금융보안원
FINANCIAL SECURITY INSTITUTE

2025 금융 AI Challenge : 금융 AI 모델 경쟁

모델 경쟁 부문

TEAM	뛰어

2025.10.23.

문제 설정

2025-01-31 PrecedenceResearch

“금융 AI, 10년 내 10배 이상 폭발적인 성장 전망”

2024-02-05 STARTUPTODAY

”할루시네이션 잡고 신뢰도 높인다” 금융, 범죄, 공공기간 등 RAG 적극 활용

2025-09-20 ZDNET Korea

“정보보호 인력 13만 4300명, 인력 부족 심각”

2023-12-15 연합뉴스

“AI는 급부상하는 취약점…, 미국 정부 금융 시스템에 위험 첫 경고”

문제 설정

01 기존 생성형 AI의 치명적 단점 : 신뢰성 부재

- 할루시네이션(Hallucination): 사실과 다른 정보를 생성하여 오류 발생
- 정보 출처 불분명: 생성된 정보의 근거를 확인할 수 없음.
- 최신 정보 미반영: 내부 데이터나 실시간 시장 상황을 반영하지 못함.

02 급증하는 금융 리스크와 인력의 부족

- 지능화되는 금융 범죄: AI를 악용한 신종 금융 사기 및 사이버 공격이 급증
- 정보보호 전문 인력 부족: 폭발적으로 증가하는 보안 위협에 비해, 인력이 부족
- 과도한 업무 부담: 수많은 데이터를 수동으로 검토하고 분석해야함.

03 방대한 내부 데이터 활용의 한계

- 산재된 정보: 여러 데이터들이 다양한 포맷으로 흩어져 있음.
- 비정형 데이터 분석의 어려움: 다양한 비정형 데이터에 담긴 정보를 활용하지 못함.
- 문맥 이해의 부족: 단순 키워드 검색으로는 복잡한 무서 내의 핵심적인 의미와 맥락을 파악할 수 없음.

01 정확성과 신뢰도를 위한 검증 시스템

- 정밀한 정보 선별(RAG Reranker): 10개의 문서를 추출 한 뒤, Reranker 모듈을 활용
- 엄격한 품질 관리(Score Threshold): 문서의 신뢰도 점수가 0.75 미만일 경우, 사용하지 않음.
- 자기 검증 생성(Iterative Self-Consistency): 5회 반복적으로 생성, 교차 검증하여 도출

02 금융 데이터에 최적화된 지능형 문서 처리

- Preprocessing: 정규식필터링과 plumber를 활용한 추출 및 JSON 형식의 데이터 활용
- Chunking: 긴 문서를 의미와 문맥이 유지되는 최적의 단위로 분할함.
- Vector DB: Cosine Similarity 방식의 FAISS 벡터 DB를 사용하여 사용자 질의와 연관성 높은 정보를 검색.

03 사용자를 위한 편의성과 활용성

- 맞춤형 답변 형식: 객관식, 주관식에 맞는 적절한 프롬프트 제공.
- 명확한 근거 제시: 모든 답변은 내부 문서(>0.75)를 참고하여 작성.
- 보안 인력 학습 및 역량 강화: 반복적인 분석 혹은, 교육자료로 활용 가능.

문제 정의

객관식 + 주관식

금융, 보안 도메인에 특화된 FSKU 평가셋

핵심 포인트

1. 금융, 보안에 따른 도메인 특이지식과 시기성 있는 사실 정보를 정확히 반영
2. 기관령/법령명등 핵심 정보의 손실 없이 자연스럽고 간결한 답변 산출
3. 실용성을 위해 제한된 시간의 추론 안에서 해결

문제 정의

객관식 + 주관식

금융, 보안 도메인에 특화된 FSKU 평가셋

핵심 목표

제한시간 내에
정확성, 명확성, 강건성을
최대한 가지는 것을 목표

1. 금융, 보안에 따른 도메인 특이지식과 시기성 있는 사실 정보를 정확히 반영
2. 기관령/법령명등 핵심 정보의 손실 없이 자연스럽고 간결한 답변 산출
3. 실용성을 위해 제한된 시간의 추론 안에서 해결

Method 소개

1. Base Model 선정
2. RAG 시스템 구현
3. Inference 파이프라인 구현

Base Model 선정 실험

Public LB Score 비교

한국어 이해 성능 리더보드인 Horangi: Korean LLM Leaderboard3,
dnotitia LLM 한국어 리더보드를 참고하여 Zero-shot 성능을 비교

모델	파라미터	Public LB
DAPADA/krx_qwen2.5	7B	0.5063
SEOKDONG/qwen2.5	14B	0.5759
Midm-2.0-Base-Instruct	11.5B	0.6334
EXAONE-4.0	32B	0.6238
Qwen3	14B	0.5940
gemma2	9B	0.5691
DNA-2.0	14B	0.6278

Base Model - Mi:dm의 학습 데이터 셋

K intelligence

Mi:dm(믿음) 2.0 Base

모델이 갖추어야 할 핵심 역량별 데이터셋 구성과
이를 내제하기위한 사후 학습 전략

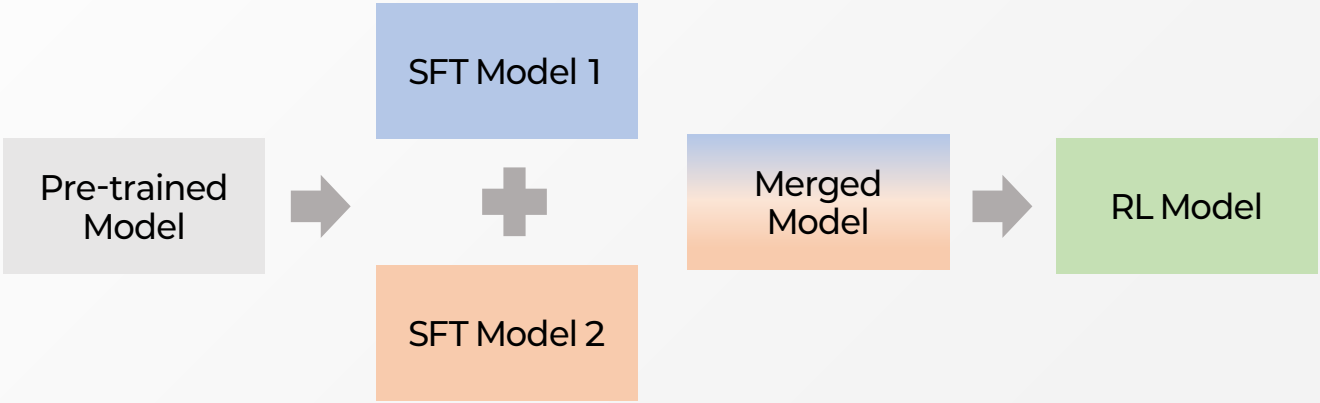
- Instruction-Following
- Reasoning
- RAG
- Agent Ability
- Safety
- Long Context Handling

Base Model - Mi:dm의 학습 전략

K intelligence

Mi:dm(믿음) 2.0 Base

모델이 갖추어야 할 핵심 역량별 데이터셋 구성과
이를 내제하기위한 사후 학습 전략



Base Model - Mi:dm의 벤치마크 성능

- 벤치마크에서의 강점
  - 역사, 법, 정치, 경제, 금융과 같은 전문 지식을 요구하는 고난이도 평가에서 20% 더 큰 규모의 SOTA 공개 모델인 Qwen을 능가하는 성능
  - 한국어 LLM 이해 KMMLU, 한국형 벤치마크 Ko-Sovereign 이외 자체구축 벤치마크 3종 등에서 우수성을 입증

Society & Culture, General Knowlege,and Instruction Following										
Model	Society& Culture					Reasoning			Instruction Following		
	K-Refer*	K-Refre -Hard*	Ko-Sovereign*	HAERAE	Avg.	KMMLU	Ko-Sovereign*	Avg.	Ko-IFEval	Ko-MTBench	Avg.
Qwen3-4B	53.6	42.9	35.8	50.6	45.7	**50.6**	**42.5**	**46.5**	**75.9**	63.0	69.4
Exaone-3.5-2.4B-inst	64.0	**67.1**	**44.4**	61.3	**59.2**	43.5	42.4	43.0	65.4	**74.0**	68.9
Mi:dm 2.0-Mini-inst	**66.4**	61.4	36.7	**70.8**	58.8	45.1	42.4	43.8	73.3	**74.0**	**73.6**
Qwen3-14B	**72.4**	**65.7**	49.8	68.4	**64.1**	**55.4**	**54.7**	55.1	**83.6**	71	77.3
Llama-3.1-8B-inst	43.2	36.4	33.8	49.5	40.7	33.0	36.7	34.8	60.1	57	58.5
Exaone-3.5-7.8B-inst	71.6	69.3	46.9	72.9	65.2	52.6	45.6	49.1	69.1	79.6	74.4
Mi:dm 2.0-Base-inst	**89.6**	**86.4**	**56.3**	**81.5**	**78.4**	**57.3**	**58.0**	**57.7**	82	**89.7**	**85.9**

Base Model - Mi:dm의 토크나이저

한국어 특화 토크나이저

한국어와 영어 모두에 효과적인 Bi-lingual BBPE 방식 기반의 토크나이저 설계

한국어 고유의 문장 구조를 유지하면서도 토크 나이저의 압축 효율을 높임

모델	요약	기사 생성	QA	공공 문서	뉴스	위키	구어체	블로그	논문
믿:음-2.0	2.283	2.1838	2.1115	2.2635	2.2174	2.2943	2.0527	2.2835	2.3518
Llama-3-8B	1.6529	1.5383	1.5504	1.5612	1.5182	1.8789	1.4884	1.6049	1.6490
Gemma-2-9B	1.4683	1.3504	1.3614	1.3669	1.3409	1.6581	1.3214	1.4321	1.4613
GPT-4	1.0661	0.9937	1.0377	0.9823	0.9715	1.3504	0.959	1.0096	1.0129
EXAONE-3.0-7.8B	1.9114	1.7714	1.7291	1.8354	1.7951	1.8959	1.7394	1.8521	1.8655

동일한 문장을 얼마나 적은 수의 토큰으로 표현할 수 있는지에 대한 지표. -> 모델의 학습 효율과 성능에 직접적인 영향을 미침.
따라서 복잡한 문법 구조나 긴 문맥을 포함한 한국어 텍스트에서도 보다 안정적이고 정확한 학습, 추론이 가능.

RAG system - 데이터셋

금융보안 법률 PDF

ISMS-P 안내서 PDF

Trendyol/cybersecurity
주관식 보안 데이터 셋

CyberMetric-10000
객관식 보안 데이터 셋

RAG system - 데이터 전처리

	특징	방법
	제 N 조의 N 항()으로 text 의 형태가 정형화 되어있음	제 N 조의 N 항()을 기준으로 모든 항 분리 후 앞에 법의 이름을 붙임
	안내서로 표, 그림, 아이콘 등 다양한 요소들이 있는 PDF, text 에 noise 가 존재	PDFplumber 를 통해 각 페이지의 텍스트를 추출한 뒤 기호와 필요없는 문자 필터링
	따로 노이즈는 없는 깔끔한 영어 text 데이터 셋	주관식, 객관식 모두 질문 + 정답의 형태로 질문과 정답을 하나의 text 로 이어서 제작

RAG system - Model & DB

Embedding Model

gtemultilingual-base

다국어(영어/한국어)
지원 임베딩 모델

Vector DB

FAISS

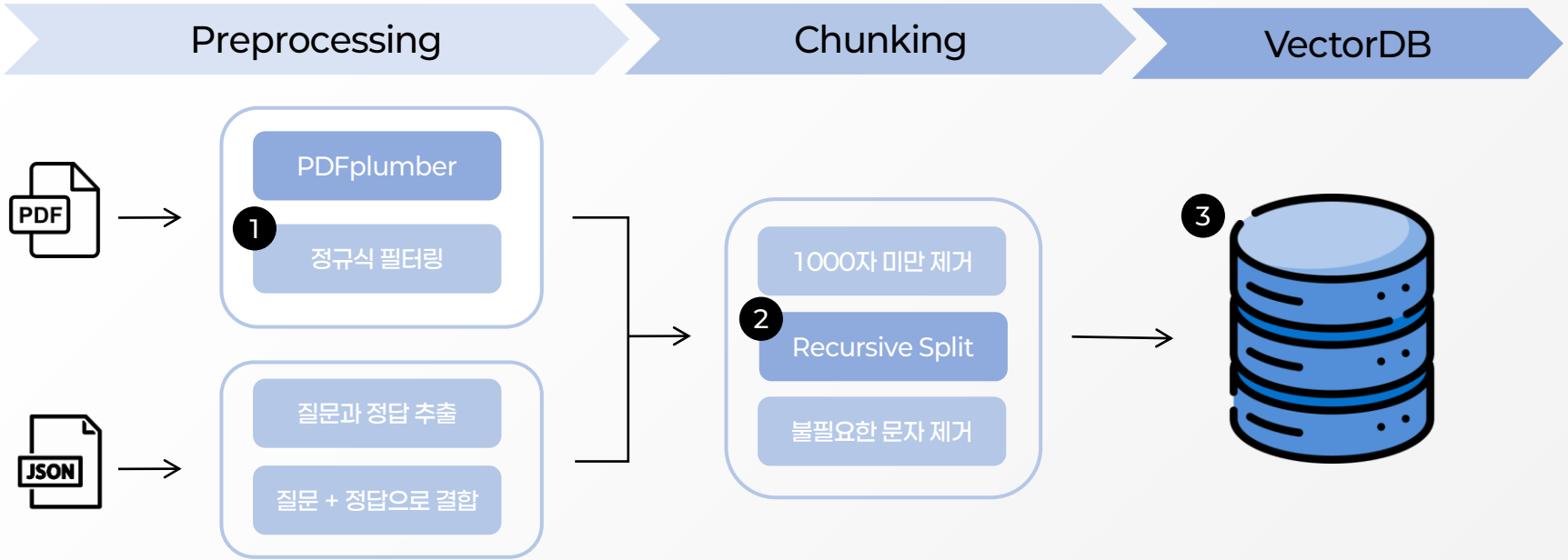
검색 속도가 빠른 백터
DB 라이브러리

Reranker Model

bge-reranker-v2-m3

다국어(영어/한국어)
지원 리랭커 모델

RAG system - 시스템 전체 구조



1

법은 제 N조의 N항 (키워드) 를 기준으로 split 후 앞에 법 이름 붙이기

ISMS-P는 한글,영어,숫자,일부 기호를 제외한 모든 문자 제거

2

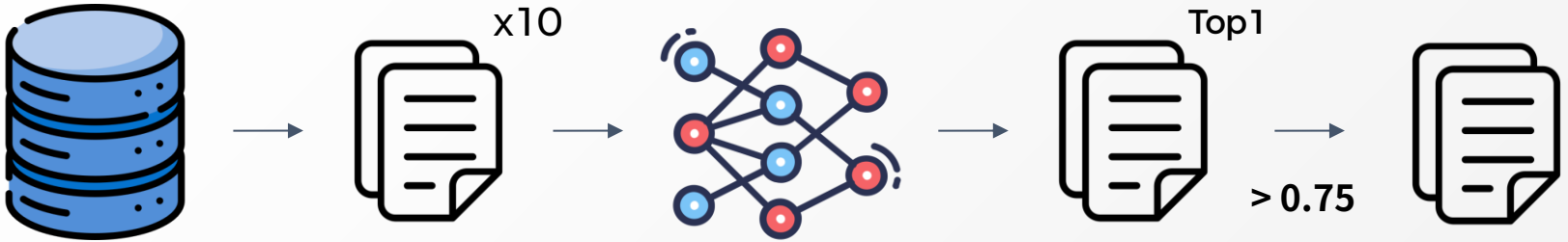
RecursiveCharactorTextSpitter 사용

Chunksplit = 2000,
Overlap = 500,
Separators = [“\n\n”, “\n”, “”, “ “]

3

Distance strategy : Cosine
Search Type : similarity
Search kwargs : {k : 10}

RAG system - Reranker System



1. query 에 대해 retriever 로 나온 10 개의 chunk 에 대해 모두 Reranking

2. 0~1 사이의 score 가 나오고 이 때 가장 높은 score 를 달성한 하나의 chunk 만 사용

3. score threshold 0.75 를 넘지 못한다면 RAG Chunk 를 사용하지 않음

Inference Pipeline - Overview

- 순서도

Question 입력

RAG 데이터 베이스 접근, Retriever

Reranking, Top-1 chunk 추출

기존 Prompt + Top-1 chunk

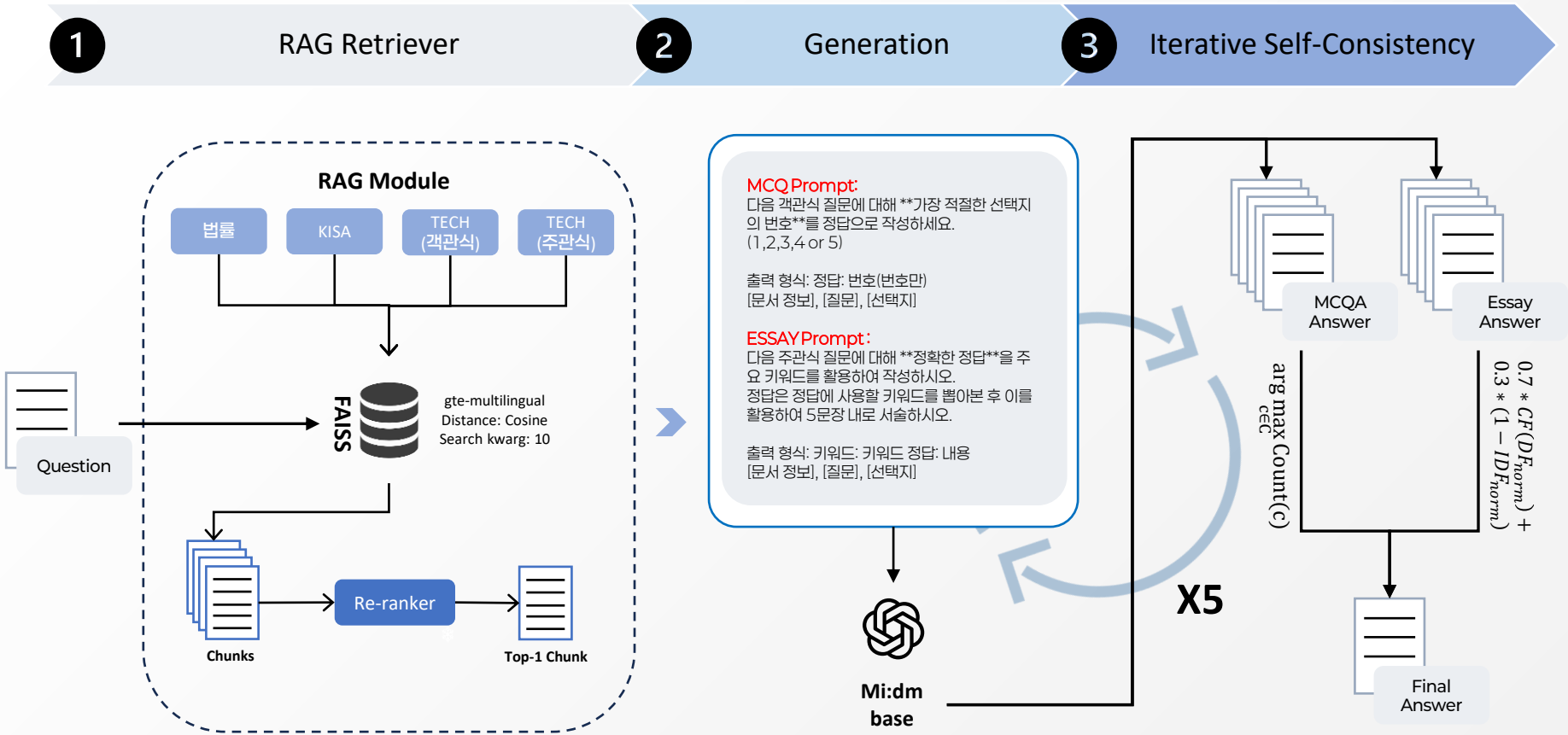
LLM 답변 생성

Self Consistency 진행(5회)

최종 답변 선택

총 7단계의 파이프라인

Inference Pipeline - Architecture



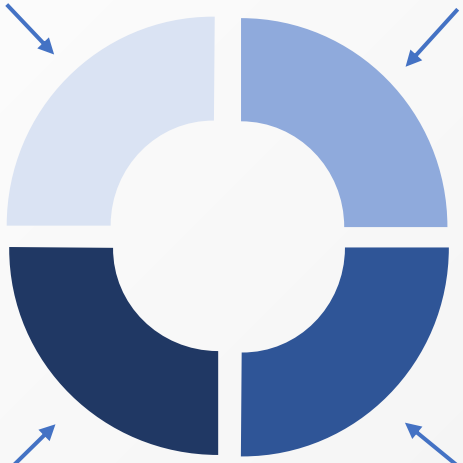
Key Strengths

High Accuracy

- High accuracy 를 통해 금융보안 분야의 전문 지식을 폭넓게 이해 및 대응 가능
- Public score : 0.664 (3nd) / Private score : 0.67 (1st)

Low latency

- CoT(Chain-of-Thought) 없이 Self-Consistency만으로 추론 수행
- 복잡한 사고 과정을 제거하여 추론 속도 대폭 개선
- 평균 응답 시간 / 객관식: 4.54초, 주관식: 95.07초



Train-free domain adaptation

- 별도의 파인튜닝 없이 RAG 지식 베이스 확장만으로 도메인 특화 가능
- 지식 확장 기반 접근으로 도메인 적응력 및 확장성 확보

→ High-accuracy 를 통해 Train-free 구조로도 도메인 성능 향상 입증

Visualization with Streamlit

- Streamlit 기반 시각화 인터페이스 구현
- 객관식 / 주관식 질문에 대한 모델 응답 결과를 실시간 시각화

적용가능성

2025-03-30 전자신문
금융 특화 '한글 말뭉치' 공유…생성형 AI 개발 돕는다.

2025-03-26 한국금융신문
"금융 AI 가이드라인, 명확성·구체성·시의성 제고 필요" 한 목소리

현재 이미 금융권에서 생성형 AI 를 통한 어시스턴스와
이에 대한 명확성 등의 중요성이 많이 알려짐

명확성 뿐만 아니라 확정성과 안정성을 적용시킨 개선된 QA 파이프라인은
실제 산업 활용, 금융 교육활용 도구로써 큰 기여를 할 수 있을 것

#Appendix - Reference

[1] 연합뉴스. (2023, December 15). *AI는 급부상하는 취약점… 美 "금융시스템에 위험" 첫 경고,* https://www.yna.co.kr/view/AKR20231215035900009

[2] 중소기업투데이. (2024, January 12). *"생성AI, 금융권 체제 통째로 바꿔."*, https://www.sbiztoday.kr/news/articleView.html?idxno=21248

[3] STARTUPTODAY. (2024, February 5). *"할루시네이션 잡고 신뢰도↑"… 금융·법조·공공기관 등 RAG 적극 활용 '눈길'*, https://www.startuptoday.kr/news/articleView.html?idxno=48273

[4] Precedence Research. (2025, January 13). *Generative AI in Financial Services Market Size, Share, and Trends 2025–2034.*, https://www.precedenceresearch.com/generative-ai-in-financial-services-market?ref=blog-ko.allganize.ai

[5] 한국금융신문. (2025, March 26). *"금융 AI 가이드라인, 명확성·구체성·시의성 제고 필요" 한 목소리.* , https://www.fntimes.com/html/view.php?ud=2025032621000656291179ad43907_18

[6] 전자신문. (2025, March 30). *"금융 특화 '한글 말뭉치' 공유… 생성형 AI 개발 돕는다."*, https://www.etnews.com/20250330000002

[7] ZDNET Korea. (2025, September 20). *"정보보호 인력 13만 4300명… 인력 부족 심각."*, https://zdnet.co.kr/view/?no=20250920162810

[8] KODE KT 블로그. (n.d.). *믿:음 2.0 언어모델 개발기: 한국적인 AI를 위한 여정.* , https://kode.kt.com/blog/article/3935