JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 8, pp. 29-39, August 2025
https://doi.org/10.9708/jksci.2025.30.08.029

KRAFT³-QA: Korean financial text-table benchmark for evaluating tool-augmented agents on QA tasks

Seungjae Park*, Sung-Bae Cho**, Ha Young Kim***

*Student, Dept. of Artificial Intelligence, Yonsei University, Seoul, Korea
**Professor, Dept. of Computer Science, Yonsei University, Seoul, Korea
***Associate Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

Periodic corporate filings are structured documents combining text and tables. Practical use of these documents requires comprehensive reasoning to integrate and interpret information across multiple sections. However, current large language models (LLMs) struggle with such reasoning, and existing financial benchmarks are insufficient for evaluating practical skills like tool usage. To address this gap, we develop KRAFT³-QA, a new benchmark based on Korean corporate filings. KRAFT³-QA consists of multiple-choice tasks that require integrating information across various sections. Model performance is evaluated using both accuracy and valid response rate. Experiments with major open LLMs demonstrate that model scale and reasoning architecture can affect performance. This study presents a real document--based, tool-augmented QA benchmark and an evaluation framework, establishing a technical foundation for quantitatively assessing the real-world problem-solving capabilities of LLM agents.

▸ **Key words**: Large Language Model, Benchmark Dataset, Text-Table Question Answering, Tool-augmented Agent, Financial Domain

[요 약]

기업의 정기 사업보고서는 텍스트와 테이블이 혼합된 복합 문서로, 실무적 활용을 위해서는 분산된 정보를 유기적으로 통합해 해석하는 종합적인 추론 능력이 요구된다. 그러나, 현행 거대 언어 모델(LLM)은 이러한 복합 추론에서 한계를 보이며, 기존 금융 벤치마크는 도구 활용 측면과 같은 실무적 역량을 충분히 평가하지 못한다. 본 연구는 이러한 한계를 극복하기 위해, 한국 상장 기업의 사업보고서를 기반으로 한 새로운 벤치마크 KRAFT³-QA를 제안한다. KRAFT³-QA는 문서 내 다양한 항목을 통합해야 정답에 도달할 수 있는 사지선다형 태스크로 구성되며, 정확도와 유효 응답률을 통해 성능을 평가한다. 주요 공개 LLM을 대상으로 한 실험 결과, 모델의 규모와 추론 구조가 성능에 영향을 줄 수 있음을 확인했다. 본 연구는 문서 기반 도구 활용형 QA 벤치마크와 평가 프레임워크를 제안함으로써, LLM 에이전트의 실질적인 문제 해결 능력을 정량적으로 평가할 수 있는 기술적 기반을 마련했다는 데 의의가 있다.

▸ **주제어**: 거대 언어 모델, 벤치마크 데이터셋, 텍스트-테이블 질의응답, 도구 증강 에이전트, 금융 도메인

• First Author: Seungjae Park, Corresponding Author: Ha Young Kim
*Seungjae Park (seungjae.park@yonsei.ac.kr), Dept. of Artificial Intelligence, Yonsei University
**Sung-Bae Cho (sbcho@yonsei.ac.kr), Dept. of Computer Science, Yonsei University
***Ha Young Kim (hayoung.kim@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2025. 07. 07, Revised: 2025. 07. 23, Accepted: 2025. 07. 30.

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

I. Introduction

기업의 정기 사업보고서는 투자자, 애널리스트, 규제 기관 등 다양한 이해관계자에게 기업의 핵심 정보를 제공하는 공식 문서로, 기업 분석과 의사결정의 기반이 된다. 이 보고서는 서술형 텍스트와 계량적 데이터가 혼합된 구조를 가지며, 이를 실무적으로 활용하기 위해서는 단순한 정보 검색을 넘어서 종합적인 해석과 판단 능력이 필수적이다[1]. 특히, 텍스트의 맥락과 표 형식의 수치를 유기적으로 연결하고, 문서 전반에 분산된 정보를 통합하는 과정에서는 고차원적 추론이 요구된다.

최근 거대 언어 모델(Large Language Model, LLM)의 비약적인 발전은 자연어 처리 분야 전반에 걸쳐 문서 처리 작업의 자동화 가능성을 크게 확장시켰다[2]. 그러나 이러한 발전에도 불구하고, LLM은 여전히 구조화된 테이블과 비정형 텍스트를 동시에 이해하고 추론하는 데에 어려움을 겪고 있다[3]. 이는 문맥 오해, 논리적 추론 실패, 그리고 복수 데이터 소스를 효과적으로 연계하지 못하는 구조적 제약에서 비롯되며, 정확성과 전문성이 요구되는 금융 도메인에서는 이러한 한계가 더욱 뚜렷하게 드러난다[4][5]. 따라서 LLM이 복합적인 문서 환경에서 효과적으로 작동하려면, 문서의 구조를 파악하고 필요한 정보를 능동적으로 탐색하고 추론하는 역량에 대한 체계적 평가와 기술적 발전이 병행되어야 한다.

기존 금융 LLM 벤치마크 연구는 주로 모델 내부 지식과 주어진 문맥을 기반으로 한 자연어 질의응답(QA) 작업에 집중되었다. 하지만, 이러한 접근은 실제 금융 문서 분석에서 필수적인 수치 연산, 외부 도구 기반 정보 탐색 등의 복합 처리 과정을 충분히 반영하지 못하기 때문에, LLM의 실질적 역량을 종합적으로 평가하기에 한계가 존재한다[6]. 따라서 금융 에이전트로서 LLM의 실무적 잠재력을 검증하기 위해서는, 도구 활용을 포함하여 복합 추론 능력을 체계적으로 측정할 새로운 벤치마크가 필요하다.

본 연구는 한국 상장 기업의 실제 사업보고서에서 발췌한 텍스트와 테이블을 기반으로, ReAct[7]를 활용한 도구 기반 복합 추론 QA 태스크를 설계하고, 한국어 금융 벤치마크 KRAFT³-QA(**K**o**R**e**A**n **F**inancial **T**ext-**T**able benchmark for evaluating **T**ool-augmented agents on **QA** tasks)를 구축했다. 이 벤치마크는 단순 정보 검색을 넘어서 재무 분석, 전략 해석, 시장 동향 파악 등 실제 기업 경영 분석에 필요한 실무적 추론 역량을 평가한다. 모든 문항은 사지선다 객관식 형식으로 구성되며, 정답을 도출하려면 모델이 사업보고서의 특정 항목을 조회하는 도구를 호출하여 필요한 정보를 확보하고, 이를 바탕으로 다단계 논리적 추론을 수행해야만 하도록 정답에 도달할 수 있도록 고안됐다.

본 연구의 주요 기여는 다음과 같다.

1. 실제 문서 기반의 한국어 금융 벤치마크 구축: 한국 상장 기업의 사업보고서를 기반으로, 도구 활용이 필요한 추론 중심의 한국어 QA 데이터셋을 구축했다. 질문은 '이사의 경영진단 및 분석의견' 항목에서 생성하고, 정답은 '사업의 내용' 또는 '재무제표' 등을 참조해야 도출되도록 설계하여 문서 간 정보 연계를 유도했다. 모든 문항은 LLM을 활용해 자동 생성되었으며, 2단계 필터링 및 선택지 무작위화를 통해 문항의 품질을 개선했다. 이로써 실무에 가까운 환경에서 LLM 에이전트의 문제 해결 능력을 정량적으로 측정할 수 있는 기반을 마련했다.

2. 도구 활용 신뢰성을 통합한 평가 구조 제안: 단순 정답 정확도(Accuracy)만을 평가하는 방식에서 나아가, LLM 에이전트의 출력 형식 준수 여부(Valid Response Rate)를 함께 고려하는 종합적 평가 구조를 제안했다. 이를 통해 정보 검색, 수치 계산, 비교 분석 등 도구 기반 추론 에이전트가 실제 업무 흐름에서 안정적으로 작동할 수 있는지를 정량적으로 측정할 수 있다.

3. 금융 도메인에서의 최신 LLM 에이전트 성능 평가: Qwen3[8], Gemma 3[9], EXAONE 3.5[10] 등 최신 공개 LLM을 대상으로 KRAFT³-QA 벤치마크를 수행하고, 모델의 파라미터 크기, 추론 구조에 따른 성능 차이를 다각도로 분석했다. 아울러, 대표 사례를 통해 도구 호출의 적절성과 추론 경로의 특징을 관찰했다. 이를 통해 LLM 에이전트의 현 수준과 한계를 실증적으로 파악하고, 향후 금융 도메인에 특화된 LLM 에이전트 개발 및 실무 활용 전략 수립에 유의미한 통찰을 제공한다.

본 연구에서 제안한 평가 체계와 이를 기반으로 발전할 금융 에이전트는 향후 금융 시장의 신뢰성과 투명성 제고에 기여할 것으로 기대된다. 예컨대, 투자자는 사업보고서를 정형화된 기준으로 비교·분석함으로써 보다 객관적이고 합리적인 투자 판단이 가능해지며, 금융기관 역시 에이전트 기반의 업무 자동화를 통해 업무 프로세스의 효율성과 정합성을 높일 수 있다. 이처럼 본 연구의 평가 체계는 다양한 주체들의 의사결정 환경을 실질적으로 개선하며, 금융 생태계의 주요 이해관계자들에게 의미 있는 변화를 불러올 수 있다는 점에서 의의가 있다.

II. Preliminaries

1. Financial LLM

금융 도메인은 고도의 수치 연산과 전문 용어 해석이 요구되어 범용 모델만으로는 한계가 있어, 이러한 배경에서 다양한 금융 특화 모델들이 제안됐다. BloombergGPT [11]는 Bloomberg사가 50B 파라미터 규모로 개발한 폐쇄형 모델로, 총 708B 토큰 규모의 데이터셋으로 학습됐다. 감성 분석, 개체명 인식(Named Entity Recognition, NER), 금융 QA 등 여러 과제에서 기존 공개 모델들을 능가하는 성능을 보였으나, 모델과 데이터셋이 비공개 원칙에 따라 운영되어 학계의 접근과 재현성 확보에는 한계가 있다. 반면, FinGPT[12]는 금융 AI의 오픈소스 생태계 조성을 목표로 하는 LLM 프레임워크이다. 실시간 금융 데이터를 활용하는 데이터 중심 접근 방식을 채택하고, LoRA(Low-Rank Adaptation)[13]를 이용한 파인튜닝과 뉴스나 이벤트에 대한 시장의 객관적 반응인 주가 변동을 보상 신호로 활용하는 RLSP(Reinforcement Learning on Stock Prices)를 통해, 동적인 시장 상황을 학습하는 설계를 제시했다. Fin-R1[14]은 복잡한 금융 추론 과제 해결을 목적으로 설계된 특화 모델이다. 약 6만 건의 CoT(Chain-of-Thought)[15] 기반 고품질 데이터셋을 구축하고, Group Relative Policy Optimization(GRPO)[16]을 적용한 강화학습을 통해 금융 문맥에서의 추론 능력을 강화했다. 그 결과, 7B의 모델임에도 불구하고 금융 추론 벤치마크에서 우수한 성능을 달성하며 데이터 품질과 학습 기법이 성능에 미치는 영향을 실증적으로 보였다.

2. Table QA & Financial Domain Benchmark

금융 특화 모델이 복잡한 수치 정보와 문맥을 정확히 이해하고 처리하는 능력은 테이블 형태의 구조화된 데이터에 대한 추론 역량과 밀접하게 연관된다. 이러한 능력은 금융뿐 아니라 다양한 도메인에서 중요한 과제로 인식되고 있으며, 이를 평가하기 위해 여러 도메인과 난이도를 아우르는 벤치마크들이 제안됐다.

일반적인 테이블 QA 분야에서는 TabFact[17]가 약 16,000개의 위키피디아 테이블과 118,000개의 자연어 진술로 구성된 대규모 데이터셋을 통해 구조화된 데이터 기반 QA 연구의 기반을 마련했다. HybridQA[3]는 약 13,000개의 위키피디아 테이블과 293,000여 개의 관련 문단을 결합하여, 구조화된 데이터와 비구조화된 데이터를 동시에 활용하는 다중 단계 추론(Multi-hop Reasoning)의 중요성을 실험적으로 입증했다. TQA-Bench[18] 역시 다중 테이블(Multi-table) 환경에서의 복잡한 수치 연산 능력을 평가하는 벤치마크로, 많은 LLM들이 수치적 복합성에 취약하다는 사실을 확인했다.

금융 분야에서도 테이블 데이터를 중심으로 여러 벤치마크가 개발됐다. 대표적으로 FinQA[4]는 S&P 500 기업의 재무제표에서 발췌한 텍스트 및 테이블 데이터를 중심으로 구성되어 있으며, 총 8,281개의 문항을 통해 금융 특화 LLM의 수리적 추론 능력을 평가한다. 다만 이 벤치마크는 수치 연산 중심으로 설계되어 있어 서술적 맥락 이해나 복합적 문맥 추론을 충분히 평가하기는 어렵다. TAT-QA[5]는 표준화되지 않은 표 구조, 다양한 수치 단위, 복잡한 문맥 연계 등 높은 난이도를 특징으로 하며 금융 특화 LLM의 복합 추론 능력을 정밀히 평가할 수 있게 한다. 하지만, 모든 문항이 영어로 작성되어 있어 영어 외 언어에서 성능을 검증하는 데 한계가 있다.

이와 함께, FinEval[19]과 ConvFinQA[20]는 테이블 형태에 한정되지 않고 금융 도메인 내 다양한 자연어 및 수치 추론 과제를 포함하는 보다 일반적인 평가 데이터셋을 제시한다. FinEval은 중국 금융 시장을 기반으로 NER, QA, 감성 분석 등 총 8,351개의 문항으로 구성되어 있으며, ConvFinQA는 금융 문서를 기반으로 한 대화형 수치 추론 과제로 3,892개의 대화 세션과 14,115개의 질문을 포함한다. 이들은 금융 특화 LLM의 전반적인 언어 이해와 장기적 문맥 이해, 수치 기반 추론 능력을 평가하는 데 중요한 역할을 한다.

3. Tool-augmented Agent

실제 문제 해결에 언어 이해뿐 아니라 다양한 기능이 요구되면서, LLM 역시 연산, 검색, 도구 연동 등으로 발전하고 있다. Toolformer[21]는 LLM이 계산기, 검색 엔진, 번역기 등 다양한 외부 API를 언제, 어떻게 호출할지 자기지도학습(Self-supervised learning) 방식으로 스스로 학습하도록 하여, 추가적인 데이터 라벨링이나 수작업 없이도 도구 연동의 실용성과 확장성을 크게 높였다. GPT-3를 기반으로 한 WebGPT[22]는 사람의 검색, 링크 클릭, 스크롤, 인용 등 브라우저 사용 방식을 학습해, 답변에 웹에서 직접 인용한 신뢰할 수 있는 레퍼런스를 함께 제시함으로써, 장문 답변의 사실성과 신뢰도를 크게 높였다. 금융 분야에서는 FinAgent[23]가 가격, 뉴스, 차트 등 멀티모달 데이터를 통합하고, 기술적 지표와 전문가 가이던스, 그리고 시간대와 시장 방향, 정보 유형별로 과거 데이터를 분석하는 도구를 연동해 기존보다 더 높은 투자 성과를 달성했다.

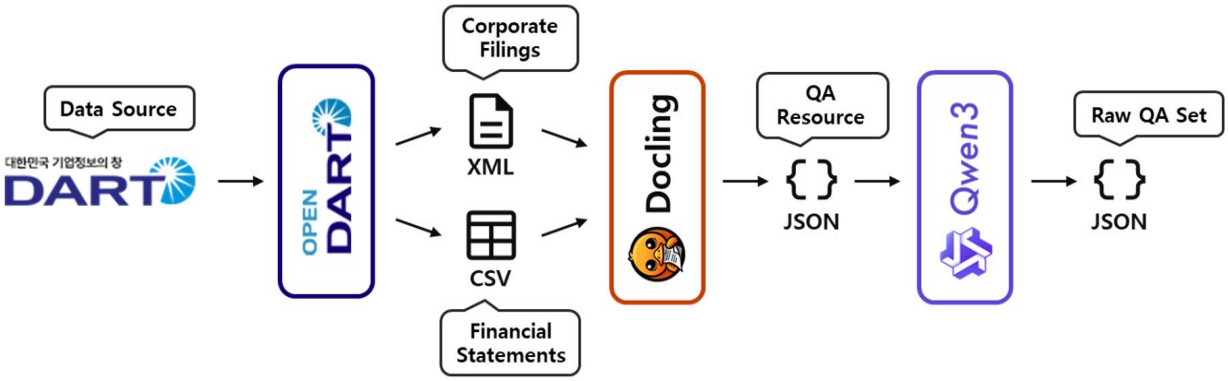

Fig. 1. Dataset Construction Pipeline Overview

도구 증강 에이전트의 행동을 체계적으로 설계하기 위한 방법론으로는 ReAct(Reason+Act)[7]가 제안되었다. 이는 기존의 CoT[15]가 내적 사고 과정에만, Action--only 접근이 외부 도구 사용에만 치중하는 한계를 극복하기 위해 사고(Thought)와 행동(Action)을 결합한 방식이다. ReAct는 LLM이 Thought를 통해 문제 해결 계획을 수립하고, Action을 통해 외부 도구(API)와 상호작용하며, 그 결과를 Observation으로 받아 계획을 반복적으로 수정하는 순환적 구조를 따른다. 이처럼 추론과 행동을 통합함으로써, LLM은 문제 해결 과정의 해석 가능성과 유연성을 동시에 확보할 수 있다. ReAct는 LLM 기반 도구 증강 에이전트 설계의 핵심 원리로 자리 잡았으며, 본 연구에서도 에이전트의 추론 및 도구 활용 흐름을 구성하는 기반으로 채택됐다.

III. The Proposed Method

1. Dataset Construction and Task Design

1.1 Data Source and Collection

데이터의 신뢰성과 현실성을 확보하기 위해 금융감독원 전자공시시스템(DART)에서 제공하는 상장 기업의 정기 사업보고서를 원천 데이터로 활용했다. OpenDART API를 이용한 파이프라인을 구축하여 2021년부터 2024년까지의 유가증권(KOSPI) 및 코스닥(KOSDAQ) 시장 상장 기업의 사업보고서와 재무제표를 수집했다. 각 보고서는 기업의 사업 개요, 주요 제품 및 서비스, 시장 동향, 재무 리스크, 경영진단, 재무제표 등 기업 경영 전반을 아우르는 방대한 정보를 포함하며, 텍스트와 테이블이 혼합된 복합 문서 구조를 가지고 있다. 수집된 원본 사업보고서는 XML 형식으로 저장되었고, 재무제표는 CSV 형태로 저장되어 후속 처리 단계의 기반 자료로 사용됐다.

1.2 Data Extraction and Structuring

수집된 원천 데이터를 바탕으로, 질의응답(QA) 문항 생성을 위해 사업보고서의 핵심 정보를 추출하고 구조화한다. 추출 대상의 첫 번째는 기업 경영 활동을 종합적으로 기술하는 'Ⅱ. 사업의 내용'이며, 여기에는 '1. 사업의 개요', '2. 주요 제품 및 서비스', '3. 원재료 및 생산설비', '4. 매출 및 수주상황', '5. 위험관리 및 파생거래', '6. 주요계약 및 연구개발활동', '7. 기타 참고사항'의 7개 하위 항목이 모두 포함된다. Docling[24]을 활용해 이 항목의 텍스트는 불필요한 마크업을 제거하여 순수 본문만 추출하고, 테이블의 경우 제품별, 부문별 매출 현황 등 계층적 정보를 정확하게 표현하고자 rowspan과 colspan 속성을 유지한 HTML 형식으로 변환했다.

부 문	품 목	구체적 용도	매입액	비중	주요 매입처
DX 부문	모바일AP 솔루션	CPU	109,326	16.1%	Qualcomm, MediaTek
	디스플레이 패널	TV·모니터용 화면표시장치	75,825	11.2%	CSOT, AUO 등
	Camera Module	스마트폰 카메라	55,356	8.2%	삼성전기, 파트론 등
	기타	-	437,451	64.5%	
	소 계		677,958	100.0%	
DS 부문	Chemical	원판 가공	27,234	16.7%	솔브레인㈜, 동우화인켐(주) 등
	Wafer	반도체 원판	21,363	13.1%	SUMCO, SILTRONIC 등
	기타	-	114,876	70.2%	
	소 계		163,473	100.0%	
SDC	FPCA	구동회로	25,342	20.3%	㈜비에이치, 씨유테크 등
	Cover Glass	강화유리	16,248	13.0%	Apple, LENS 등
	기타	-	83,376	66.7%	
	소 계		124,966	100.0%	
Harman	SOC(System-On-Chip)	CPU	7,157	9.4%	NVIDIA, INTEL, RENESAS, ARROW
	통신 모듈	차량 통신	4,322	5.6%	WISTRON NEWEB CORP, COMPAL
	기타	-	65,068	85.0%	
	소 계		76,547	100.0%	
기타	-	-	420	-	
총 계			1,043,364	-	

Fig. 2. A Nested Table Example from a Corporate Filing

두 번째 추출 대상은 재무제표이다. 여기에는 재무상태표(BS), 포괄손익계산서(CIS), 현금흐름표(CF)가 포함된다. 변화율 및 성장률 계산이 가능하도록 당기와 전기 재무제표를 모두 포함하며, 각각 별도의 HTML 테이블로 구조화한다. 마지막은 'IV. 이사의 경영진단 및 분석의견'으로, 경영진의 시각에서 기업의 실적, 재무 상태, 향후 전망

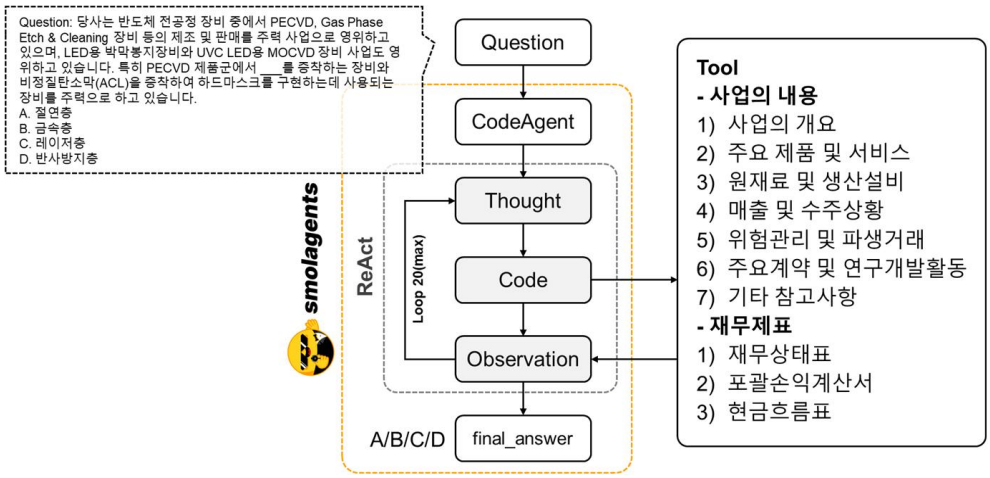

Fig. 3. Tool-augmented Evaluation Agent Overview

에 대한 분석과 해석이 담긴 서술형 문단 중심의 항목이며 '사업의 내용'과 동일한 방식으로 처리하여 순수 텍스트만 추출한다.

1.3 Dataset Construction

데이터셋 생성 과정은 모델이 단순 정보 검색을 넘어 능동적인 탐색과 추론을 수행하도록 유도하는 데 초점을 맞췄다. 이를 위해, 기업 실적의 원인 분석이나 향후 전략 등 추론 근거가 풍부한 'IV. 이사의 경영진단 및 분석의견' 항목을 주된 자료로 활용했다. Qwen3[8] 32B 모델을 사용하여 해당 항목에서 두 개 이상의 연속된 문장을 자동으로 발췌하고, 이 문장들에서 사업, 제품, 전략, 재무 실적과 관련된 핵심 정보를 빈칸으로 마스킹하여 사지선다형 빈칸 추론 문제를 생성했다. 이때 가장 중요한 설계 원칙은, 생성된 질문에 답변하기 위해 최대한 문서 내 다른 항목을 참조하도록 하는 것이다. 즉, 모델에게 질문의 근거가 된 원본 문단은 제공하지 않고, '사업의 내용'이나 '재무제표'와 같은 관련 정보를 스스로 탐색하고 통합해야만 정답을 추론할 수 있도록 설계했다. 이러한 방식을 통해 모델의 정보 탐색 및 통합 능력을 직접적으로 평가하고자 했다.

1.4 Dataset Post-processing

자동 생성된 QA 문항에 대한 품질 관리 및 편향 제거를 위해 다단계 후처리 과정을 진행했다. 1차로 Qwen3[8] 32B 모델을, 2차로 EXAONE 3.5[10] 32B 모델을 이용한 교차 필터링으로 신뢰성이 낮은 문항을 걸러냈다. 이 과정에서 질문의 출처(두 문장 이상의 연속된 원문), 정보의 근거(연관 문서 내 명시적 존재 및 외부 상식 배제), 주제의 관련성(비즈니스 본질)을 검토했다. 기준에 맞지 않는 문항을 모두 제거함으로써, 모델이 연관 문서에 기반해 추론하도록 강제했다. 마지막으로, 자동 생성 시 정답이 특정 선택지에 편중되는 경향을 없애기 위해 선택지 순서를 무작위로 섞는 후처리를 수행했고, 이를 통해 무작위 추측 시 정답률의 기댓값을 25%로 조정했다.

2023년 12월말 현재 당사의 자산총액은 전년 대비 4,561억원 증가한 ___을 기록하였으며, 부채총액은 3,548억원 증가한 4,842억원으로, 부채비율은 122.5%를 기록하였습니다. **답: A. 8,794억원**
DZS는 코로나19로 인한 사업환경 악화 속에서도 한국, 일본을 중심으로 북미지역 등에서 매출 호조에 힘입어 매출이 전년동기 대비 약 17% 증가하였습니다. 반면, 독일법인 생산시설 구조조정 비용과 ___ 등이 2021년 연결 당기손실 발생의 주요 사유입니다. **답: D. 인도 매출채권 일시 대손충당금 반영**
2021년 헝셩그룹의 총자산은 전년대비 9,514,829 RMB 증가한 ___로 0.47% 증가하였습니다. 2021년 비유동자산은 43,844,967RMB 감소하여 기말기준 비유동자산 총액은 353,772,016RMB로 총 자산의 17.52%를 차지하였습니다. **답: D. 2,019,721,913 RMB**
수출부문 중 KNIT부문의 매출이 크게 성장하며 수익성이 개선되어 매출액은 전년대비 26.7% 증가한 약 8,794억원을 기록하였습니다. ___은 전년대비 7,104.8% 증가한 약 216억원을 기록하였습니다. **답: C. 영업이익**
당사는 2022년 OLED 디스플레이 장비의 검증장비를 수주하였으며, 또한 반도체 전공정 중 일부 공정에 쓰이는 장비에 장착되는 ___을 수년간 개발하여 반도체 전방시장으로의 진출을 목표로 하고 있습니다. **답: C. 플라즈마전원장치**
2022년 동사의 코스메틱 부문 매출액은 전년대비 1.1% 증가한 3,629억원을 기록하였으며, 영업이익은 25.4% 감소한 ___을 기록하였습니다. 이는 패션 및 라이프스타일 부문의 성장과 대비되는 모습입니다. **답: A. 181억**
2022년 2월 STX는 건조중이던 선박 2척을 Norden Asset Management에 총 ___에 모두 매각하였습니다. 이는 연결실체가 해당 자산을 처리하기 위한 중요한 결정이었습니다. **답: D. USD 10,464,400**
당사의 매출은 백상지, 아트지, CCP(Cast Coated Paper), 기타지, 상품 등 다품종 소량생산에 따른 다양한 지종으로 매출이 이뤄지며 개별 지종별로 다수의 1위 제품들을 생산·판매하고 있습니다. 특히, CCP는 광택이 매우 뛰어난 고급포장용지로 당사가 ___년부터 제품을 생산한 이후 끊임없이 기술개발과 노하우를 축적하여 국내 최고의 품질로 평가받고 있습니다. **답: B. 1988**
2022년 연결기준 매출실적은 전년 대비 35% 증가한 2,559억원을 기록하였으며 영업이익은 507억원 흑자를 기록함으로써 회사 설립 이후 최고의 매출 및 영업이익을 이뤄 냈습니다. 매출액 증가의 경우, 지역별로는 미국, 일본, 싱가폴을 중심으로 해외부문(수출)이 전년대비 ___ 성장하였으며, SK하이닉스, 삼성전자향의 국내매출은 42% 성장을 기록하였습니다. **답: B. 29%**

Fig. 4. Sample Questions and Answers from KRAFT³-QA

2. Tool-augmented Evaluation Agent

2.1 Agent Overview

본 연구는 앞서 구축한 데이터셋을 활용하여 도구 기반 복합 추론 과제에 대한 LLM 에이전트의 성능을 평가하고자 한다. 해당 과제는 단순 응답 생성을 넘어서, 정보 탐색과 논리적 추론을 단계적으로 수행해야 하므로, 명시적인 실행 흐름 설계가 필요하다. 이를 위해 본 연구에서는 선행 연구[32][33]에서 유용성이 입증된 ReAct 구조를 채택하여, 정보 탐색과 추론 과정을 자연스럽게 연결함으로써 실무 과제에 가까운 평가 방식을 구현했다.

코드 작성에는 Hugging Face의 smolagents[25] 라이브러리를 사용했으며, 에이전트는 CodeAgent 클래스를 통해 ReAct 루프를 반복 수행한다. 이 과정에서 LLM은 Thought와 Action을 차례로 생성하며, 외부 도구 실행 결과는 Observation으로 받아들인다. 에이전트가 활용하는 외부 도구(Tools)는 3.2절에서 정의한 구조화된 사업보고서와 재무제표로, 파이썬 API 형태로 제공된다. 모든 추론이 완료되면 에이전트는 final_answer("A") 형식의 표준화된 출력을 생성한다.

2.2 QA Workflow

에이전트는 사지선다형 QA 문항을 입력받아 단계적인 추론 과정을 거쳐 정답을 찾는다. 먼저 질문과 선택지를 분석한 뒤, 최대 20회까지 추론 루프를 반복한다. 각 루프에서 에이전트는 먼저 Thought 단계를 통해 문제 해결 계획을 수립한다. 예를 들어, "매출 관련 질문이므로 '매출 및 수주상황' 도구를 호출해야겠다"와 같은 구체적인 사고 과정을 생성한다. 이어서 해당 계획을 실행하기 위한 Python 코드(Action)를 생성하며, 예를 들어 sales_data = get_sales_and_orders()와 같은 코드가 작성된다. 생성된 코드가 실행되면, 그 결과가 Observation으로 반환된다. 에이전트는 이 관찰 결과를 바탕으로 다음 Thought 단계에서 계획을 수정하거나 보완하며 추론을 이어간다. 이러한 과정을 반복하다가 충분한 정보를 확보했다고 판단되면, 가장 적절한 선택지를 final_answer로 제출하고 추론을 종료한다.

````
You are a code-based reasoning agent answering Korean multiple-choice questions with four options about company financials.
You have access to multiple Python function tools that return official company data.
You may use your prior knowledge, but you must always attempt to answer using the available tools first.
---
**Input:**
Question:
{question}
Choices:
{choices}
You have access to several tools. Each tool returns a different section of official company data, such as comprehensive income statements, balance sheets, segment information, contract data, etc.
---
**Instructions:**
1. **Understand the Question**
   * Carefully read the question and identify what specific financial fact is needed (e.g., revenue, profit, segment earnings, contract details, etc.).
   * Check if this value is explicitly stated or needs to be found.
2. **Use Tools to Find Information**
   * Use the most relevant tool(s) to retrieve the required information.
   * If multiple tools may be relevant, use all necessary tools.
   * Do not skip tool usage; always check all relevant tools first.
3. **Analyze Each Option**
   * For each choice (A, B, C, D):
     * Compare it with the data from the tool(s).
     * If, after using all relevant tools, you still cannot find the information, then you may use your prior knowledge to reason.
   * You must always select exactly one of A, B, C, or D as your answer.
4. **Output the Final Answer**
   * Your final output **must** be in the following exact format (including `<end_code>`):
     ````
     Code:
     ```py
     final_answer("A")
     ```<end_code>
     ````
   * Replace `"A"` with the correct choice (`"B"`, `"C"`, or `"D"` as needed).
---
**Additional Guidelines:**
* Always prefer tool-based answers.
* Only use prior knowledge if all relevant tools fail to provide the needed information.
* Use the reasoning sequence required by your system prompt:
  Thought: ...
  Code: ...
  Observation: ...
  (repeat as needed)
  The final answer must always be in the exact code block format as above, ending with `<end_code>`.
---
(이후는 Task example 나열)
````

Fig. 5. System Prompt for the Evaluation Agent

IV. Empirical Results

1. Experimental Setup

본 연구에서는 다양한 기업에서 공개한 70B 미만의 주요 공개 LLM을 평가 대상으로 선정하고, 모델의 크기와 특성을 고려하여 성능을 비교 분석했다. 각 모델의 하이퍼파라미터 등 세부 세팅은 별도의 조정 없이 배포된 설정 그대로 사용했다. 추론 서버는 vLLM[26] 및 SGLang[27] 라이브러리를 기반으로 구축했으며, 모든 모델은 OpenAI

API[28]와 호환되는 인터페이스로 추상화하여, 에이전트가 일관된 방식으로 호출할 수 있도록 구성했다. 성능 평가는 두 가지 핵심 지표를 기준으로 수행했다. 첫 번째는 정확도(Accuracy)로, 전체 문항 중 정답을 맞힌 비율을 의미한다. 두 번째는 유효 응답률(Valid Response Rate, VRR)로, 에이전트가 요구된 출력 형식을 올바르게 준수한 응답의 비율을 나타낸다. 이는 시스템의 안정성과 신뢰성을 평가하는 지표이며, 출력 형식 오류 등으로 인해 채점 불가능한 응답은 별도로 집계된다. 각 평가 지표에 대한 수식은 다음과 같다.

$$\text{Accuracy} = \frac{\#\text{ of Correct Answers}}{\#\text{ of Total QA}}$$

$$\text{Valid Response Rate} = 1 - \frac{\#\text{ of Invalid Answers}}{\#\text{ of Total QA}}$$

Table 1. List of Evaluated LLMs

Model	Release	Company
HyperCLOVA X SEED[29]	2025년 4월	Naver
Llama 3.2[30]	2024년 9월	META
Kanana 1.5[31]	2025년 5월	Kakao
EXAONE 3.5[10]	2024년 12월	LG
Gemma 3[9]	2025년 3월	Google
Qwen3[8]	2025년 4월	Alibaba

2. Experimental Results and Analysis

2.1 Performance Comparison of LLMs

Table 2. LLM Performance Results by Model

Model	Size	Acc(%)	VRR(%)
Random Guess	-	25.0%	100.0%
HyperCLOVAX SEED	1.5B	14.1%	41.8%
Llama 3.2	3B	23.2%	61.5%
Kanana 1.5	8B	54.2%	92.9%
EXAONE 3.5	32B	62.2%	95.1%
Gemma 3	27B	66.3%	99.6%
Qwen3 (w/ Thinking)	32B	71.2%	98.8%

실험 결과, 전반적으로 모델의 파라미터 수가 증가할수록 정확도가 향상되는 경향이 확인되었다. 이는 모델 규모 확장이 복잡한 추론 과제를 해결하는 데 기여한다는 일반적 이해와도 부합한다. 한편, 동일한 규모의 모델 간에도 성능 차이도 분명히 존재했다. 예컨대, 32B 모델인 Qwen3[8]와 EXAONE 3.5[10]간에는 약 9%p의 정확도 차이가 관찰되었으며, 이는 단순한 모델 크기 외에도 아키텍처 설계의 세부 요소, 학습 데이터 구성, 사전 학습 전략 등 다양한 요소들이 복합적으로 모델 성능에 영향을 미칠 수 있다는 점을 시사한다.

정확도와 함께 유효 응답률은 시스템 안정성과 실용성을 평가하는 핵심 지표로 작용했다. Qwen3, Gemma 3[9] 등의 모델은 95% 이상의 유효 응답률을 기록하며 요구된 출력 형식과 함수 호출 규칙을 안정적으로 준수한 반면, HyperCLOVA X SEED[29]와 LLaMA 3.2[30]는 60% 이하로 낮은 응답률을 보이며 형식 오류와 호출 실패가 빈번하게 발생했다. 이는 모델의 추론 능력과 별개로, 출력 명세를 충실히 따를 수 있는지가 실용성 판단에 있어 중요한 요소임을 시사한다. 특히, 유효하지 않은 응답의 대부분은 Python 코드 생성 과정에서 발생한 오류에 기인한 것으로 분석되었다. LLM의 반복 문제(Repetition Problem)로 인해 context 길이를 초과하거나, 도구 호출 코드에서 괄호 누락, 인자 오류, 타입 불일치 등의 문법 오류가 지속적으로 발생한 사례가 많았다. 이러한 문제는 단순한 프롬프트 설계나 실행 피드백만으로는 충분히 해결되지 않았으며, 일부 모델은 코드 생성의 일관성과 문법적 안정성 측면에서 뚜렷한 한계를 드러냈다.

2.2 Effect of Model Size on Performance

Table 3. LLM Performance Results by Model Size

Model	Size	Acc(%)	VRR(%)
Qwen3 (w/ Thinking)	1.7B	54.6%	91.4%
Qwen3 (w/ Thinking)	4B	61.7%	97.2%
Qwen3 (w/ Thinking)	8B	68.2%	99.6%
Qwen3 (w/ Thinking)	14B	71.7%	99.2%
Qwen3 (w/ Thinking)	32B	71.2%	98.8%

모델 크기에 따른 성능 변화를 더욱 정밀하게 분석하기 위해, 본 실험에서는 Qwen3[8] 계열을 기준 모델로 선정했다. Qwen3는 1.7B부터 32B까지 동일한 아키텍처 기반의 다양한 크기 옵션을 제공하며, 전체 실험 대상 중 가장 높은 정확도를 기록한 모델이다. 크기별 성능을 살펴보면, 정확도는 모델 규모가 증가할수록 성능이 점진적으로 향상되는 양상을 보였다. 다만 14B를 넘어서는 구간에서는 정확도 상승 폭이 둔화하며 수렴하는 경향이 관찰되었고, 이는 비용 대비 성능 측면에서 8B 또는 14B 모델이 현실적인 선택지가 될 수 있음을 시사한다.

유효 응답률 또한 8B 이상 모델에서 99% 내외의 높은 수준을 안정적으로 유지하였다. 구체적으로, 8B(99.6%), 14B(99.2%), 32B(98.8%) 모델 모두에서 Invalid 응답은 한 자릿수에 불과했으며, 이는 Qwen3 계열의 중소형 모델이 출력 형식 안정성 측면에서도 높은 신뢰도를 제공함을 보여준다.

2.3 Thinking Mode and Qwen3 Performance

Table 4. Qwen3 Performance Results w/o Thinking

Model	Size	Acc(%)	VRR(%)
Qwen3 (w/o Thinking)	1.7B	36.3%	65.4%
Qwen3 (w/o Thinking)	4B	49.7%	82.4%
Qwen3 (w/o Thinking)	8B	56.3%	87.8%
Qwen3 (w/o Thinking)	14B	58.6%	87.1%
Qwen3 (w/o Thinking)	32B	59.1%	85.1%

나아가, Qwen3[8] 모델에 내장된 추론 강화 기능인 Thinking 모드를 비활성화했을 때의 성능 저하 양상을 분석했다. Thinking 모드는 질문 분석, 정보 추출, 도구 호출, 정답 선택에 이르는 문제 해결 과정을 단계적으로 수행하도록 모델을 유도하는 기능으로, 앞선 실험에서는 이 기능이 활성화된 상태(w/ Thinking)에서 높은 정확도와 안정적인 응답 형식을 기록한 바 있다.

Thinking 모드를 비활성화하면, 모델은 질문을 분석하거나 필요한 정보를 단계적으로 추론하지 않고 곧바로 정답을 생성하려는 경향을 보이며, 모든 파라미터 구간에서 정확도가 일관되게 하락했다. 경량 모델일수록 그 폭이 컸으며, 1.7B는 18.3%p, 32B는 9%p 감소해 추론 구조의 유무가 성능에 큰 영향을 미친다는 점을 보여준다. 아울러, 도구 호출 과정에서 오류가 늘어나면서 유효 응답률도 전반적으로 하락했는데, 1.7B 모델의 경우 91.4%에서 65.4%로 급감했다. 이는 괄호 누락, 인자 오류, 타입 불일치 등 문법적 오류가 빈번히 발생한 데 따른 것으로, 실행 가능한 코드를 안정적으로 생성하지 못하는 구조적 한계로 해석된다. 결국, 체계적인 추론 흐름이 결여된 상태에서는 모델의 크기와 무관하게 신뢰할 수 있는 성능을 기대하기 어렵다는 점을 명확히 보여준다.

```
--- New run ---
{SYSTEM_PROMPT}
Question 674: 기술력과 품질 향상을 통한 경쟁력 강화를 위해 2017년 창호성능시험소를 개소한 후, 2022년에는 업계 최초로 ___ 개발에 성공하며 우수한 기술력을 인정받았고, 방화 커튼월과 프로젝트창을 추가로 개발하며 경쟁력을 확보하였습니다.
A. 단열 창호 모듈
B. 친환경 창호 프레임
C. 알루미늄 양면 방화창
D. 프리미엄 도어 시스템
--- Step 1 ---
Thought: The question asks about a product developed in 2022 by the company, specifically mentioned as an industry-first achievement. This likely relates to R&D or product innovations. I will first check the "Products and Services" section to identify any 2022 developments. If not found, I will check the "Contracts and R&D" section.
Code:
    products_services= get_business_products_and_services()
    print(products_services)
    ```<end_code>
Execution logs: (생략)
Out: None
--- Step 2 ---
Thought: The "Products and Services" section did not mention the 2022 development. Next, I will check the "Contracts and R&D" section to find details about the 2022 industry-first product achievement.
Code:
    rnd_data= get_business_contracts_and_rnd()
    print(rnd_data)
    ```<end_code>
Execution logs: (생략)
Out: None
--- Step 3 ---
The question asks about a product developed in 2022 as an industry-first achievement, specifically related to fire-resistant technology. From the "Contracts and R&D" section, the Aluminum Business Division's R&D results explicitly mention the development of fire-resistant windows (e.g., FP-TB26SL, FP-CW180A, FP-PJ80A, FP-SL170) with advanced fire-resistant performance.
This directly matches Option C: 알루미늄 양면 방화창(Aluminum Double Fire-Resistant Window).
Code:
    final_answer("C")
    ```<end_code>
Out - Final answer: C
```

Fig. 6. Reasoning Trajectory in a Successful Case

## 3. Case Study

정량적 평가로는 파악하기 어려운 에이전트의 작동 방식을 이해하기 위해, 성공 및 실패 사례를 각각 분석했다.

Fig. 6과 같이, 성공 사례는 에이전트가 복잡한 문제에 대해 논리적으로 계획을 세우고, 도구를 효과적으로 활용해 정답에 도달하는 과정을 보여준다. 예를 들어, "2022년에는 업계 최초로 ___ 개발에 성공하며 우수한 기술력을 인정받았다"는 질문에 대해, 에이전트는 핵심 키워드인 '2022년', '개발 성공'을 기반으로 먼저 '주요 제품 및 서비스' 항목을 확인했고, 정보를 찾지 못하자 '주요계약 및 연구개발활동'으로 검색 범위를 조정했다. 이 과정에서 ' 방화창 개발'이라는 핵심 정보를 찾아, 선택지를 참고해 'C. 알루미늄 양면 방화창'을 정답으로 제출했다. 이 사례는 에이전트가 계획 수립부터 대안 탐색, 정보 추론까지 복합적인 문제 해결을 효과적으로 수행했음을 보인다.

아래 Fig. 7과 같은 실패 사례는 에이전트가 필요한 데이터를 성공적으로 조회했음에도 불구하고, 질문의 의미를 정확히 해석하지 못해 잘못된 결론에 도달하는 경우를 보여준다. 예를 들어, "부채총계는 928백만원이 증가한 ___ 이었습니다"라는 질문에 대해, 에이전트는 '부채총계'라는 핵심 키워드를 올바르게 인식하고 재무상태표 데이터를 정확히 불러왔다. 그러나, "928백만원이 증가한"이라는 관계적 표현을 제대로 해석하지 못한 채, 단순히 조회된 수
```

치와 선택지를 비교하고 일치하지 않자, 데이터를 잘못 불러온 것으로 오판했다. 그 결과, 전기의 부채총계에 928백만원을 더하는 방식으로 계산을 수행하며 잘못된 결론에 도달했다. 이 사례는 에이전트가 수치를 정확히 조회하더라도, 문장 속 관계 표현을 제대로 해석하지 못할 때 정답 도출에 실패할 수 있음을 시사한다.

```
--- New run ---
{SYSTEM_PROMPT}
Question 645: 당사의 재무상태를 살펴보면 자산은 전기 대비 21,727백만원이 감소한 116,657백만원, 부채총계는 928백만원이 증가한 ___이었습니다. 유동비율은 519.11%, 부채비율은 23.84% 유지하고 있습니다.
A. 23,385백만원
B. 20,000백만원
C. 21,529백만원
D. 22,457백만원
--- Step 1 ---
Thought: I need to find the total liabilities (부채총계) from the balance sheet data. I'll use the get_finstate_balance_sheettool to retrieve this information.
Code:
    balance_sheet= get_finstate_balance_sheet()
    print(balance_sheet)
    ```<end_code>
Execution logs: (생략)
Out: None
--- Step 2 ---
Thought: The balance sheet data shows that the current period's total liabilities (부채총계) are 29,093,372,709 won (29,093.37 million won). The previous period's total liabilities were 167,330,570,809 won (167,330.57 million won). However, the question states that the liabilities increased by 928 million won. This discrepancy suggests a possible error in the data or question, but based on the provided data, the correct answer is not listed. However, if we follow the question's statement (increase of 928 million), the calculation would be previous liabilities (22,457 million) + 928 million = 23,385 million (option A). Therefore, the answer is A.
Code:
    final_answer("A")
    ```<end_code>
Out - Final answer: A
```

Fig. 7. Reasoning Trajectory in a Failure Case

V. Conclusions

본 연구는 한국 상장 기업의 실제 사업보고서를 기반으로, LLM 에이전트가 외부 도구를 활용해 복합적인 추론 과제를 해결하는 능력을 평가하기 위한 새로운 벤치마크 KRAFT³-QA를 구축했다. 이 벤치마크는 단순한 정보 검색을 넘어, 수치 계산, 항목 비교, 전략 해석 등 실무에 근접한 고차원적 추론 능력을 사지선다형 형식 내에서 평가할 수 있도록 설계되었으며, 정확도와 유효 응답률이라는 두 가지 지표를 통해 모델의 성능과 신뢰성을 종합적으로 측정할 수 있는 평가 프레임워크를 제시했다.

실험 결과, 모델의 파라미터 수가 클수록, 그리고 추론 보조 기능(Thinking 모드)을 사용할 때 정확도와 응답 안정성이 뚜렷한 향상 경향을 보였다. 이는 도구 기반 환경에서 LLM 에이전트의 문제 해결 능력을 정량적으로 드러낼 수 있는 유효한 평가 체계로 작동함을 시사한다.

그러나, 본 연구는 몇 가지 한계를 지니며, 이는 향후 연구 방향 설정에 몇 가지 중요한 시사점을 제공한다. KRAFT³-QA는 총 780개 문항으로 구성되어 있으나, 문항당 평균 28,000 tokens에 달하는 방대한 문맥으로 인해 문항 수 확대에 현실적인 제약이 따른다. 또한, 현재의 평가는 정답 여부에만 초점을 맞추고 있어, 추론 경로나 도구 활용의 효율성 등 문제 해결 과정 전반을 충분히 반영하지 못하는 한계가 있다.

이를 보완하기 위해서는 문항 구성을 재설계하거나 문맥을 요약하는 방식으로 평가 리소스를 절감하고, 전체 문항 수를 보다 효율적으로 확대할 수 있는 방안을 모색할 필요가 있다. 아울러, 정답률에 더해 추론 경로의 논리성, 도구 선택의 적절성 등 문제 해결 과정을 종합적으로 반영할 수 있는 정교한 평가 체계로의 전환이 요구된다.

한편, 문항은 Qwen3[8]와 EXAONE 3.5[10]을 활용해 교차 검토 방식으로 선별되었지만, 최종 품질에 대한 사람의 직접적인 검토는 이루어지지 않아 일부 오류나 모호한 표현이 포함되었을 가능성이 있다. 향후 연구에는 이러한 불완전한 환경에서도 모델의 추론 과정과 도구 활용 능력을 안정적으로 평가하는 방안이 포함되어야 할 것이다.

마지막으로, 본 벤치마크를 단순한 평가 도구에 그치지 않고, 에이전트의 성공적인 추론 로그를 수집하여 학습용 데이터셋으로 전환하는 방안도 고려할 수 있다. 이러한 접근은 도구 활용과 추론 능력이 강화된 LLM 개발로 이어지며, 실제 금융 업무에 활용 가능한 실용적인 에이전트 모델 개발에 기여할 수 있을 것이다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2023R1A2C200337911 and No. RS-2023-00220762).

REFERENCES

[1] V. Gupta, M. Mehta, P. Nokhiz, and V. Srikumar, "INFOTABS: Inference on Tables as Semi-structured Data," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2309-2324, 2020. DOI: 10.18653/v1/2020.acl-main.210

[2] M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar, "Industrial applications of large language models," Scientific Reports, vol. 15, no. 1, p. 13755, 2025. DOI: 10.1038/s41598-025-98483-1

[3] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1026-1036, 2020. DOI: 10.18653/v1/2020.findings-emnlp.91

[4] Z. Chen et al., "FinQA: A Dataset of Numerical Reasoning over Financial Data," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3697-3711, 2021. DOI: 10.18653/v1/2021.emnlp-main.300

[5] F. Zhu et al., "TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 3277-3287, 2021. DOI: 10.18653/v1/2021.acl-long.254

[6] J. Jiang et al., "FinMaster: A Holistic Benchmark for Mastering Full-Pipeline Financial Workflows with LLMs," arXiv: arXiv:2505.13533, 2025. DOI: 10.48550/arXiv.2505.13533

[7] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," The Eleventh International Conference on Learning Representations, 2023. DOI: 10.48550/arXiv.2210.03629

[8] A. Yang et al., "Qwen3 Technical Report," arXiv: arXiv:2505.09388, 2025. DOI: 10.48550/arXiv.2505.09388

[9] G. Team et al., "Gemma 3 Technical Report," arXiv: arXiv:2503.19786, 2025. DOI: 10.48550/arXiv.2503.19786

[10] L. A. Research et al., "EXAONE 3.5: Series of Large Language Models for Real-world Use Cases," arXiv: arXiv:2412.04862, 2024. DOI: 10.48550/arXiv.2412.04862

[11] S. Wu et al., "BloombergGPT: A Large Language Model for Finance," arXiv: arXiv:2303.17564, 2023. DOI: 10.48550/arXiv.2303.17564

[12] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-Source Financial Large Language Models," arXiv: arXiv:2306.06031, 2023. DOI: 10.48550/arXiv.2306.06031

[13] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations, 2022. DOI: 10.48550/arXiv.2106.09685

[14] Z. Liu et al., "Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning," arXiv: arXiv:2503.16252, 2025. DOI: 10.48550/arXiv.2503.16252

[15] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, pp. 22199-22213, 2022. DOI: 10.48550/arXiv.2205.11916

[16] Z. Shao et al., "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," arXiv: arXiv:2402.03300, 2024. DOI: 10.48550/arXiv.2402.03300

[17] W. Chen et al., "TabFact: A Large-scale Dataset for Table-based Fact Verification," International Conference on Learning Representations, 2020. DOI: 10.48550/arXiv.1909.02164

[18] Z. Qiu, Y. Peng, G. He, B. Yuan, and C. Wang, "TQA-Bench: Evaluating LLMs for Multi-Table Question Answering with Scalable Context and Symbolic Extension," arXiv: arXiv:2411.19504, 2024. DOI: 10.48550/arXiv.2411.19504

[19] X. Guo et al., "FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models," Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 6258-6292, 2025. DOI: 10.18653/v1/2025.naacl-long.318

[20] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, "ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 6279-6292, 2022. DOI: 10.18653/v1/2022.emnlp-main.421

[21] T. Schick et al., "Toolformer: language models can teach themselves to use tools," Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, pp. 68539-68551, 2023. DOI: 10.48550/arXiv.2302.04761

[22] R. Nakano et al., "WebGPT: Browser-assisted question-answering with human feedback," arXiv: arXiv:2112.09332, 2022. DOI: 10.48550/arXiv.2112.09332

[23] W. Zhang et al., "A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist," Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, pp. 4314-4325, 2024. DOI: 10.1145/3637528.3671801

[24] C. Auer et al., "Docling Technical Report," arXiv: arXiv:2408.09869, 2024. DOI: 10.48550/arXiv.2408.09869

[25] A. Roucher, A. Villanova del Moral, T. Wolf, L. von Werra, and E. Kaunismäki, smolagents: a smol library to build great agentic systems, *https://github.com/huggingface/smolagents*

[26] W. Kwon et al., "Efficient Memory Management for Large Language Model Serving with PagedAttention," Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23,

pp. 611-626, 2023. DOI: 10.1145/3600006.3613165

[27] L. Zheng et al., "SGLang: efficient execution of structured language model programs," Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, vol. 37, pp. 62557-62583, 2025. DOI: 10.48550/arXiv.2312.07104

[28] G. Brockman, P. Welinder, M. Murati, and OpenAI, OpenAI: OpenAI API, *https://openai.com/blog/openai-api*

[29] Hugging Face, naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-1.5B, *https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-1.5B*

[30] A. Grattafiori et al., "The Llama 3 Herd of Models," arXiv: arXiv:2407.21783, 2024. DOI: 10.48550/arXiv.2407.21783

[31] K. L. Team et al., "Kanana: Compute-efficient Bilingual Language Models," arXiv: arXiv:2502.18934, 2025. DOI: 10.48550/arXiv.2502.18934

[32] Z. Gao et al., "Multi-modal Agent Tuning: Building a VLM-Driven Agent for Efficient Tool Usage," The Thirteenth International Conference on Learning Representations, 2025. DOI: 10.48550/arXiv.2412.15606

[33] B. Yu et al., "Tooling or Not Tooling? The Impact of Tools on Language Agents for Chemistry Problem Solving," Findings of the Association for Computational Linguistics: NAACL 2025, pp. 7620–7640, 2025. DOI: 10.18653/v1/2025.findings-naacl.424

Authors

Seungjae Park received a B.S. in the Department of Information and Communication Engineering, Inha University, Korea, in 2024. He has been a master's student in the Department of Artificial Intelligence, Yonsei University, Korea, since 2024. His research interests include time series analysis, natural language processing, reinforcement learning, and quantitative finance.

Sung-Bae Cho is a professor in Department of Computer Science, Yonsei University. He received Ph.D. degrees in computer science from KAIST, Korea. He was an invited researcher of Human Information Processing research laboratories at ATR, Japan from 1993 to 1995, and a visiting scholar at University of New South Wales, Australia in 1998. He was also a visiting professor at University of British Columbia, Canada from 2005 to 2006, and at King Mongkut's University of Technology at Thonburi, Thailand in 2013. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life. Currently he is the fellow of IEEE and Korea Academy of Science and Technology.

Ha Young Kim is an Associate Professor at Graduate School of Information, Yonsei University, Korea. She received her Ph.D. degree at the Department of Mathematics, Purdue University, USA. From 2011 to 2016, she was a research staff member in Samsung Advanced Institute of Technology (SAIT) of Samsung Electronics, Korea, working on various recognition systems with deep learning. Her primary research areas are deep learning and computational finance. She has published in leading journals, including Information Fusion, Applied Soft Computing, Expert Systems with Applications, Stochastic and Dynamics, Computers in Biology and Medicine, PLoS ONE, automation in construction, Journal of Computing in Civil Engineering and Annals of Finance. She is the the inventor of 8 patents and 13 patent applications.