

**SK네트웍스 Family AI 과정 14기  
 데이터 전처리 학습된 인공지능 모델**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 모델별 자동 채점 결과표, 자유문항 정성 및 정량평가, 재현 스크립트 |
| **제출 일자** | 2025-09-29 |
| **깃허브 경로** | https://github.com/skn-ai14-250409/SKN14-Final-3Team |
| **작성 팀원** | 조성재 |

1. **모델 목적**:  
    KB 국민은행 여신 업무에서  
   - 금리, DSR, 중도상환수수료, 예금보호, ECL 등 정량 계산 질의에 정확히 응답,  
   - 내규/약관 절차, 판정 로직을 일관된 형식으로 설명,  
   - 멀티턴 상담 맥락(예: 소득, 기존상환등)을 기억하고, 필요시 툴 호출(계산기/규정조회/RAG)을 수행하는 sLLM
2. **모델 아키텍처**

* 선정 모델(베이스 후보): *Qwen-2.5-7B / Qwen-2.5-7B-instruct / Qwen-3-8B-Base*
* 아키텍처 개요: Qwen-2.5-7B

| **계층** | **구성 요소** | **역할** |
| --- | --- | --- |
| 입력층 | **Tokenizer (byte-level BPE)**, Vocab ≈ **151,646**, **Token Embedding (dim=3,584)**, **RoPE**(base≈1e6) | 텍스트를 토큰화하고 임베딩으로 변환. RoPE로 위치 정보를 부여해 긴 문맥에서도 안정적인 주의집중(Attention) 수행 |
| 디코더(본체) | **Transformer Decoder Blocks × 28**, **GQA**(Query Heads=28, **KV Heads=4**, Head size=128), **SwiGLU FFN**(intermediate=18,944), **RMSNorm(Pre-Norm)**, **QKV bias**, **Dual Chunk Attention + YARN** | **다음 토큰 예측**을 위한 자기회귀적(autogressive) 언어 모델링. GQA로 **KV 캐시 효율**과 추론 처리량을 개선, DCA+YARN으로 **롱컨텍스트** 안정성 향상 |
| 출력층 | **LM Head (Linear → Softmax over Vocab)  Task Head(Dense + Softmax)** | LM Head를 통해 각 시점의 토큰 로짓을 산출해 확률 분포(Softmax)로 변환, 텍스트를 생성/완성. 혹은 Task Head를 통해 다중 분류 수행 |

* 선정 근거  
   Qwen 계열의 모델은 최대 256K 토큰(최대 1M까지 확장가능) 폭넓은 컨텍스트 윈도우로 멀티  
   턴을 수월하게 하며, 한국어가 잘 학습 되어있고 멀티턴과 tool calling이 가능하여 프로젝트에  
   적합하며, “인공지능 학습 결과서” 에서 볼 수 있다 싶이 실험을 해보았을때 프로젝트에 가장 적합한 성능을 보여줌.

## 베이스 모델 선정 과정

* 1. 기존에 많이 사용하는 Qwen 2.5 와 가장 최근에 발표된 Qwen3 중 프로젝트에 가장 적합한 모델을 선정하기 위해 평가
  2. 업계에서 가장 많이 사용하는 평가 방법인 LLM-as-a-judge와 lm-eval-harness를 통해 평가

### 3.1 LLM as a Judge

* 문항 세트 (총 5문항)
* **하-1** 원리금균등 월 상환액(툴콜 권장): JSON { "monthly\_payment": number }
* **하-2** 상품 분류: JSON { "label": "신용대출" }
* **중-1** DSR(스트레스 +2.0%p, 멀티턴): JSON { "dsr\_pct": number, "eligible": boolean, "new\_monthly\_payment": number }
* **중-2** 중도상환수수료(스텝다운+일할): JSON { "fee": number }
* **자유질문-1** '스트레스 DSR’이란 뭐고 현재 한국에선 어떻게 시행하고 있어? output은 이제 상관없이 자유롭게 말해줘

### 3.1.1 정답 기준(요약)

* Q1: **1,060,655.15원** (±200원)
* Q2: **“신용대출”**
* Q3: **월상환 1,696,785.40원**, **DSR 38.5920%**, **eligible=true** (±200원 / ±0.5%p)
* Q4: **556,666.67원** (±200원)
* Q5: **정의 정확성, 구체성, 한국 맥락 충실도, 독자 친화성**

### 3.1.2 실험 환경

* **Ollama**: GUI(대화) + API(툴콜/스키마/temperature=0 고정)
* 공통 옵션: temperature=0, JSON-only, 동일 컨텍스트 길이
* 멀티턴 검증: 동일 세션에서 턴A→턴B 순차

**3.2 lm-eval-harness (한국어 및 금융 벤치마크 기준 평가)**

* **HAERAE**: 한국어 LLM 종합 벤치마크. 한국어 독해, 상식 추론, 수학, 코드 등 여러 태스크로 한국어 모델 성능을 전반적으로 평가.
* **Kobest**: 한국어용 MMLU 스타일 벤치마크. 법, 경제, 역사, 의학 등 다양한 분야의 전문지식을 묻는 객관식 문제로 **지식·추론 능력** 평가.
* **HRM8k\_GSM8k**: GSM8K(초등 수준 수학 단답형 벤치마크)의 한국어 번역/적용 버전. **연산·문제 해결 능력**을 한국어 맥락에서 검증.
* **HRM8k\_MATH**: MATH 벤치마크(고등·대학 수준 수학 문제)의 한국어 버전. 복잡한 수학적 **추론·증명 능력** 평가.
* lm-eval-harness 벤치마크 결과

|  | HAERAE | Kobest | HRM8k\_GSM8k | HRM8k\_MATH |
| --- | --- | --- | --- | --- |
| Qwen2.5-7B | 0.37 | 0.48 | 0.54 | 0.44 |
| Qwen3-8B | 0.22 | 0.5 | 0.87 | 0.78 |

1. **성능 평가 결과**

**[LLM as a Judge 평가 결과]**

* Q2(라벨 분류)\*\*는 세 모델 모두 무난(3/3 통과).
* 수치·규정 계산형(Q1, Q3, Q4)에서 성패가 갈림.
* **Qwen-3-8B**가 **5/8**로 가장 낫지만, **Q1 오차 +495원**(허용±200)으로 아깝게 탈락, Q3(3점짜리 핵심)\*\*는 계산/퍼센트 스케일 오류로 실패.
* **Qwen-2.5-7B**는 **2/8**이지만 Dense 구조 + FT 레시피가 풍부해 **파인튜닝/툴콜로 상향 여지 큼**.
* **자유질문-1:  
  - Qwen-3-8B: 금융당국 시각에서 개괄적으로 요약하는 데는 강점이 있으나, 다소 추상적.  
  - Qwen-2.5-7B: 구체적 수치, 시뮬레이션 가정, 한국 도입 시기 등을 제시하여 독자가 실제 제도의 모습과 정책 취지를 직관적으로 이해할 수 있게 함.**

**[lm-eval-harness를 통한 평가 결과]**

* 크기가 1B 정도 차이가 나고 다음 세대의 모델임에도 불구하고 HARAE의 성능이 2.5가 3보다 뛰어남
* 수학적 능력은 Qwen-3가 압도적으로 좋은 것을 확인할 수 있었음

1. **베이스 선정 및 학습 계획**

### 5.1 베이스 선정

* **최종 선정**: **Qwen-2.5-7B**
* **선정 사유**
  1. **학습 불확실성**: Qwen-3-8B는 **Thinking/Non-thinking** 구조로 추론모드의 복잡성이 존재하므로 fine-tuning의 어려움이 있을것 같으며, 공개 FT 레시피/성공사례가 상대적으로 적고, 라우팅/게이팅 변화에 따른 **미세튜닝 일관성 리스크** 존재.
  2. **생태계/재현성**: Qwen-2.5-7B는 **LoRA/QLoRA, TRL, PEFT** 기반 FT 자료와 커뮤니티 사례가 풍부하여 **재현성·문제 해결 속도**가 빠름.
  3. 프로젝트에서 sLLM이 맡게 될 업무는 금융 QA이며 수학문제를 계산하는 부분은 Tool calling으로 커버 가능.
  4. **실험 결과 해석**: 제로샷에서 Qwen-3-8B가 수치형 일부 우위였으나, \*\*우리 업무(여신 규정·계산·JSON 형식·툴콜/RAG)\*\*는 FT와 툴 체인 표준화로 **2.5-7B도 충분히 상향** 가능.

**결론**: **학습/운영 리스크 최소화 + 도메인 특화 후 성능 상향 가능성**을 종합해 **Qwen-2.5-7B**를 베이스로 선정.

### 5.2 파인튜닝(Full-fine-tuning) 계획

* **목표**:  
  + 금융지식 및 KB 국민은행 내규 및 상품 지식 학습 (SFT)
  + **형식 준수**(상세한 설명, 넘버링과 강조구문) 정렬
  + \*\*선호/정렬 데이터(DPO/ORPO)\*\*로 안전·정확 응답 강화
* **데이터**:  
  + (사내) 내규·심사 체크리스트·상품설명서·FAQ → **지시–응답(SFT)**
  + **합성+AI 피드백(distilabel)** → 다양성·품질 확보, **임베딩/MinHash**로 중복 제거
  + **휴먼 스팟체크**로 환각/용어/책임 문구 정제
* **학습 스펙(예시)**:  
  + **QLoRA**: r=16, α=16, dropout=0.05, target=q\_proj,k\_proj,v\_proj,o\_proj
  + **seq len** 4k, **bf16**, **effective batch tokens**≈128k, **epoch** 1~2
  + **스케줄**: cosine, warmup 2% / **정규화**: gradient clipping 1.0
* **출력 통제**:  
  + **JSON 스키마 강제**(테스트: Ollama format, 운영: 서버 측 validation)
  + **툴콜 우선**: 원리금균등/DSR/수수료/ECL 계산 함수를 호출 → 수치 오차 최소화
  + **금지/완화 토큰**: 개인정보·단정적 법률 자문 등 제한

**6. 종합 평가 및 활용방안**

* **모델 안정성**: 동일 프롬프트·옵션에서 저장/로드 후 결과 재현 가능. 수치 문제는 툴콜(계산 함수) + JSON 스키마 강제로 오차·형식 흔들림 최소화.
* **일반화**: 여신 도메인 내 규정·절차는 SFT + 선호정렬(ORPO/DPO)로 안정화, 최신성은 RAG(사내 문서/내규 개정본)로 보완. 멀티턴 대화 기준 정보(소득·기존상환 등) 유지.
* **재사용성**: QLoRA 어댑터로 경량 배포(수백 MB 수준) 및 증분 업데이트 용이. 실험은 Ollama, 운영은 vLLM(동적 배칭·KV 캐시)로 비용/지연 최적화.
* **향후 활용**:  
  1. **여신 상담 보조(DSR/적격성·서류 안내)**
  2. **심사 체크 자동화(체크리스트 질의·누락 탐지)**
  3. **약관/상품 Q&A(근거 조항 인용)**
  4. **사내 검색·요약(공지/지침 요약)**
* 운영 가드레일: PII 마스킹, 금지 토큰/표현 필터, 근거 인용 의무화, 실패 케이스 주간 점검(어댑터/프롬프트 핫픽스).