JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 6, pp. 65-76, June 2025
https://doi.org/10.9708/jksci.2025.30.06.065

Quantitative Assessment of OCR for Complex Documents on Retrieval-Augmented Generation Performance

Minchae Song*, Jaeyoung Park**

*Research Fellow, Finance Research Institution, Nonghyup Financial Group, Seoul, Korea
**Professor, Dept. of Digital Finance, Pukyong National University, Busan, Korea

[Abstract]

Retrieval-Augmented Generation (RAG) enhances the accuracy of generative AI services by allowing Large Language Models (LLMs) to reference external knowledge bases rather than relying solely on pre-trained knowledge. This study analyzes various types of financial document images to examine the impact of document image structure on RAG effectiveness. The results reveal that, although OCR achieves high recognition accuracy even with handwritten text, the overall performance of RAG remains suboptimal. This suggests that increased structural complexity in original document images hinders contextual understanding, which in turn degrades performance across the retrieval, chunking, and generation stages of the RAG pipeline. Therefore, assuming OCR text quality exceeds a certain threshold, structuring input data into a format that is more readily interpretable by machines through post-processing plays a more critical role in enhancing RAG performance.

▸ **Key words**: Generative Artificial Intelligence, Finance, OCR, RAG, Word Error Rate

[요 약]

RAG(Retrieval-Augmented Generation) 기술은 LLM(Large Language Model)이 사전 학습한 지식에만 의존하지 않고, 외부 지식 데이터베이스를 참조함으로써 생성형 AI(Artificial Intelligence) 서비스의 정확도를 향상시킨다. 본 연구는 다양한 유형의 금융 문서 이미지를 분석하여, 문서 이미지 구조의 복잡도가 RAG 성능에 미치는 영향을 살펴본다. 분석 결과, 문서 내 손글씨에 대한 OCR(Optical Character Recognition) 인식률은 높은 수준을 보였으나, RAG의 전반적인 성능은 기대에 미치지 못하는 것으로 나타났다. 이는 원본 문서 이미지의 구조가 복잡할수록 문맥 이해를 어렵게 하여, RAG 파이프라인의 분할, 검색, 생성 과정 전반에서 성능 저하를 초래할 수 있음을 시사한다. 따라서 OCR로 추출된 텍스트의 품질이 일정 수준 이상인 경우, 입력 데이터를 기계가 더 쉽게 해석할 수 있도록 구조화하는 후처리 과정이 RAG 성능 향상의 핵심 요인이 될 수 있다.

▸ **주제어**: 검색 증강 생성, 광학 문자 인식, 금융업, 단어 오류율, 생성형 AI

• First Author: Minchae Song, Corresponding Author: Jaeyoung Park
*Minchae Song (nicenara84@naver.com), Finance Research Institution, Nonghyup Financial Group
**Jaeyoung Park (jyp2503@pknu.ac.kr), Dept. of Digital Finance, Pukyong National University
• Received: 2025. 04. 24, Revised: 2025. 05. 25, Accepted: 2025. 06. 10.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

Chat GPT(Generative Pre-trained Transformer), Gemini와 같은 거대 언어모델(Large Language Model, LLM)이 자연어처리 분야에서 괄목할 성과를 내며 생성형 AI(Artificial Intelligence) 기술의 대중화를 촉발하였다. LLM은 방대한 양의 텍스트와 이미지 등의 데이터를 학습한 후 언어모델의 매개변수에 내재된 암시적 지식을 활용해 외부 정보에 접근하지 않아도 복잡한 자연어처리 작업을 수행하게 한다[1]. LLM에 기반한 생성형 AI는 특정 작업에서는 인간의 능력을 능가했으며, 제조, 금융, 법률, 의료 등 다양한 산업 분야의 생산성 향상을 가져오고 있다[2].

그러나 LLM의 기술 진보에 불구하고 정확한 정보가 아님에도 응답을 지어내거나, 최신 정보와 상충하는 답변을 생성하는 환각(Hallucination) 현상은 생성형 AI의 실무 도입 및 활용에 있어 큰 장애물로 작용하고 있다[3]. 이러한 문제는 생성형 AI 기술의 기반이 되는 LLM이 학습을 마친 정보를 토대로 답변을 생성하기 때문에 데이터가 희소해 학습이 어려운 특수한 도메인이나 학습되지 않은 새로운 정보는 답변하기 어렵기 때문이다[4].

현재 LLM이 가진 한계점인 환각 현상을 줄일 수 있는 대표적 방법으로 RAG(Retrieval Augmented Generation)가 있다. RAG 기술은 LLM이 사전에 학습한 지식에만 의존하지 않고, 최신 정보나 금융과 같은 특정 도메인에 대한 세부 정보를 보강하여 답변하게 함으로써 생성형 AI 서비스의 정확도와 신뢰도 향상을 가능하게 한다[5]. RAG 기술을 구성하는 핵심 요소 중 하나는, 기존 사전 학습에 사용된 데이터베이스 외에 추가된 외부 지식 데이터이다. 따라서 RAG 시스템이 접근하는 데이터베이스의 범위와 품질이 RAG의 성능을 크게 좌우한다.

그러나 최근 RAG 분야의 활발한 연구에도 불구하고, 스캔된 PDF나 문서 이미지에 대해 OCR(Optical Character Recognition) 기술로 외부 지식 데이터베이스를 사용한 경우 원본 이미지의 구조가 RAG 시스템의 성능에 미치는 영향의 연구는 찾아보기 어렵다[6]. 이는 OCR 기술이 RAG 시스템의 성능에 미치는 영향을 평가할 수 있는 다양한 벤치마크 데이터가 부재하기 때문이다. 그러나 RAG 기술은 RAG 시스템을 구성하는 각 요소의 개별 성능에 크게 의존하기 때문에, RAG의 입력 데이터에 해당하는 OCR의 결과물이 RAG 성능에 미치는 영향은 중요한 연구주제이다[7]. 또한, 산업계의 많은 데이터가 스캔된 PDF나 이미지 형태의 문서로 존재하는 경우가 많기 때문에 RAG의 실무 도입 및 활용에 있어 이는 매우 중요한 문제이다. 특히 최근 OCR 기술의 개선으로 텍스트의 인식률이 높아졌지만 OCR로 추출된 데이터를 외부 데이터베이스로 활용했을 때 RAG의 성능이 좋지 않아 LLM이 생성한 응답의 품질 저하로 이어진다면, 어떠한 요인이 이를 초래하는지 파악하는 것이 선행되어야 한다.

이러한 문제의식에서 본 연구는 다양한 유형의 금융문서 이미지를 대상으로 문서 이미지 구조의 복잡도가 RAG 성능에 미치는 영향을 분석했다. 구체적으로 OCR 기술이 높은 인식률을 보여도 원본 문서의 구조가 복잡할 경우 RAG의 성능이 저하될 수 있다고 보았다. 이 가정이 옳다면 OCR 문서를 외부 데이터베이스로 사용하는 RAG 시스템에서 문서의 인식률도 중요하지만 어느 정도의 품질이 보장된다면, 인식률의 개선보다 컴퓨터가 이해하기 쉬운 형태로 구조화하여 입력 데이터의 정보를 후처리하는 것이 RAG 성능 개선에 더 중요한 요인일 수 있다. 본 연구의 가정을 확인하기 위해 문서의 구조가 단순한 것부터 복잡한 유형까지 다양한 금융문서 이미지를 분석에 활용했다.

본 논문의 구성은 다음과 같다. II 장에서는 관련 연구를 살펴보고, III 장에서는 본 연구에서 사용한 OCR과 RAG를 구성하는 핵심 요소 및 그 과정을 기술하였다. IV 장에서는 분석 데이터 및 OCR과 RAG의 성능 평가 지표, 분석결과를 설명하였다. 마지막으로 V 장에서는 연구결과를 요약하며, 본 연구의 한계점 및 향후 연구 방향을 제시하였다.

II. Preliminaries

1. Optical Character Recognition

OCR이란 문서를 스캔하거나 촬영된 이미지에서 기계가 읽을 수 있는 형태의 디지털 정보로 변환하는 기법이다[8]. 따라서 OCR을 적용하면 이미지에 담긴 정보를 자동으로 추출하여 다양한 NLP(Natural Language Processing) 분석이 가능하다. NLP와 딥러닝 기술 발전으로 OCR 분야 역시 문서 이미지 인식률이 크게 향상되어 여러 산업 영역에서 OCR 기술이 활용되고 있다[9]. 그럼에도 복잡하고 다양한 레이아웃을 가진 문서 이미지에서 오류 없이 정확하게 정보를 추출하는 것은 여전히 어려운 과제이다.

선행연구에 따르면 OCR의 품질이 낮으면 추출된 정보의 오류와 불일치로 인해 품사 태깅, 텍스트 요약, 질의응답(Question Answering, QA) 등 OCR의 결과물을 입력 데이터로 사용하는 NLP의 다운스트림 태스크(Downstream task)의 성능 저하로 이어지는 것으로 나타

났다[10-11]. 특히 많은 NLP 태스크가 텍스트에 내포된 노이즈에 민감하므로 기존 연구들은 NLP 태스크의 성능 향상을 위한 OCR 기법의 고도화 및 텍스트의 품질 개선에 집중해 왔다[12].

이러한 OCR은 이미지 내 텍스트 검출(Text detection) 과정과 텍스트 인식(Text recognition) 과정으로 구성되며[13], 크게 세 가지 접근 방식으로 분류할 수 있다. 파이프라인(Pipeline) 기반의 접근 방식[14], 엔드투엔드(End-to-end) 모델[15], 비전-언어 모델(Vision-Language Model, VLM)을 활용한 방식이다[16].

파이프라인 기반 시스템은 OCR을 문서의 레이아웃 검출, 텍스트와 수식, 표 인식 등 문서의 대상을 여러 하위 작업으로 나눠 접근하여 대상의 정밀한 추출을 가능하게 한다. 그러나 이러한 접근 방식은 각 단계의 최적화(Local optimum)에 집중할 수 있으며, 하위 과정별 별도의 모델이 필요하다는 점에서 일반화가 어렵다는 한계가 있다[17].

반면, 엔드투엔드 모델은 문서 이미지를 입력으로 받아 인식 결과를 직접 출력함으로써 각 과정을 하나의 단계로 통합하여 동작한다. 이들은 텍스트 검출과 인식에 동일한 특징 추출기(Feature extractor)를 공유하며, 하나의 최적화 프레임워크 하에 학습된다. 그러나 이 접근 방식 역시 현실에서 생성되는 다양한 문서의 구조와 텍스트 형태를 추출하는 데 어려움이 있어 일반화 성능을 보장하기 어렵다[18].

최근 등장한 VLM은 이미지와 텍스트를 동시에 모두 처리할 수 있는 멀티-모달(Multi-modal) 기법이다[19]. 멀티-모달 모델은 텍스트, 이미지, 오디오, 비디오 등 다양한 유형의 데이터를 함께 고려하여 서로의 관계성을 학습하여 처리한다[20]. 최근 LLM이 멀티-모달 분야에 적용되면서 주목받고 있다[21]. 특히 VLM은 텍스트와 이미지 정보로부터 문맥적 의미를 동시에 추출하는 것이 가능하여 OCR 분야에도 뛰어난 성능을 보이고 있다[22]. 최근 연구에 따르면, OCR 벤치마크 데이터에 대해 전통적인 OCR 기법보다 VLM에 기반한 접근 방식이 더 좋은 성능을 보이며 문서 이미지 연구로 확장되고 있다[23].

2. Retrieval Augmented Generation

RAG가 LLM의 생성 기능과 외부 지식 데이터베이스의 확장성을 결합하여 생성형 AI의 정확도와 신뢰도를 개선하는 효과적인 방법으로 주목받고 있지만 몇 가지 한계점이 존재한다. 대표적으로 RAG 성능이 RAG 시스템을 구성하는 하위요소의 개별 성능에 좌우되고, 특히 외부 데이터베이스의 품질에 크게 의존한다는 점이다[24]. 예를 들어, 데이터 분할 시 청크(Chunk) 크기가 적절하게 설정되지 않으면 벡터 데이터베이스(Vector database)에서 연관 검색이 제대로 작동하지 않을 수 있다. 특히 입력 데이터의 구조가 복잡할 경우 데이터 분할 과정에 의미론적 유사성을 파악하기 어려운 형태로 청크가 생성되므로 분할 이후 단계인 임베딩(Embedding)과 검색을 수행하는 RAG 시스템 내 개별 구성요소의 성능이 저하된다. 일례로 텍스트가 등장하는 문맥 정보에 기반해 텍스트의 의미를 숫자로 변환하는 임베딩 모델의 결과물도 앞 단계의 데이터 분할이 의미 파악이 어려운 형태로 청크가 생성되면, 성능이 좋은 임베딩 모델을 사용하더라도 의미론적 유사성 포착이 어려워진다. 검색 단계에서 연관 정보가 제대로 검색되지 않으면, 관련 없는 정보가 LLM에 전달되어 부정확한 답변이 생성된다. 결과적으로 LLM의 생성 성능이 떨어지면, 전달된 정보의 품질에 상관없이 잘못된 답변 생성으로 이어진다. RAG 시스템의 각 단계에 대한 설명은 III 장 2 절에 상세히 서술하였다.

현재 대부분의 LLM은 오류가 없는, 정제된 텍스트 데이터를 활용해 학습되었기 때문에 OCR 적용 후 문자 인식 오류가 포함된 문서를 입력 데이터로 사용할 경우 RAG 시스템의 하위 구성요소들이 최적의 성능을 내지 못할 수 있다. 또한, 서론에 언급하였듯이 문서 이미지의 구조가 OCR의 품질에 영향을 미칠 수 있으며[13], OCR 기술이 높은 인식률을 보여도 원본 문서 구조가 복잡할 경우 RAG 시스템의 구조상 성능 저하로 이어질 수 있다. 최근 VLM과 RAG 분야에 활발한 연구가 진행되고 있지만, 문서 이미지 구조의 복잡도에 따라 이것이 RAG 성능에 미치는 영향을 분석한 연구는 찾아보기 어렵다.

III. Methodology

1. Document OCR-based on Vision-Language Model

VLM은 일반적으로 이미지 인코더(Image encoder)와 텍스트 인코더(Text encoder), 두 인코더의 정보를 결합하는 방식으로 구성된다. VLM을 구성하는 세 가지 핵심 요소는 서로 밀접하게 연관되어 시각적 특징과 텍스트의 의미론적 특징을 효과적으로 추출 및 학습한다. Fig. 1과 같이 VLM 내에서 시각적 특징과 텍스트의 의미론적 특징 정보가 상호작용하고 융합되어 최종 출력 시퀀스가 생성된다[25].

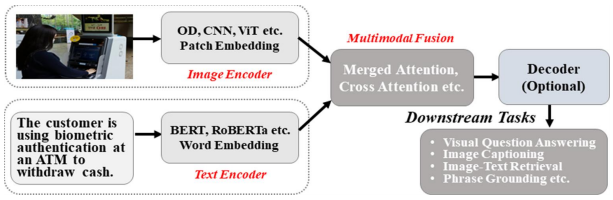

Fig. 1. General Framework of VLMs

본 연구는 OCR의 인식률을 높이는 방법을 제안하는 데 목적이 있지 않기 때문에 한국어에 대해 그 기술력이 검증되어 상용화된 OCR 솔루션을 활용하였다. 구체적으로, 본 연구에서는 업스테이지(Upstage)의 Document OCR 솔루션을 사용하였다. 동 솔루션은 기존 VLM 구조에 마스크 언어모델(Masked language model)과 Cross Attention 기법을 활용하고 있다. 마스크 언어모델이란 부분적으로 마스킹된 설명문이 주어졌을 때, 이 설명문과 일치하는 이미지에 대해 마스킹된 단어를 예측하는 방법이다. 이런 방식의 마스크 언어모델은 바운딩 박스가 포함된 멀티-모달 데이터세트, 또는 입력 텍스트의 부분에 대한 Region Proposal이 가능한 객체 검출 모델에 활용된다[26]. Cross Attention 기법은 직접 이미지 데이터를 언어모델의 디코더에 Cross Attention을 활용하여 시각 정보와 결합한다[27]. 이러한 통합적 접근으로 다양한 OCR 관련 벤치마크 데이터에 대해 높은 성능을 가져왔다[28].

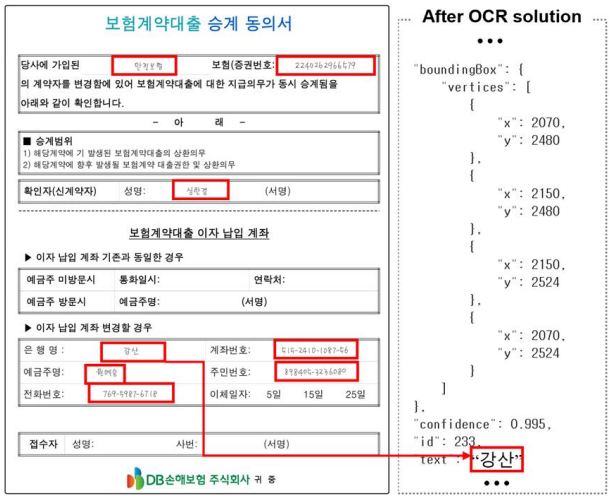

Fig. 2. Document OCR Output

Fig. 2는 본 연구에서 사용한 실제 원본 이미지와 OCR 솔루션 적용 후 추출된 정보의 일부 예시이다. Fig. 2에서 빨간색 네모로 표시된 영역은 손글씨로 작성된 문자에 해당한다. OCR 솔루션을 적용하면, Fig. 2와 같이 x, y로 해당 문자의 위치가 표시된 바운딩 박스(Bounding box)의 고유 식별자와 해당 위치의 문자(Text)가 json 형태로 추출된다.

2. General Retrieval Augmented Generation System

Fig. 3은 일반적인 RAG의 작동 방식을 도식화한 것으로, 크게 데이터 인덱싱(Indexing)[1]과 문서 검색(Retrieval), 답변 생성(Generation)으로 구성된다[5],[29].

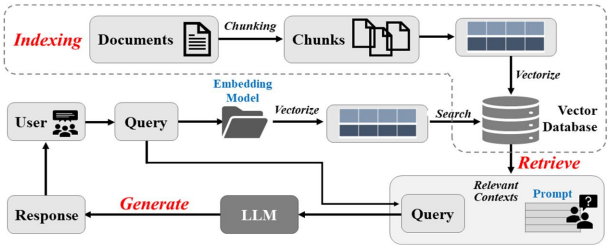

Fig. 3. General RAG Workflow

구체적으로 각 단계를 살펴보면, 데이터 분할 단계는 시스템에 입력된 데이터를 청크 단위의 작은 조각으로 나누는 과정이다. 여기서 청크란 문단, 문장, 또는 더 작은 단위의 텍스트 조각을 의미한다. 데이터 임베딩 단계에서는 청크 단위로 분할된 텍스트를 고차원의 숫자 벡터로 전환한다. 이러한 임베딩 과정을 통해 텍스트의 의미론적 특성을 보존하면서 LLM이 처리할 수 있는 데이터 형태로 변경된다. 문서 검색의 결과가 좋기 위해서는 원천 데이터에 사용된 언어에 대한 성능이 좋은 임베딩 모델을 선택하는 것이 중요하다.

다음으로 데이터 저장 단계에서는 임베딩된 청크와 메타 정보 등을 저장하는 작업이 이뤄진다. 이때의 해당 데이터에 대한 저장소는 주로 벡터 데이터베이스를 사용한다. 벡터 데이터베이스는 임베딩된 청크를 기반으로 구축되며, 일반적으로 최근접 이웃(Approximate nearest neighbor) 알고리즘을 사용해 고차원 벡터가 청크에 부착된 인덱싱을 통해 검색된다. 따라서 벡터 데이터베이스 내 인덱스를 통해 각 청크가 저장된 위치를 확인할 수 있으며, 질문과 청크 벡터 간 유사성 계산 후 관련성이 높은 청크를 빠르게 검색하는 것이 가능해진다[30].

문서 검색 단계에서는 사용자의 질문에 대한 관련 정보를 검색하는 과정이 일어난다. 프롬프트로 질문이 입력되면, 임베딩 과정을 거쳐 벡터화한 후 코사인 유사도(Cosine similarity) 등의 방법을 이용해 벡터 데이터베이

1) 본 연구에서는 데이터 분할과 임베딩, 데이터 저장 단계를 합쳐서 인덱싱으로 표현하였다.

스에서 관련성 높은 청크 순으로 검색한다. 검색된 청크는 원래 텍스트 데이터로 디코딩하여 정보를 추출하고, 이 정보는 프롬프트의 컨텍스트(Context)와 결합해 생성기(Generator)로 전달된다.

마지막 답변 생성 단계에서는 질문과 문서 검색 단계에서 조회된 연관 정보를 토대로 최종 답변을 생성한다. 생성할 텍스트 종류나 길이, 언어적인 스타일 등을 지정할 수 있으며, 기타 특화된 프롬프트 템플릿 등을 사용하면 답변의 품질 개선이 가능하다[31].

본 연구에서는 데이터 분할에 LangChain을, 문서 검색 수행에는 FlashRank를 활용하였다. 검색의 정확도를 높이기 위해 2단계 검색(Re-rankers and two-stage retrieval) 기법을 적용했다. 동 기법은 Fig. 4와 같이 초기 검색 결과를 재정렬(Re-ranking)한 뒤 다시 관련성이 높은 문서를 선별하여 이를 LLM의 입력으로 활용하는 방식이다[32]. 벡터 데이터베이스로는 ChromaDB를 사용했으며, 유사도 측정에는 코사인 유사도 방식을 채택하였다. 생성형 AI 서비스 호출과 임베딩 모델에는 OpenAI가 최근 발표한 GPT-4o와 text-embedding-3-large 모델을 사용하였다.

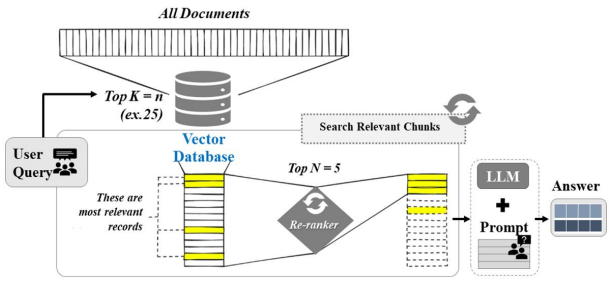

Fig. 4. Re-Rankers and Two-Stage Retrieval Method

한편 RAG 시스템 구성을 위해 몇 가지 하이퍼 파라미터 설정이 필요하다. 본 연구는 문서 이미지 구조의 복잡도에 초점이 있으므로 하이퍼 파라미터의 조건에 따라 RAG 성능이 달라지는 영향을 통제하는 것이 필요하다. 따라서 하이퍼 파라미터를 모든 문서에 같은 값을 적용하였다. 문서 이미지에 등장한 텍스트의 길이가 짧다는 점을 고려해 텍스트 분할의 청크 크기는 80자, 청크 중첩(Chunk overlap) 크기는 30자로 작게 설정하였다. 마지막으로 검색기의 검색 수(Retriever research)는 5를 적용하였다.

IV. Experiments and Results

1. Experimental Design

1.1 Dataset Description

본 연구는 AI-Hub[2]에서 제공하는 금융 분야의 한글 문서 이미지 데이터를 원시 데이터로 사용했다. 동 데이터는 OCR 분야의 연구 및 활용을 위해 구축된 것으로, png 형태의 원본 문서 이미지와 문서에 손글씨로 작성된 문자에 대한 바운딩 박스의 식별자인 x, y 및 이에 대응되는 문자, 즉 정답인 라벨링 데이터를 json 형태로 제공하고 있다.[3] 전체 데이터는 금융과 물류 분야에서 작성되는 다양한 문서 서식을 포함하고 있는데, 본 연구에서는 이 중 은행과 보험 분야의 10가지 유형을 선택했다. 본 연구에서 금융 분야를 선택한 것은 금융업에서 여러 형태의 서식이 작성되어 문서 구조의 복잡성을 다양한 측면에서 검토해볼 수 있기 때문이다. 구체적으로 문서 이미지 안에 사용되는 표의 개수와 형태도 다양하며, 영어와 한자, 한글, 숫자 등 여러 형태의 손글씨 문자에 대한 분석이 가능하다.

한편, Document OCR 솔루션을 사용하면 문서 이미지 내 컴퓨터로 작성된 문자와 손글씨로 작성된 문자를 포함한 모든 정보가 추출된다. 따라서 본 연구에서는 손글씨로 작성된 부분만 OCR의 품질 측정과 RAG의 질문 대상으로 포함하였다. 이는 본 연구에서 사용한 OCR의 솔루션이 컴퓨터로 작성된 문자에 대한 인식 정확도가 100%이며, 현실에서 생성형 AI을 통해 질문하는 의미 있는 정보는 손글씨 정보에 포함되기 때문이다. 예를 들어 보험대출계약승계동의서는 계약자 이름과 계약자가 소유한 보험종류, 증권번호, 은행명 등이 자필로 작성되어 있다. 따라서 계약자 이름(질문 예시: 아주환)을 중심으로 다음의 (1)부터 (3)과 같이 손글씨에 해당하는 부분을 질문의 대상으로 포함한 뒤 프롬프트에 질문을 넣고 답변을 생성했다.

*(1) 보험계약대출승계동의서의 계약자인 **아주환**이 가진 **보험종류**는 무엇입니까?*
*(2) 보험계약대출승계동의서의 계약자인 **아주환**이 거래하는 **은행명**은 무엇입니까?*
*(3) 보험계약대출승계동의서의 계약자인 **아주환**의 **전화번호**는 무엇입니까?*

2) https://www.aihub.or.kr/

3) 분석 활용도를 높이기 위해 원시 데이터는 단순 동의 표시만을 요구하는 간단한 양식을 제외하고, 필기 입력이 가능한 공란 형식의 문서들로 구성되어 있다.

다음으로, 본 연구는 비교적 측정이 단순하고 객관적인 비교가 가능한 세 가지 지표를 문서의 복잡도로 정의하였다: (a) 줄 바꿈으로 구분된 라인의 수, (b) 문서 내 단어 수, (c) 손글씨로 작성된 단어 수. 이러한 세 가지 지표를 문서 복잡도의 기준으로 설정한 이유는, 문서 내 단어 수가 많고 표가 여러 개 포함된 경우 OCR로 추출된 결과에서 라인(Line)[4]의 수가 증가하면서 추출된 정보의 구조가 복잡해지기 때문이다. Fig. 5가 이를 보여준다. 왼쪽이 원본 이미지이며, 오른쪽이 OCR 솔루션 적용 후 추출된 실제 결과이다. Fig. 5(a)는 문서 구조가 단순한 반면, Fig. 5(b)는 복잡한 형태를 예시한다.

본 연구에서 문서 이미지의 복잡도를 직관적으로 가장 잘 나타내는 지표가 라인이며, 라인이 길어질수록 RAG 성능 저하가 클 것으로 예상하였다. 이는 생성형 AI의 답변 생성을 위해 RAG의 데이터 인덱싱과 문서 검색 과정을 거쳐야 하는데, 문서 내 정보들이 라인으로 구분될 경우 인덱싱과 검색 단계를 거치면서 이를 서로 연관된 정보로 파악하여 유사성이 높은 문서로 검색되기 어렵기 때문이다. Fig. 5(a)의 추출 결과를 보면, 보험계약대출승계동의서의 계약자명(아주환)이 9번째 라인에서 등장하나, 계약자가 소유한 보험종류(자동차보험)는 2번째, 은행명은 15번째, 전화번호는 17번째 라인에서 나타난다. 그러나 (1)부터 (3)의 질문 예시를 보면 계약자 이름을 기준으로 보험종류와 은행명, 전화번호를 질문하기 때문에 OCR로 추출된 문서 정보를 가공 없이 RAG 시스템에 입력하면 일반적인 문장으로 구성된 문서 텍스트와 비교해 RAG의 검색 성능이 좋지 않을 것이라 예상할 수 있다.

한편 Fig. 5(a)와 Fig. 5(b)에서 보듯이 표나 단어 수가 많아질수록 자필로 작성되는 단어도 많아지고, 이에 따라 OCR로 텍스트를 추출하는 과정에 라인이 길어지게 된다. 따라서 세 가지 지표는 서로 높은 관련성을 가지게 된다. Fig. 6은 세 가지 지표를 문서별로 평균하여 시각화한 것이다. 그림에서 보듯 세 가지 지표가 서로 관련되어 있음을 알 수 있다. 예를 들어, 보험대출계약승계동의서(Consent form for insurance policy loan transfer)는 다른 유형에 비해 세 가지 지표 모두 낮은 반면 보험청구서(Insurance claim form)는 높게 나타났다. 스피어만 상관계수에서도 라인과 단어의 수는 서로 0.76, 라인과 손글씨로 작성된 단어 수는 0.55로 계산되었다.

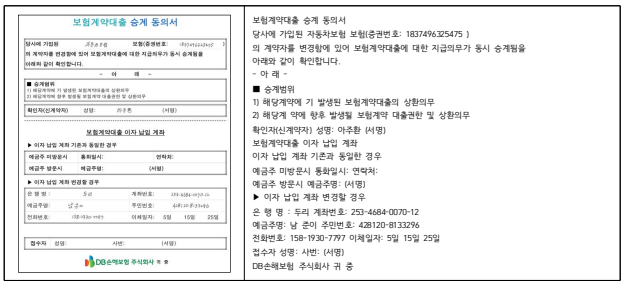

Fig. 5(a). Consent Form for Insurance Policy Loan Transfer

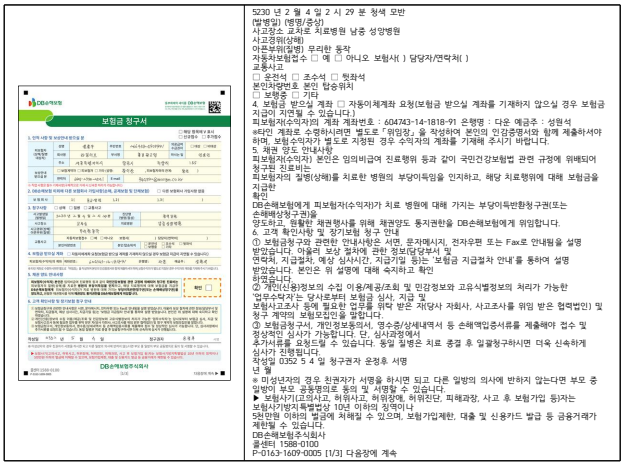

Fig. 5(b). Insurance Claim Form

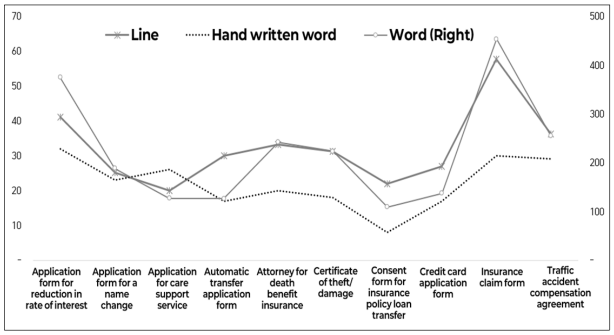

Fig. 6. Document Complexity Metrics by Document Type

본 연구의 분석 데이터 구성은 Table 1과 같다. 각 유형별 문서의 수는 (A)이며, 유형별 질문의 수는 (B)이다. 따라서 각 문서 유형에 대한 총 질문의 수는 A와 B를 곱해서 산출된다. 각 질문에 대한 응답은 은행명과 같이 텍스트로 표현된 형태와 전화번호, 증권번호 등 숫자 형태를 모두 포함하고 있다.

Table 2는 IV.1.1 절에서 정의한 문서의 복잡도 지표에 따라 라인의 수와 단어의 수, 손글씨로 작성된 단어의 수를 문서 유형별로 평균한 값이다.

4) OCR 솔루션을 적용하면 추출된 데이터가 자동으로 줄바꿈(줄 개행으로 처리)되는데, 본 연구는 이 줄 바꿈의 개수를 본문에서 라인으로 표현하였다.

Table 1. Number of Samples in the Experimental Dataset

Document type	Number of documents (A)	Number of questions per document (B)	Number of total questions (A×B)
Application form for reduction in rate of interest	393	6	2,358
Application form for a name change	279	9	2,511
Application for care support service	129	9	1,161
Automatic transfer application form	153	18	2,754
Attorney for death benefit insurance	133	14	1,862
Certificate of theft/damage	110	16	1,760
Consent form for insurance policy loan transfer	141	7	987
Credit card application form	157	16	2,512
Insurance claim form	105	18	1,890
Traffic accident compensation agreement	107	16	1,712
Total	1,707	129	220,203

1.2 Question-Answering Construction

본 연구에서 사용한 원시 데이터는 OCR 분야의 활용 목적으로 설계되었기 때문에 동 데이터를 활용해 RAG 성능 분석을 진행하기 위해서는 별도의 질의응답 데이터 세트가 필요하다[33]. 이에 따라 본 연구에서는 추가적으로 질의응답 데이터세트를 구축하였다. 각 문서 유형별로 손글씨로 작성된 정보가 상이해 모든 문서에 동일하게 적용할 수 있는 질문을 생성하기는 어려웠다. 따라서 모든 질문에 적어도 하나 이상의 특정 사람 또는 기관을 포함하며, 해당 정보를 중심으로 손글씨에 해당하는 부분을 질문하였다. IV.1.1 절의 세 가지 질문 예시처럼 '보험계약대출 승계동의서의 계약자인 아주환이 가진 보험종류는 무엇입니까?'로 계약자명과 보험종류를 연관하여 질문하였다. 질문에 대한 정답은 원시 데이터에서 제공하고 있는 손글씨에 대한 라벨링 데이터에서 가져왔다.

Table 2. Document Complexity Metrics (Average)

Document type	Number of Lines	Number of Total Words	Number of Hand Written Words
Application form for reduction in rate of interest	41	374	32
Application form for a name change	25	188	23
Application for care support service	20	127	26
Automatic transfer application form	30	126	17
Attorney for death benefit insurance	33	242	20
Certificate of theft/damage	31	224	18
Consent form for insurance policy loan transfer	22	109	8
Credit card application form	27	137	17
Insurance claim form	58	453	30
Traffic accident compensation agreement	36	255	29

1.3 Performance Evaluation

OCR의 품질은 OCR 기술을 적용해 문서 이미지 데이터를 디지털로 변환한 뒤 손글씨로 작성된 문자만을 원시 데이터의 라벨링 정보와 비교해 계산하였다. OCR의 품질 측정에 사용된 지표는 NLP와 OCR 성능 평가에서 많이 사용되는 CER(Character Error Rate)과 WER(Word Error Rate)이다[34]. CER은 문자 단위에서의 인식 오류율로, OCR로 추출된 문자 중 잘못 인식된 문자 수를 기준 문자 수로 나눈 값을 의미한다. 반면 WER은 단어 단위의 인식 오류율로, 전체 기준 단어 수 대비 잘못 인식된 단어 수의 비율을 나타낸다[35]. 두 지표 모두 문자열 간의 삽입, 삭제, 치환 등의 최소 편집 횟수를 기반으로 하는 레벤슈타인 거리(Levenshtein distance)를 통해 계산된다[36].

- CER = (Number of incorrect characters / Total number of characters) × 100
- WER = (Number of incorrect words / Total number of words) × 100

한편 RAG의 성능은 ROUGE(ROUGE-1, ROUGE-2, ROUGE-L), BLEU 등의 정량적 지표뿐만 아니라, 신뢰성(Faithfulness)과 일관성(Consistency) 측면에서도 평가할 수 있다[5] [37]. 본 연구에 사용된 질의응답 데이터는

연락처나 은행명과 같이 답변이 짧고, 사실적인 응답을 요구하도록 설계하였다. 따라서 신뢰성 측면에서 생성된 응답이 외부 지식 기반으로 활용된 원본 문서 이미지에 포함된 정보를 정확하게 반영하고 있는지가 생성 응답의 품질 측정에서 가장 중요한 요인이다. 이에 따라 본 연구에서는 OCR과 동일하게 RAG 성능 측정에도 CER과 WER을 적용했다. 이때 OCR로 추출된 결과물이 RAG의 성능에 미치는 영향을 분석하기 위해 OCR로 추출된 정보를 그대로 RAG 내 입력 데이터로 사용하였다. 즉, OCR의 인식 오류가 포함된 상태이며, 별도의 후처리 없이 줄 바꿈으로 추출된 형태를 사용했다. 그 뒤 본 연구에서 구축한 질문을 생성형 AI의 프롬프트에 입력하여 답변을 생성하였다. 마지막으로 생성 모델의 응답에서 질문에 대응하는 단답형 응답만 추출하여 해당 정보를 원본 문서 이미지에서 수집한 정답 레이블과 비교하여 CER와 WER를 계산하였다. 이러한 과정을 통해 OCR의 결과와 비교해 RAG의 생성 응답 정확도를 비교했다.

2. Experiment Results

Table 3은 10개 유형에 대한 OCR 품질의 결과이다. 이전 연구들과 일치하게 WER이 CER보다 높게 나타났다[38-39]. 이는 CER이 잘못 인식된 문자만을 고려하는 반면, WER은 단 하나의 문자가 잘못 인식되더라도 해당 단어를 오류로 처리하기 때문이다. 다음으로 OCR 성능을 벤치마크 데이터셋을 기반으로 비교한 선행연구 결과[40]와 비교해 금리인하신청서(Application form for reduction in rate of interest)를 제외한 모든 문서 유형에서 CER과 WER 모두 오류율이 낮은 것을 알 수 있었다. 이는 본 연구의 문서 이미지 데이터에 대해 OCR 기술의 인식률이 전반적으로 우수함을 보여준다.

Table 4는 인식 오류와 산출물의 구조를 가공하지 않고, 즉 OCR의 산출물 그대로 RAG의 입력 데이터로 사용하여 RAG의 성능을 평가한 결과이다. Table 4에서 확인할 수 있듯이 OCR의 품질이 좋음에도 CER와 WER 모두 크게 높아져 RAG의 성능이 좋지 않음을 알 수 있다. 이는 OCR의 인식률이 우수하더라도 RAG 시스템의 성능이 저하될 수 있다는 본 연구의 가정을 뒷받침하는 결과이다.

Table 3. CER and WER Results for OCR Output

Document type	CER	WER
Application form for reduction in rate of interest	3.84	41.75
Application form for a name change	8.00	36.24
Application for care support service	3.32	27.97
Automatic transfer application form	9.10	33.70
Attorney for death benefit insurance	0.90	5.25
Certificate of theft/damage	4.51	16.53
Consent form for insurance policy loan transfer	0.57	4.55
Credit card application form	5.73	27.25
Insurance claim form	1.99	9.35
Traffic accident compensation agreement	4.77	16.62

Table 4. CER and WER Results for RAG Output

Document Type	CER	WER
Application form for reduction in rate of interest	37.54	40.84
Application form for a name change	25.95	25.09
Application for Care Support Service	83.79	73.94
Automatic transfer application form	28.74	38.63
Attorney for Death Benefit Insurance	83.44	80.99
Certificate of Theft/Damage	54.12	59.82
Consent Form for Insurance Policy Loan Transfer	64.73	71.43
Credit card application form	107.64	79.46
Insurance Claim Form	113.92	86.67
Traffic Accident Compensation Agreement	65.49	70.14

다음으로 본 연구는 OCR로 추출한 결과와 RAG의 결과 간 차이가 있는지 보기 위해 독립 표본 t-test(Independent sample t-test)를 수행했다. 그 결과, Table 5에서 확인할 수 있듯이 CER와 WER에 대해 1% 유의수준에서 귀무가설을 기각해 문서 유형별로 OCR과 RAG의 결과가 다른 것으로 나타났다.

Table 5. Independent Sample t-Tests for CER and WER Results

Document Type	t-statistics[1)]	
	CER	WER
Application form for reduction in rate of interest	-44.15	-28.52
Application form for a name change	-70.82	-37.98
Application for care support service	-16.16	-13.40
Automatic transfer application form	-68.69	-41.53
Attorney for death benefit insurance	-27.30	-32.22
Certificate of theft/damage	-19.15	-16.11
Consent form for insurance policy loan transfer	-15.06	-16.07
Credit card application form	-149.01	-61.29
Insurance claim form	-10.51	-17.69
Traffic accident compensation agreement	-20.65	-20.39

[1)] All t-statistics are statistically significant at the 1% level, with p-values less than 0.001.

5) 여기서 신뢰성이란, 생성된 응답이 RAG를 통해 검색된 문서에 포함된 정보를 얼마나 정확하게 반영하는지를 의미하며, 사실이 아닌 내용을 포함하는ㅁ 정도를 나타낸다. 반면 일관성은 생성된 응답이 검색된 근거와 논리적으로 얼마나 잘 부합하는지를 평가하며, 내부적 논리성과 사실적 정합성을 포함한다.

본 연구는 이러한 결과가 외부 지식 데이터, 즉 문서 자체의 구조적 특성과 관련이 높다고 해석하였다. Fig. 5에서 볼 수 있듯이 다양한 구조를 가진 문서로부터 OCR 기술로 문자 및 숫자 정보를 추출하는데, 그 구조가 복잡할수록 원본 결과에서 각 정보 간 연관 관계를 파악하기가 어렵기 때문이다. 즉, OCR로 추출된 결과는 줄 바꿈 형태로 정보가 구분되어 출력되기 때문에, RAG의 데이터 분할 단계에서 문맥 정보를 이해할 수 있는 방식으로 청크를 생성하는 데 어려움을 초래한다. 이렇게 구조적으로 단절되고, 문맥이 결여된 데이터가 입력 데이터로 사용되면 OCR의 품질과 무관하게 RAG 시스템을 구성하는 인덱싱, 임베딩, 문서 검색, 생성 단계에서의 개별 컴포넌트의 성능 저하로 이어질 수 있다.

이를 확인하기 위해 IV. 1절에서 정의한 문서의 복잡도와 OCR의 품질 및 RAG의 성능 간 관계를 살펴보았다. 구체적으로 문서의 복잡도와 성능 지표 간 상관관계를 계산하였다. 먼저 문서의 라인 수와 단어 수로 표현한 복잡도와 OCR의 품질을 나타내는 CER과 WER 간 상관계수를 보면 0에 가까워 문서의 구조가 복잡하더라도 인식률과는 관련이 낮은 것으로 나타났다. 일반적으로 문서의 구조가 복잡하면 정확한 문자 인식이 어려운 것으로 알려져 있으나 본 연구에서 사용한 OCR 기술이 우수하고, 손글씨 대비 컴퓨터로 작성한 문자의 수가 많아 기존 OCR의 한계점은 제한적으로 나타났다. 즉, 복잡한 문서라도 그 구조와 무관하게 모든 유형에서 높은 인식률을 보였다. 그러나 OCR의 품질이 좋더라도 문서의 구조가 복잡할 경우 RAG의 성능은 좋지 않았다. Table 6-8에서 보듯 라인 수의 길수록(문서가 복잡할수록) 오류율은 높아져 양의 상관관계를 보였다. 이는 문서 이미지의 문맥 정보 파악이 어려운 형태로 OCR 결과가 추출되었고, 이러한 정보가 RAG의 입력 데이터로 사용됨에 따라 RAG의 성능이 저하된 것으로 해석된다.

Table 6. Spearman Correlation between Number of Document Line and Error Rate

Error rate		Correlation coeff.	p-value
Original OCR output	CER	-0.047	0.073
	WER	-0.072	0.006
RAG output	CER	0.395	0.000
	WER	0.370	0.000

Table 7. Spearman Correlation between Number of Document Words and Error Rate

Error rate		Correlation coeff.	p-value
Original OCR output	CER	0.065	0.013
	WER	-0.043	0.102
RAG output	CER	0.235	0.000
	WER	0.174	0.000

Table 8. Spearman Correlation between Number of Hand Written Words and Error Rate

Error rate		Correlation coeff.	p-value
Original OCR output	CER	-0.461	0.000
	WER	-0.266	0.000
RAG output	CER	0.248	0.000
	WER	0.353	0.000

한편 전체 단어의 수와 자필로 작성된 단어의 수는 라인 수와 비교해 오류율 간 양의 상관관계가 다소 떨어졌는데, 이는 IV.1 절에서 언급한 것과 같이 라인이 문서 내 등장하는 표 형태의 정보, 즉 문맥 정보 추출이 어려운 형태를 가장 잘 나타내기 때문이다. 예를 들어 Fig. 5(b)와 같이 OCR로 추출된 데이터를 보면, 계약자명과 계약자의 보유한 보험종류가 서로 다른 라인으로 구분되기 때문에 계약자명에 대응해 보험종류를 질문하면 RAG 시스템에서 이를 유사 문서로 검색하기가 어렵다. 따라서 이미지 서식이 라인이 길게 추출되는, 즉 복잡한 구조를 가진 문서라면 RAG의 검색 성능 저하가 더 크게 나타날 수 있다. 실제로 Table 3과 Table 4를 보면, 보험대출계약승계동의서와 같이 라인과 단어의 수, 자필로 작성된 단어의 수로 클수록(문서 구조가 복잡할수록) OCR과 RAG의 성능 차이는 더 크게 나타났다. 반면 카드신청서(Credit card application form)나 명의변경신청서(Application form for a name change)처럼 세 가지 지표 모두 낮은 값을 가지는 즉, 구조가 단순한 문서의 경우에는 OCR과 RAG 간의 성능 차이가 감소하였다. 이러한 결과는 OCR 품질이 일정 수준 이상을 유지한 이후에는 이를 이용한 RAG 시스템 구축 시 입력정보의 구조화가 RAG 성능 개선에 더 중요한 요인이 될 수 있음을 시사한다.

V. Conclusions

본 연구에서는 VLM을 활용한 OCR 솔루션을 사용하여 금융 분야의 문서 이미지에 대한 데이터베이스를 구축하였다. OCR 성능 측정을 위해 CER과 WER을 사용하였으며, 그 결과 평균적으로 두 지표 모두 낮게 나타나 OCR의 인식률은 우수한 것으로 나타났다. 그러나 OCR의 품질이

좋음에도 OCR의 산출물을 가공 없이 RAG의 입력 데이터로 적용할 경우 RAG 성능은 악화되었다. 이러한 결과는 OCR 품질이 일정 수준 이상을 유지한 이후에는 컴퓨터가 이해하기 쉬운 형태로 입력 데이터의 정보를 구조화하는 것이 RAG 성능 개선에 더 중요한 요인이 될 수 있음을 보여준다.

본 연구의 결과는 RAG 적용에 최적화된 OCR 활용 방안을 모색하는 데 있어 유의미한 통찰을 제공한다. 또한, OCR과 RAG 간의 상호작용 방식을 이해하고 개선의 방향성을 제안하여 학술적인 기여점 외에도 실무적으로도 유용할 것으로 기대한다.

본 연구는 OCR 인식률이 높아도 RAG 성능이 좋지 않을 수 있으며, 문서 구조의 복잡도가 이러한 영향을 가져오는 주된 요인임을 확인하는 데 초점을 두었다. 반면 어떻게 RAG의 성능을 개선할 수 있을지는 다루지 못했으며, 이는 본 연구의 한계점이다. 향후 복잡한 문서 이미지에 대해 OCR 기술로 데이터베이스를 구축할 경우 RAG의 성능을 높이는 방법이 제시된다면 OCR-RAG 통합 분야에 중요한 시사점을 제공할 수 있을 것이다.

REFERENCES

[1] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, “Retrieval-Augmented Generation for AI-Generated Content: A Survey,” arXiv preprint, Feb. 2024. DOI: 10.48550/arXiv.2402.19473

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain of Thought Prompting Elicits Reasoning in Large Language Models,” Proceedings of the Part of Advances in Neural Information Processing Systems 35(NeurIPS), pp. 1-13, Dec. 2022.

[3] J. Kasai, K. Sakaguchi, Y. Takahashi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui, “Realtime QA: What’s the Answer Right Now?,” arXiv preprint, Jul. 2022. DOI: 10.48550/arXiv.2207.13332

[4] B. Saha, U. Saha, and M. Z. Malik, “QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance,” IEEE Access, Vol. 12, pp. 185401-185410, Dec. 2024. DOI: 10.1109/ACCESS.2024.3467890

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint, Dec. 2023. DOI: 10.48550/arXiv.2312.10997

[6] J. Zhang, Q. Zhang, B. Wang, L. Ouyang, Z. Wen, Y. Li, K. H. Chow, C. He, and W. Zhang, “OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation,” arXiv preprint, Dec. 2024. DOI: 10.48550/arXiv.2412.02592

[7] E.-S. Choi, “Development of an Automated ESG Document Review System Using Ensemble-Based OCR and RAG Technologies,” Journal of The Korea Society of Computer and Information, Vol. 29, No. 9, pp. 25-37, Sep. 2024. DOI: 10.9708/jksci.2024.29.09.025

[8] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, “Survey of Post-OCR Processing Approaches,” ACM Computing Surveys, Vol. 54, No. 6, pp. 1-37, Jul. 2021. DOI: 10.1145/3453476

[9] D. Fleischhacker, R. Kern, and W. Göderle, “Enhancing OCR in Historical Documents with Complex Layouts Through Machine Learning,” International Journal on Digital Libraries, Vol. 26, No. 3, pp. 1-15, Sep. 2025. DOI: 10.1007/s00799-025-00413-z

[10] K. Ghosh, A. Chakraborty, S. K. Parui, and P. Majumder, “Improving Information Retrieval Performance on OCRed Text in the Absence of Clean Text Ground Truth,” Information Processing & Management, Vol. 52, No. 5, pp. 873-884, Sep. 2016. DOI: 10.1016/j.ipm.2016.03.006

[11] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, “Survey of Post-OCR Processing Approaches,” ACM Computing Surveys, Vol. 54, No. 6, pp. 1-37, Jul. 2021. DOI: 10.1145/3453476

[12] O. Vitman, Y. Kostiuk, P. Plachinda, A. Zhila, G. Sidorov, and A. Gelbukh, “Evaluating the Impact of OCR Quality on Short Texts Classification Task,” Proceeding of the Mexican International Conference on Artificial Intelligence (MICAI), pp. 171-185, Oct. 2022. DOI: 10.1007/978-3-031-19496-2_13

[13] J. Zhang, Q. Zhang, B. Wang, L. Ouyang, Z. Wen, Y. Li, K. H. Chow, C. He, and W. Zhang, “OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation,” arXiv preprint, Dec. 2024. DOI: 10.48550/arXiv.2412.02592

[14] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, B. Zhang, L. Wei, Z. Sui, W. Li, B. Shi, Y. Qiao, D. Lin, and C. He, “Mineru: An Open-Source Solution for Precise Document Content Extraction,” arXiv preprint, Sep. 2024. DOI: 10.48550/arXiv.2409.18839

[15] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, “LLaVA-Next: Improved Reasoning, OCR, and World Knowledge,” arXiv preprint, Jan. 2024. DOI: 10.48550/arXiv.2401.17103

[16] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution,” arXiv preprint, Sep. 2024. DOI: 10.48550/arXiv.2409.12191

[17] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, and Y. Xiao, “Towards

Unconstrained End-to-End Text Spotting," Proceeding of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4704-4714, Oct. 2019. DOI: 10.1109/ICCV.2019.00481

[18] L. Li, F. Gao, J. Bu, Y. Wang, Z. Yu, and Q. Zheng, "An End-to-End OCR Text Re-Organization Sequence Learning for Rich-Text Detail Image Comprehension," Proceeding of the European Conference on Computer Vision (ECCV), pp. 85-101, Aug. 2020. DOI: 10.1007/978-3-030-58595-2_6

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," arXiv preprint, Apr. 2023. DOI: 10.48550/arXiv.2304.08485

[20] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang, Q. Jin, L. He, and X. A. Lin, "UReader: Universal OCR-Free Visually-Situated Language Understanding with Multimodal Large Language Model," arXiv preprint, Oct. 2023. DOI: 10.48550/arXiv.2310.05126

[21] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, T. Gao, E. Li, K. Tang, Z. Cao, T. Zhou, A. Liu, X. Yan, S. Mei, J. Cao, Z. Wang, and C. Zheng, "A Survey on Multimodal Large Language Models for Autonomous Driving," Proceeding of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 958-979, Jan. 2024. DOI: 10.1109/WACVW 60836.2024.00106

[22] B. Lamm, and J. Keuper, "Can Visual Language Models Replace OCR-Based Visual Question Answering Pipelines in Production? A Case Study in Retail," arXiv preprint, Jun. 2024. DOI: 10.48550/arXiv.2406.10357

[23] Y.-Q. Yu, M. Liao, J. Zhang, and J. Wu, "TextHawk2: A Large Vision-Language Model Excels in Bilingual OCR and Grounding with 16x Fewer Tokens," arXiv preprint, Jul. 2024. DOI: 10.48550/arXiv.2407.16858

[24] Y.-C. Lin, Y. Chen, Z. Wang, and X. Liu, "Novel Preprocessing Technique for Data Embedding in Engineering Code Generation Using Large Language Model," arXiv preprint, Nov. 2023. DOI: 10.48550/arXiv.2311.16267

[25] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, "Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends," Foundations and Trends in Computer Graphics and Vision, Vol. 14, No. 3-4, pp. 163-352, Nov. 2022. DOI: 10.1561/0600000105

[26] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra, "An Introduction to Vision-Language Modeling," arXiv preprint, May 2024. DOI: 10.48550/arXiv.2405.17247

[27] K. Chen, D. Shen, H. Zhong, H. Zhong, K. Xia, D. Xu, W. Yuan, Y. Hu, B. Wen, T. Zhang, C. Liu, D. Fan, H. Xiao, J. Wu, F. Yang, S. Li, and D. Zhang, "EVLM: An Efficient Vision-Language Model for Visual Understanding," arXiv preprint, Jul. 2024. DOI: 10.48550/arXiv.2407.14177

[28] B. Na, Y. Kim, and S. Park, "Multi-Modal Text Recognition Networks: Interactive Enhancements Between Visual and Semantic Features," Proceeding of the European Conference on Computer Vision (ECCV), pp. 448-463, Oct. 2022. DOI: 10.1007/978-3-031-19815-1_26

[29] H. N. Patel, A. Surti, P. Goel, and B. Patel, "A Comparative Analysis of Large Language Models with Retrieval-Augmented Generation-Based Question Answering System," Proceeding of the 8th International Conference I-SMAC (IoT Social, Mobile, Analytics Cloud), pp. 1-8, Oct. 2024.

[30] M. Nam, and N. Kim, "Sentence-Based Extraction Methodology from External References to Enhance Performance in RAG," Journal of The Korea Society of Computer and Information, Vol. 29, No. 12, pp. 29-39, Dec. 2024. DOI: 10.9708/jksci.2024.29.12.029

[31] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction," Proceeding of the 5th ACM International Conference on AI in Finance, pp. 608-616, Nov. 2024.

[32] Y. Huang, and J. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," arXiv preprint, Apr. 2024. DOI: 10.48550/arXiv.2404.10981

[33] G.-W. Yi, and S. K. Kim, "Design of a Question-Answering System Based on RAG Model for Domestic Companies," Journal of The Korea Society of Computer and Information, Vol. 29, No. 7, pp. 81-88, Jul. 2024. DOI: 10.9708/jksci.2024.29.07.081

[34] A. Tsimpiris, D. Varsamis, and G. Pavlidis, "Tesseract OCR Evaluation on Greek Food Menus Datasets," International Journal of Computer Optimization, Vol. 9, No. 1, pp. 13-32, Jan. 2022. DOI: 10.12988/ijco.2022.9829

[35] C. Neudecker, K. Baierer, M. Gerber, C. Clausner, A. Antonacopoulos, and S. Pletschacher, "A Survey of OCR Evaluation Tools and Metrics," Proceeding of the 6th International Workshop on Historical Document Imaging and Processing, pp. 13-18, Sep. 2021. DOI: 10.1145/3476887.3476888

[36] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text Recognition in the Wild: A Survey," ACM Computing Surveys, Vol. 54, No. 2, pp. 1-35, Mar. 2021. DOI: 10.1145/3440756

[37] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," Communications in Computer and Information Science, vol. 2301, 2025. DOI:

10.1007/978-981-96-1024-2_8

[38] S. Drobac, and K. Lindén, “Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods,” International Journal on Document Analysis and Recognition, Vol. 23, No. 4, pp. 279-295, Dec. 2020. DOI: 10.1007/s10032-020-00359-9

[39] C. Neudecker, K. Baierer, M. Gerber, C. Clausner, A. Antonacopoulos, and S. Pletschacher, “A Survey of OCR Evaluation Tools and Metrics,” Proceeding of the 6th International Workshop on Historical Document Imaging and Processing, pp. 13-18, Sep. 2021. DOI: 10.1145/3476887.3476888

[40] M. Nazeem, R. Anitha, and S. Navaneeth, “Open-Source OCR Libraries: A Comprehensive Study for Low Resource Language,” Proceeding of the 21st International Conference on Natural Language Processing (ICON), pp. 416-421, Dec. 2024. URL: https://aclanthology.org/2024.icon-1.48

Authors

Minchae Song received her B.A. degree in Economics from Ewha Womans University in 2007, her M.A. degree in Economics from the same university in 2010, and her Ph.D. degree in Big Data Analysis from Ewha Womans University in 2019. Dr. Song has been working at NongHyup Financial Group’s Financial Research Institution. She has expertise in artificial intelligence, natural language processing, and financial data science. Her research interests include NLP, LLMs, financial analysis, and digital finance.

Jaeyoung Park received his B.S. degree in Engineering from Soongsil University, Korea, in 2012, and his M.S. and Ph.D. degrees in Information Systems from Yonsei University, Korea, in 2017 and 2021, respectively. Dr. Park joined the faculty of the Department of Digital Finance at Pukyong National University, Busan, Korea, in 2025. He is currently an Assistant Professor in the same department. His research interests include FinTech and privacy.