[LG U+ Why Not SW Camp 5기]
커리어 설계 AI 에이전트 - CH 팀

성능 평가 결과서

프로젝트명: 커리어 설계 AI 에이전트

팀명: Challenge is Happiness

작성자: 김민지

목 차

Ⅰ. 개요

1. 평가 목적

- 사용자 이력 정보를 기반으로 채용공고를 추천하는 AI 시스템의 각 구성 요소(임베딩, 클러스터링, 재랭킹 LLM)의 정량적·정성적 성능을 평가

- 추천 시스템이 의미 기반 유사도, 직무 분류 적합성, 추천 사유 생성 품질 등에서 정의된 목표에 부합하는지 검증

- 모델 간 성능 비교를 통해 최적의 구성을 선정하고 향후 개선점을 도출

2. 평가 대상

- 서비스 : JOB자 추천 시스템
- 주요 모델 및 기능 :
    - 임베딩 기반 유사도 모델 (SentenceTransformer: multilingual-e5-large)
    - 직무군 클러스터링 모델 (KMeans + 리매핑)
    - LLM 기반 재랭킹 및 추천사유 생성 (llama-4-maverick API)

- 운영 환경: AWS EC2, FastAPI, PostgreSQL, MongoDB, Redis, Airflow

Ⅱ. 성능 평가 기준 및 방법

1. 임베딩 모델

- 기준 및 목표:
    - 본문 평균 유사도: 0.90 이상
    - Top-N 평균 유사도: 0.90 이상
    - 유사도 분산: 0.00005 이하 (낮을수록 추천 일관성 우수)
    - Recall (Top-20): 80% 이상

- 평가 방법:
    - 3개 모델 (BAAI/BGE, Text3S, E5) 간 동일한 사용자/공고 쌍에 대해 유사도 측정
    - 항목별 유사도, Top-N 품질, 분산, Recall을 기준으로 모델 비교

2. 클러스터링 모델

- 기준 및 목표:
    - 평균 Intra-cluster 거리: 낮을수록 좋음
    - 평균 Inter-cluster 거리: 높을수록 좋음
    - Silhouette Score: 0.02 이상

- 평가 방법:
    - K=20 vs K=200 리매핑 비교

- Intra/Inter 거리, 각 군집 간 직무 텍스트 유사도 시각화
- 클러스터 기반 추천 정합성 분석

3. LLM

- 기준 및 목표:
  - 응답 시간: 15초 이내
  - 일관성 점수: 동일 입력에 대해 90% 이상 동일 응답
  - Reasoning 점수: 명시되지 않은 조건 추론 포함률 > 70%
  - 추천 사유 포함률: 주요역량, 기술, 제외 조건 언급 포함률 > 80%

- 평가 방법:
  - lama-4-maverick vs qwen 등 비교
  - 동일 Prompt에 대해 5회 반복 테스트
  - 생성된 추천 사유의 평가 항목 체크리스트 기반 분석

III. 평가 결과 및 분석

1. 평가 결과 요약

항목	기준	결과	Pass 여부
본문 평균 유사도	≥ 0.90	0.903	✅
Top-20 평균 유사도	≥ 0.90	0.9027	✅
유사도 분산	≤ 0.00005	0.000023	✅
Recall (Top-20)	≥ 80%	80.0%	✅
Intra Distance	↓	0.0736 (vs 0.0910)	✅
Inter Distance	↑	0.0864 (vs 0.0331)	✅
Silhouette Score	≥ 0.02	0.0235	✅
LLM 응답 시간	≤ 15초	11초	✅
LLM 추론력	≥ 70%	80% 이상	✅
추천 사유 포함률	≥ 80%	92%	✅

2. 결과 분석

임베딩 모델

- E5-large는 모든 지표에서 가장 높은 정량 성능 기록
- 유사도 분산이 가장 낮아 추천 결과의 일관성 우수
- Recall 또한 80%를 넘어 실질적 추천 품질 확보

클러스터링 모델

- K=200 리매핑 후 군집 품질이 크게 향상
- 직무 세분화 + 리매핑을 통해 의미 기반 직무 분류 가능

LLM 모델

- llama-4-maverick이 일관성과 추론력에서 가장 우수
- 유저 조건, 제외사항 반영 품질 높고, 추천 사유 생성 문장이 자연스러움
- 평균 응답 시간도 서비스 적합성 확보 (11초)

IV. 결론

JOB자 추천 시스템은 임베딩, 클러스터링, LLM 재랭킹 등 3단계 모델 통합 구조를 통해 의미 기반 커리어 추천과 사용자 인사이트 제공에 적합하다.

각 모델은 정량 평가 기준을 모두 충족하며, 정성 평가에서도 추천 품질 및 해석 가능성에서 우수함을 보인다.

향후 사용자 피드백 기반 온라인 평가 구조, 도메인 특화 모델 파인튜닝, LLM RAG 구조 개선 등을 통해 지속적 성능 개선 가능하다.

Ⅴ. 별첨

1. LLM 후보군 비교 정량 평가표

모델명	시간(초)	비용(원)	답변품질① (추천 이유)	답변품질② (스킬 적용)	답변품질③ (제외사항)	모델 내 공통	모델 간 중복	모델 일관성 (0~1)	모델 추론력 (0~1)
GPT-4.0 mini	16.5	20	✅	✅	✅	3	5	1.0	1.0
GPT-4.1 nano	5.5	2.1	❌	❌	❌	4	2	1.0	0.0
GPT-4.1 mini	15.5	9.6	✅	✅	✅	4	4	1.0	1.0
Gemini 1.5 flash	9.5	3.7	❌	✅	✅	2	5	1.0	1.0
Qwen	80.0	0	✅	✅	✅	3	3	0.8	0.5
Maverick (무료)	6.0	0	✅	✅	✅	4	4	0.6	0.5
Maverick (유료)	11.0	3.2	✅	✅	✅	3	5	1.0	1.0

2. LLM 프롬프팅 및 응답 예시

[프롬프트]

```
# 프롬프트 생성 함수
def make_prompt(user, job_postings):
    user_summary = summarize_user_for_prompt(user)
    jobs_text = make_job_list_text(job_postings)

    prompt = (
        "다음은 신입 데이터 분석가의 상세 이력과 채용 공고 30개입니다.₩n₩n"
        "[지원자 정보]₩n"
        f"{user_summary}₩n₩n"
        "[요청사항]₩n"
        "- 아래 채용 공고 목록은 **지원자와의 유사도 순으로 정렬된 상태**입니다. 상위 공고일수록 우선적으로 검토해 주세요.₩n"
        "- 아래 공고 중 **지원자가 현실적으로 지원할 수 없는 자격 요건**(예: 전문연구요원, 병역특례, 경력 1년 이상 등)을 포함한 공고는 제
        "  단, '지원 가능' 등의 예외 문구가 명시된 경우는 포함 가능합니다.₩n"
        "- 지원자의 기술 스택은 **숙련도(상/중/하)** 정보가 포함되어 있습니다. 이를 기반으로 다음과 같이 **가중치를 적용해 매칭**해 주세요
        "  - 숙련도: 상 = 3점 / 중 = 2점 / 하 = 1점₩n"
        "- 단순 조건 충족이 아니라, **실제로 잘 맞는 공고** 5개를 추천해 주세요.₩n"
        "- 각 추천 공고에 대해 아래 3가지를 작성해 주세요:₩n"
        "  1) 적합한 이유 (지원자의 정보와 일치, 예: 지원자 정보, 기술스택 등 )₩n"
        "  2) 부족하거나 충족하지 못한 부분이 있다면 구체적으로 작성₩n₩n"
        "[채용 공고 목록]₩n"
        f"{jobs_text}₩n₩n"
        "위 목록에서 가장 적합한 5개 공고를 아래 형식으로 출력해 주세요:₩n"
        "- 공고 ID₩n"
        "- 직무명₩n"
        "- 적합 이유₩n"
        "- 부족한 부분₩n"
    )

    return prompt
```

[응답 예시]

✅ 응답 완료!
김민지님께 적합한 채용 공고 5개를 추천합니다. 병역특례 및 전문연구요원 관련 공고는 제외했습니다.

1. 공고ID: 10 직무명: 주니어 데이터 사이언티스트

* **적합 이유:** 김민지님의 주요 경험(캐치테이블 사용자 분석, 이직률 예측, 토스 AARRR 분석)과 기술 스택(Python, SQL, Pandas, Tableau)이 공고의 요구사항과 매우 잘 부합합니다. 특히, 다양한 데이터 분석 프로젝트 경험은 "대규모 데이터셋의 수집, 정제 및 전처리", "데이터 분석 및 패턴 탐색을 통해 인사이트 도출", "머신러닝 및 통계 모델을 활용한 예측 분석 및 성능 평가" 등의 주요 업무와 직접적으로 연결됩니다. ADsP 자격증도 데이터 분석 능력을 증명하는 좋은 자료가 될 것입니다.

* **부족한 부분:** 우대사항에 언급된 "영어 커뮤니케이션 능력", "TensorFlow, PyTorch 등 머신러닝 및 딥러닝 모델 개발 경험", "데이터 시각화 도구 활용 경험 (Tableau, Power BI 등)", "빅데이터 처리 기술 경험 (Hadoop, Spark 등)" 등은 김민지님의 이력서에 명시적으로 드러나지 않습니다. Tableau 경험은 있으나 Power BI 경험은 없고, 빅데이터 처리 경험 또한 부족합니다.

* **부족한 점 보완을 위한 커리어 로드맵 및 학습 방향:** 영어 공부와 더불어 TensorFlow/PyTorch를 활용한 머신러닝/딥러닝 프로젝트 경험을 쌓는 것이 중요합니다. 가능하다면 Kaggle 경진대회 참여나 개인 프로젝트를 통해 실력을 향상시키고 포트폴리오를 구축해야 합니다. Power BI 학습을 통해 데이터 시각화 툴 활용 범위를 넓히고, Hadoop/Spark 관련 온라인 강의나 프로젝트를 수행하여 빅데이터 처리 경험을 추가하는 것이 좋습니다.

2. 공고ID: 20 직무명: Generative AI Engineer

* **적합 이유:** 김민지님은 Python과 Pandas를 능숙하게 다루며, 다양한 데이터 분석 프로젝트를 수행한 경험이 있습니다. Generative AI Engineer는 데이터 처리 및 분석 능력이 중요한데, 김민지님의 경험이 이 부분을 충족합니다. 특히, AARRR 분석 경험은 "비정형 데이터(문서, 이메일, 채팅 등)에서 구조화된 인사이트를 추출하고 분석"하는 능력과 연관되어 있습니다.

* **부족한 부분:** 공고에서 우대하는 "LangChain, LlamaIndex 등 LLM 애플리케이션 프레임워크 활용 경험", "RAG(Retrieval Augmented Generation) 시스템 설계 및 구현 경험", "벡터 데이터베이스 활용 경험" 등은 김민지님의 이력서에 없습니다. 생성형 AI 관련 경험이 부족합니다.

* **부족한 점 보완을 위한 커리어 로드맵 및 학습 방향:** LangChain, LlamaIndex, 벡터 데이터베이스(Pinecone, Weaviate 등)를 활용한 프로젝트를 진행해야 합니다. LLM과 RAG를 활용한 개인 프로젝트를 통해 실무 경험을 쌓고, 이를 포트폴리오로 제시하는 것이 좋습니다. 관련 온라인 강의 수강도 도움이 될 것입니다.

3. 클러스터별 직무 분표 시각화