미래사회
Journal of Future Society
2026. Vol. 17, No. 2, 204~213.
https://doi.org/10.22987/jifso.2026.17.2.204

망분리 금융 환경에서의 프라이빗 LLM 구축 및 성능 평가

임 기 범* **오 종 훈****

서울과학종합대학원대학교

본 연구는 망분리 환경에서의 금융기관을 위한 프라이빗 대규모 언어모델을 개발하고 이를 FinGPT와 비교 평가한다. 제안 모델은 LLaMA-2 7B 모델을 기반으로 Low-Rank Adaptation (LoRA)을 통해 미세조정을 하고 경량화된 인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback, RLHF) 파이프라인을 통해 추가 정렬하였다. 학습에는 5억 개의 금융 도메인 토큰을 사용하였다. 실험 결과 프라이빗 모델은 FinGPT 대비 상당히 높은 BLEU 점수(47.03%대 13.34%)를 보였으며, 한국 금융 규제 맥락에도 더 밀접하게 부합하였다. 이러한 결과는 망분리 환경에서도 프라이빗 LLM이 경쟁력 있는 성능을 발휘할 수 있음을 실증적으로 확인하였으며, 금융기관이 안전하게 AI를 활용할 수 있는 구체적 가이드라인을 제공한다. 다만 본 연구는 LLaMA-2 7B 기반 단일 모델과 제한된 금융 질의응답·요약·감성분석 평가를 중심으로 수행되었으므로, 모든 금융기관의 망분리 환경이나 대규모 상용 LLM 환경으로 일반화하는 데에는 한계가 있다.

주요어 : 프라이빗 대규모 언어모델(LLM), 금융 AI, LoRA, 망분리, 규제 준수

* 주저자: 임기범/서울과학종합대학원대학교 AI융합공학 박사과정/서울시 서대문구 이화여대2길 46
/Tel: 070-7012-2700/E-mail: love4her2@naver.com
** 교신저자: 오종훈/서울과학종합대학원대학교 AI전문대학원 원장/서울시 서대문구 이화여대2길 46
/Tel: 070-7012-2700/E-mail: jhoh@assist.ac.kr

Ⅰ. 서론

본 연구의 목적은 다음과 같다.

금융산업은 데이터 집약적 업무와 복잡한 의사결정 과정으로 특징지어지며, 운영 효율성과 신뢰성을 제고하기 위한 인공지능(Artificial Intelligence, AI) 도입 수요가 빠르게 증가하고 있다. 그러나 2013년 전국적인 전산망 마비 사고 이후 한국의 금융 규제는 내부 업무망과 외부 인터넷망 사이의 물리적 망분리를 요구하고 있으며, 이에 따라 핵심 금융 업무는 망분리 환경에서 수행되어야 한다(금융위원회, 2013; 전자금융감독규정, 2013).

이러한 망분리 환경은 외부 침입을 차단하는 데 효과적이지만, 최신 AI 기술 도입에는 중대한 진입 장벽으로 작용한다. 현재 AI 혁신을 주도하는 OpenAI의 GPT 계열과 같은 클라우드 기반 상용 LLM은 외부 API 호출을 필요로 하나, 망분리가 적용된 금융권에서는 이러한 서비스에 대한 직접 접근이 금지되기 때문이다. 최근 규제 샌드박스를 통해 Software-as-a-Service(SaaS) 활용을 부분적으로 허용하려는 움직임이 나타나고 있으나, 민감한 개인신용정보를 처리하는 핵심 업무에는 여전히 데이터 유출 우려를 방지할 수 있는 폐쇄망 환경이 요구된다.

따라서 본 연구는 보안 규제를 준수하면서도 AI의 활용 가치를 극대화하기 위한 방안으로, 외부 네트워크와 단절된 망분리 환경에서의 온프레미스 서버에서 완전히 구동될 수 있는 프라이빗 LLM 개발 접근법을 제안한다. 특히 개별 기관이 처음부터 모델을 사전학습하는 비효율성을 극복하기 위해, 검증된 오픈소스 모델을 채택하고 공개된 금융 데이터와 내부 보안 데이터를 함께 활용하여 미세조정하는 방법론의 유용성에 주목한다.

본 연구는 LLaMA-2 모델을 기반으로 Low-Rank Adaptation(LoRA) 기법을 적용하여 금융 도메인 특화 프라이빗 LLM을 구축하고, 이를 오픈소스 금융 모델인 FinGPT와 비교 분석한다. 이를 통해 폐쇄망 환경에서 수행되는 효과적인 미세조정이 각 국가 또는 기관의 관할권별 규제 요구사항을 반영하는 데 얼마나 효과적인지를 실증적으로 제시한다. 본 연구는 한국뿐 아니라 데이터 주권과 보안이 중시되는 전 세계의 다양한 규제 산업에도 적용 가능한 실무적 방법론을 제시한다는 점에서 의의가 있다.

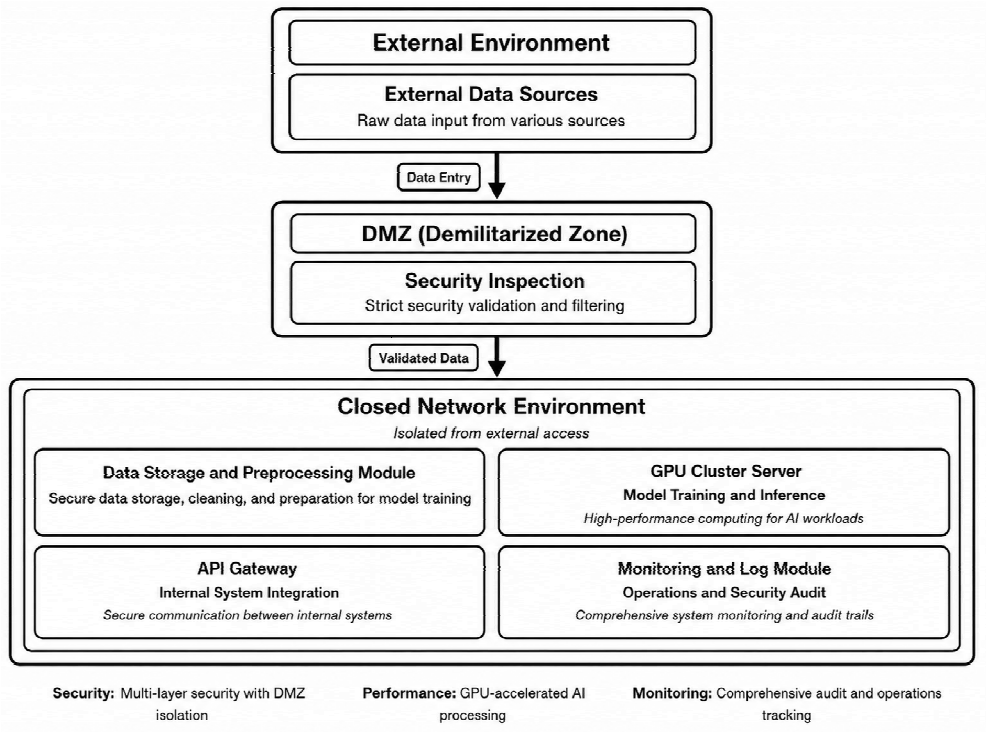
Ⅱ. 이론적 배경

규제를 받는 금융 환경에서 효과적인 대규모 언어 모델(Large Language Model, LLM)을 개발하기 위해서는 데이터 접근성, 계산 효율성, 모델 투명성 사이의 균형이 요구된다. 독점 금융 데이터로 학습된 BloombergGPT와 같은 초기 모델은 도메인 특화 정확도에서 상당한 개선을 보였으나, 복제나 비교·분석을 위한 접근이 제한되어 있다(Wu et al., 2023). 반면 FinGPT와 같은 오픈소스 대안은 웹 스크래핑 데이터세트와 Low-Rank Adaptation(LoRA)과 같은 경량 미세조정 방법을 활용하여 저비용으로 금융 LLM을 구축할 수 있게 한다(Yang et al., 2023). Meta의 LLaMA-2 계열(Meta AI, 2023)은 7B에서 70B 매개변수 규모에 이르는 오픈 기반 모델을 제공하며, 도메인 특화 맞춤화를 위한 유연성을 제공한다.

한편 국내 금융권에서는 망분리 규제의 보안 효과와 함께 클라우드·AI 활용 제약에 대한 논의가 지속되고 있으며, 최근에는 금융 분야의 생성형 AI 활용과 보안성 검증 체계를 병행하려는 정책적 논의가 확대되고 있다(금융위원회, 2024; 금융보안원, 2025). 이러한 흐름은 금융기관이 외부 API 기반 LLM을 그대로 도입하기보다, 내부 데이터 통제와 규제 준수를 보장할 수 있는 프라이빗 LLM 구축 방안을 검토해야 할 필요성을 높이고 있다.

LoRA는 사전학습된 가중치를 고정한 상태에서 추가적인 저랭크 행렬만 학습하는 효율적인 미세조정 접근법으로, 전체 미세조정과 유사한 성능을 유지하

면서도 자원 요구량을 줄일 수 있다(Hu et al., 2022). 이러한 모델의 평가는 BLEU와 ROUGE와 같은 자동 평가 지표에 의존하는 경우가 많다(Papineni et al., 2002; Lin, 2004). BLEU는 엔그램 정밀도에 기반한 지표로서 말뭉치 수준 비교에 표준적으로 활용되는 반면, 문장 수준 BLEU(Sentence-BLEU)는 소규모 또는 과업 특화 평가에서 더 세밀한 분석 단위를 제공한다(Bird et al., 2009). 본 연구에서는 제약된 계산 환경에서 금융 텍스트 생성 품질의 미묘한 차이를 포착하기 위해 후자를 채택하였다.



[그림 1] 프라이빗 LLM의 시스템 아키텍처

주: 망분리 환경 내 데이터 반입, 모델 학습·추론, 내부 API 연동, 보안 로깅 흐름을 나타냄.
출처: 저자 작성, 생성형 AI 도구를 활용하여 시각화.

이상의 선행연구와 평가 방법론 논의를 종합하면, 기존 금융 특화 LLM 연구는 주로 금융 데이터 기반의 성능 향상과 오픈소스 모델 활용 가능성에 초점을 두어 왔다. 그러나 BloombergGPT는 대규모 독점 데이터와 폐쇄형 인프라를 기반으로 하여 재현성과 접근성에 한계가 있으며, FinGPT는 공개 금융 데이터와 경량 미세조정의 가능성을 제시하였으나 인터넷 기반 데이터 수집과 글로벌 금융 맥락을 전제로 한다. 이에 비해 본 연구는 한국 금융권의 망분리 환경, 국내 금융 규제, 내부 데이터 보호 요구를 함께 고려하여 온프레미스 기반 프라이빗 LLM의 구축과 평가 절차를 제시한다는 점에서 차별성을 갖는다.

최근 금융 특화 LLM 연구에서는 단순 자동평가 지표만으로는 모델의 금융 지식, 추론 능력, 규제 적합성, 실제 업무 활용 가능성을 충분히 설명하기 어렵다는 점이 강조되고 있다. Dou et al.(2025)는 금융 LLM의 성능을 단일 과업 점수만으로 평가하

기 어렵다는 점을 지적하고, 금융 지식과 추론 능력을 분리하여 측정하는 FinEval-KR 평가 프레임워크를 제안하였다. 이러한 최근 연구 흐름은 본 연구가 Sentence-BLEU를 주요 지표로 활용하되, ROUGE-L, 감성분석 정확도, 전문가 평가를 함께 사용하여 단일 자동평가 지표의 한계를 보완하고자 한 방법론적 근거를 제공한다.

Ⅲ. 연구 방법론

1. 시스템 아키텍처

망분리 금융 환경의 운영상 제약을 충족하기 위해, 제안하는 프라이빗 LLM 아키텍처는 네 가지 핵심 모듈을 통합한다. 구체적으로 (1) 보안 데이터 저장소 및 전처리 유닛, (2) 모델 학습과 추론을 위한 GPU 클러스터 서버, (3) 시스템 통합을 위한 내부 API 게이트웨이, (4) 보안 모니터링 및 로깅 프레임워크로 구성된다. 모든 프로세스는 데이터 기밀성과 규제 기준 준수를 보장하기 위해 폐쇄망 내부에서 실행된다. 외부 데이터는 비무장지대(demilitarized zone, DMZ)를 통한 보안 검증 이후에만 반입된다. [그림 1]은 엄격한 데이터 보호 요구와 계산 효율성 사이의 균형을 통해 재현 가능한 모델 배포를 가능하게 하는 전체 아키텍처를 보여준다.

2. 모델 학습 전략

기반 모델은 LLaMA-2 7B(Meta AI, 2023)이며, Low-Rank Adaptation(LoRA) 기법(rank=8)을 활용하여 미세조정하였다(Hu et al., 2022). 내부 금융 문서와 공개 데이터셋을 결합하여 약 5억 개의 토큰으로 학습 말뭉치를 구성하였다. 학습 데이터는 공개 금융 데이터와 비식별화된 내부 금융 문서로 구성하였다. 공개 데이터에는 금융 뉴스, 공시 자료, 금융 용어 자료, 공개 금융 질의응답 자료가 포함되며, 내부 데이터에는 비식별화된 상담 질의응답, 금융상품 설명 자료, 내부통제 및 보안 관련 문서가 포함된다. 모든 데이터는 중복 제거, 개인정보 및 기관 식별자 마스킹, 문장 단위 분리, 토큰화, 저품질 문서 제거 절차를 거쳤다. 내부 데이터는 보안상 원문 공개가 제한되므로, 본 연구에서는 데이터 유형, 전처리 절차, 학습 규모를 중심으로 재현 가능성을 확보하였다. 학습은 지도 미세조정(Supervised Fine-Tuning, SFT)과 경량 RLHF 단계(Christiano et al., 2017)로 이루어진 2단계 파이프라인을 채택하였다. 여기서 전문가 평가는 단순 보상 신호를 구성하는 데 사용되었으며, 이후 선호 기반 최적화에 적용되었다. 학습은 네 대의 NVIDIA A100 40GB GPU

<표 1> LoRA 기반 미세조정을 위한 모델 구성

항목	설명
기반 모델	Meta LLaMA-2 7B
미세조정 방법	LoRA(Low-Rank Adaptation)
학습 데이터	약 5억 토큰(내부 및 공개 금융 데이터 결합)
학습 환경	4 × NVIDIA A100 40GB GPU, 3 에포크
지도 미세조정	약 10,000개의 Q&A 쌍
RLHF 적용	금융 전문가 피드백 기반(약 500개 응답, 전문가 3인, 5점 척도, 고득점 응답 중심 튜닝)

주: LoRA 기반 미세조정에 사용된 기반 모델, 학습 데이터, 학습 환경, 지도 미세조정 및 RLHF 적용 조건을 요약함.
출처: 저자 작성.

에서 24시간 동안 수행되었으며, 과적합을 방지하기 위해 세 번째 에포크 이후 조기 종료하였다. <표 1>은 실험의 재현성과 비교 가능성을 확보하기 위한 하이퍼파라미터 구성을 요약한다.

3. 평가 지표 및 설계

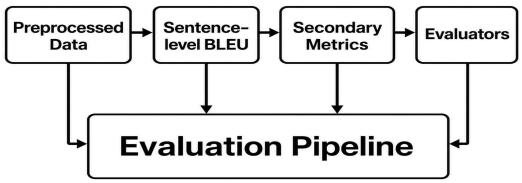
모델 성능은 제한된 자원 조건에서 세밀한 텍스트 생성 품질을 평가하기 위해 Sentence-BLEU를 주요 지표로 사용하여 평가하였다. BLEU-1부터 BLEU-4까지의 점수는 동일 가중치를 적용하여 NLTK 라이브러리로 산출하였다(Bird et al., 2009). FinGPT (fingpt-llama2-7b-chat)(Yang et al., 2023)를 벤치마크 모델로 사용하였다. 보완 지표로는 요약 성능 평가를 위한 ROUGE-L(Lin, 2004), 감성분석 정확도, 5점 척도의 전문가 평가를 포함하였다. Sentence-BLEU 평가는 FiQA-QA 기반 금융 질의응답 표본 300개를 대상으로 수행하였다. 각 문항에 대해 모델이 생성한 답변과 기준 답변을 비교하였으며, 개별 질의응답 단위로 산출한 Sentence-BLEU 점수의 산술평균을 최종 평균 점수로 사용하였다. 본 연구에서 Sentence-BLEU는 모델 응답과 기준 답변 간의 수정 엔그램 정밀도를 바탕으로 핵심 금융 용어와 정답 표현의 재현성을 비교하기 위한 지표로 활용하였다. 특히 금융 질의응답은 핵심 용어, 수치 표현, 규제 관련 표현의 일관성이 중요하므로 Sentence-BLEU는 제한된 평가 환경에서 금융 텍스트 생성 품질을 정량적으로 비교하는 데 유용하다. 다만 Sentence-BLEU는 의미적으로 동등하지만, 표현이 다른 답변, 복합 추론의 타당성, 규제 적합성을 완전히 포착하기 어렵기 때문에 본 연구에서는 이를 단독 지표로 사용하지 않고 보완 지표와 함께 해석하였다.

ROUGE-L은 생성 요약문과 기준 요약문 간의 최장 공통 부분수열을 기반으로 산출하였으며, 금융 문서 요약에서 핵심 정보의 흐름과 주요 표현이 얼마나 보존되는지를 확인하기 위한 지표로 사용하였다. 감성분석 정확도는 금융 텍스트 분류 과업에서 모델이 예측한 감성 라벨과 기준 라벨이 일치한 표본의 비율로 산출하였다. 보완 평가는 요약, 감성분석, 규제 적합성 판단 문항을 포함하여 구성하였다. 전문가 평가는 자동평가 지표가 포착하기 어려운 사실 정확성, 국내 금융 규제 적합성, 질문 맥락 충실성, 응답의 실무 유용성, 민감정보 및 부적절 응답 회피 여부를 확인하기 위해 수행하였다. 평가는 금융 도메인 지식과 규제 이해도를 갖춘 전문가 3인이 수행하였으며 각 항목은 5점 척도로 평가하고 최종 점수는 평가자별 점수의 평균값으로 산출하였다. [그림 2]는 정량적·정성적 측정이 결합되어 제안 모델의 견고성을 검증하는 평가 파이프라인을 제시한다.

<표 2> Sentence-BLEU 통계 요약

지표	프라이빗 LLM	FinGPT(fingpt-llama2-7b-chat)
평균	47.03%	13.34%
표준편차	6.48%	9.72%
최솟값	31.21%	0.41%
최댓값	65.70%	48.04%

주: Sentence-BLEU는 FiQA-QA 기반 금융 질의응답 표본 300개를 대상으로 NLTK sentence_bleu 함수를 활용하여 산출하였으며, BLEU-1부터 BLEU-4까지 동일 가중치를 적용함.
출처: 저자 실험 결과를 바탕으로 저자 작성.



[그림 2] 평가 파이프라인 개요

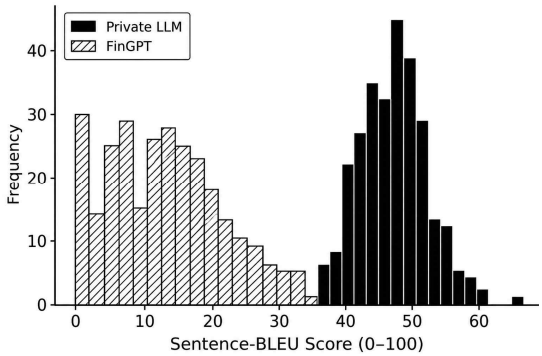
주: 정량 평가와 전문가 평가를 결합한 본 연구의 평가 절차를 나타냄.
출처: 저자 작성, 생성형 AI 도구를 활용하여 시각화.

Ⅳ. 결과 및 논의

1. 정량적 성능 비교

제안된 프라이빗 LLM은 평균 Sentence-BLEU 점수 47.03%를 달성하여 FinGPT의 평균 13.34%(표준편차= 9.72%)를 크게 상회하였다. <표 2>에 요약된 바와 같이, 본 연구의 모델은 더 낮은 분산과 중앙에 안정적으로 집중된 점수 분포를 보이며 더 높고 일관된 성능을 나타냈다.

[그림 3]은 프라이빗 LLM의 출력이 중·고성능 구간에 군집되어 있음을 보여주며, 이는 과업 전반에 걸쳐 안정적인 생성 품질을 시사한다. 이러한 결과는 자원이 제한된 망분리 환경에서도 도메인 적응형 미세조정이 참조 답변과의 엔그램 중복률을 향상시킬 수 있음을 나타낸다.



[그림 3] Sentence-BLEU 점수 분포

주: 평가 표본별 Sentence-BLEU 점수 분포를 나타냄.
출처: 저자 실험 결과를 바탕으로 저자 작성.

2. 보완 평가

보완 지표 또한 정량 평가 결과를 뒷받침한다. <표 3>에 제시된 바와 같이, 프라이빗 LLM은 ROUGE-L(41.0%대 39.0%), 감성분석 정확도(94.5%대 93.8%), 전문가 평가(5점 척도 기준 4.3 대 4.0)에서 FinGPT보다 우수한 성능을 보였다. 이러한 결과는 LoRA 기반 미세조정이 사실 정확성과 인지된 답변 품질을 모두 효과적으로 향상시키며, 다양한 평가 차원에서 제안 접근법의 견고성을 검증한다.

3. 정성적 및 해석적 분석

정성적 비교는 프라이빗 LLM이 맥락 충실성과 규제 준수성 측면에서 갖는 장점을 추가적으로 부각

<표 3> ROUGE-L, 감성분석 정확도 및 전문가 평가 결과

평가 지표	프라이빗 LLM	FinGPT(fingpt-llama2-7b-chat)
ROUGE-L(요약)	41.0%	39.0%
감성분석 정확도	94.5%	93.8%
전문가 평가(5점 척도)	4.3 / 5	4.0 / 5

주: ROUGE-L은 생성 요약문과 기준 요약문 간의 최장 공통 부분수열 기반 요약 성능 평가 결과이며, 감성분석 정확도는 기준 라벨 대비 모델 예측 라벨의 일치 비율임. 전문가 평가는 금융 분야 전문가 3인의 5점 척도 평가 평균값임.
출처: 저자 실험 결과를 바탕으로 저자 작성.

한다. <표 4>에 제시된 바와 같이, 본 연구의 모델은 전자금융감독규정 및 내부통제기준과 같은 한국 금융 관련 법령을 정확히 참조하면서 법적으로 일관되고 맥락에 부합하는 답변을 제공하였다. 반면 FinGPT는 언어적으로 유창한 응답을 생성하였음에도 불구하고, 국내 법제 맥락을 충분히 포착하지 못하고 국제 기준이나 일반적인 서구 금융 관행에 의존하는 경우가 빈번하였다. 예컨대 <표 4>에서 확인할 수 있듯이, FinGPT는 한국 법령상 요구되는 구체적인 국내 규정 대신 글로벌 ISO 표준을 인용하였다. 이러한 증거는 규제 산업에서 범용 LLM을 별도의 적응 없이 배포할 경우 중대한 규제 준수 위험을 초래할 수 있음을 시사한다. 따라서 폐쇄망 환경에서의 도메인 특화 미세조정은 모델 성능 향상 수단일 뿐 아니라, 관할권별 금융 규제 준수를 보장하기 위한 전제 조건으로 기능한다.

모델 비교를 넘어, 본 연구의 결과는 금융 부문 AI 시스템 도입에 대한 실무적 함의도 제공한다. 구체적으로 규제 적합성을 고려한 언어모델은 규제 환경 내에서 안전한 자동화를 지원할 수 있으며, AI 혁신과 기관 거버넌스 사이의 간극을 완화할 수 있음을 보여준다.

<표 4> 대표 질의응답 사례 비교

질문	프라이빗 LLM 응답	FinGPT 응답	비고 / 해석
ISA 계좌의 세제 혜택은 무엇인가?	ISA 계좌는 연간 최대 2천만 원까지 납입할 수 있다. 5년 이상 유지할 경우 이자 및 배당소득은 비과세되며, 투자소득 중 추가 200만 원까지 비과세 혜택이 적용된다.	ISA 계좌의 세제 혜택은 10%에서 20% 사이의 환급 불가능한 세액공제이다.	프라이빗 LLM은 한국의 ISA 규정과 한도를 정확히 반영한 반면, FinGPT는 모호하고 부정확한 설명을 제공하였다.
금융의 내부통제 기준은 무엇인가?	내부통제기준법 제24조에 따르면, 금융기관은 준법감시인을 임명하고 이사회 산하의 내부감사기구를 운영해야 한다.	금융의 내부통제 기준은 ISO 31000:2018이다.	프라이빗 LLM은 한국 법적 체계를 구체적으로 인용한 반면, FinGPT는 국내에 직접 적용되지 않는 국제 표준을 잘못 참조하였다.
정기예금을 중도 해지하면 세제 혜택은 어떻게 되는가?	예금자보호법에 따라 원금은 보호되지만, 중도 해지 시 이자는 지급되지 않으며 세제 혜택도 상실된다.	정기예금의 중도 해지 위약금은 일반적으로 은행 및 계약 조건에 따라 3~6개월분 이자에 해당한다.	프라이빗 LLM은 한국 규정에 부합하는 응답을 제공한 반면, FinGPT는 일반적인 해외 은행 관행을 반영하였다.

주: 대표 질의응답 사례는 국내 금융 규제 및 금융상품 맥락 반영 여부를 비교하기 위한 예시임.
출처: 저자 실험 결과를 바탕으로 저자 작성.

Ⅴ. 결론

본 연구는 엄격한 망분리 환경에서도 프라이빗 LLM이 경쟁력 있는 성능을 달성할 수 있음을 실증적으로 제시하였다. Sentence-BLEU 지표를 기준으로 제안 모델은 정량적 지표(BLEU, ROUGE-L)와 정성적 지표(감성분석 정확도, 전문가 평가) 모두에서 오픈소스 FinGPT보다 일관되게 우수한 성능을 보였다. 이러한 결과는 LoRA 기반 미세조정이 자원 제약이 존재하는 금융 환경에서 도메인 특화 모델 역량을 효과적으로 향상시킬 수 있음을 확인해 준다.

이론적으로 본 연구는 안전하게 폐쇄형 네트워크 환경에 배포 가능한 AI 시스템의 실현 가능성을 제시함으로써 도메인 적응형 LLM에 관한 문헌에 기여한다. 또한 특수 도메인 전반에서 LLM을 평가할 때 평가 지표 선택이 갖는 방법론적 중요성을 강조한다.

실무적으로 본 연구는 엄격한 규제 및 보안 제약 하에서 AI를 구현하려는 금융기관에 재현 가능한 아키텍처를 제공하며, 금융 분야에서 생성형 모델을 책임 있게 도입할 수 있는 경로를 제시한다. 이는 엄격한 데이터 주권 및 네트워크 분리 요건을 충족해야 하는 기관의 경우 독립적인 온프레미스 모델 구축이 필요할 수 있음을 부각한다.

그럼에도 불구하고 본 연구는 다음과 같은 한계를 갖는다. 첫째, 주요 정량 지표로 Sentence-BLEU를 활용하였으므로 의미적 정확성, 복합 추론 능력, 규제 적합성을 완전히 포착하기에는 한계가 있다. 둘째, 실험 모델은 LLaMA-2 7B 단일 모델을 기반으로 하므로 더 큰 규모의 최신 모델이나 다른 기반 모델로 결과를 일반화하는 데에는 주의가 필요하다. 셋째, 본 연구는 특정 망분리 금융 환경을 전제로 수행되었으므로 모든 금융기관의 인프라와 운영 환경에 동일하게 적용된다고 보기 어렵다. 향후 연구에서는 corpus-BLEU, 의미 기반 평가, 금융 규제 준수 평가, 다기관 검증을 포함하여 평가 범위와 모델 규모를 확장할 필요가 있다.

종합하면, 본 연구는 금융과 같은 규제 산업에서 안전하고 고성능의 AI 시스템을 개발하기 위한 실증적 근거와 방법론적 지침을 함께 제공한다.

참고문헌

금융보안원 (2025). 금융권의 안전한 AI 활용 환경 조성을 위한 보안성 평가 본격 실시[보도자료]. 게시일: 2025.02.19. https://www.fsec.or.kr/bbs/detail?bbsNo=11629&menuNo=69

금융위원회 (2013). 금융전산 망분리 가이드라인 배포 [보도자료]. 게시일: 2013.09.16. https://www.fsc.go.kr/no010101/70834

금융위원회 (2024).「금융분야 망분리 개선 로드맵」 발표 [보도자료]. 게시일: 2024.08.13. https://www.fsc.go.kr/no010101/82885

전자금융감독규정 [시행 2013. 12. 3.]. (2013. 12. 3. 금융위원회고시 제2013-27호).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS), 30,* 4299-4309.

Dou, S., Shen, Y., Chen, M., Wang, Z., Xu, J., Guo, Q., Shao, K., Chen, C., Hu, H., Shi, H., Min, M., & Zhang, L. (2025). FinEval-KR: A financial domain evaluation framework for large language models' knowledge and reasoning. *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing,* 47-69.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations (ICLR).*

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop on Text Summarization Branches Out,* 74-81.

Meta AI. (2023). *LLaMA-2: Open foundation and fine-tuned chat models*(Technical Report).

Meta AI Research.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 311-318.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv pre-print arXiv:2303.17564.*

Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. *arXiv preprint arXiv: 2306.06031.*

투고일자: 2026. 5. 1.
심사일자: 2026. 5. 26.
게재확정일자: 2026. 6. 4.

미래사회
Journal of Future Society
2026. Vol. 17, No. 2, 204~213.
https://doi.org/10.22987/jifso.2026.17.2.204

Private Large Language Model Development and Performance Evaluation under Network Isolation in the Financial Sector

Ki-Bum Lim Jong-Hoon Oh

Seoul AI School, Assist University

We develop a private large language model (LLM) for network-isolated financial institutions and benchmark it against FinGPT. Our model uses a LLaMA-2 7B model fine-tuned through Low-Rank Adaptation and aligned via lightweight reinforcement learning from human feedback pipeline. We trained our model on 500 million financial tokens and evaluate it using sentence-BLEU, ROUGE-L, sentiment accuracy, and expert ratings. The private model achieves a considerably higher BLEU score (47.03% vs. 13.34%) and closer alignment with Korean financial regulations. These findings suggest, rather than establish, the feasibility of developing secure and compliance-aware LLMs under air-gapped constraints. However, this study is limited by its reliance on a single LLaMA-2 7B base model, restricted evaluation sample, and experimental setting of a specific network-isolated financial environment.

Key words: private large language model (LLM), financial AI, LoRA, network separation, regulatory compliance