

**SK네트웍스 Family AI 과정 14기  
 데이터 전처리 인공지능 학습 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 학습 결과서 |
| **제출 일자** | 2025-09-29 |
| **깃허브 경로** | https://github.com/skn-ai14-250409/SKN14-Final-3Team |
| **작성 팀원** | 김의령, 조성재 |

| 1. 모델 비교 및 선정 이유 | 1-1. 비교 대상 모델  1-2. 고려한 조건 3가지 |
| --- | --- |
| 2. 모델 구조 | 2. 모델 아키텍쳐 |
| 3. 평가 설정 | 3. 평가 설정 및 하이퍼파라미터 |
| 4. 평가 결과 및 성능 평가 | 4-1. 학습 결과 요약  4-2. 해석 및 분석 |
| 5. 결론 | 5. 결론 |
| 6. 향후 계획 | 6. 향후 계획 |
| 7. 부록 | 7.부록 및 test에 사용한 로그(일부) 및 test셋 깃헙링크 |
| 8. 추가 | 1. 작업환경  2. vLLM 서버 설정 |

**1. 모델 비교 및 선정 이유**

**1-1. 비교 대상 모델 :**

| **모델명** | **종류** | **선정 이유** |
| --- | --- | --- |
| Gemma-7B-it | Decoder-only Transformer | 안정성/지원 강점점 |
| Qwen2.5-7B-instruct | Transformer 기반 모델 | 7B 중 최고 성능·효율성 |
| Mistral-7B -instruct-0.3 | Transformer 기반 모델 | 한국어어 강점, 범용 RAG/ 코드에 유리 |
| FinShibaiu | Qwen2.5-7B-instruct 기반 모델 | 7B 중 최고 성능·효율성 + 금융지식 |

**1-2. 고려한 조건 3가지**

* 1. 멀티턴 지원 유무
* 2. 긴 context size 가능 여부
* 3. Tool calling 지원 여부

| **모델명** | **Context size** | **Multi-turn 대화** | **Tool calling** | **특이사항** |
| --- | --- | --- | --- | --- |
| Gemma 7B-it | 기본 8K, 확장 시 32K까지 지원 (환경 의존) | 안정적이나 길 게 이어질수록 coherence 약간 저하 | 직접적인 Tool Calling 지원은 미약. | 안정성 적합 |
| Qwen2.5-7B-instruct | 최대 32K(기 본), 일부 버전 은 128K 지원 | 멀티턴 유지력 가장 강력 (지 시 따름 잘함) | OpenAI-style function calling 과 유사한 구조지원. 툴 연계 최적. | 한국어 강점 |
| Mistral-7B -instruct-0.3 | 8K (최근 LongChat 변형 은 최대 32K) | 짧은 멀티턴은 강하지만 길게 가면 coherence 떨어짐 | Tool Calling 미지원 | 속도나 추론강 점 |
| FinShibaiu | 최대 32K(기 본), 일부 버전 은 128K 지원 | 멀티턴 유지력 가장 강력 (지 시 따름 잘함) | OpenAI-style function calling 과 유사한 구조지원. 툴 연계 최적. | 한국어 강점, 금융 특화 |

**2. 모델 구조 및 아키텍처**

선정모델: transformer기반계열 Qwen2.5 7B Instruct

[Input Text]

↓

Tokenizer & Embedding

↓

80층 Transformer

(SwiGLU, RoPE, RMSNorm, Attention QKV Bias)

↓

Dense Layer + Softmax

↓

[Output: 텍스트 생성 혹은 분류

**3. Base 모델 평가**

**3-1. Base 모델 성능지표 평가에 사용한 task**

| **Dataset** | **Description** | **Language** |
| --- | --- | --- |
| **kbl** | Korean Benchmark for Legal Language Understanding. | Korean |
| **kmmlu** | Knowledge-based multi-subject multiple choice questions for academic evaluation. | Korean |
| **kobest** | A collection of tasks designed to evaluate understanding in Korean language. | Korean |
| **kormedmcqa** | Medical question answering tasks in Korean to test specialized domain knowledge. | Korean |

"kbl",-법률 지식[한국어]

"kmmlu",-수리·추론 [한국어]

"kobest",-QA/독해 [한국어]

Kormedmcaqa 평가 제외 이유 : 영어로 이루어지는 test나 한국 의학쪽 지식은 현 주제 모델에 필요없는 task이므로 테스트에서 제외

**3. 평가 설정 및 하이퍼파리미터**

| **파라미터** | **값** | **설명** |
| --- | --- | --- |
| temperature | 0 | 결정적 응답 |
| Max\_tokens | 512 | 최대 출력 토큰 |
| Batch\_size | 1 | 안정적평가를 위해 1 사용 |
| seed | 42 | 재현성 확보 |
| limit | 10 | 테스크당 평가 샘플 제한 |

**4. 평가 결과 및 성능 평가**

4-1. 평가 결과 요약

| **모델** | **KBL Exact Match** | **KMMlU Accuracy** | **KOBEST Accuracy** | **Acc\_norm** | **F1 Score** | **특징** |
| --- | --- | --- | --- | --- | --- | --- |
| Gemma-7B-it | 10~30% | 8.22% | 42% | 30% | 0.325 | 한국 법률/학문 지식 부족, 일반 상식 중간 |
| Qwen2.5-7B-instruct | 35.5% | 36.2% | 64% | 60% | 0.587 | HUMSS강점, 감정/상식 추론 강함, 최신 법률 문제 대응 중간 |
| Mistral-7B -instruct-0.3 | 21.2% | 28.4% | 54% | 60% | 0.453 | KBL 최근년 정확도 낮음, 한국어 일반 지능 중간 |

4-2. 해석 및 분석

| **모델** | **테스크** | **Accuracy / F1** | **특징 / 분석** |
| --- | --- | --- | --- |
| Gemma-7B-it | KBL | 10~30% | 법률 지식 부족, 최신 기출(2023~2024) 완전히 불가. 정답 포맷에 민감. |
| Gemma-7B-it | KMMLU | 8.22% | 사회과목(HUMSS) 약간 가능. |
| Gemma-7B-it | KOBEST | Acc 42%, Acc\_norm 30%, F1 0.3253 | 일부 추론형 문제에서만 성능 있음. 일반 상식은 중간 수준. STEM/Applied Science 거의 불가 |
| Gemma-7B-it | KorMedMCQA | (평가 제외) | 영어 기반 테스트이거나 한국 의학 지식이 필요 없는 태스크라 테스트에서 제외 |
| Mistral-7B-0.3-instruct | KBL | 21% | 평균 수준. 최근(2024~2025) 문제에서 정확도 하락. 포맷·정답 일치 중요. |
| Mistral-7B-0.3-instruct | KMMLU | 28.4% | HUMSS 강점, 과학·공학·기타 20~30%대. |
| Mistral-7B-0.3-instruct | KOBEST | Acc 54%, Acc\_norm 60%, F1 0.4534 | 참/거짓·추론 유형에서 안정적. 선택지형 문제는 acc\_norm 적용 시 개선. |
| Qwen2.5-7B-instruct | KBL | 35.5%(최신 2024 문제는 23% → RAG 연계 필요) | 평균 35% 수준. 법률·금융 문서 이해 가능(문서 연계 권장). |
| Qwen2.5-7B-instruct | KMMLU | 36.2%(HUMSS 50.9%, STEM/Applied 30~32%) | HUMSS 강점, STEM/Applied Science 중간 수준. |
| Qwen2.5-7B-instruct | KOBEST | Acc 64%, Acc\_norm 60%, F1 0.587 | 감정·참거짓 유형에서 강함. 선택지형 문제는 정규화(acc\_norm) 시 개선. 금융/규정 질문에 적합. |

4.3 해석 및 분석

* 1.법률(KBL) 성능:   
  - Gemma-7B-it < Mistral-7B-0.3-instruct < Qwen2.5-7B-instruct   
  - 최신 법률/규정 반영이 중요한 경우 RAG 연계 필요
* 2.학문 지식(KMMLU) 성능:   
  - 전반적으로 HUMSS > Other > STEM/Applied Science   
  - Qwen2.5-7B-instruct가 가장 균형 잡힌 성능
* 3.한국어 일반 지능(KOBEST) 성능:   
  - Qwen2.5-7B-instruct > Mistral-7B-0.3-instruct > Gemma-7B-it  
  - 감정·참거짓/추론 문제에서 Qwen2.5-7B-instruct 안정적
* 4.금융업무 적합성 관점 :   
  - Qwen2.5-7B-instruct: 한국어 문서 이해 및 RAG 기반 금융 QA에 가장 적합  
  - Mistral-7B-0.3-instruct: 중간 수준, 일부 참거짓/일반 추론 가능   
  - Gemma-7B-it: 금융/법률 최신 문서 이해 어려움, 단순 상식·기초 추론 가능
* 5. RAG 라우팅 관점:   
  - Qwen2.5-7B-instruct: RAG로 최신 법규/금융 문서 참조 시 정확도 향상 가능   
  - Mistral-7B-0.3-instruct: 기본 RAG 가능, 최신 포맷 민감   
  - Gemma-7B-it: RAG 보완 필요, 출력 포맷 민감

**5. 결론 및 향후 계획**

Qwen2.5-7B-instruct는 한국어 금융 업무 문서 이해 및 RAG 기반 질의응답 테스트에 가장 적합하며, 일반 지능과 법률 지식 모두 중간 이상 수준으로 안정적인 성능을 보임.

* **최종 선정 모델:** Qwen2.5-7B-instruct
* **활용 방안:** 금융업무 Q&A 및 RAG 기반 정보 라우팅

**6. 향후 계획**

1. SFTTrainer 로 금융 합성성 데이터를 SFT 진행
2. DPOTrainer로 Humance preference 성능 향상
3. 웹 연동 및 REST API 서비스화(배포/운영 파이프라인 구축)
4. 최신 금융·법률·정책 문서를 RAG로 연동하여 정확도 개선

**7 .부록 및 test에 사용한 로그(일부) 및 test셋 깃헙링크**

* **Eval GitHub (lm-evaluation-harness — tasks 목록)** <https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks>
* **Test log (Notion: gemma-test.ipynb)** [file.notion.so/f/f/ba62e9a5-7e5a-4764-9900-cb23dcf70cfe/95f9c1ce-dca0-4f91-beef-53fed9b623b0/gemma-test.ipynb?table=block&id=25789456-94a7-8024-9a93-cac60b08cc58&spaceId=ba62e9a5-7e5a-4764-9900-cb23dcf70cfe&expirationTimestamp=1756080000000&signature=qVwS37mboYze8T8YmW1DyK2NfWCVlZqQ-wunX3zVXyQ&downloadName=gemma-test.ipynb](http://file.notion.so/f/f/ba62e9a5-7e5a-4764-9900-cb23dcf70cfe/95f9c1ce-dca0-4f91-beef-53fed9b623b0/gemma-test.ipynb?table=block&id=25789456-94a7-8024-9a93-cac60b08cc58&spaceId=ba62e9a5-7e5a-4764-9900-cb23dcf70cfe&expirationTimestamp=1756080000000&signature=qVwS37mboYze8T8YmW1DyK2NfWCVlZqQ-wunX3zVXyQ&downloadName=gemma-test.ipynb)

**추가 1 ) 작업환경**

* **클라우드 환경:** RunPod
* **GPU:** NVIDIA A40 (40GB VRAM)
* **OS / 시스템:** Ubuntu 20.04 / 22.04 기반 컨테이너
* **디스크 용량:** 40GB (유틸리티 포함, 학습/평가 데이터 및 캐시 저장 가능)
* **Python 환경:** 3.10 이상 권장
* **라이브러리:**
  + vLLM >= 0.6.3 → 로컬 GPU 기반 LLM 추론 서버
  + lm-eval[api] >= 0.4.5 → 표준 벤치마크 평가
  + transformers, accelerate, bitsandbytes → 모델 로딩/추론 최적화
  + requests, huggingface\_hub → HF Hub 모델 다운로드 및 API 호출

**추가 2 ) vLLM 서버 설정**

| **항목** | **설정** |
| --- | --- |
| 모델 | google/gemma-7b-it 등 |
| GPU 메모리 사용 제한 | 0.85 (85%) |
| 최대 토큰 길이 | 4096 |
| 호스트 / 포트 | 0.0.0.0:8000 |
| 로그 | 최소화 (--disable-log-requests) |
| Stop 토큰 | `<` |
| 데이터 타입 | auto |

# 서버는 subprocess로 실행되며 stdout/stderr를 실시간 스트리밍한다. API 준비 여부(/v1/models)를 확인한 다음 평가를 시작한다.