멀티모달 LLM기반 표/차트 데이터 직렬화를 이용한 검색/생성 능력 개선

손 현 규[1)].강 창 묵[2)].김 한 울[*3)]

현대자동차[1)].한양대학교 전기공학과[2)].서울과학기술대학교 인공지능응용학과[3)]

Enhanced Retrieval and Generation Performance via Multimodal LLM-based Serialization of Tabular and Chart Data

Hyun Kyu Shon[1)] · Changmook Kang[2)] · Hanul Kim[*3)]

Hyundai Motor Company[1)], Hanyang University[2)], Seoul National University of Science and Technology[3)]

Key words : Multimodal LLM (멀티모달 대규모 언어모델), Visual Data Serialization (시각적 데이터 직렬화), Table & Chart Serialization (시각적 표 · 차트 직렬화), Retrieval-Augmented Generation (RAG) (검색 증강 생성), Document Preprocessing for Retrieval (검색을 위한 문서 전처리), Document Layout Analysis (문서 구조 분석)

Corresponding author, E-mail: hukim@seoultech.ac.kr

1. Introduction

Retrieval-Augmented Generation (RAG) represents pivotal advancement in question-answering systems that operate over large-scale document corpora. It has demonstrated considerable success in Natural Language Processing (NLP) tasks across diverse applications. However, despite these achievements, RAG systems exhibit substantial performance degradation when processing tables, charts, and other visually structured information.

Traditional processing pipelines, Optical Character Recognition (OCR) followed by straightforward text chunking strategies, proved inadequate for the challenge. While OCR can extract raw textual content from visual elements, it fails to preserve the critical structural and semantic relationships—row-column associations and hierarchical organizations—intrinsically encoded within the visual layout. Structural blindness not only compromises retrieval accuracy but also significantly increases the probability of generating hallucinated responses, thereby undermining the reliability of the entire system.

To address these fundamental limitations, we introduce a two-stage pipeline that harnesses the capabilities of Multimodal Large Language Models (MLLMs) to perform intelligent serialization of visual data. Our approach enables MLLMs to interpret the visual structure of tables and charts, transforming into descriptive, semantically enriched JSON representations. Serialization substantially enhances RAG performance by making structured visual information fully accessible to text-based retrieval mechanisms.

2. METHODOLOGY

2.1 System Architecture

Our methodology employs modular architecture that utilizes MLLMs as semantic arrangers for document processing, (Figure 1). The system incorporates an adaptive refinement iterative loop designed to verify and correct extracted contents. Architecture diverges from conventional approaches that depend on cascaded auxiliary models for OCR and document structure analysis. Instead, our pipeline directly processes both page screenshots and raw text through MLLM, thereby enabling comprehensive document understanding through simultaneous visual and textual interpretation—effectively allowing the model to "view the document structure while reading its content."

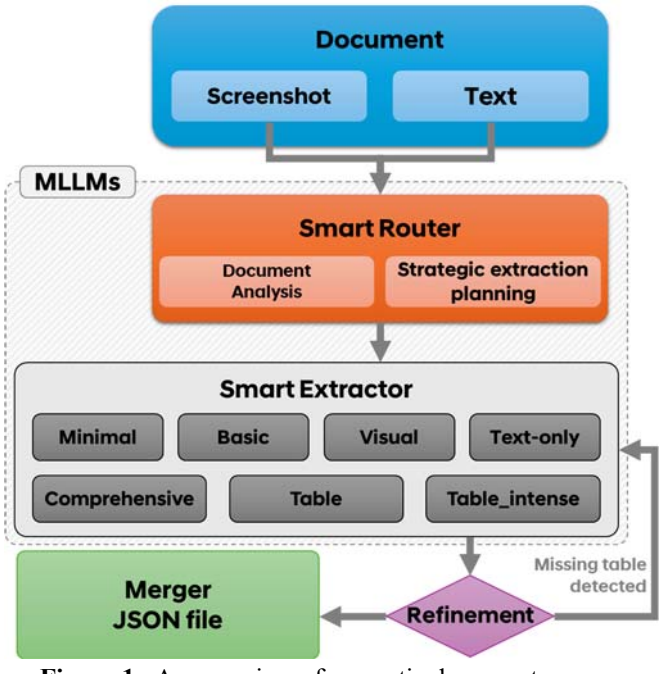

Figure 1 : An overview of semantic document arranger

2.2 Processing Pipeline

Processing workflow implements a multi-stage adaptive pipeline. Router module is initiated when raw screenshots and text are processed, performing rapid and comprehensive layout analysis (Figure 1). Router generates an initial coarse-grained document structure encompassing titles, summaries, and paragraphs, while simultaneously categorizing content sections into seven distinct classes (e.g., text-only, table, visual). To enhance classification accuracy, we employ exhaustive **few-shot prompting** methodology that ensure comprehensive pattern coverage within the prompt designed framework.

Subsequently, each categorized section undergoes processing by specialized extractors tailored to specific content types. Extractors utilize **template-based few-shot prompting** methodology to transform data into JSON objects optimized for each section category. Notably, sections identified as "table" undergo sophisticated serialization. These sections are explicitly transformed according to a structured schema incorporating "table_title," "headers," and "rows" fields, derived from the visually interpreted table structure.

2.3 Refinement and Validation

Refinement Analyzer module validates extraction accuracy through statistical verification (Figure 1). Analyzer implements multiple heuristic checks, to identify missing tabular structures, including analysis of : (1) Numeric character density (2) Rows exhibiting similar length patterns, and (3) Repeated spacing and separators. Upon detection of missing tables, the system triggers re-extraction by the appropriate extractor. Iterative refinement process culminates in the consolidation of extracted data from each page, which is subsequently merged with relevant metadata into the final JSON representations.

Notably, pipeline eliminates the computational overhead associated with the usage of dedicated Visual Language Models (VLMs) or standalone OCR engines. Instead, it leverages the visual-processing capabilities inherent in a MLLMs, thereby achieving both architectural simplicity and reduced operational complexity. Experimental evaluation demonstrates the clear superiority of our proposed method over conventional approaches.

3. EXPERIMENTAL EVALUATION

3.1 Experimental Setup

To evaluate the effectiveness of our proposed approach, we conducted experiments using two distinct corpora: the Alphabet Q1 2025 earnings report (English) and the Hyundai IONIQ 9 Owners' Manual (Korean). These datasets were selected to assess cross-lingual robustness and domain coverage. For the control cases, vector database was constructed from standard text extraction. The experimental condition employed a vector database populated with the serialized JSON representations generated through our proposed pipeline.

Models. We employed the following model configuration: (1) Router module: Claude Opus 4.1, (superior reasoning) (2) Extractor module: Gemini 2.5 Pro, (advanced visual processing)

Evaluation Protocol. We formulated twenty quantitative questions per document (e.g., "How much did Alphabet spend on research and development during Q1 2024?"). Both control and experimental systems utilized identical RAG architectures to eliminate confounding variables. Each question was queried against both vector stores, with comprehensive results presented in Table 1.

Remarks	Control (Baseline)	Experimental (Suggested pipeline)
Alphabet Q1 '25 earnings report		
- Retrieval Precision / %	75%	95%
- Factual Accuracy	1.3 / 2.0	1.8 / 2.0
IONIQ 9 Owner's Manual		
- Retrieval Precision / %	65%	90%
- Factual Accuracy	1.0 / 2.0	1.7 / 2.0

Table 1. Comparative evaluation between baseline and proposed pipeline across English and Korean corpora.

3.2 Results and Analysis

Experimental results demonstrate substantial performance improvements across all metrics. Baseline system exhibited challenges in context retrieval, achieving a retrieval precision of 70.0% with a factual accuracy score of 1.15 on a 2.0 scale. In comparison, our proposed system achieved a retrieval precision of 92.5% and factual accuracy of 1.75, representing improvements of 32.1% and 52.2% respectively. These findings indicate that the enriched metadata generated through our pipeline significantly enhances retrieval precision, while structured serialization of visual data substantially improves factual accuracy.

Qualitative analysis revealed pronounced differences in system behavior. The baseline system frequently generated incomplete responses or produced numerically inaccurate outputs. Conversely, the experimental system consistently generated precise, contextually grounded responses derived from the retrieved structured data. Notably, our approach demonstrated performance parity with, and in certain cases exceeded, several commercial LLM services.

4. CONCLUSION

Our findings establish that structure-preserving serialization represents a critical preprocessing component for enhancing RAG system reliability in data-intensive domains. The proposed two-stage pipeline successfully bridges the semantic gap between visual document elements and text-based retrieval through semantic transformation of tables and charts. Transformation into information-preserving JSON representations has shown measurable improvements in both indexing efficiency and question-answering accuracy within text-based RAG systems. Furthermore, achieved significant reduction in systematic complexity has been achieved by eliminating dependencies on VLM/OCR.

Future research directions encompass: (i) investigation of LLM selection criteria for router and extractor modules and their impact on extraction quality, and (ii) exploration of scalability toward automated knowledge graph construction from unstructured document collections.

Multimodal LLM-based serialization methodology presented significant advancement toward developing comprehensive and reliable document intelligence systems capable of handling heterogeneous information formats.