미래사회
Journal of Future Society
2026. Vol. 17, No. 3, 222~246.
https://doi.org/10.22987/jifso.2026.17.3.222

망분리 한국 금융기관을 위한 Private LLM 책임운영 프레임워크

임 기 범*　　오 종 훈**

서울과학종합대학원대학교

금융기관의 생성형 AI 도입은 문서 처리, 준법지원, 리스크 분석, 고객 설명자료 작성 등에서 활용 가능성이 크지만, 한국 금융권의 망분리, 보안통제, 설명의무, 소비자보호, 감사추적성 요구는 공개형 클라우드 LLM의 단순 도입을 제한한다. 본 연구는 망분리 한국 금융기관에서 Private LLM을 책임 있게 운영하기 위한 책임운영 프레임워크를 설계과학연구(DSR) 접근에 따라 제안하고 정교화한다. 프레임워크는 경영진 책임구조, 위험기반 사용사례 분류, 승인된 모델·데이터 레지스트리, 인간참여형 검토 및 승인 기록, 설명 및 근거 제시, 불변 로그와 감사추적성, 운영자 역량관리의 7개 요소로 구성된다. 또한 산출물을 최종/중간, 구조화/비구조화, 내부/대고객/규제판단 보조로 구분하고, 감사가능성을 사후 재구성 가능성으로 재정의한다. 문헌·정책 매핑, 시나리오 통제 매트릭스, FMEA, 전문가의 형성적 피드백을 통해 설계 취약점을 식별하고 프레임워크를 보완하였다. 연구 결과는 핵심 과제가 모델 성능 자체보다 산출물 유형별 통제, 근거 제시, 인간 검토 및 사후 재구성 가능성을 확보하는 운영 설계에 있음을 시사한다. 다만 본 연구는 설계 수준의 형성적 검토에 머물며, 현장 검증이 이루어지지 않았고 전문가 피드백도 소표본(n=3)에 기반한다는 한계가 있다.

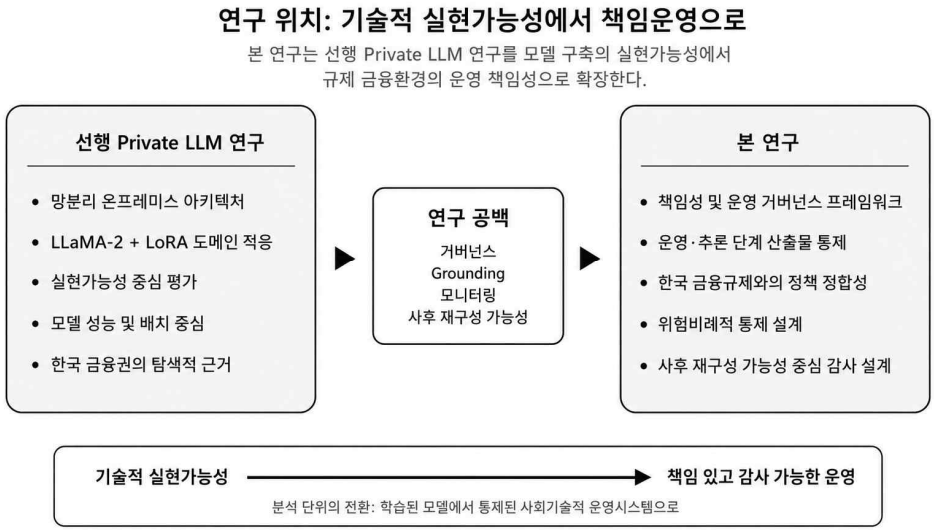
주요어: Private LLM, 금융 AI 거버넌스, 망분리, 책임운영, 감사추적성

* 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: jhonho.assist@gmail.com

Ⅰ. 서론

금융산업은 대규모 데이터 처리, 복잡한 규제 해석, 문서 중심 업무, 고객응대와 내부통제의 동시 수행이라는 특성 때문에 생성형 AI, 특히 대규모 언어모델(Large Language Model, 이하 LLM)의 도입 잠재력이 큰 분야로 평가된다. 금융기관은 질의응답, 문서 요약, 리스크 분석 보조, 준법지원, 고객 설명자료 작성 등 다양한 업무에서 생성형 AI를 활용할 수 있다. 그러나 한국 금융권의 경우, 이러한 기술적 잠재력이 곧바로 운영 가능성을 의미하지는 않는다. 핵심 금융업무는 망분리, 강한 보안통제, 설명의무, 소비자보호, 승인 절차, 감사가능성 같은 규율 환경 안에서 수행되어야 하므로, 공개형 클라우드 LLM의 단순 도입은 구조적으로 제한된다(OECD, 2023; U.S. Department of the Treasury, 2024; 금융위원회, 2024a).

기존 연구는 망분리 금융환경을 위한 Private LLM의 기술적 구축 가능성을 제시하였다. 예컨대 LLaMA-2 기반 모델에 LoRA(Low-Rank Adaptation) 미세조정과 경량 RLHF(인간 피드백 기반 강화학습) 정렬을 적용하고, 온프레미스 폐쇄망 아키텍처와 내부 API 게이트웨이, 보안 모니터링 및 로깅 계층을 포함한 Private LLM 구조를 제안한 연구는 한국 금융 규제 문맥에 특화된 LLM 구축 가능성을 탐색적 근거(exploratory evidence) 수준에서 보여주었다(임기범, 오종훈, 2026; Touvron et al., 2023; Hu et al., 2021). 그러나 해당 연구 역시 후속 과제로 거버넌스 프레임워크, 인용 및 근거 연결(citation/grounding), 정책 위반율 측정, 비교 기준선 확대, 운영 모니터링의 필요성을 제시한다. 이는 금융권 Private LLM 연구가 기술 구현과 성능 비교를 넘어 운영 책임성의 문제로 확장되어야 함을 시사한다.



[그림 1] 연구 위치: 기술적 구축 가능성에서 책임운영으로

주: I장과 II장의 기술적 구축 논의와 V장의 책임운영 프레임워크 간 연구 위치를 나타냄.
출처: 저자 작성, 생성형 AI 도구를 활용하여 시각화.

[그림 1]은 기존 Private LLM 연구가 폐쇄망 아키텍처, 도메인 적응, 성능평가 등 기술적 구축 가능성에 초점을 두었던 반면, 본 연구는 이를 책임 있는 운영 가능성의 문제로 확장한다는 연구 위치

를 보여준다. 특히 본 연구는 Private LLM을 단일 모델이 아니라 정책 정합성, 위험기반 통제, 근거 제시, 인간 검토, 감사추적성이 결합된 운영 시스템으로 파악한다.

최근 한국 금융정책의 흐름도 이러한 문제 전환과 맞닿아 있다. 한국 금융권의 AI 규율은 2021년 「금융분야 인공지능(AI) 가이드라인」의 책임성, 데이터 정확성 및 안전성, 투명성 및 공정성, 소비자 권리 보호에서 출발해, 2024년 「금융권 생성형 AI 활용 지원 방안」과 2025년 가이드라인 개정방향을 거쳐 2026년 시행된 「금융분야 인공지능 가이드라인」의 7대 원칙으로 구체화되었다. 2026년 가이드라인은 거버넌스, 합법성, 보조수단성, 신뢰성, 금융안정성, 신의성실, 보안성을 핵심 원칙으로 제시한다(금융위원회, 2021, 2024a, 2025, 2026). 동시에 「전자금융감독규정」 제15조는 내부망과 외부망의 분리를 전제로 하며, 예외적 외부 활용에도 대체 정보보호통제와 승인 절차를 요구한다(금융위원회, 2024b; 「전자금융감독규정」 제15조). 따라서 현재 한국 금융권에서 중요한 질문은 “LLM을 도입할 수 있는가”가 아니라 “어떤 범위에서, 어떤 조건 하에, 누구의 책임으로, 어떤 증거를 남기며 운영할 것인가”이다.

본 연구의 목적은 망분리 한국 금융기관을 위한 Private LLM의 책임성(accountability) 및 운영 거버넌스 프레임워크를 제안하고, 이를 정책 정합성, 통제 완결성, 사후 재구성 가능성의 관점에서 설계 수준으로 정교화하는 데 있다. 본 연구는 Private LLM을 하나의 생성모델이 아니라 승인된 모델·데이터 체계, 위험기반 사용사례 분류, 인간참여형 검토(Human-in-the-Loop, 이하 HITL) 및 승인 기록, 설명 및 근거 제시, 불변 로그와 감사추적성, 운영자 역량관리가 결합된 운영 시스템으로 재정의한다(NIST, 2024; OECD, 2024; Ojewale et al., 2026).

본 연구의 적용 범위는 Private LLM의 개발 생애주기 전체가 아니라, 주로 운영 및 추론 단계에서 생성되는 업무 산출물의 통제에 있다. 특히 승인·배포·고객 안내·규제판단 보조·내부통제 의사결정에 영향을 미칠 수 있는 산출물을 주요 대상으로 한다. 단순한 아이디어 탐색, 비공식 초안, 내부 사고 보조 과정에서 발생하는 중간 산출물은 동일한 강도의 통제 대상이 아니라 위험수준에 따라 별도 관리되어야 한다.

본 연구는 다음 세 가지 연구 질문을 다룬다.

RQ1. 망분리 한국 금융기관에서 Private LLM 운영에 요구되는 핵심 책임성 및 거버넌스 요소는 무엇인가?

RQ2. 한국 금융권의 규제·가이드라인 요구사항을 충족하기 위해 Private LLM 운영 프레임워크는 어떤 통제 구조를 가져야 하는가?

RQ3. 문헌·정책 매핑, 시나리오 분석, 실패 모드 및 영향 분석(Failure Mode and Effects Analysis, 이하 FMEA), 소규모 전문가 피드백에 기초한 설계 분석은 제안 프레임워크의 어떤 보완 필요사항을 드러내며, 프레임워크는 어떻게 정교화될 수 있는가?

본 연구의 기여는 세 가지로 정리된다. 첫째, 기존 Private LLM 연구의 기술·배치 중심 논의를 한국 금융권의 운영·추론 단계에서 발생하는 산출물 책임성과 위험기반 통제 문제로 확장한다. 둘째, 한국 금융정책과 법령의 요구를 문헌·정책 기반 요구사항으로 추출하고, 이를 경영진 책임구조, 사용사례 위험등급화, 승인된 모델·데이터 체계, 인간 검토 및 승인 기록, 설명 및 근거 제시, 감사추적성, 운영자 역량관리의 7개 구성요소로 구조화한다. 셋째, 현장 파일럿이 제한된 상황에서 적용 가능한 설계과학연구(Design Science Research, 이하 DSR) 기반 절차로서 요구사항 도출, 정책 매핑, 시나리오별 통제 매트릭스, FMEA, 소규모 전문가의 형성적 피드백을 결합하여 프레임워크를 정교화하는 방식을 제시한다. 요컨대 본 연구의 기여는 단순한 통제 항목 목록의 제시가 아니라, 산출물의 3축 분류(최종/중간, 구조화/비구조화, 내부/대고객/규제판단 보

조)와 감사가능성의 사후 재구성 가능성으로의 재정의를 중심으로, 망분리 금융기관의 Private LLM 운영에서 산출물 유형과 사용사례 위험수준을 결합한 책임운영 구조를 정식화한 데 있다.

Ⅱ. 선행연구

1. 금융 특화 LLM과 Private LLM

금융 특화 LLM 연구는 대규모 금융 데이터와 도메인 특화 사전학습을 활용한 BloombergGPT 계열 접근과, 오픈소스 모델에 금융 데이터를 결합하여 도메인 적응을 수행하는 FinGPT 계열 접근으로 구분할 수 있다(Wu et al., 2023; Yang et al., 2023). LLaMA-2와 LoRA의 결합은 제한된 계산 자원에서도 금융 특화 모델을 구축할 수 있는 경량 미세조정 경로를 제공하였다(Touvron et al., 2023; Hu et al., 2021). 기존 Private LLM 연구 역시 이러한 흐름 위에서 한국 금융 규제 문맥에 맞춘 폐쇄망 LLM의 구축 가능성을 제시하였다(임기범, 오종훈, 2026).

그러나 기존 연구의 무게중심은 모델 구축, 도메인 적합화, 자동평가 지표에 놓여 있다. 이는 중요한 출발점이지만, 망분리·설명의무·소비자보호·승인절차가 동시에 작동하는 한국 금융권에서는 기술적 가능성만으로 충분하지 않다. 실제 운영에서는 승인된 문서 범위, 근거 제시, 정책 위반 방지, 인간 검토, 사후 재구성 가능성이 핵심 과제가 된다(Mökander et al., 2024; NIST, 2024).

2. RAG, 근거 제시, 평가

RAG(Retrieval-Augmented Generation)는 생성 시점에 외부 근거를 검색해 응답에 결합함으로써 파라메트릭 지식의 한계를 보완하는 방식이다(Lewis et al., 2020). 그러나 RAG 시스템은 검색 품질(retrieval quality), 근거 충실성(grounding fidelity), 관련성, 정확성, 사실 충실성(faithfulness)과 같은 새로운 평가 문제를 동반한다(Gao et al., 2023; Yu et al., 2024). 금융기관처럼 규정과 상품문서의 버전, 효력일, 승인 여부가 중요한 환경에서는 단순 생성품질보다 승인 코퍼스 기반 검색, 근거 부족 시 응답 유보, 문서 버전 고정 같은 운영 통제가 더 중요하다.

3. 책임성, 감사, 설명가능성

최근 책임성·감사 문헌은 LLM을 하나의 모델이 아니라 거버넌스, 모델, 애플리케이션이 결합된 사회기술적 시스템으로 이해해야 한다고 본다. LLM 감사는 거버넌스 감사, 모델 감사, 애플리케이션 감사의 다층 구조로 설명될 수 있으며, 내부 알고리즘 감사는 개발 전 과정에 걸친 문서화와 감사 보고 체계를 요구한다(Mökander et al., 2024; Raji et al., 2020). 또한 LLM 감사기록(audit trail) 연구는 모델, 데이터, 평가, 배포, 모니터링, 승인, 예외처리, 책임기록을 연결하는 변조 흔적이 드러나는 원장(tamper-evident ledger) 개념을 제시한다(Ojewale et al., 2026).

국내 금융 AI 연구에서도 설명가능성과 책임성의 중요성이 제기되어 왔다. AI 기반 신용평가 연구에서는 금융 분야에서 예측 정확도뿐 아니라 결과에 대한 설명가능성을 함께 고려할 필요성이 제시되었으며(천예은 외, 2021), AI를 활용한 금융감독에 관한 연구에서는 데이터의 정확성·무결성, 알고리즘 편향과 함께 책임성과 투명성이 주요 법적·제도적 고려사항으로 논의되었다(허유경, 2022). 이러한 연구는 금융 AI의 책임성 문제가 모델 성능이나 설명기법에 한정되지 않고, 결과에 대한 설명과 책임 구조 및 운영 통제의 문제로 확장될 필요가 있음을 보여준다.

국외에서도 AI 책임성을 조직 운영 차원에서 다

룬 실증연구가 이루어졌다. AI를 개발하거나 활용하는 19개 조직의 실무자 26명을 대상으로 한 반구조화 인터뷰 연구는 책임 있는 AI의 구현이 기술적 도구나 원칙의 제시만으로 이루어지기보다 조직 구조, 책임 배분, 협업 방식과 이를 지원하는 운영 프로세스의 영향을 받는다는 점을 보여주었다(Rakova et al., 2021). 이는 알고리즘 감사와 책임성 확보가 기술적 검증뿐 아니라 조직 차원의 역할과 절차 설계가 필요하다는 기존 논의를 실무 차원에서 뒷받침한다.

이러한 논의는 금융권 Private LLM에서 설명가능성이 모델 내부 해석기법에 그칠 수 없음을 보여준다. 금융기관의 설명가능성은 어떤 문서가 검색되었고, 누가 검토했으며, 어떤 근거로 승인·반려·수정되었는지를 재구성할 수 있는 운영 증거와 결합되어야 한다.

4. 연구 공백

기존 연구는 금융 특화 모델의 구축과 도메인 적응, RAG 평가, LLM 감사의 일반론을 다루는 데 의미 있는 진전을 보였다. 국제 기준 차원에서도 NIST의 생성형 AI 프로파일, ISO/IEC 42001의 AI 경영시스템 표준, EU AI Act의 고위험 AI 시스템 요건이 위험기반 관리, 인간 감독, 기록 유지를 공통적으로 요구한다(NIST, 2024; ISO/IEC, 2023; European Parliament and Council, 2024). 그러나 이러한 기준들은 조직 일반 수준의 관리체계 요건 또는 역외 법제의 규제 요건을 제시할 뿐, 한국 금융권의 망분리, 설명의무, 소비자보호, 대체 통제, 승인 절차를 하나의 운영 프레임으로 통합한 연구는 부족하다. 또한 운영 및 추론 단계에서 생성되는 산출물을 최종/중간, 구조화/비구조화, 내부/대고객/규제판단 보조 산출물로 구분해 통제수준을 차등화하는 논의도 충분하지 않다. <표 1>은 주요 기준 틀과 본 연구의 초점을 비교한 것이다. 본 연구는 이 공백을 보완하기 위해 한국 금융권 Private LLM의 책임성 및 운영 거버넌스를 정책 정합성, 위험기반 통제, 사후 재구성 가능성의 관점에서 구조화한다.

<표 1> 주요 AI 거버넌스 기준 틀과 본 연구의 비교

기준 틀	성격	주요 초점	본 연구와의 차별점
NIST AI RMF 생성형 AI 프로파일(2024)	자율적 위험관리 프레임워크	생성형 AI 위험 식별과 거버넌스·측정·관리 기능	위험 식별은 포괄적이나 국내 망분리·설명의무 등 제도 특수성은 다루지 않음
ISO/IEC 42001(2023)	인증 가능한 AI 경영시스템 표준	조직 단위 AI 관리체계(AIMS)의 수립·운영	조직 일반 요건 중심으로, 산출물 단위 통제와 증거 구조는 구체화하지 않음
EU AI Act(2024)	법적 구속력 있는 역외 규제	고위험 AI의 위험관리, 기록 유지, 인간 감독, 투명성	위험기반 접근은 유사하나 한국 금융 법제·가이드라인과의 정합성은 별도 매핑 필요
금융분야 인공지능 가이드라인(2026)	국내 자율규제 원칙	거버넌스, 합법성, 보조수단성 등 7대 원칙	원칙 수준 규범으로, 산출물 유형별 통제 절차와 증거 객체는 제시하지 않음
본 연구 프레임워크	설계 수준 운영 프레임워크	산출물 3축 분류, 위험비례적 통제, 사후 재구성 가능성	국내 규제 요구를 산출물 단위의 통제 절차와 증거 구조로 운영화

출처: NIST(2024), ISO/IEC(2023), European Parliament and Council(2024), 금융위원회(2026)를 바탕으로 저자 작성.

Ⅲ. 한국 금융권의 규제 및 정책 맥락

한국 금융권의 AI 규율은 단일 규정이 아니라 가이드라인, 보안지침, 망분리 규정, 설명의무 관련 법률이 중첩되는 다층 구조로 형성되어 있다. 금융위원회는 2021년 「금융분야 인공지능(AI) 가이드라인」에서 AI 윤리 원칙, AI 조직 구성, 위험관리정책의 3중 내부통제 장치를 제시했고, 데이터 정확성 및 안전성, 투명성 및 공정성, 금융소비자 권리 보장을 핵심 가치로 강조했다(금융위원회, 2021). 이어 2022년 「금융분야 AI 개발·활용 안내서」는 이러한 원칙을 개발·활용 단계별 실무 기준으로 구체화하였다(금융위원회, 2022).

이후 2024년 「금융권 생성형 AI 활용 지원 방안」은 상용 AI 활용 허용과 내부망 설치형 오픈소스 AI 활용 지원을 포함한 이원(two-track) 체계를 제시하였다. 2025년 공개된 개정방향은 기존 AI 가이드라인을 생성형 AI와 금융권 AX 환경에 맞게 보완할 필요성을 제기하였고, 2026년 시행된 「금융분야 인공지능 가이드라인」은 거버넌스, 합법성, 보조수단성, 신뢰성, 금융안정성, 신의성실, 보안성의 7대 원칙을 중심으로 금융권 AI 활용의 자율규제 체계를 제시하였다. 본 연구에서는 2025년 문서를 정책 개정방향으로, 2026년 문서를 시행 가이드라인으로 구분하여 해석한다(금융위원회, 2024a, 2025, 2026).

망분리의 직접적인 법적 기반은 「전자금융감독규정」 제15조와 시행세칙에 있다. 해당 규정은 내부통신망과 연결된 내부 업무용 시스템을 원칙적으로 외부통신망과 분리·차단하고 접속을 금지하면서도, 연구·개발 목적 등 일정한 예외에 대해서는 자체 위험성 평가, 망분리 대체 정보보호통제 적용, 정보보호위원회 승인 등을 요구한다(「전자금융감독규정」 제15조; 금융위원회, 2024b).

금융 분야에서는 설명가능성과 소비자 보호도 별도의 법적 요구로 작동한다. 「신용정보의 이용 및 보호에 관한 법률」 제36조의2는 자동화평가 설명·정정·재산출 관련 권리를 인정하고, 「금융소비자 보호에 관한 법률」 제19조는 일반금융소비자가 이해할 수 있도록 중요한 사항을 설명해야 한다고 규정한다. 다만 신용정보법 제36조의2는 개인신용정보에 기초한 자동화평가에 적용되는 조항이므로 Private LLM 산출물 전반에 곧바로 적용되는 것은 아니며, 본 연구는 이를 설명·정정·재산출 절차 설계의 준거점으로 참조한다. 따라서 대고객 요약이나 규제판단 보조에 Private LLM을 사용할 경우, 단순 응답 생성만으로는 부족하고 설명 패키지, 재산출 흐름, 이의제기 가능성, 검토 및 승인 기록이 함께 설계되어야 한다.

보안 측면에서도 한국의 정책 맥락은 생성형 AI를 별도 통제대상으로 본다. 「금융분야 AI 보안 가이드라인」은 학습데이터 수집, 전처리, 모델 설계 및 학습, 검증 및 평가 전 단계에 걸친 보안을 강조하고, 한국인터넷진흥원의 제로트러스트 가이드라인과 AI 보안 안내서는 최소권한, 지속 검증, 접근통제, 모델 자산 무결성을 강조한다(금융보안원, 2023; 한국인터넷진흥원, 2024, 2025). 개인정보 측면에서는 생성형 AI 개발·활용 단계별 개인정보 처리의 적법성과 안전조치를 요구하는 안내서가 마련되어 있다(개인정보보호위원회, 2025).

국제 문헌 역시 금융 분야 생성형 AI의 데이터 품질, 보안, 편향, 환각, 제3자 의존, 모델 위험관리의 중요성을 강조한다(OECD, 2023, 2024; NIST, 2024; U.S. Department of the Treasury, 2024). 특히 금융기관의 생성형 AI 활용은 환각, 독성, 모델 검증, 운영 통제 등 기존 모델위험관리 체계가 충분히 포괄하지 못했던 새로운 위험을 동반한다는 지적도 제기된다(Bhattacharyya et al., 2025). 따라서 한국 금융권의 Private LLM 프레임워크는 기술 확산의 속도를 높이는 장치라기보다, 허용 가능한 자동화의 범위를 경계 짓고 책임 가능한 운영 조건을 명확히 하는 장치로 이해하는 것이 타당하다.

한편, 2024년 망분리 개선 로드맵 이후 규제 샌

드박스를 통한 생성형 AI 활용 허용 등 망분리 규제의 단계적 완화가 진행되고 있어, 망분리를 전제로 한 Private LLM 논의의 유효성에 의문이 제기될 수 있다. 그러나 현재까지의 완화는 자체 위험성 평가, 대체 정보보호통제, 승인 및 기록 의무를 전제로 한 조건부 허용의 성격을 가지며(금융위원회, 2024b), 2026년 가이드라인 역시 보안성과 거버넌스를 핵심 원칙으로 유지한다(금융위원회, 2026). 즉 규제 완화는 통제의 소멸이 아니라 통제 방식의 전환을 의미한다. 어떤 산출물에 어떤 증거와 책임구조를 요구할 것인지를 다루는 본 연구의 책임운영 프레임워크는 완화 이후의 환경에서도 유효하며, 외부 연동이 확대될수록 그 필요성은 오히려 커진다.

Ⅳ. 연구 방법

1. 연구설계

본 연구는 설계과학연구(DSR) 접근을 따른다. DSR은 문제 인식, 목표 정의, 설계·개발, 시연, 평가, 소통의 반복 절차를 통해 인공물(artifact)을 구축하고 정교화하는 방법론으로(Hevner et al., 2004; Peffers et al., 2007), 현장 실증이 제한된 규제 환경에서 프레임워크형 인공물을 다루는 본 연구에 적합하다. 본 연구의 인공물은 Private LLM 책임운영 프레임워크이며, 평가는 실증적 검증(empirical validation)이 아니라 설계 취약점을 식별하고 보완하는 형성적(formative) 설계 평가로 위치한다. 즉, 본 연구는 프레임워크의 현장 효과를 입증하려는 것이 아니라, 문헌·정책 근거, 시나리오 분석, FMEA, 소규모 전문가 피드백을 통해 설계를 정교화하는 과정을 제시한다(Raji et al., 2020; Mökander et al., 2024; NIST, 2024).

초기 설계안(v1)은 문헌·정책 요구사항에 기초한 6개 구성요소로 구성되었고, 설계 분석과 전문가 피드백을 반영한 정교화를 거쳐 7개 구성요소의 최종안(v2)으로 보완되었다. 5장은 정교화가 반영된 최종안을 제시하고, 6장은 그 도출 및 보완 경과를 보고한다.

2. 자료 및 분석 절차

설계 분석은 다섯 단계로 수행하였다. 첫째, 한국 금융권 공식 정책·법령 문서와 국제 감독·표준 문헌에서 Private LLM 운영과 직접 관련된 요구사항 후보를 추출하였다. 둘째, 요구사항 후보를 의미 단위로 코딩하고 유사·중복 항목을 통합하여 초기 요구사항 R1~R14로 정리하였다. 셋째, 요구사항을 프레임워크의 구성요소와 대응시켜 정책–프레임워크 매핑을 수행하였다. 넷째, 내부 준법 Q&A, 상품 설명 요약, 고위험 규제판단 보조의 세 시나리오에 프레임워크를 적용하여 위험수준에 따른 통제 차등성과 누락 가능성을 검토하였다. 다섯째, FMEA와 소규모 전문가 피드백을 통해 설계상 우선 보강해야 할 실패 모드와 구성요소를 식별하고, 그 결과를 반영하여 요구사항 R15~R17과 일곱째 구성요소를 추가하는 정교화를 수행하였다.

분석 대상 문서는 본 연구의 적용 범위와 직접 관련성이 있는 금융 AI 거버넌스, 생성형 AI 보안, 망분리, 설명의무, 소비자보호, 감사추적성 관련 자료로 한정하였으며, 그 목록은 부록 <표 C1>에 제시한다. 일반적인 AI 윤리 선언이나 기술 성능평가 연구 중 Private LLM의 운영 통제와 직접 연결되지 않는 문헌은 배경문헌으로만 활용하고 요구사항 도출 대상에서는 제외하였다. 요구사항 추출과 코딩은 연구진이 수행하였으며, 원문 조항이 요구사항으로 전환되는 과정의 예시는 부록 <표 C2>에 제시한다. 코딩이 연구진에 의해 수행됨에 따라 발생할 수 있는 해석 편향의 가능성은 연구의 한계에 명시한다. 이러한 절차는 현장 효과성 검증을 대체하지 않지만, 설계 연구 단계에서 프레임워크 구성의 추적가능성과 방법론적 투명성을 높이기 위한 것이다.

3. FMEA 적용 방식

FMEA는 설계 단계에서 우선적으로 보강해야 할 위험을 식별하기 위해 사용하였다. 심각도(S), 발생가능성(O), 탐지난이도(D)를 각 1~5점으로 설정하고, 위험우선순위지수 RPN = S × O × D로 우선순위를 계산하였다. 심각도는 고객영향, 규제위반 가능성, 내부통제 훼손 정도를 기준으로 판단하였고, 발생가능성은 해당 실패 모드가 Private LLM 운영 과정에서 발생할 수 있는 개연성을 기준으로 보았다. 탐지난이도는 사전 차단 또는 사후 발견의 어려움을 기준으로 판단하였다. S/O/D의 세부 점수 기준은 부록 <표 A2>에 제시하였다.

S/O/D 점수는 연구진이 문헌·정책 요구와 시나리오 영향도를 근거로 부여하였으며, 전문가 합의 점수나 현장 사고 데이터에 기반한 정량평가가 아니다. IEC 60812는 분석 목적에 따른 척도의 조정을 허용하므로, 본 연구는 설계 비교 목적에 맞게 5점 척도를 사용하였다(IEC, 2018). 따라서 FMEA 점수는 실제 사고 빈도나 손실 규모의 추정치가 아니라 설계 단계에서 상대적으로 중요한 실패 모드를 식별하기 위한 휴리스틱(heuristic)으로 해석해야 하며, 연구진 단독 채점에 따른 신뢰성의 한계는 7장에서 논의한다.

4. 전문가 검토 절차

제안 프레임워크의 설계를 보완하기 위해 금융권 실무 경험을 보유한 전문가를 대상으로 구조화 서면 검토를 수행하였다. 패널은 금융 IT 운영, IT 거버넌스·준법, 현업·디지털 추진의 세 관점을 포괄하도록 금융권 경력 20년 이상의 실무 전문가를 기준으로 저자의 실무 네트워크를 통해 총 5인에게 검토를 요청하였고, 3인이 회신하였다(회신율 60%). 검토는 2026년 5월, 5점 리커트 척도와 자유 의견을 결합한 구조화 설문지(부록 <표 B1>)를 이용해 서면으로 수행하였다.

본 검토는 통계적 일반화나 전문가 합의 도출이 아니라 형성적 피드백을 통한 설계 보완을 목적으로 하며, 소표본(n=3)의 특성을 고려하여 본문에는 평균 대신 응답 분포를 보고한다. 검토는 개인 식별 정보를 수집하지 않는 전문가 자문 성격의 서면 검토로 수행하였고, 소속 기관명과 구체적 직위는 수집·보고 단계에서 익명화하였다. 패널이 IT·현업 관점 중심으로 구성되어 법률·감독당국 관점이 제한적이라는 점은 본 연구의 한계로 명시한다. <표 2>는 전문가 패널의 익명화된 구성을 요약한 것이다.

<표 2> 전문가 패널의 익명화된 구성

구분	주요 경력 분야	경력 수준	관련 전문성 또는 검토 관점
E1	금융권 IT 개발·운영 및 IT 총괄	40년 이상	금융 IT 운영 통제, 망분리 환경, 시스템 적용 가능성 검토
E2	금융 IT 거버넌스 및 경영	30년 이상	금융 IT 거버넌스, 준법/감사, Private LLM 도입 전략 검토
E3	금융 현업·디지털 추진 및 영업관리	20년 이상	현업 업무 적용성, 대고객 산출물 활용, 사용자 수용성 검토

주: 전문가의 소속 기관명, 구체적 직위, 특정 프로젝트명 등 개인 식별 가능 정보는 익명성 보호를 위해 제외한다.
출처: 본 연구의 전문가 검토 자료를 바탕으로 저자 작성.

V. Private LLM 책임운영 프레임워크

1. 프레임워크의 목적

본 장은 4장의 절차에 따른 설계 분석과 정교화가 반영된 최종 프레임워크(v2)를 제시하며, 도출과 보완의 경과는 6장에 보고한다. 본 연구가 제안하는 프레임워크의 목적은 망분리 한국 금융기관에서 Private LLM을 단순한 생성모델이 아니라 책임 가능한 업무보조 시스템으로 운영하기 위한 통합 구조를 제시하는 데 있다. 본 연구는 실제 금융기관에서 중요한 질문인 “누가 검토하는가”, “어떤 산출물에 통제를 적용하는가”, “어떤 근거와 기록이 남아야 하는가”, “사후에 무엇을 재구성할 수 있어야 하는가”에 답하고자 한다(Mökander et al., 2024; Ojewale et al., 2026).

본 프레임워크는 Private LLM의 학습·개발·배포·운영 전 과정을 모두 동일한 깊이로 다루는 것이 아니라, 망분리 한국 금융기관에서 Private LLM이 실제 업무에 사용되는 운영 및 추론 단계를 중심으로 설계된다. 특히 AI가 생성한 답변, 요약, 추천안, 설명문 등 업무 산출물이 내부통제, 대고객 설명, 규제판단 보조, 감사 대응에 영향을 줄 수 있는 경우를 주요 대상으로 한다.

2. 설계 원칙

본 프레임워크는 네 가지 설계 원칙에 기반한다.

첫째, 보조수단성 원칙이다. Private LLM은 금융기관의 의사결정을 자동 대체하는 시스템이 아니라, 승인된 정보에 근거해 초안, 추천안, 요약안을 제시하는 보조수단으로 제한된다(금융위원회, 2026; OECD, 2024).

둘째, 위험비례적 통제 원칙이다. 모든 사용사례에 동일한 통제를 적용하는 대신, 위험수준이 높아질수록 검색 범위는 좁아지고, 생성 자유도는 줄어들며, 인간 검토와 승인 기록 요건은 강화된다(NIST, 2024; U.S. Department of the Treasury, 2024).

셋째, 근거우선성 원칙이다. 지식집약 업무는 승인된 코퍼스 기반 RAG로 운영되어야 하며, 답변은 반드시 근거문서와 연결되어야 한다. 근거가 부족하거나 상충 문서가 존재할 경우 시스템은 응답을 유보하거나 인간 검토 단계로 전환되어야 한다(Lewis et al., 2020; Gao et al., 2023; Yu et al., 2024).

넷째, 감사추적성 원칙이다. Private LLM은 누가 어떤 질의를 했고, 어떤 문서가 검색되었으며, 어떤 모델 버전이 사용되었고, 누가 검토·승인·반려·오버라이드(override) 했는지를 재구성할 수 있어야 한다(Raji et al., 2020; Ojewale et al., 2026).

3. 통제 대상 산출물의 범위

본 프레임워크는 Private LLM 산출물을 세 가지 기준으로 구분한다. 첫째, 최종 산출물과 중간 산출물을 구분한다. 최종 산출물은 고객, 감독기관, 내부 의사결정자 또는 감사 절차에 제출되거나 영향을 미치는 문서·요약·추천안이다. 중간 산출물은 아이디어 탐색, 표현 개선, 초안 작성 등 내부 사고 보조 과정에서 생성되는 임시 결과물이다.

둘째, 구조화 산출물과 비구조화 산출물을 구분한다. 데이터, 표, 차트, 수치 비교와 같은 구조화 산출물은 원천 데이터와 산출 로직을 통해 상대적으로 명확한 감사추적이 가능하다. 반면 문구, 문단, 설명문과 같은 비구조화 산출물은 완전한 감사가능성을 보장하기 어렵다. 따라서 비구조화 산출물은 근거문서, 문서 버전, 작성 경로, 검토자 의견, 승인·수정 이력을 함께 보존하여 사후 재구성 가능성을 높이는 방식으로 접근한다(Mökander et al., 2024; Ojewale et al., 2026).

셋째, 내부 참고 산출물, 대고객 산출물, 규제판단 보조 산출물을 구분한다. 내부 참고 산출물은 제

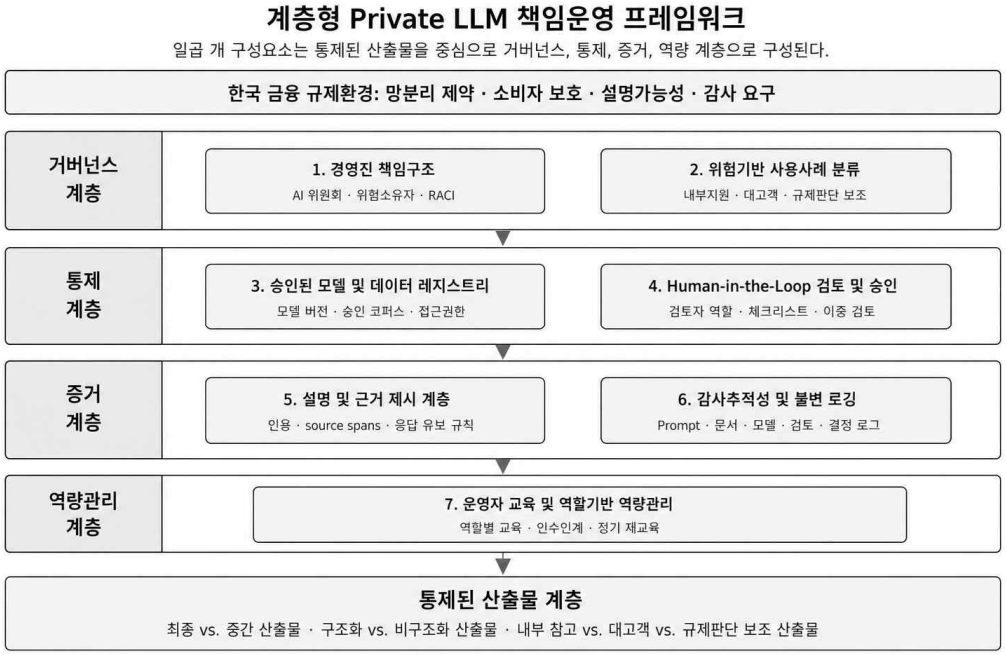
한된 자동화 보조가 가능하지만, 대고객 산출물은 설명의무와 소비자 오인 가능성을 고려해 강화된 검토와 안내문이 필요하다. 규제판단 보조 산출물은 AI의 자동결정이나 자동조치를 금지하고, 반대 근거와 불확실성을 함께 제시하며, 독립적 이중 검토와 감사로그를 요구한다(NIST, 2024; OECD, 2024). 이 절의 산출물 범위 구분은 초기 설계안에는 명시되지 않았던 것으로, 정교화 과정에서 전문가 피드백을 반영하여 신설되었다(6.5절 참조).

4. 핵심 구성요소

본 연구는 Private LLM 책임운영 프레임워크가 다음 일곱 요소로 구성되어야 한다고 본다.

1. 경영진 책임구조
2. 위험기반 사용사례 분류
3. 승인된 모델·데이터 레지스트리
4. HITL 검토 및 승인 기록
5. 설명 및 근거 제시 계층
6. 감사추적성과 불변 로그
7. 운영자 교육 및 역할기반 역량관리

[그림 2]는 본 연구가 제안하는 Private LLM 책임운영 프레임워크를 계층형 구조로 제시한다. 프레임워크는 경영진 책임구조와 사용사례 위험분류를 포함하는 거버넌스 계층, 승인된 모델·데이터와



[그림 2] 계층형 Private LLM 책임운영 프레임워크

주: 4개 계층은 V.4절의 7개 핵심 구성요소(①~⑦)에 대응함.
출처: 저자 작성, 생성형 AI 도구를 활용하여 시각화.

HITL 검토를 포함하는 통제 계층, 설명·근거 제시와 불변 로그를 포함하는 증거 계층, 그리고 운영자 교육과 역할기반 역량관리를 포함하는 역량관리 계층으로 구성된다. 이 구조는 Private LLM이 단순 모델이 아니라 정책·운영·검토·감사·교육 요소가 결합된 책임운영 시스템임을 보여준다.

초기 설계안(v1)은 위의 1~6에 해당하는 여섯 요소로 구성되었다. 설계 분석 과정의 전문가 검토에서 운영자와 검토자의 접근수준, 이해도, 역할 변경 시 학습 프로그램의 필요성이 제기됨에 따라(6.5절 참조), 정교화 단계에서 운영자 교육 및 역할기반 역량관리를 일곱째 구성요소로 추가하였다. 운영자, 검토자, 승인권자, 감사 담당자는 각자 다른 관점에서 모델 산출물을 이해해야 하므로, 역할별 교육, 변경 시 인수인계, 정기 재교육, 검토 체크리스트 숙지가 통제 항목에 포함되어야 한다.

5. HITL 검토 규칙

HITL 검토는 승인 버튼을 누르는 형식적 절차가 아니라, 검토자가 무엇을 확인해야 하는지 명확히 정의된 실질적 통제행위여야 한다. 본 프레임워크는 위험수준에 따라 단일 검토 또는 독립적 이중 검토 절차를 적용한다. 독립적 이중 검토는 동일한 결재라인의 반복 승인이 아니라, 업무 책임자와 준법·리스크 담당자 등 서로 다른 책임 관점의 검토자가 산출물의 근거, 사용 맥락, 고객 또는 규제 영향, 불확실성을 확인하는 절차를 의미한다.

특히 고객영향 또는 규제영향이 있는 산출물은 담당자의 명시적 검토와 승인 기록을 남겨야 하며, 자동 조치나 자동 의사결정으로 연결되어서는 안 된다(OECD, 2023; NIST, 2024).

6. 운영 통제 흐름

본 프레임워크의 운영 통제 흐름은 입력 → 검색 → 생성 → 검토 → 승인 또는 반려 → 로그의 구조를 기본으로 한다. 다만 고위험 시나리오에서는 생성 직후 인간 검토 단계로 전환되고, 근거 부족, 권한 위반, 정책 위반이 탐지되면 승인 단계로 가지 못한 채 반려되도록 설계된다.

[그림 3]은 Private LLM의 위험비례적 운영 통제 흐름을 보여준다. 이 흐름은 사용자의 질의 입력에서 시작해 인증 및 역할기반 접근통제, 사용사례 위험분류, 산출물 유형분류, 승인 코퍼스 검색, 시나리오별 생성, 근거 및 정책 검증, 인간 검토, 승인·반려·수정, 배포, 불변 로그 보존으로 이어진다. 각 단계에서는 프롬프트 로그, 검색 문서 해시, 모델 버전, 검토자 결정, 승인·반려 기록, 배포 기록과 같은 증거 객체가 생성·보존되며, 이는 사후 재구성 가능성을 확보하기 위한 핵심 기반이 된다.

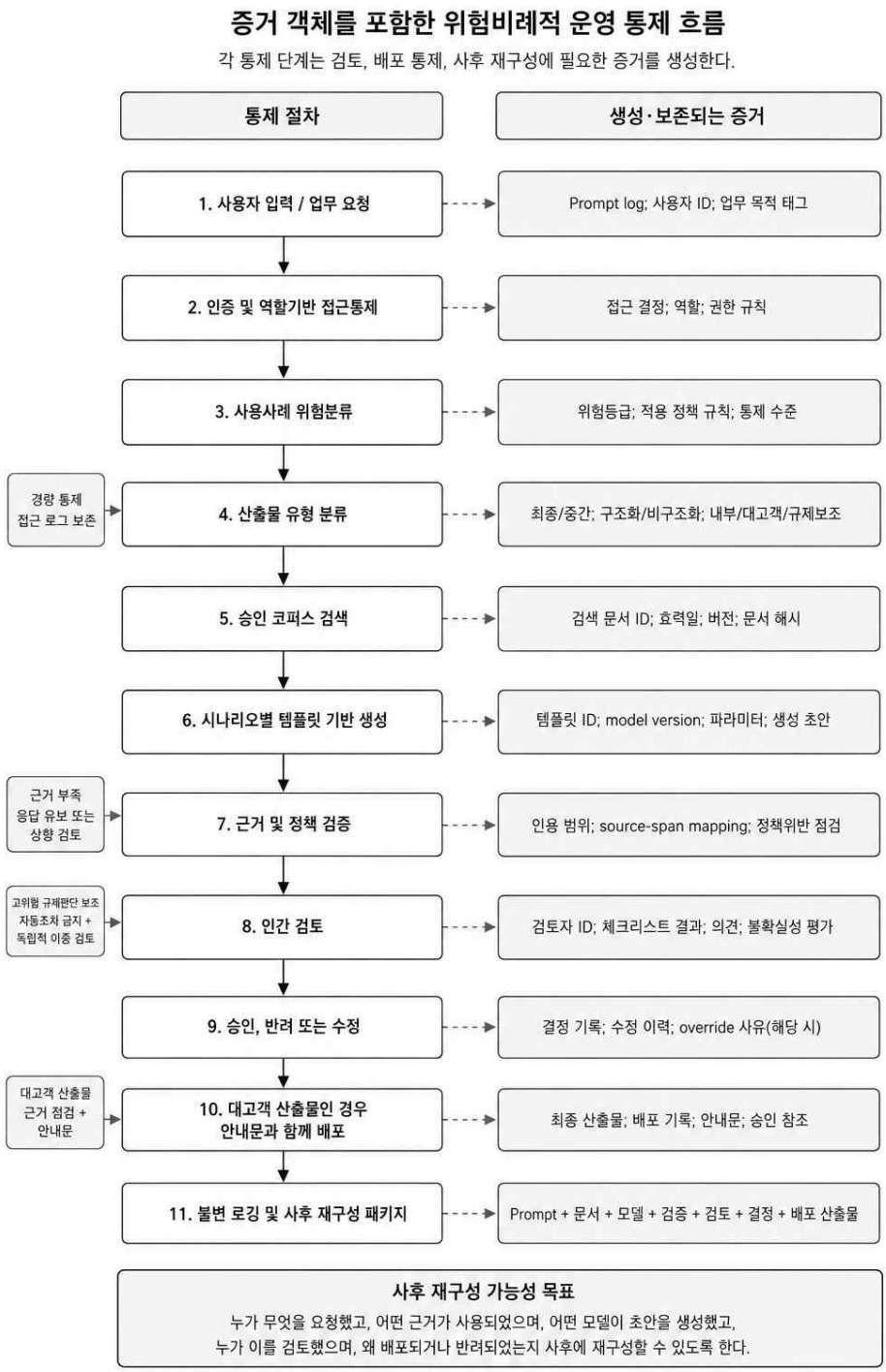
운영 통제는 입력 단계에서 통합인증(SSO), 역할기반 접근권한, 질의 목적 태깅, 민감정보 패턴 차단을 적용하고, 검색 단계에서는 승인 코퍼스만 검색하며 문서 효력일과 버전을 확인한다. 생성 단계에서는 시나리오별 템플릿과 응답 유보 규칙을 적용하고, 검토 단계에서는 위험등급에 따라 단일 검토 또는 독립적 이중 검토로 분기한다. 마지막으로 프롬프트, 검색 문서, 모델 버전, 검토자, 승인·반려, 오버라이드, 배포 기록을 저장해 사후 재구성 가능성을 확보한다(한국인터넷진흥원, 2024, 2025; Ojewale et al., 2026).

7. 대표 시나리오별 적용

내부 초안·아이디어 보조는 저위험 시나리오로 분류되며, 중간 산출물에 해당하는 한 접근통제와 자동 로그 중심의 경량 통제만 적용된다. 다만 중간 산출물이 외부에 회람되거나 최종 산출물로 전환되는 시점에는 해당 산출물 유형의 상위 위험등급 통제가 적용된다.

내부 준법 Q&A는 중위험 시나리오로 분류되며, 승인된 내부 준법규정, 공시된 법규 및 가이드라인, 승인 FAQ만 검색 대상으로 허용한다. 근거 부족 시 답변을 유보하고, 외부 회람이나 정책 확정 용도라면 담당자의 추가 검토가 필요하다.

상품설명 요약은 고위험 시나리오로 분류되며, 최종 승인된 상품설명서, 약관, 핵심설명서, 투자설



[그림 3] 증거 객체를 생성하는 위험비례적 운영 통제 흐름

주: 운영 단계는 V.3절의 산출물 분류와 V.6절의 위험비례적 운영 통제 흐름에 대응함.
출처: 저자 작성, 생성형 AI 도구를 활용하여 시각화.

명서만 허용된다. 문서 해시 검증, 최신본 확인, 근거 구간 연결(source span mapping), 미확인 수치 자동 마스킹을 적용하며, 상품부서와 준법 또는 법무 담당자의 독립적 이중 검토와 승인 기록을 요구한다. 대고객 산출물에는 AI 보조 여부, 근거문서, 최종 확인 필요성을 알리는 안내문(disclaimer)을 포함할 수 있다(「금융소비자 보호에 관한 법률」 제19조).

고위험 규제판단 보조는 최고위험 시나리오로 분류되며, 의심거래보고(STR) 및 자금세탁방지(AML), 제재대상 탐지, 거래제한 검토 등 규제상 영향이 큰 업무를 상정한다. AI는 추천안만 제시하고 최종 판단은 인간이 수행한다. 생성 결과는 판단 근거, 반대 근거, 불확실성, 추가 확인 필요사항 구조만 허용하며, 자동 조치 호출은 금지된다(NIST, 2024; OECD, 2024).

Ⅵ. 설계 분석 결과 및 프레임워크 정교화

1. 문헌 기반 요구사항 도출

문헌 기반 요구사항 도출은 한국 금융권의 규제·정책 문서와 국제 감독·표준 문헌에서 반복적으로 요구되는 통제 요소를 추출하고, 이를 프레임워크 요건으로 재구성하는 방식으로 수행하였다. 먼저 각 문헌에서 책임성, 보안, 설명가능성, 소비자보호, 모델·데이터 관리, 인간 검토, 감사추적성, 운영 복원력과 관련된 요구사항 후보를 추출하였다. 이후 의미가 유사한 항목을 통합하고, 한국 금융권 Private LLM 운영에서 직접 통제항목으로 전환 가능한 항목을 중심으로 초기 요구사항 R1~R14를 정리하였으며, 정교화 단계에서 전문가 피드백을 반영한 R15~R17이 추가되었다. <표 3>은 Private LLM 책임운영을 위한 핵심 요구사항과 실무 구현 포인트를 요약한 것이다.

2. 정책-프레임워크 매핑

정책 매핑의 목적은 제안 프레임워크가 한국 금융권의 제도적 요구를 실제로 포괄하는지 확인하는 데 있다. 매핑은 각 정책·법령의 핵심 요구를 프레임워크 구성요소와 연결하고, 해당 요구가 운영 단계에서 어떤 증거로 보존되어야 하는지를 함께 식별하는 방식으로 수행하였다. 추적가능성을 확보하

<표 3> Private LLM 책임운영을 위한 요구사항

ID	요구사항	실무 구현 포인트	근거 범주
R1	최고 의사결정기구와 독립 위험관리 조직 구축	AI 운영위원회, 위험소유자, 역할책임배분(RACI)	정책/거버넌스
R2	사용사례 위험등급화	내부지원/대고객/고위험 규제판단	정책/운영
R3	보조수단성 및 인간 최종책임 보장	자동조치 금지 목록, 필수 검토자 게이트	정책/거버넌스
R4	승인된 모델 및 데이터만 사용	모델 레지스트리, 출처 검증, 반입 승인	보안/거버넌스
R5	승인 코퍼스 기반 RAG 운영	효력일, 버전 잠금, 화이트리스트 검색	RAG/준법
R6	근거 연결 응답 또는 응답 유보	인용 필수 템플릿, 응답 유보 임계값	RAG/감사추적
R7	고객영향 및 규제영향 산출물 인간 검토	단일 검토 또는 독립적 이중 검토	거버넌스/준법
R8	설명, 정정, 재산출 절차 제공	설명 패키지, 근거 구간 연결, 이의제기 대응	법/소비자보호
R9	전 생애주기 보안 적용	프롬프트 필터링, 오염 점검, 평가 보안	보안

ID	요구사항	실무 구현 포인트	근거 범주
R10	사전·사후 시험과 정기 모니터링	시험·평가·검증·확인(TEVV), 레드팀 점검, 드리프트 점검	리스크관리
R11	제3자 및 오픈소스 공급망 리스크 관리	라이선스 검토, 업데이트 승인, 무결성 스캔	공급망/보안
R12	불변 로그와 추적성 확보	프롬프트·문서·버전·검토자·오버라이드 로그	감사추적
R13	비상정지 및 수동 대체 절차	비상정지(kill switch), 수동 대체 절차	운영/복원력
R14	편향 및 불공정 위험 점검	공정성 시험세트, 영향 검토	설명가능성/공정성
R15	산출물 유형별 통제수준 구분	최종/중간, 구조화/비구조화, 내부/대고객/규제판단	전문가 피드백/운영
R16	운영자 및 검토자 역량관리	역할별 교육, 재교육, 인수인계, 체크리스트	전문가 피드백/거버넌스
R17	대고객 AI 산출물 안내문 관리	AI 보조 안내, 근거, 최종 확인 필요성	전문가 피드백/소비자보호

주: R1~R14는 문헌·정책 문서에서 도출한 초기 요구사항이며, R15~R17은 6.5절의 전문가 피드백을 반영하여 정교화 단계에서 추가된 요구사항이다.
출처: 관련 정책·문헌 및 전문가 피드백을 바탕으로 저자 작성.

기 위해 대응 요소는 5.4절의 구성요소 번호(①~⑦)로 표기하였다. 하나의 정책 요구가 복수의 구성요소에 걸쳐 나타나는 경우에는 가장 직접적인 통제 책임을 갖는 구성요소를 우선 대응 항목으로 표시하였고, 감사·설명·보안처럼 중첩되는 요구는 보존 증거 항목을 통해 연결성을 보완하였다. <표 4>는 주요 정책·법령 요구와 프레임워크 구성요소의 대응 관계를 요약한다.

<표 4> 한국 금융권 정책·법령과 프레임워크 구성요소 매핑

정책/법령	핵심 요구	대응 프레임워크 요소	보존해야 할 증거
금융분야 AI 가이드라인(2021)	책임성, 데이터 안전성, 투명성, 소비자 권리	① 경영진 책임구조, ② 위험기반 사용사례 분류	AI 거버넌스 문서, 책임자 지정
금융분야 AI 보안 가이드라인	데이터·학습·검증·챗봇 보안	③ 승인된 모델·데이터 레지스트리, ⑥ 감사추적성과 불변 로그(접근통제 포함)	보안 체크리스트, 접근 로그
금융권 생성형 AI 활용 지원 방안	상용 AI와 내부 설치형 AI 활용 방향	② 위험기반 사용사례 분류, ③ 승인된 모델·데이터 레지스트리(모델 반입 승인)	모델 반입 승인, 코퍼스 목록
금융분야 인공지능 가이드라인(2026)	거버넌스, 합법성, 보조수단성, 신뢰성, 금융안정성, 신의성실, 보안성	①~⑦ 프레임워크 전체	원칙–통제 매핑표, 최종책임 수행체계
전자금융감독규정 제15조	망분리, 예외 승인, 대체 통제	프레임워크 운영 전제(폐쇄망 배치), ① 경영진 책임구조(예외 승인), ⑥ 불변 로그(예외 승인 기록)	망구성도, 예외 승인 기록
신용정보법 제36조의2	자동화평가 설명·정정·재산출	⑤ 설명 및 근거 제시 계층	설명 양식, 재산출 기록

정책/법령	핵심 요구	대응 프레임워크 요소	보존해야 할 증거
금융소비자보호법 제19조	이해 가능한 설명의무	④ HITL 검토(대고객 이중 검토), ⑤ 설명 및 근거 제시(안내문)	고객 설명자료, 승인 기록
개인정보위 AI 안내서	데이터 처리 적법성, 안전조치	③ 승인된 모델·데이터 레지스트리	데이터 출처, 비식별 처리 기록
KISA 제로트러스트/AI 보안	최소권한, 지속 검증, 무결성	⑥ 감사추적성과 불변 로그, ⑦ 운영자 교육 및 역량관리	IAM 정책, 산출물 해시, 교육 이력

출처: 표에 제시한 정책·법령 및 가이드라인을 바탕으로 저자 작성.

3. 시나리오 기반 통제 매트릭스

프레임워크의 적용 가능성은 내부 준법 Q&A, 상품설명 요약, 고위험 규제판단 보조라는 세 가지 대표 시나리오에 대한 통제 매트릭스를 통해 검토하였다. 세 시나리오는 금융권 Private LLM 운영에서 산출물의 영향 범위가 단계적으로 커지는 전형적 사용 맥락을 반영한다. 내부 준법 Q&A는 내부 참고용 비구조화 산출물, 상품설명 요약은 대고객 최종 산출물, 고위험 규제판단 보조는 규제·내부통제 의사결정에 영향을 줄 수 있는 최고위험 보조 산출물을 대표한다. <표 5>는 저위험 참조 시나리오를 포함하여, 위험수준이 높아질수록 허용 데이터 범위, 생성 자유도, 검토 및 승인 요건, 로그 요구가 강화되는 구조를 보여준다. 저위험 시나리오는 경량 통제만 적용되므로, 설계 분석은 통제 설계가 실질적으로 문제 되는 중위험 이상의 세 시나리오를 중심으로 수행하였다.

<표 5> 시나리오 기반 통제 매트릭스

시나리오	위험수준	산출물 유형	허용 데이터	주요 통제
내부 초안·아이디어 보조	저위험	중간 산출물, 비구조화 텍스트	일반 내부 업무문서, 공개 자료	접근통제와 자동 로그 중심의 경량 통제, 외부 회람·최종 전환 시 상위 등급 통제 적용
내부 준법 Q&A	중위험	내부 참고 산출물, 비구조화 텍스트	승인된 내부 준법규정, 공시 규정, 승인 FAQ	승인 코퍼스 검색, 근거 표시, 근거 부족 시 유보, 외부 회람 시 검토
상품설명 요약	고위험	대고객 최종 산출물, 혼합형 산출물	최종 승인 상품설명서, 약관, 핵심설명서	문서 해시 검증, 근거 구간 연결, 수치 마스킹, 독립적 이중 검토, 안내문
고위험 규제판단 보조	최고위험	규제판단 보조 산출물	승인 케이스 파일, 내부 규정, 규제 기준	자동조치 금지, 반대근거 표시, 독립적 이중 검토, 불변 감사로그

주: 저위험 시나리오는 비교 기준으로 제시하며, 설계 분석은 중위험 이상의 세 시나리오를 중심으로 수행함.
출처: 저자 작성.

4. FMEA 기반 실패 모드 분석

FMEA 분석 결과, 고위험 규제판단 보조 시나리오의 실패 모드가 가장 높은 우선순위로 나타났다. 근거가 약한 추천안이 최종 판단에 영향을 미치는 경우와 검토자 과신에 따른 형식적 승인이 각각 RPN 60으로 가장 높았고, 설명·근거 패키지의 부재(RPN 50)와 비구조화 텍스트에 대한 감사가능성 과대 주장(RPN 48)이 그 뒤를 이었다. 본문에는 가독성을 위해 <표 6>과 같이 RPN 상위 위험을 중

심으로 축약표를 제시하고, 상세 FMEA는 부록 <표 A1>에 제시한다.

<표 6> FMEA 기반 주요 위험과 통제 시사점

위험영역	핵심 실패 모드	RPN	주요 통제
고위험 규제판단 보조	근거 약한 추천안이 최종 판단에 영향	60	자동결정 금지, 반대근거 표시, 독립적 이중 검토
고위험 규제판단 보조	검토자 과신과 형식적 승인	60	검토의견 필수, 역할 분리, 검토자 교육
설명·근거 패키지	설명·근거 패키지 부재	50	구조화 설명 번들, 재현 가능한 로그
비구조화 텍스트	감사가능성 과대 주장	48	완전 감사가능성 대신 사후 재구성 패키지
상품설명 요약	수수료·위험고지 누락	45	필수 템플릿, 근거 구간 연결, 독립적 이중 검토
공급망 보안	외부/오픈소스 구성요소 취약점	40	모델 레지스트리, 업데이트 승인, 무결성 검증

주: 본 표의 RPN은 현장 사고 빈도나 손실 규모에 근거한 정량적 추정치가 아니라, 연구진이 문헌·정책 요구와 시나리오 영향도를 바탕으로 부여한 설계 단계의 상대적 우선순위 지표임. S/O/D의 세부 기준은 부록 <표 A2>에 제시함.
출처: 본 연구의 FMEA 분석 결과를 바탕으로 저자 작성.

5. 전문가 피드백 기반 프레임워크 정교화

4.4절의 절차에 따라 수행한 공통 문항 4개에 대한 응답 결과, 위험등급별 차등 통제 구조는 3인 모두 5점으로 평가하여 가장 강한 지지를 받았다. 규제 및 운영 환경 반영성은 5점·4점·3점 각 1인, 설계 분석 접근의 합리성은 5점 2인·4점 1인으로 전반적으로 긍정적인 평가를 받았다. 반면 설명·근거 제시, 불변 로그, 승인 절차를 결합한 감사가능성 항목은 4점 2인·3점 1인으로 상대적으로 낮게 평가되었으며, 이는 LLM 산출물의 유형에 따라 감사가능성 확보 수준이 달라질 수 있음을 시사한다. 이 결과는 정량적 타당성 검증의 근거가 아니라 후속 설계 보완을 위한 형성적 피드백으로 해석한다.

전문가 피드백은 단순한 의견 수렴에 그치지 않고, 프레임워크의 구성요소와 통제수준을 조정하는 근거로 활용되었다. 설문 문항과 익명화된 응답 요약은 각각 부록 <표 B1>과 <표 B2>에 제시하였다. <표 7>은 주요 전문가 피드백 주제와 그에 따른 프레임워크 수정 반영 내용을 정리한 것이다.

<표 7> 전문가 피드백과 프레임워크 수정 반영

피드백 주제	응답 근거	수정 반영
위험기반 차등 통제 지지	세 응답자 모두 위험등급별 차등 통제 항목을 5점으로 평가	위험비례적 통제를 중심 설계 원칙으로 유지
적용 범위 명확화 필요	최종 산출물과 중간 산출물 구분 필요 제기	통제 대상 산출물 범위 절(5.3절) 신설
감사가능성 수위 조정 필요	구조화 산출물과 문구·문단형 산출물의 감사가능성 차이 지적	감사가능성을 사후 재구성 가능성으로 재정의
HITL 형식화 우려	검토 대상과 승인 항목 불명확성 지적	검토자 역할, 검토 항목, 독립적 이중 검토, 오버라이드 기록 추가
운영자 역량관리 필요	역할별 이해도, 접근수준, 담당자 변경 시 교육 필요 지적	운영자 교육 및 역할기반 역량관리를 일곱째 구성요소로 추가
대고객 산출물 보호장치 필요	AI 보조 여부와 최종 확인 필요성 안내 제안	대고객 안내문 및 근거 제시 통제 추가

주: 본 표는 본 연구의 전문가 설문 응답 3건을 연구자가 주제별로 정리한 것이다.
출처: 본 연구의 전문가 설문 응답을 바탕으로 저자 작성.

6. 연구 질문별 분석 결과와 기여의 정렬

<표 8>은 본 연구의 연구질문, 분석 절차, 주요 결과, 기여가 어떻게 연결되는지를 요약한다. 이는 제안 프레임워크가 단순한 정책 요약표가 아니라, 요구사항 도출과 시나리오 적용, 실패 모드 분석, 전문가 피드백을 거쳐 정교화된 설계 산물임을 보여주기 위한 것이다.

<표 8> 연구질문, 분석 절차, 주요 결과 및 기여의 정렬

연구질문	분석 절차	주요 결과	기여
RQ1. 핵심 책임성 및 거버넌스 요소	문헌·정책 기반 요구사항 도출(R1~R14) 및 전문가 피드백 반영(R15~R17)	경영진 책임구조, 위험기반 분류, 모델·데이터 레지스트리, HITL, 설명·근거, 감사로그, 역량관리의 7개 요소 도출	Private LLM을 단일 모델이 아니라 책임 가능한 운영 시스템으로 재정의
RQ2. 규제·가이드라인 충족을 위한 통제 구조	정책–프레임워크 매핑, 시나리오 기반 통제 매트릭스	위험수준이 높아질수록 허용 데이터, 생성 자유도, 인간 검토, 승인기록, 로그 요건이 강화되는 통제 구조 제시	망분리 금융환경에 맞는 위험비례적 산출물 통제 모델 제안
RQ3. 설계 보완 필요사항과 정교화	FMEA, 소규모 전문가 피드백	감사가능성 수위 조정, HITL 형식화 방지, 산출물 범위 명확화, 운영자 역량관리, 대고객 안내문 필요성 확인 및 반영	설계 연구 단계에서 후속 파일럿과 실증 검증으로 연결 가능한 보완 항목 제시

주: RQ1~RQ3은 각각 요구사항 도출, 정책·시나리오 분석, FMEA 및 전문가 피드백을 통해 프레임워크의 구성, 통제 구조 및 정교화 결과와 연결됨.
출처: 본 연구의 분석 결과를 바탕으로 저자 작성.

Ⅶ. 논의

본 연구의 가장 중요한 시사점은 망분리 금융환경에서 Private LLM의 도입 문제가 더 이상 “모델을 만들 수 있는가”에 머물지 않는다는 점이다. <표 8>에서 정리한 바와 같이, 본 연구의 분석 결과는 핵심 거버넌스 요소의 도출(RQ1), 위험비례적 통제 구조의 설계(RQ2), 그리고 FMEA와 전문가 피드백을 통한 보완 필요사항의 식별과 반영(RQ3)으로 연결된다. 기술적 실현가능성만으로는 실제 금융기관에서의 책임 있는 활용을 보장할 수 없다. 한국 금융권에서는 망분리, 설명의무, 소비자보호, 승인 절차, 외부 서비스 제한, 문서 최신성 통제와 같은 요소가 동시에 작동하므로, Private LLM은 단일 모델이 아니라 운영 통제와 감사 구조를 포함한 사회기술적 시스템으로 이해되어야 한다.

본 연구는 위험비례적 통제 원칙이 한국 금융권 Private LLM 운영에서 특히 중요하다고 본다. 모든 사용사례에 동일한 수준의 통제를 적용하면 운영 효율이 과도하게 저하되고, 반대로 모든 사용사례를 동일하게 자동화하면 규제위반과 소비자 피해 가능성이 커진다. 따라서 내부 준법 Q&A, 상품설명 요약, 고위험 규제판단 보조처럼 업무 맥락에 따른 위험등급화는 단순 편의적 분류가 아니라 프레임워크의 핵심 전제다(NIST, 2024; OECD, 2024).

또한 본 연구는 Private LLM의 핵심 통제를 모델 내부 정교화보다 근거 연결(grounding), 설명 제시(explanation), 감사추적성(auditability)에 둔다. 다만 전문가 피드백은 감사가능성을 단일 개념으로 다루기보다 산출물 유형에 따라 구분해야 함을 시

사한다. 데이터, 표, 차트와 같은 구조화 산출물은 원천 데이터와 산출 로직을 통해 상대적으로 명확한 추적이 가능하지만, LLM이 생성한 문구·문단형 산출물은 의미 해석과 표현 선택이 개입되므로 완전한 감사가능성을 확보하기 어렵다. 따라서 본 연구에서 말하는 감사가능성은 모든 생성과정을 완전히 설명한다는 의미가 아니라, 승인된 근거문서, 문서 버전, 모델 버전, 검토자 의견, 승인·반려·수정 이력을 보존하여 사후 재구성 가능성을 높이는 운영적 감사가능성을 의미한다(Mökander et al., 2024; Ojewale et al., 2026). 국제 기준과의 비교(<표 1>)에서 본 프레임워크의 차별성 역시 원칙이나 관리체계 일반 요건을 산출물 단위의 통제 절차와 증거 객체로 운영화한 데 있다.

HITL 역시 단순히 결재 단계를 추가하는 것으로는 충분하지 않다. 검토자가 확인해야 할 근거, 불확실성, 고객 또는 규제 영향, 반대 근거를 명확히 제시하고, 검토자의 판단과 수정 이력을 기록하는 방식으로 설계되어야 한다. 또한 검토자 교육과 담당자 변경 시 인수인계 절차가 병행되지 않으면 동일한 프레임워크라도 조직별로 상이한 결과를 낳을 수 있다(NIST, 2024).

본 연구의 의미는 특정 금융기관의 내부통제 설계를 넘어, 고위험 규제산업이 생성형 AI를 사회적으로 수용 가능한 방식으로 도입하기 위한 운영 모델을 제시한다는 데 있다. 향후 AI가 금융, 공공, 의료, 교육 등 책임성이 요구되는 영역으로 확산될수록, 단순한 성능 중심 접근보다 산출물 책임성, 인간 검토, 감사추적성을 결합한 운영 프레임워크의 중요성은 더욱 커질 것이다.

본 연구에는 몇 가지 한계가 있다. 첫째, 검토는 설계 타당성에 초점을 두고 있어 실제 금융기관 운영 환경에서의 효과성과 효율성을 직접 입증하지 못한다. 둘째, 정책 매핑과 요구사항 도출은 한국 금융권 공식 문서와 일부 국제 기준을 중심으로 수행되었으므로, 다른 규제권역에 그대로 일반화하기에는 제한이 있다. 셋째, FMEA 점수는 연구진이 부여한 설계 단계의 상대적 우선순위 판단으로, 실제 사고 빈도나 손실 규모를 정량적으로 반영한 것이 아니며 단독 채점에 따른 신뢰성의 한계가 있다. 넷째, 요구사항 추출과 코딩 역시 연구진에 의해 수행되어 해석 편향의 가능성을 배제할 수 없다. 다섯째, 전문가 피드백은 회신된 3인의 의견을 중심으로 수행된 소규모 형성적 검토로, 패널이 IT·현업 관점 중심으로 구성되어 법률·감독당국 관점이 제한적이며, 금융권 전체 전문가 집단의 일반화된 의견으로 해석해서는 안 된다.

향후 연구는 법률·감독 관점을 포함하도록 전문가 패널을 확대하고, 제한적 현장 파일럿을 통해 산출물 유형별 통제수준, HITL의 실제 작동 방식, 감사로그의 활용성, 운영자 교육체계의 효과를 검증할 필요가 있다. 또한 근거 충실성(grounding fidelity), 정책 위반율(violation rate), 인용 완전성(citation completeness), 인간 검토 품질, 감사기록 활용성(audit trail usability) 지표를 포함한 위험정렬 평가 벤치마크(risk-aligned evaluation benchmark)를 정교화할 필요가 있다.

Ⅷ. 결론

본 연구는 망분리 한국 금융기관에서 Private LLM을 책임 있게 운영하기 위한 프레임워크를 설계과학연구 접근에 따라 제안하고, 이를 정책 정합성, 통제 완결성, 사후 재구성 가능성의 관점에서 설계 수준으로 정교화하였다. 기존 연구가 폐쇄망 환경에서의 LLM 구축 가능성, 도메인 적응, 성능 비교에 초점을 맞추었다면, 본 연구는 그 위에 책임성 계층을 추가하여 경영진 책임구조, 위험기반 사용사례 분류, 승인된 모델 및 데이터 체계, 인간 검토 및 승인 기록, 설명 및 근거 제시, 불변 로그와 감사추적성, 운영자 역량관리를 통합한 운영 프레임워크를 제시하였다. 특히 산출물의 3축 분류와 사용

사례 위험수준을 결합하여 통제 강도를 차등화하고, 완전한 설명가능성보다 사후 재구성 가능성에 초점을 둔 감사 구조를 제안했다는 점에서 단순한 체크리스트와 구별된다.

문헌 기반 요구사항 도출, 한국 금융권 정책 및 법령 매핑, 시나리오별 통제 매트릭스, FMEA 기반 실패 모드 분석, 소규모 전문가 피드백을 통해 본 연구는 한국 금융권 Private LLM의 핵심 과제가 단순 모델 정확도 향상이 아니라, 허용 가능한 자동화 범위를 설정하고, 근거 중심 응답과 실질적 HITL, 그리고 사후 재구성이 가능한 로그 체계를 확보하는 것임을 시사하였다.

다만 본 연구는 설계 연구 단계에 머물러 실제 금융기관 현장에서의 운영 효과를 직접 검증하지 못했으며, 전문가 피드백도 3인의 소표본을 대상으로 한 형성적 검토에 한정되었다. 그러나 규제산업에서 Private LLM을 도입하기 위해 필요한 책임운영 구조를 정식화하고, 후속 실증 연구와 파일럿 검증의 기반을 제시했다는 점에서 의의가 있다. 향후 제한적 현장 파일럿과 추가 전문가 검토가 병행될 때, Private LLM은 한국 금융권에서 단순한 생성 도구가 아니라 책임 있고 감사 가능한 업무보조 인프라로 자리 잡을 수 있을 것이다.

참고문헌

개인정보보호위원회 (2025). 생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서.

금융보안원 (2023). 금융분야 AI 보안 가이드라인.

「금융소비자 보호에 관한 법률」 제19조.

금융위원회 (2021). 금융분야 인공지능(AI) 가이드라인.

금융위원회 (2022). 금융분야 AI 개발·활용 안내서.

금융위원회 (2024a. 12. 12.). 금융권의 AI 활용을 적극 지원하겠습니다: 금융권 생성형 AI 활용 지원 방안 [보도자료]. https://www.fsc.go.kr/no010101/83594

금융위원회 (2024b. 8. 13.). 금융분야 망분리 개선 로드맵 발표 [보도자료]. https://www.fsc.go.kr/no010101/82885

금융위원회 (2025). 금융분야 통합 AI 가이드라인 (안): 금융분야 AI 가이드라인 개정방향.

금융위원회 (2026. 6. 18.). 정부와 금융권이 함께 금융권 AX를 가속화해 나가겠습니다: 금융분야 인공지능 가이드라인 [보도자료]. https://www.fsc.go.kr/no010101/87142

「신용정보의 이용 및 보호에 관한 법률」 제36조의2.

임기범, 오종훈 (2026). 망분리 금융 환경에서의 프라이빗 LLM 구축 및 성능 평가. 미래사회, 17(2), 204-213. https://doi.org/10.22987/jifso.2026.17.2.204

「전자금융감독규정」 제15조.

천예은, 김세빈, 이자윤, 우지환 (2021). 설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구. 한국데이터정보과학회지, 32(2), 283-295. https://doi.org/10.7465/jkdi.2021.32.2.283

한국인터넷진흥원 (2024). 제로트러스트 가이드라인 2.0.

한국인터넷진흥원 (2025). 인공지능(AI) 보안 안내서.

허유경 (2022). 인공지능을 활용한 금융감독의 법적 함의. 증권법연구, 23(1), 221-250. https://doi.org/10.17785/kjsl.2022.23.1.221

Bhattacharyya, A., Yu, Y., Yang, H., Singh, R., Joshi, T., Chen, J., & Yalavarthy, K. (2025). *Model risk management for generative AI in financial institutions.* arXiv preprint arXiv: 2503.15668.

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey.* arXiv preprint arXiv: 2312.10997.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models.* arXiv preprint arXiv:2106.09685.

IEC. (2018). *IEC 60812:2018 Failure modes and effects analysis (FMEA and FMECA).* International Electrotechnical Commission.

ISO/IEC. (2023). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system.* International Organization for Standardization.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33,* 9459-9474.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics, 4,* 1085-1115.

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* NIST AI 600-1.

OECD. (2023). *Generative artificial intelligence in finance.* OECD Artificial Intelligence Papers, No. 9.

OECD. (2024). *Regulatory approaches to artificial intelligence in finance.* OECD Business and Finance Policy Papers.

Ojewale, V., Suresh, H., & Venkatasubramanian, S. (2026). *Audit trails for accountability in large language models.* arXiv preprint arXiv: 2601.20727.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 33-44.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-23. https://doi.org/10.1145/3449081

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models.* arXiv preprint arXiv: 2307.09288.

U.S. Department of the Treasury. (2024). *Artificial intelligence in financial services: Uses, oppor-*

tunities, and risks.

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 preprint arXiv:2303.17564.

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

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). *Evaluation of retrieval-augmented generation: A survey.* arXiv preprint arXiv:2405.07437.

투고일자: 2026. 8. 13.
심사일자: 2026. 9. 29.
게재확정일자: 2026. 10. 1.

미래사회
Journal of Future Society
2026. Vol. 17, No. 3, 222～246.
https://doi.org/10.22987/jifso.2026.17.3.222

Responsible Operations Framework for Private Large Language Models in Network-Isolated Korean Financial Institutions

Ki-Bum Lim Jong-Hoon Oh

Seoul AI School aSSIST University

The adoption of generative artificial intelligence (AI) by financial institutions offers substantial potential for optimizing document processing, compliance support, risk analysis, and customer communication. However, the network separation, security controls, explanation obligations, consumer protection, and auditability requirements in the Korean financial sector constrain the direct adoption of public cloud-based large language models (LLMs). Therefore, this study proposes and refines a responsible operations framework for private LLMs in network-isolated Korean financial institutions using a design science research approach. The proposed framework comprises seven components: executive accountability, risk-based use-case classification, approved model and data registries, human-in-the-loop review and approval records, explanation and evidence provision, immutable logging and audit trails, and operator competency management. Further, it classifies outputs along three dimensions: final versus intermediate, structured versus unstructured, and internal versus customer-facing versus regulatory decision support, and redefines auditability as post-hoc reconstructability. Design weaknesses were identified and addressed through literature and policy mapping, a scenario-based control matrix, failure mode and effects analysis (FMEA), and formative expert feedback. The findings suggest that responsible private LLM operations depend less on the performance of the model than on output-specific controls, evidence provision, human review, and reconstructability. However, the study was a formative design-level evaluation without field validation, and expert feedback was based on a small sample size (n=3).

Key words: Private LLM, financial AI governance, network separation, responsible operations, auditability

부록

<표 A1> FMEA 상세표

번호	시나리오	실패 모드	S/O/D/RPN	주요 통제
A1	내부 준법 Q&A	효력이 지난 규정 검색	4/3/2/24	효력일 필터, 버전 잠금, 출처일 표시
A2	내부 준법 Q&A	근거 없는 답변 또는 인용 환각	4/3/3/36	근거 필수화, 응답 유보 규칙, 검토자 상향 보고
A3	상품설명 요약	핵심 수수료·위험·안내문 누락	5/3/3/45	필수 템플릿, 근거 구간 연결, 독립적 이중 검토
A4	상품설명 요약	수치 또는 조건 오류	4/3/3/36	추출 규칙, 원천 비교, 변경분 검토
A5	고위험 규제판단 보조	근거 약한 추천안이 최종 판단에 영향	5/3/4/60	자동결정 금지, 독립적 이중 검토, 반대 근거 표시
A6	고위험 규제판단 보조	검토자 과신과 형식적 승인	5/3/4/60	검토의견 필수, 역할 분리, 검토자 교육
A7	고영향 산출물 전반	중간·최종 산출물 간 통제 범위 불명확	4/3/3/36	생성·배포 전 산출물 유형분류
A8	비구조화 텍스트 산출물	생성 텍스트의 감사가능성 과대 주장	4/3/4/48	완전 감사가능성 대신 사후 재구성 패키지
A9	고영향 운영 전반	운영자 또는 검토자 역량 격차	4/3/3/36	역할별 교육, 재교육, 인수인계 체크리스트
A10	고위험 규제판단 보조	설명·근거 패키지 부재	5/2/5/50	구조화 설명 번들, 재현 가능한 로그
A11	고위험 규제판단 보조	취약한 외부/오픈소스 구성요소	5/2/4/40	모델 레지스트리, 업데이트 승인, 무결성 검증

출처: 본 연구의 FMEA 분석 결과를 바탕으로 저자 작성.

<표 A2> FMEA S/O/D 점수 기준

점수	심각도(S)	발생가능성(O)	탐지난이도(D)
1	고객영향·규제위반 가능성이 낮고 내부 참고 수준에 그침	예외적 상황에서만 발생 가능	자동 필터·정형 검증으로 쉽게 탐지
2	제한적 내부 수정 또는 재검토 필요	낮은 빈도로 발생 가능	기본 로그·검토 절차로 탐지 가능
3	업무 지연, 추가 검토, 제한적 고객영향 가능	일반 운영 중 발생 가능	검토자가 주의하면 탐지 가능하나 자동 탐지는 제한
4	중대한 내부통제 훼손 또는 규제 리스크 가능	고위험 업무에서 반복 발생 가능	사후 로그 분석 없이는 탐지 곤란
5	소비자 피해, 중대 규제위반, 자동결정 오남용 가능	통제 미비 시 높은 개연성	사전 탐지가 어렵고 사고 후에야 확인 가능

주: 본 기준은 현장 사고 빈도 또는 손실 규모의 정량 추정이 아니라, 설계 단계에서 실패 모드의 상대적 우선순위를 판단하기 위한 휴리스틱 기준이다.

출처: IEC 60812(2018)를 참고하여 본 연구의 설계 비교 목적에 맞게 저자 작성.

<표 B1> 전문가 설문 문항

문항	평가 방식	검토 초점
Q1	5점 리커트 + 의견	프레임워크가 한국 금융권의 규제 및 운영 환경을 적절히 반영하는가
Q2	5점 리커트 + 의견	위험등급별 차등 통제 구조가 실무적으로 타당한가
Q3	5점 리커트 + 의견	설명·근거 제시, 불변 로그, 승인 절차의 결합이 감사 대응에 충분한가
Q4	5점 리커트 + 의견	문헌·정책 매핑, 시나리오 분석, FMEA를 결합한 설계 분석 접근이 합리적인가
Q5	자유 의견	추가하거나 보완해야 할 통제항목, 적용 범위, 운영상 유의사항은 무엇인가

주: 실제 설문은 각 문항별 점수와 자유 의견을 함께 수집하는 구조화 설문 방식으로 수행하였다.
출처: 저자 작성.

<표 B2> 전문가 응답 요약

항목	응답 분포	주요 의견	프레임워크 반영
규제 및 운영 환경 반영성	5점·4점·3점 각 1인	망분리·승인절차·소비자보호 요구를 대체로 반영하나 적용 범위 명확화 필요	산출물 범위와 시나리오별 통제 구분 보강
위험등급별 차등 통제	5점 3인	모든 응답자가 위험비례적 통제 구조를 강하게 지지	프레임워크의 중심 설계 원칙으로 유지
감사 대응 가능성	4점 2인·3점 1인	비구조화 문구·문단형 산출물은 완전한 감사가능성보다 사후 재구성이 현실적	감사가능성을 사후 재구성 가능성으로 재정의
설계 분석 접근의 합리성	5점 2인·4점 1인	정책 매핑·시나리오·FMEA 결합은 설계 취약점 식별에 유용	부록을 통해 S/O/D 기준과 설문 문항 제시

주: 본 표는 전문가 3인의 응답을 익명화해 연구자가 주제별로 요약한 것이며, 소표본 특성을 고려해 평균이 아닌 응답 분포로 제시한다. 통계적 일반화의 근거가 아니라 프레임워크 보완을 위한 형성적 피드백으로 해석한다.
출처: 본 연구의 전문가 설문 응답을 바탕으로 저자 작성.

<표 C1> 분석 대상 정책·법령·기준 문서 목록

구분	문서	발행 주체(연도)
국내 정책	금융분야 인공지능(AI) 가이드라인	금융위원회(2021)
국내 정책	금융분야 AI 개발·활용 안내서	금융위원회(2022)
국내 정책	금융분야 AI 보안 가이드라인	금융보안원(2023)
국내 정책	금융권 생성형 AI 활용 지원 방안	금융위원회(2024a)
국내 정책	금융분야 망분리 개선 로드맵	금융위원회(2024b)
국내 정책	금융분야 통합 AI 가이드라인(안): 개정방향	금융위원회(2025)
국내 정책	금융분야 인공지능 가이드라인	금융위원회(2026)
국내 정책	생성형 AI 개발·활용을 위한 개인정보 처리 안내서	개인정보보호위원회(2025)
국내 정책	제로트러스트 가이드라인 2.0, AI 보안 안내서	한국인터넷진흥원(2024, 2025)
법령	전자금융감독규정 제15조, 신용정보법 제36조의2, 금융소비자보호법 제19조	—
국제 기준	NIST AI 600-1, OECD(2023, 2024), U.S. Treasury(2024)	각 기관
국제 기준	ISO/IEC 42001:2023, EU AI Act(Regulation 2024/1689)	ISO/IEC, EU(2023, 2024)

출처: 각 기관의 공식 정책·법령·기준 문서를 바탕으로 저자 작성.

<표 C2> 요구사항 코딩 예시

원문 요구(요지)	출처	코딩(의미 단위)	도출 요구사항
내부통신망 연결 시스템의 외부 분리·차단, 예외 시 자체 위험성 평가·대체 정보보호통제·승인	전자금융감독규정 제15조	예외 승인 절차, 대체 통제, 승인 기록	R4, R12
일반금융소비자가 이해할 수 있도록 중요한 사항을 설명	금융소비자보호법 제19조	대고객 설명, 이해 가능성, 확인 절차	R7, R8, R17
생성형 AI 위험의 식별과 인간 감독, 시험·평가·검증·확인	NIST AI 600-1	인간 감독, TEVV, 모니터링	R3, R10

출처: 표에 제시한 원문 자료를 바탕으로 저자 작성.