ChatGPT

금융·보안 QA 성능 개선을 위한 데이터셋 조사

배경

해커톤에서 평가되는 FSKU 테스트 데이터는 약 515개 문제로 구성되며 법령·규정과 기술적 보안이 큰 비중을 차지한다. 분석 결과 질문의 **28.2%**가 개인정보보호법 • 전자금융거래법 • 정보통신망법 등 **법률 및 규정**에 관한 내용이고, 금융/투자(10.9%), 접근통제(9.7%), 개인/신용정보 보호(8.2%), 정보보안/거버넌스(7.4%), 네트워크/시스템 보안(7.2%), 악성코드·사이버 위협(6.6%), 암호·인증(6.4%) 등 다양한 금융·보안 도메인으로 구성되어 있다 [1].
튜닝 과정에서는 한국어 데이터와 법령 텍스트에 익숙한 LLM이 필요하며, **공개·비상업적 이용이 가능한 데이터**만 사용할 수 있다는 규정을 준수해야 한다【572749263160623 † L0-L7】.

데이터셋 조사

아래 표는 테스트 데이터의 도메인과 라이선스 조건을 고려하여 조사한 데이터셋을 요약한 것이다. 테이블이 길어지지 않도록 핵심 정보만 표로 정리하고 세부 설명은 본문에서 다루었다.

데이터셋	주요 내용/출처	라이선스 및 특징
₩ON-Instruct (KRX-Data)	한국거래소·금융위원회·금감원 공시·교육 자료 등 **200 k+** 문서를 수집 후 GPT-4o와 Qwen으로 재작성한 **80 k**여개의 다지선다·Instruction–Response 쌍. 금융 규정, 주식시장, 회계 등을 폭넓게 포함【572749263160623 † L0-L7】.	**MIT** 라이선스 [2]. 금융 도메인에 특화된 instruction-tuning 데이터로, 해커톤 평가 항목과 가장 밀접하다.
₩ON-Reasoning / KRX LLM Competition 데이터	₩ON LLM에서 사용된 **5.5k** 문제 및 오픈 QA 데이터. 금융·회계·시장 분석·회사분석·주식 예측 등 6개 부문에 대한 MCQA와 오픈형 질문으로 구성【572749263160623 † L0-L7】.	**MIT** 라이선스로 공개되어 있고, 경쟁 참가자들이 제출한 200 k+ QA를 필터링한 고품질 reasoning 응답을 포함【572749263160623 † L0-L7】. 금융 시장의 계산/추론 문제를 다루므로 LLM의 reasoning 성능 향상에 유용.
FinShibainu	한국은행 경제·금융 용어, 금융감독원 금융 용어사전, 한국거래소 규정, 기업 공시 등에서 수집한 문서를 GPT-4o로 변환한 **42.5 k** 다지선다 질문과 **44.9 k** QA/추론 쌍. 질문마다 근거와 체계적인 추론 과정을 제공 [3] [4].	**Apache-2.0** 라이선스 [3]. 검증된 출처에서 수집된 데이터로 금융 규제와 경제 용어에 강하다. 추론 과정이 포함돼 있어 모델의 설명가능성 개선에 도움.
LBOX OPEN	한국 법원 판결문과 판례를 수집해 만든 분류·판결 예측·요약 데이터셋. 케이스명 분류, 법률 조항 분류, 판결 예측 등 다양한 태스크를 포함한다 [5].	**CC BY-NC 4.0** 라이선스 [6]. 법률 전문 용어 습득에 유용하며 비상업적 목적에 한해 자유롭게 활용 가능.

데이터셋	주요 내용/출처	라이선스 및 특징
KorMedLawQA	한국 의료법·의사 면허 시험에서 추출한 법률 조항과 사례를 GPT-4o를 이용해 질문·선택지·정답·추론을 생성한 MCQA 데이터셋. `law_title`, `article`, `question`, `options`, `answer`, `reasoning` 필드로 구성 [7].	**Apache-2.0** 라이선스 [8]. 의료법 중심이지만 행정법·규정 이해 및 추론 연습에 적합.
KorMedMCQA	2012–2023년 의사·간호사·약사 국가시험 문제에서 추출한 **7 k+** 다지선다형 질문. HuggingFace에 공개된 데이터의 라이선스는 **CC BY-NC 2.0**으로 비상업적 연구에 사용 가능 [9].	의료 분야 시험 문제이지만 의약품 관련 법규와 의료행정 규정도 포함되어 있어 법령 데이터 보강에 활용 가능.
HAI (HIL-based Augmented Industrial Control Security) dataset	한국 전력 계통을 모사한 HIL 기반 산업제어시스템 테스트베드에서 수집한 보안 데이터. 정상 운영과 38종 이상의 공격 시나리오가 기록된 시계열 데이터 [10].	**CC BY-SA 4.0** 라이선스 [11]. 네트워크/시스템 보안과 침입 탐지 연구에 적합. 해커톤 테스트 데이터에 있는 악성코드·시스템 보안 문제를 보완하는 실제 데이터.
SQuARe & KoSBi (Korean safety benchmarks)	NAVER가 공개한 **SQuARe**(민감 질문–허용 답변)와 **KoSBi**(사회적 편향) 데이터셋. 질문과 응답에 대한 주석 및 원본 주석을 포함하고 모델 안전성 연구를 위해 제작됐다.	**MIT** 라이선스로 재배포 및 수정 가능 [12]. 안전하고 책임감 있는 금융 QA 시스템 구축 시 도움.
CLIcK (Cultural & Linguistic Intelligence in Korean)	한국 CSAT·공무원시험·한국어능력시험 등에서 추출한 1,995개의 QA로 한국 문화·언어 지식을 평가하는 벤치마크 [13].	**CC BY 4.0** 라이선스 [13]. 금융/보안과 직접 관련은 없지만 한국어 독해와 문화적 맥락 이해를 향상시킬 수 있다.
한국 공공데이터 포털 (data.go.kr)	정부·공공기관이 제공하는 공공데이터를 누구나 기계가 읽을 수 있는 형태로 재사용할 수 있도록 한 포털. 재사용 시 별도 신청 절차 없이 사용할 수 있으며 [14], 저작자 표기·비영리 사용·동일조건변경허용 등의 Creative Commons 라이선스 조건에 따라 활용 가능하다 [15].	특정 데이터셋이 아닌 포털 전체. 금융/보안 관련 통계·보고서·공시자료 등 공개된 데이터는 법령에 따라 자유롭게 활용 가능해 데이터 수집원으로 적합하다.

주요 데이터셋 세부 설명

1. ₩ON-Instruct 및 ₩ON-Reasoning

- **₩ON-Instruct**는 한국거래소와 금융감독기관 등에서 수집한 200 k+ 문서를 바탕으로 약 80 k개의 다지선다형 질문과 Instruction-Response 쌍을 생성한 데이터셋으로, **MIT 라이선스**로 공개되어 있다【572749263160623†L0-L7】 [2]. 질문은 금융 시장 운영, 회계, 상품 거래, 금융법령 등 폭넓은 주제를 포함한다.
- **₩ON-Reasoning/경진대회 데이터**는 KRX Financial LLM Competition에서 사용한 약 5.5 k개의 MCQA와 오픈형 질문을 포함한다. 참가자들이 제출한 200 k+ 문제를 필터링하여 고품질 응답과 추론을 제공하며 역시 **MIT 라이선스**로 공개된다【572749263160623†L0-L7】.
- **활용 방안**: FSKU 테스트의 금융/투자(10.9 %), 법령·규정(28 %) 영역과 직결되는 질문이 많아 튜닝 시 가장 우선적으로 활용할 수 있다. 금융 도메인의 용어와 규정을 학습하고, reasoning 데이터로 추론 능력을 강화한다.

2. FinShibainu

FinShibainu는 한국은행·금감원·한국거래소 규정, 기업 공시 자료 등을 바탕으로 GPT-4o로 재작성된 42.5 k개의 다지선다 질문과 44.9 k개의 QA/추론 쌍으로 구성되어 있다 [3] [4]. 각 질문에는 정답 선택지와 함께 **추론 과정**이 포함돼 있어 모델이 논리적 근거를 제시하는 능력을 기르는 데 유리하다. 라이선스는 **Apache-2.0**으로 기업 내 연구 프로젝트에도 제한 없이 사용할 수 있다 [3].

3. LBOX OPEN

LBOX OPEN은 한국 법원 판결문 데이터로 이루어진 대규모 법률 AI 벤치마크로, 판례 분류, 법률 조항 분류, 판결 예측, 요약 등 여러 태스크를 포함한다 [5]. 라이선스는 **CC BY-NC 4.0**으로 비상업적 연구에 사용할 수 있다 [6]. 금융·보안 QA 모델은 법령 텍스트를 이해해야 하므로, 이 데이터로 법률 용어와 판결 구조를 학습할 수 있다.

4. KorMedLawQA

KorMedLawQA는 한국 의료법과 관련 규정에서 추출한 기사를 기반으로 다지선다형 질문·선택지·추론을 생성한 데이터셋이다. `law_title`와 `article` 필드가 있어 법령 조문에 대한 이해와 적용을 학습할 수 있다 [7]. 라이선스는 **Apache-2.0** [8]. 의료법 중심이지만 개인정보보호·전자서명 등 보건 분야 규정도 포함돼 있어 법률 QA 모델에 도움이 된다.

5. KorMedMCQA

KorMedMCQA는 2012-2023년 의사·간호사·약사 국가시험에서 추출한 약 **7,489**개의 시험 문제를 수록한 다지선다형 QA 데이터로, HuggingFace에서 **CC BY-NC 2.0** 라이선스로 제공된다 [9]. 의료법규, 환자정보 보호, 약사법 등의 법적 지식이 포함돼 있어 법령 대응 학습 자료로 활용 가능하다.

6. HAI 보안 데이터셋

HAI(HIL-based Augmented Industrial Control System) 데이터셋은 한국의 산업제어시스템 테스트베드에서 정상 상태와 38종 이상의 공격 시나리오를 수집한 시계열 데이터이다. 라이선스는 **CC BY-SA 4.0** [11]. 네트워크/시스템 보안과 악성코드 탐지와 관련된 문제를 학습하는 데 유용하며, FSKU 테스트의 네트워크 보안(7.2 %)과 사이버 위협(6.6 %)을 보완한다.

7. SQuARe & KoSBi

NAVER의 **SQuARe**와 **KoSBi** 데이터셋은 민감 질문과 허용 가능한 답변(SQuARe), 사회적 편향과 안전한 응답(KoSBi)을 포함한 한국어 안전성 벤치마크이다. 두 데이터 모두 **MIT 라이선스**로 공개되어 자유롭게 수정·배포할 수 있다 [12]. 금융 챗봇 개발 시 적절한 응답을 생성하고 편향을 줄이는 데 활용할 수 있다.

8. CLIcK

CLIcK는 한국 문화·언어 지식을 평가하기 위해 CSAT(수능)·공무원시험·TOPIK 등에서 1,995개의 질문을 추출한 벤치마크로, **CC BY 4.0** 라이선스가 적용된다 [13]. 금융/보안과 직접적 연관성은 낮지만, 한국어 독해 능력을 강화하고 문화적 맥락을 이해하는 데 도움이 된다.

9. 공공데이터 포털

한국 공공데이터 포털(data.go.kr)은 공공기관이 보유한 데이터를 기계가 읽을 수 있는 형태로 제공하며 누구나 별도 신청 절차 없이 재사용할 수 있다 [14]. 재사용 시에는 저작자 표기(CC BY), 비영리(CC BY-NC), 변경 허용 동일조건(CC BY-SA) 등 Creative Commons 라이선스 조건을 준수해야 한다 [15]. 금융위원회·금융감독원·개인정보보호위원회 등에서 제공하는 금융통계, 전자서명 가이드, 보안 가이드라인 자료를 수집해 QA 데이터로 가공하는 데 활용할 수 있다.

튜닝 전략 및 데이터 결합 제안

1. **금융 규제·거래 지식 강화** – ₩ON-Instruct와 FinShibainu를 기본 교육 데이터로 사용한다. 두 데이터 모두 금융 시장 규정, 회계, 증권 거래 등 실무 지식을 포함하여 법령·투자 영역 문제를 학습시키는 데 적합하다.

2. **추론 능력 및 설명 가능성 향상** – ₩ON-Reasoning과 FinShibainu의 추론 과정을 활용하고, 의료법규 기반의 KorMedLawQA에서 제공하는 상세 이유를 통해 LLM이 답변 근거를 명시하도록 학습시킨다.

3. **법률 언어 및 판례 이해** – LBOX OPEN에서 판례 분류·판결 예측 데이터를 학습하여 판결문 구조와 법률 용어를 습득한다. KorMedLawQA, KorMedMCQA를 함께 사용하여 법령 해석과 규정 적용에 대한 경험을 넓힌다.

4. **보안·악성코드 사례 학습** – HAI 데이터셋을 이용해 산업제어시스템 공격·정상 데이터를 학습하여 네트워크/시스템 보안과 악성코드 탐지에 대한 이해를 높인다. 추가로, 공공데이터 포털에서 한국인터넷진흥원(KISA)·금융보안원(FSI)의 공개된 보안 가이드라인, 악성코드 분석 보고서를 수집해 QA 형태로 재가공한다.

5. **안전성 및 편향 제어** – SQuARe와 KoSBi를 활용해 민감 질문에 대한 안전한 응답과 사회적 편향을 최소화하는 방법을 학습시킨다. 해커톤 시스템에서 금융 상담 시 부적절한 조언이나 차별적 언어를 방지할 수 있다.

6. **한국어 독해와 문화적 맥락** – CLIcK와 같은 일반 언어·문화 벤치마크를 섞어서 기본 한국어 이해 능력을 강화하고, 공공데이터의 정부 보고서·법령 해설서 등을 추가로 수집하여 모델의 배경지식을 넓힌다.

7. **데이터 재사용 법적 준수** – 모든 데이터 사용 시 라이선스 조건을 준수해야 한다. 공공데이터 포털의 데이터는 저작자 표기·비영리·변경허용 등의 조건을 확인 후 활용한다 [15]. KLRI가 제공하는 법령 원문은 무단 복제·배포가 금지되어 있으므로 참고용으로만 사용하고 데이터셋에는 포함하지 않는다.

결론

해커톤에서 제시된 테스트 데이터는 법률, 금융 거래, 네트워크 보안 등 다양한 지식을 요구한다. **₩ON-Instruct/Reasoning**과 **FinShibainu**가 금융 규정과 실제 투자 상식을 포함하여 가장 중요한 학습 자료이며, **LBOX OPEN**과 **KorMedLawQA**, **KorMedMCQA**는 법률 이해와 추론을 강화한다. **HAI**는 산업 보안 데이터로서 시스템·네트워크 보안 영역을 보강하며, **SQuARe/KoSBi** 및 **CLIcK**는 모델의 안전성과 한국어 이해 능력을 높인다.

또한, 한국 공공데이터 포털의 개방된 금융·보안 자료를 수집해 QA 형식으로 가공하면 테스트 데이터의 미세 영역(전자금융거래법·개인정보보호법·악성코드 등)을 보강할 수 있다. 라이선스 규정을 꼼꼼히 확인하여 **MIT, Apache, CC BY-NC 등 허용된 라이선스** 내에서 데이터셋을 구성하는 것이 필수적이다.

---

[1] KFinEval-Pilot: A Comprehensive Benchmark Suite for Korean Financial Language Understanding
https://arxiv.org/html/2504.13216v1

[2] KRX-Data/Won-Instruct · Datasets at Hugging Face
https://huggingface.co/datasets/KRX-Data/Won-Instruct

[3] [4] aiqwe/FinShibainu · Datasets at Hugging Face
https://huggingface.co/datasets/aiqwe/FinShibainu

[5] [6] lbox/lbox_open · Datasets at Hugging Face
https://huggingface.co/datasets/lbox/lbox_open

[7] [8] snuh/KorMedLawQA · Datasets at Hugging Face
https://huggingface.co/datasets/snuh/KorMedLawQA

9 sean0042/KorMedMCQA · Datasets at Hugging Face
https://huggingface.co/datasets/sean0042/KorMedMCQA

10 11 raw.githubusercontent.com
https://raw.githubusercontent.com/icsdataset/hai/master/README.md

12 GitHub - naver-ai/korean-safety-benchmarks: Official datasets and pytorch implementation repository of SQuARe and KoSBi (ACL 2023)
https://github.com/naver-ai/korean-safety-benchmarks

13 CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean
https://arxiv.org/html/2403.06412v1

14 15 Use Policies on the Open Government Data Portal | OPEN DATA PORTAL
https://www.data.go.kr/en/ugs/selectPortalPolicyView.do