ChatGPT

한국어 금융/보안 분야 데이터셋 제안

금융보안 해커톤에서는 공개된 데이터를 사용해야 하며, 모델과 데이터는 2025년 7월 31일 이전에 공개돼야 하고 MIT, Apache2.0, CC BY 등 **오픈 라이선스** 또는 최소한 **비상업적 이용**을 허용하는 라이선스를 가져야 한다【775931346822941†L161-L167】. 다음은 이러한 조건을 고려해 찾은 한국어 데이터셋과 이용 방안이다.

1. 한국어 법률·재판 데이터

데이터셋	요약	라이선스/근거	활용 방안
LBOX Open	15만 건의 한국 법원 판결문, 8만 건은 법원 오픈데이터, 7만 건은 LBox 데이터베이스에서 추출한 대규모 전례 코퍼스로, 2개 분류(task)와 2개 판결 예측(task)을 포함한다 [1].	오픈리뷰 설명에서 라이선스가 **CC BY-NC-ND 4.0**(저작자 표시·비영리·변경금지)임을 명시 [2].	한국 법령과 판례 기반 질문에 답변하는 RAG 시스템이나 법률 분류 모델을 만들 때 사용 가능. 비상업적 목적이어야 하며, 2차 저작물 금지 조항을 준수해야 한다.
국가법령정보센터 법령/행정규칙	law.go.kr에서 제공되는 모든 법령과 행정규칙은 공공저작물로 자유 이용허락 또는 저작권 보호 범위 밖에 해당(대한민국 정부 저작물)한다. 국내 법률·금융 규제 관련 데이터를 확보할 수 있으며, 일반적으로 출처를 표시하면 사용 가능.	법제처 홈페이지의 공공저작물 이용 지침에 따라 공공기관의 저작물은 자유롭게 사용할 수 있으나 출처 표시가 요구된다.	한국 금융 규제와 개인정보보호법, 전자금융거래법 등 전문 법령을 수집하여 RAG 용도로 활용. Q&A 데이터로 변환하여 한국어 금융 보안 LLM 파인튜닝에도 활용할 수 있다.
대한민국 대법원 공개 판례	대법원이 공개한 판례 검색 서비스에서 판결 요지·판결 이유 등을 제공한다. 사이트 이용약관에서 판례에 대한 저작권은 미표시(공공자료)이며, 대부분 자유롭게 활용 가능(출처 표기 권장).	구체적 라이선스는 명시되지 않았지만 대법원 판례는 공공기록으로 간주된다.	한국 금융 사건·사기 사건 판례를 수집해 요약/질문-답변 형태로 변환하여 데이터셋으로 사용할 수 있다.

2. 한국어 금융·경제 텍스트

데이터셋	요약	라이선스/근거	활용 방안
KB-ALBERT-KO 학습 데이터	국민은행이 ALBERT 구조로 구축한 **경제·금융 특화 한국어 PLM**. 설명에 따르면 일반 도메인 텍스트(위키, 뉴스) 약 25 GB와 금융도메인 뉴스·리포트 약 15 GB가 사용됐다. GitHub 저장소는 **Apache 2.0** 라이선스를 명시 [3].	Apache2.0 라이선스가 부여된 코드와 모델을 통해 동일 데이터 출처(금융 뉴스/리포트) 이용 가능성을 보여주지만, 원본 데이터는 저작권 문제로 포함되지 않는다.	이미 공개된 모델의 파라미터를 LoRA로 미세조정하거나, 동일한 금융 기사 출처(증권사 리포트, 뉴스)의 공공저작물을 새로 수집해 파인튜닝 데이터를 만들 수 있다.

데이터셋	요약	라이선스/근거	활용 방안
한국어 위키백과 (금융/보안)	위키백과 한국어판의 모든 문서는 **CC BY-SA 3.0** 라이선스로 제공된다. 금융 시장, 경제학, 사이버보안 등 주제별 문서를 사용할 수 있다.	CC BY-SA 라이선스 → 2차 저작물도 동일한 라이선스로 공개해야 함.	금융 용어 정의, 해킹 사건, 개인정보 보호 등 설명형 문서를 필터링해 RAG 문서로 저장하거나 질의응답 데이터를 생성.
금융 뉴스 데이터 (AI Hub 등)	AI Hub에는 '경제·금융 뉴스 요약 데이터', '해외 고객과의 채팅 데이터(금융)' 등 여러 한국어 금융 데이터셋이 존재한다. 대부분 공공데이터 사업으로 구축돼 있으며, **출처 표기 후 자유 이용**을 허용하는 공공누리 제 1유형 라이선스가 붙어 있다.	AI Hub 사용 안내에 따르면 한국지능정보사회진흥원을 권리자로 명시하고 출처를 밝히면 자유로운 이용과 변경이 가능하다.	경제 뉴스 기사와 요약, 금융 상담 채팅 데이터를 활용해 챗봇의 대화형 데이터로 확장하거나, RAG 검색용 문서로 저장한다. 사용 시 공공데이터 출처를 표시해야 한다.
금융/보안 분야 한국어 논문·보고서	금융감독원, 한국은행, 한국인터넷진흥원(KISA) 등의 보고서는 PDF 형태로 제공되며, 공공기관 저작물로 자유 이용 가능하다.	각 기관의 공공저작물 규정을 따르면 상업적 사용도 가능하나 출처를 명시해야 한다.	최신 금융 규제, 피싱 사례, 정보보호 가이드라인 등을 RAG 문서로 활용. 또한 문서 내용을 요약하고 Q&A 형식으로 변환해 파인튜닝 데이터로 활용할 수 있다.

3. 한국어 정보보안 데이터

데이터셋	요약	라이선스/근거	활용 방안
KISA 사이버 위협 분석 보고서	한국인터넷진흥원(KISA)이 발행하는 '인터넷침해사고 대응센터 분석보고서', '발간자료' 등은 사이버 공격 동향, 악성코드 분석, 해킹 사례를 담고 있다.	KISA 발간자료는 공공기관 저작물로 자유 이용할 수 있으며, 출처와 저작권 표시를 하면 2차 활용이 가능하다.	최신 한국어 위협 사례를 RAG 소스로 추가하거나 Q&A 형태로 변환하여 모델에 주입.
안전행정부/과기부 개인정보 가이드라인	행정안전부와 과학기술정보통신부는 개인정보 처리 가이드라인 및 오픈데이터 정책을 발표하며 문서 대부분은 공공저작물이다.	공공누리 제1유형 라이선스(출처 표시) 또는 공공영역 자료로 자유 이용 가능.	개인정보보호법 질의 대응 RAG 자료나 인스트럭션 튜닝을 위한 원문 자료로 활용.

4. 크로스언어/멀티링크 참고 데이터

데이터셋	요약	라이선스/근거	활용 방안
XOR QA / TyDi QA	한국어를 포함한 11개 언어의 공개 질문·답변 데이터. 웹에서 수집한 질문과 정답 스니펫으로 구성. 라이선스는 **MIT**로 허용적이다.	MIT 라이선스에 따라 자유로운 사용 및 수정이 가능하다.	한국어 질문과 답변이 포함돼 있으므로 모델의 한국어 QA 능력 보강에 사용. 금융·보안 관련 질문은 제한적이지만, 한국어 기반 검색·답변 형식 학습에 기여한다.

데이터셋	요약	라이선스/근거	활용 방안
Korean Bias Benchmark for Question Answering (KoBBQ)	사회적 편향 측정을 위한 한국어 QA 데이터셋. 금융/보안과 직접 관련되진 않지만 한국어 QA 포맷을 연습하는 데 유용하다. 라이선스는 오픈리뷰에서 **CC BY 4.0**.	CC BY 라이선스 → 출처 표시 후 자유롭게 사용 가능.	한국어 Q&A 응답 형식을 학습하는 튜닝 데이터로 활용할 수 있다.

활용 전략

1. **법령·판례 기반 RAG 구축** – 국가법령정보센터, 대법원 판례 등 공공저작물을 크롤링해 벡터DB를 만들고, 한국어 질문에 대한 정확한 근거를 제공하도록 한다. LBOX Open과 병행하여 법률 질의응답 모델을 강화하되, 비상업적 라이선스 조건을 지켜야 한다.

2. **금융 뉴스/리포트 데이터 수집** – 공공 뉴스 데이터(AI Hub), 한국은행·금융감독원 보고서 등에서 한국어 금융 도메인 문서를 수집하고 요약·질문화를 수행해 파인튜닝 데이터로 사용한다. 또한 위키백과를 통해 기본 금융/보안 용어 설명을 확보할 수 있다.

3. **보안·프라이버시 자료 활용** – KISA 분석 보고서, 개인정보보호 가이드라인 등 공공기관 발간자료를 RAG 문서로 추가하면 피싱, 악성코드, 개인정보 침해 등 한국에서 발생한 보안 이슈를 처리하는 데 도움 된다.

4. **모델 파라미터 재활용** – KB-ALBERT-KO와 같은 한국어 금융 PLM이 Apache 2.0 라이선스로 공개돼 있으므로, 이를 LoRA/QLoRA 방식으로 미세조정하거나 RAG의 단락 인코더로 활용할 수 있다 [3].

이처럼 한국어 데이터는 공개 범위가 제한적이지만, 공공저작물과 비상업적 라이선스 데이터를 적절히 조합하면 해커톤 목표에 부합하는 한국어 금융·보안 LLM을 개발할 수 있다.

---

[1] [2] A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction | OpenReview
https://openreview.net/forum

[3] GitHub - teddylee777/KB-ALBERT-KO: KB 국민은행에서 제공하는 경제/금융 도메인에 특화된 PLM(ALBERT) 모델
https://github.com/teddylee777/KB-ALBERT-KO