking questions):
 > The system is able to analyze data and related information, helping enterprises or companies make more accurate decisions and judgments, improving overall business strategy.
 - Avoid overusing synonyms, ensure the meaning remains unchanged, and do not affect the effect of vector search.

Expands key terms with synonyms to improve keyword matching while ensuring important keywords remain intact for query relevance.

- Original:

系統可以分析數據，以提升商業決策能力。
 (The system can analyze data to enhance business decision-making.)

- Rewritten (with synonym expansion while retaining key terms):

系統能夠分析數據與相關資訊，幫助企業或公司做出更精確的決策，並提升整體策略。
 (The system is capable of analyzing data and relevant information, helping businesses or companies make more accurate decisions and improve overall strategy.)

Expansion: Added synonyms for "數據 (data)" (資訊 for information) and elaborated on "決策 (decision)" (策略 for strategy).

3.3 Chunking Strategy

Traditional chunking methods such as Recursive Chunking [15] are not optimized for non-narrative documents, often splitting critical tabular content across multiple chunks. To mitigate this issue, KAP adopts a two-step chunking strategy:

- (1) Page-level segmentation: Documents are first segmented based on page boundaries to preserve contextual relationships.
- (2) Recursive Chunking: After MLLM Post-OCR processing, text is further segmented using Recursive Chunking with optimized chunk size (8,000 tokens with 500-token overlap) to prevent excessive fragmentation.

4 Experiments

4.1 Experimental Setup

4.1.1 *Datasets and Preprocessing.* This study uses a non-public dataset from E.SUN Bank, provided through the "AI CUP 2024 E.SUN

Artificial Intelligence Open Competition." The dataset contains three types of documents: FAQ, insurance terms, and financial reports. This study primarily focuses on financial reports since they contain a large number of tables, aligning with our research goal of handling non-narrative text. In contrast, the FAQ category is stored in JSON format with simple question-answer pairs, and the insurance terms mainly consist of plain text contracts, which differ from the complex formatted texts that this study targets. Therefore, these categories are not included in the main analysis.

Table 1 provides an overview of the document formats and contents in the E.SUN Bank dataset.

Table 1: Overview of E.SUN Bank Dataset Categories

Category	Description	Docs	File Type
FAQ	Frequently asked questions	617	JSON
Insurance	Policy terms sold by E.SUN Bank	643	PDF
Finance	Financial reports of listed companies	1035	PDF

The validation set, provided by E.SUN, contains multiple questions, each associated with a document that provides the answer. Each entry includes a question ID, query content, data source, and category. For each category (FAQ, Insurance, Finance), the validation set includes 50 questions. Table 2 summarizes the format of the validation set fields.

Table 2: Validation Set Question Format

Field	Description
qid	Question ID
query	Question content
source	Candidate Document ID (pid)
category	Data type (FAQ / Insurance / Finance)

To enhance the diversity of the validation set and rigorously evaluate the retrieval model’s robustness under various conditions, we expand the original set of 50 Q&A pairs to 500 questions through systematic data augmentation. Specifically, for each original question, we generate 9 distinct reformulations using the following augmentation strategies:

- (1) Replacing all keywords in the question with their synonyms (e.g., “modify” → “change”).
- (2) Substituting half of the keywords with synonyms while keeping the other half unchanged.
- (3) Extracting the core keywords from the question and presenting them as a space-separated list.
- (4) Extracting the core keywords and replacing all of them with synonyms, formatted as a space-separated list.
- (5) Extracting the core keywords and replacing only half of them with synonyms, maintaining the rest as originally phrased, in a space-separated format.
- (6) Transforming sentence structure by altering the word order, such as swapping the subject and verb while preserving the original meaning.

1 請擔任專業的繁體中文知識改寫專家，基於 OCR 轉換後的文本進行改寫，使其滿足混合檢索 (BM25 + Dense Retrieval) 的需求。

2

3 你的任務包括以下幾點：

4

5 1. OCR 錯誤校正

6 - 修正 OCR 轉換中的錯誤 (例如：錯字、漏字、語句順序倒置等)，使文本更加流暢且符合語法規則。

7

8 2. 檢索友善改寫 (適合 Dense Retrieval)

9 - 針對「表格型 / 非純敘述型」內容，轉換為易於語意檢索的敘述句。

10 - 例如，若原文本包含發票、財務報表等表格資訊，請改寫為連貫的描述句，確保數據與背景資訊完整：

11 - 原始表格內容：

12 ` ``

13 日期：2022/03/03

14 公司：XX 公司

15 購買項目：XXX

16 金額：YY 元

17 ` ``

18 - 改寫後：

19 「2022 年 3 月 3 日，XX 公司購買了 XXX，總共花費了 YY 元。」

20 - 若文本內容雜亂，請優化段落結構，使其更適合語意理解與檢索。

21

22 3. 檢索友善改寫 (適合 BM25)

23 - 在「檢索友善改寫 (適合 Dense Retrieval)」後，針對同義詞做檢索友善改寫，在不改變原文本語意的前提下，自然融入同義詞與近義詞。

24 - 例如：

25 - 原文：

26 > 該系統可分析數據，提升企業決策能力。

27 - 改寫後 (確保原關鍵詞保留，並擴展一般民眾提問時常用的同義詞)：

28 > 該系統能夠分析數據與相關資訊，幫助企業或公司更準確地做出決策與判斷，提升整體經營策略。

29 - 避免過度使用同義詞，確保語意不變，且不影响向量搜索的效果。

30

31 ---

32

33 請基於以下 OCR 後的雜亂文本進行優化改寫：

34 {ocr_text}

35

36 ---

37

38 你可以參考附件的圖片協助你理解這段文本在原 PDF 檔案上是以什麼格式 (e.g. 表格、敘述句) 呈現，各個文字、數字又分別代表什麼、呈現在原文哪些位置。

39

40 輸出格式：

41 1. 請輸出「完整」文本，確保文本上的所有內容皆有輸出

42 2. 請不要輸出任何與文本無關的其他字元

43 3. 請用繁體中文

Figure 2: The prompt template used for post-OCR processing with MLLM, originally written in Traditional Chinese. Its English translation is provided in the subsections of Section 3.2 MLLM Post-OCR Processing.

- (7) Condensing the question into its most concise form while retaining its essential meaning.
- (8) Applying structural variations to the condensed version, such as reordering the subject and verb.
- (9) Reformulating the question into a more informal and conversational style.

To generate these augmented question variants, we leverage a large language model (Claude-3.7-Sonnet [2]) by providing structured prompt that systematically apply the above transformation techniques.

4.1.2 Model Selection. For our OCR processing, we utilize Tesseract OCR [12], the most commonly used open-source engine with robust support for Traditional Chinese. This selection ensures reliable text extraction from the PDF documents in our dataset.

In the MLLM post-OCR processing stage, we employ Claude-3.7-Sonnet [2] as our multimodal language model. This model was selected for its strong multimodal text-image understanding capabilities, which are essential for correcting OCR errors, reconstructing table structures, and optimizing text formatting. The same model was also leveraged for generating our augmented validation questions, ensuring consistency throughout our experimental pipeline.

4.2 Metrics

To comprehensively evaluate retrieval performance, this study adopts the following standard metrics. Mean Reciprocal Rank (MRR) measures the rank of the first correct answer in the retrieval results, where a higher value indicates that correct answers tend to appear higher in the model’s ranked results. Average Precision@1 measures whether the top 1 retrieved result is the correct answer, with a higher value indicating that the model can accurately locate key information.

4.3 Evaluation Methodology

To evaluate our KAP preprocessing framework, we measure improvements across three retrieval strategies:

- (1) **Sparse Retrieval (BM25):** Ranks documents based on keyword matching using term frequency-inverse document frequency and document length normalization. In our implementation, we use the Jieba [14] tokenizer with a Traditional Chinese lexicon.
- (2) **Dense Retrieval:** Encodes queries and documents into a shared embedding space—using the text-embedding-3-large [9] model—allowing for semantic similarity search rather than exact keyword matching.
- (3) **Hybrid Retrieval:** Combines both sparse (BM25) and dense retrieval methods to leverage the strengths of keyword-based and semantic-based retrieval, capturing both exact matches and deeper semantic relevance.

4.3.1 Baseline System. We use **Tesseract OCR** as the baseline system, which directly applies OCR-extracted text for retrieval without additional preprocessing.

4.3.2 Ablation Study. To evaluate the impact of different components in the preprocessing framework, we perform ablation experiments on our KAP system. The test configurations include:

- (1) **KAP w/o Vision:** Removing MLLM’s ability to parse images (tables, layout), using only OCR text results to test whether MLLM improves format reconstruction.
- (2) **KAP w/o OCR Text:** Removing OCR-converted text and relying solely on MLLM to extract content from PDF images to test the influence of text input.
- (3) **KAP w/o Rewrite:** Removing the text rewriting module in the KAP framework, which aligns text formats for Dense and Sparse retrieval, to test its impact on retrieval accuracy.
- (4) **Full KAP Framework (Ours):** Providing complete OCR text, MLLM parsing, and text rewriting to ensure optimal output for Hybrid Retrieval.

4.4 Results

Each experiment was independently repeated three times. We report the mean and standard deviation across the three runs to ensure consistency and robustness. Tables 3, 4, and 5 summarize the retrieval performance under Sparse, Dense, and Hybrid settings, respectively. As shown, our proposed KAP framework consistently outperforms the baseline across all retrieval methods.

Table 3: Performance of Sparse Retrieval (Mean ± SD)

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	53.16±0.83	41.51±1.67
KAP w/o Vision	54.84±1.24	43.66±1.45
KAP w/o OCR Text	62.32±0.43	49.39±0.73
KAP w/o Rewrite	59.60±1.03	45.45±1.56
KAP (Ours)	63.64±0.09	51.16±0.21

Table 4: Performance of Dense Retrieval (Mean ± SD)

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	48.41±0.60	32.10±0.74
KAP w/o Vision	56.62±0.39	42.98±0.65
KAP w/o OCR Text	54.00±0.63	44.41±0.47
KAP w/o Rewrite	58.46±1.32	46.11±1.65
KAP (Ours)	65.16±1.51	53.65±2.24

Table 5: Performance of Hybrid Retrieval (Mean ± SD)

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	53.23±0.57	38.98±0.88
KAP w/o Vision	58.52±0.51	47.33±0.81
KAP w/o OCR Text	65.06±0.15	56.39±0.11
KAP w/o Rewrite	66.02±1.71	55.48±2.13
KAP (Ours)	69.46±0.61	59.73±1.10

4.4.1 Comparison with Baseline. For Sparse Retrieval (BM25), KAP improves MRR from $53.16\% \pm 0.83$ to $63.64\% \pm 0.09$ and Precision@1 from $41.51\% \pm 1.67$ to $51.16\% \pm 0.21$. This improvement is attributed to better text normalization and structural refinements, which enhance keyword matching effectiveness. In contrast, the baseline struggles with OCR errors and unstructured text, resulting in lower accuracy.

In Dense Retrieval, KAP outperforms the baseline by increasing MRR from $48.41\% \pm 0.60$ to $65.16\% \pm 1.51$ and Precision@1 from $32.10\% \pm 0.74$ to $53.65\% \pm 2.24$. The significant performance gain highlights the role of text refinement in improving vector-based retrieval, as cleaner input text leads to better semantic embeddings. The baseline’s noisy OCR output negatively impacts dense retrieval performance.

Hybrid Retrieval benefits the most from KAP, achieving the highest accuracy with an MRR improvement from $53.23\% \pm 0.57$ to $69.46\% \pm 0.61$ and Precision@1 from $38.98\% \pm 0.88$ to $59.73\% \pm 1.10$. This suggests that KAP effectively enhances both keyword-based and semantic retrieval components, leading to a more robust retrieval pipeline. The baseline’s poor text quality affects both BM25 and vector search components.

4.4.2 Ablation Study. We conduct an ablation study to evaluate the impact of different components in the KAP framework. Across all retrieval strategies, removing any component results in performance degradation, with the most significant drops consistently occurring when vision processing is removed.

For Sparse Retrieval, the largest decrease occurs with KAP w/o Vision, reducing Precision@1 from $51.16\% \pm 0.21$ to $43.66\% \pm 1.45$. This indicates that vision-based structural refinements significantly enhance keyword alignment and retrieval accuracy.

In Dense Retrieval, the most significant drop is also observed with KAP w/o Vision, where Precision@1 declines from $53.65\% \pm 2.24$ to $42.98\% \pm 0.65$. Without vision-assisted layout parsing, the quality of semantic representations deteriorates, negatively impacting dense retrieval performance.

For Hybrid Retrieval, KAP w/o Vision again results in the largest performance loss, with Precision@1 decreasing from $59.73\% \pm 1.10$ to $47.33\% \pm 0.81$. The absence of vision-driven layout reconstruction undermines the integration of sparse and dense retrieval components.

These results collectively highlight that vision-based enhancements are critical across all retrieval settings. The consistent impact of vision removal confirms the importance of our multimodal approach, demonstrating that leveraging MLLMs for structural understanding substantially strengthens the overall retrieval effectiveness of KAP.

5 Conclusion

In this study, we introduced KAP, a two-stage preprocessing framework that improves text quality for Hybrid Retrieval in Traditional Chinese non-narrative documents. By integrating MLLM-based post-OCR processing, KAP effectively corrects OCR errors, restores table structures, and enhances text representation for both Sparse and Dense Retrieval.

Experimental results demonstrate that KAP significantly improves retrieval accuracy across all retrieval paradigms. The ablation study further validates the contributions of each component,

confirming that OCR correction, layout-aware format reconstruction, and retrieval-aware rewriting collectively enhance retrieval performance.

Future work includes reducing the computational cost of the MLLM and extending KAP to other document domains. These improvements will further enhance the scalability and effectiveness of text preprocessing for retrieval tasks in real-world complex document settings.

Acknowledgments

This study was supported by E.SUN Bank, which provided the dataset from the "AI CUP 2024 E.SUN Artificial Intelligence Open Competition." We sincerely appreciate E.SUN Bank for its generous data support, which has been invaluable to this research.

References

- [1] Anthropic. 2024. Introducing Contextual Retrieval. <https://www.anthropic.com/news/contextual-retrieval>
- [2] Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>
- [3] Alec Berntson. 2023. Azure AI Search: Outperforming vector search with hybrid retrieval and reranking. <https://techcommunity.microsoft.com/blog/azure-ai-services-blog/azure-ai-search-outperforming-vector-search-with-hybrid-retrieval-and-reranking/3929167>
- [4] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*. 6769–6781.
- [5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [6] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 13094–13102.
- [7] Ji Ma, Ivan Korotkov, Keith B Hall, and Ryan T McDonald. 2020. Hybrid First-stage Retrieval Models for Biomedical Literature. In *CLEF (Working Notes)*.
- [8] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-OCR processing approaches. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–37.
- [9] OpenAI. 2024. text-embedding-3-large. <https://platform.openai.com/docs/models/text-embedding-3-large>
- [10] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp 109* (1995), 109.
- [11] Devendra Kumar Sahu and Mohak Sukhwani. 2015. Sequence to sequence learning for optical character recognition. *arXiv preprint arXiv:1511.04176* (2015).
- [12] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.
- [13] Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. BART for post-correction of OCR newspaper text. In *Proceedings of the seventh workshop on noisy user-generated text (W-NUT 2021)*. 284–290.
- [14] J Sun. 2012. Jieba chinese word segmentation tool. <https://github.com/fxsjy/jieba>
- [15] Weaviate. 2024. Recursive Chunking - Weaviate Knowledge Cards. <https://weaviate.io/learn/knowledgecards/recursive-chunking>
- [16] Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020. Table fact verification with structure-aware transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1624–1629.