SK Networks Family AI Camp 9기

**모델 테스트 계획 및 결과 보고서**

가로선

**프로젝트 명: DAIS (Divergent AI with Science)**

**담당자: 조이현**

# <목차>

1. 테스트 개요

2. 테스트 항목

3. 테스트 케이스

4. 테스트 결과

5. 결론 및 개선방향

# 

# 1. 테스트 개요

가. 모델 테스트 목적: ​​본 모델 테스트의 목적은 과학 AI 인플루언서(이하 DAIS)의 한국어

과학 도메인 질의응답 및 생성 태스크에 적합한 성능을 보유했는지

확인하는 것임. 구체적으로, 본 테스트를 통해 DIAS의 성능 결과를

정량적/정성적 지표를 확인하여 실제 서비스상에 있어 효용력을

가늠하고자 함

나. 모델 테스트 정보

| 테스트 실시 일자 | 테스트 버전 | 테스트 환경 | 테스트 담당자 |
| --- | --- | --- | --- |
| 2025-06-19 | 1.0 | A100 SXM RunPod/pytorch:2.1.0-py3.10-cuda11.8.0 | 조이현 |
| 모델명 | | 모델 특징 | |
| 베이스 모델:  Qwen/Qwen3-8B | | 80억 파라미터, 32K 컨텍스트, instruction-aware, 최신 SOTA  (현 시점 최고 성능을 보이는 모델, 알고리듬, 기술) LLM | |
| 추론(생성) 모델:  SIQRIT/DAIS-Qwen3-8B-qdora | | 과학/기술 기반 3697개 rows의 jsonl 파일로 Q-DoRA 파인튜닝,  각 row당 최대 48 tokens의 input 및 최대 2000 tokens의  output으로 구성됨. additional\_special\_tokens 적용 | |
| 임베딩(검색) 모델:  nlpai-lab/KURE-v1 | | 약 1.9B 파라미터, 8192 토큰, 1024차원,  한국어+과학/전문용어 검색 최신 SOTA | |

# 

# 

# 

# 2. 테스트 항목

| 평가 방식 | 평가 항목 | 항목 설명 | 평가 형식 |
| --- | --- | --- | --- |
| 정량적 | RAG 시스템 | RAG 시스템을 이용하여 Vector DB top-1 값의 Cosine Similarity가 0.6 이상인 경우 DB에서 데이터를 가져오고, 0.6 미만인 경우, Wikipedia-api를 참조하는지 여부 | Pass / Fail |
| Perplexity | 언어 모델이 다음 단어를 얼마나 잘 예측하는지 나타내는 지표로, 파인튜닝 모델 기준 50 이하가 보편적 | Pass / Fail |
| BLEU-4 | 예측 문장과 정답 문장 사이에 4-그램(연속된 4단어) 단위로 얼마나 많이 겹치는지 평가하는 점수로, 파인튜닝 모델 기준 0.3 이상이 보편적 | Pass / Fail |
| ROUGE-L | 예측 문장과 정답 문장 사이의 최장 공통 부분수열(LCS) 길이를 기반으로 구조적 유사성을 측정하는 점수로, 파인튜닝 모델 기준 0.3 이상이 보편적 | Pass / Fail |
| Response time | 입력을 받고 결과를 내놓기까지 걸린 시간(초 단위)으로, 테스트 모델 기준 30초 이하가 보편적 | Pass / Fail |
| 정성적 | Special tokens | 프롬프트 강화를 위해 파인튜닝시 지정한 특수 토큰으로, 프롬프트의 형식이 잘 지켜지는지 여부 | Pass / Fail |
| 문장 구사력 | 온전한 단어를 이용한 문장의 형태로 구사하는지 여부 | Pass / Fail |
| 할루시네이션 | 사실에 기반한 내용으로 답변을 하는지 여부 | Pass / Fail |
| 어투 통일성 | 한 답변 내에서 동일한 어투를 지속적으로 구사하는지 여부 | Pass / Fail |

# 3. 테스트 케이스

| Test Case ID | 질문 내용 |
| --- | --- |
| TC-01-01 | 원소의 아버지 멘델레예프에 대해 설명좀 해줘 |
| TC-01-02 | 이차전지의 미래는 어떨거 같아? |
| TC-01-03 | 아인슈타인에 대해 알고 싶어 |
| TC-01-04 | 정자 은행이 대체 뭔데? |
| TC-01-05 | 퀴리 부인의 업적이 궁금해 |
| TC-01-06 | 뇌-컴퓨터 인터페이스(BCI) 기술을 간단히 알려줘 |
| TC-02-01 | 코로나 바이러스의 발생원인을 알려줘 |
| TC-02-02 | 리만 가설을 간단히 요약해봐 |
| TC-02-03 | 탈리도마이드 사건을 자세히 설명해봐 |
| TC-02-04 | 파킨슨병을 알려줘 |

# 

# 4. 테스트 결과

| Test Case ID | RAG 시스템 | Perplexity | BLEU-4 | ROUGE-L | Response Time | Special Tokens | 문장 구사력 | 할루시  네이션 | 어투 통일성 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TC-01-01 | Pass | Pass | Fail | Fail | Pass | Pass | Pass | Pass | Pass |
| TC-01-02 | Pass | Fail | Fail | Fail | Pass | Pass | Pass | Pass | Pass |
| TC-01-03 | Pass | Pass | Fail | Fail | Pass | Pass | Fail | Pass | Pass |
| TC-01-04 | Pass | Pass | Pass | Fail | Pass | Pass | Pass | Pass | Pass |
| TC-01-05 | Pass | Pass | Fal | Fail | Pass | Pass | Fail | Pass | Pass |
| TC-01-06 | Pass | Pass | Fail | Pass | Pass | Pass | Fail | Pass | Pass |
| TC-02-01 | Pass | Pass | Pass | Fail | Pass | Pass | Fail | Pass | Pass |
| TC-02-02 | Pass | Pass | Pass | Fail | Pass | Pass | Pass | Pass | Pass |
| TC-02-03 | Pass | Pass | Fail | Fail | Pass | Pass | Fail | Pass | Pass |
| TC-02-04 | Pass | Fail | Pass | Fail | Pass | Pass | Pass | Pass | Pass |

# 

# 5. 결론 및 개선방향

- Cosine similarity의 값에 따라 Agent가 RAG 시스템을 이용하여 Vector DB 및

Wikipedia-api 데이터를 적절히 이용할 수 있다는 것을 보여주었음. 모든 결과에서

ROUGE-L 척도는 상관관계가 없음이 나타남. 그러나, 테스트 결과에 따라서 BLEU-4

척도와 문장 구사력은 상관관계가 있는 것으로 보여짐. 대체적으로 Wikipedia에서는

자유로운 스타일과 높은 답변 품질을 일관되게 유지하였으나, VectorDB 데이터에서는

글자가 종종 누락되는 현상이 발생함. 여러 척도를 기준으로 파인튜닝 LLM 응답 생성에

대한 종합 평가를 진행해본 결과, 번역 스크립트의 품질을 끌어올리면, 더욱 안정적인

성능을 기대할 수 있다는 결과로 귀결됨.