LLM 평가 프레임워크 비교 분석

↳ 각 프레임워크의 특징이 뚜렷이고, 어떤 환경도 평가를 수행하는 지
파악한 다음에 사용할 평가 프레임워크를 선정하고 사용해야한다.

[비교해볼 대표 LLM 평가 프레임워크]

1) RAGAS. DeepEval 등 'LLM-as-a-judge' 기반 프레임워크
2) LangSmith. Trulens 등 Observability 프레임워크
3) Huggingface evaluate
4) ARES
5) AutoRAG

※ LLM-as-a-judge 프레임워크
↳ "LLM이 평가자가 되어 평가한다"라는 Concept
(핵심) ↳ 대부분 "프롬프트 엔지니어링"이다.

1) RAGAS
- 가장 유명한 평가 프레임워크
- Retrieve 한 단락이 질문과 연관이 있는 지를 LLM으로 평가
  ↳ 단락 내용을 모두 LLM 에게 전달해야하기 때문에 '굉장히 비싸다'
  ↳ RAGAS는 'RAG 평가'에 사용되기에는 비싼 편, LLM 평가 에서는 많이 사용됨.
- LLM 평가 뿐 아니라 여러 평가 Metric을 지원한다는 장점을 가짐
  ↳ Agent의 Tool 사용에 대한 평가도 존재 (RAGAS에만 거의 유일하게 존재)
  ↳ SQL Metric 과 전통 NLP 스코어도 존재
    ↳ Vector DB기반이 아닌 RDB기반 DB를 LLM Application에 사용 시
      LLM이 SQL Query 문을 작성하도록 하는 경우도 존재하는데,
      이 때, 이를 평가하는 지표.
- 더러 제한 X (영어만 지원), 한국어는 성능이 떨어질 수 있음.

2) DeepEval

- RAGAS, G-Eval 등 LLM-as-a-judge 메트릭 포함
- AI Safety 관련 메트릭 포함 (LLM-as-a-judge, RAGAS에는 X)
- 여러 벤치마크 데이터셋 지원, CI/CD 통합 지원
  - ↳ LLM 평가 반 복잡하거나, ↳ 자동화된 평가를 자동으로 수행하는 것도 지원
  - 직접 구현한 LLM application이 가진 데이터셋을 활용하도 충분히 General한 도메인 평가
- 다국어 지원 X
- Retrieval Metric X ⇝ RAG를 전혀 고려하고 있지 X

3) OpenAI Evals

- OpenAI에서 직접 제작한 평가 프레임워크
- OpenAI Leaderboard에서 사용 가능
- LLM 평가에 집중되어 있음.
- OpenAI 모델만 사용 가능.
  ⇝ 즉, OpenAI 모델이 아닌 모델을 파인튜닝한 경우 사용이 제한됨.

4) LangSmith

- LangChain과 연동되는 모니터링 및 테스트 플랫폼.
- 기본적으로는 모니터링 툴, 평가 기능 지원
- 'LLM-as-a-judge'를 기본적으로 제공, 커스텀 메트릭 지원
- 평가 결과를 쉽게 확인 가능한 대시보드 지원
- 치명적 단점 : 부분적 유료...

5) Trulens ; LangSmith와 비슷하지만 조금 더 '평가'에 초점을 맞춘 프레임워크

- Human-in-the-loop 등 반복적 평가에 특화
  ⇝ 평가 과정에 사람이 개입할 수 있도록 해 신뢰성↑
- 결과 확인을 위한 대시보드 지원
- AI Safety 관련 메트릭 포함
- Observility 등의 성격이 있는 평가 프레임워크

6) Hugging face Evaluate

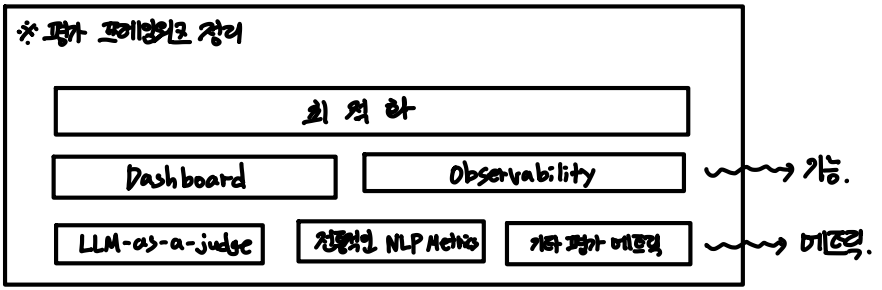
- 프레임워크X, 라이브러리 O → 평가 메트릭을 모아둔 것.
  → 평가용 ML 메트릭과 데이터셋을 로딩하는 라이브러리
- 정통 NLP 메트릭 다수를 쉽게 사용할 수 있도록 제공
  → 'LLM-as-a-judge'보다는 전통적인 NLP 평가 메트릭을 많이 쓰임
- 간단한 사용법으로 빠르게 메트릭 사용 가능
- 다른 평가용 라이브러리들에 의존 관계가 많음
  → 다른 라이브러리를 Wrapping 하기에 좋음

7) ARES

- 스탠포드 Univ.에서 만든 LLM평가 프레임워크
- 평가만을 위한 sLLM을 파인튜닝 하는 것이 특징
  → 가장 높은 연산 비용이 필요하지만, 'LLM-as-a-judge'중 정확도가 높은 편
  → 하지만 파인튜닝이 필요하다는 점이 가장 큰 단점.
  → Local 모델을 사용해야하기 때문에 GPU도 필요.

8) AutoRAG

- RAG에 특화된 최적화 프레임워크
  RAG 최적화에 초점
- NLP 메트릭, LLM-as-a-judge 등 여러 메트릭 조합
  → ARES를 제외한 많은 프레임워크가 통합되어 있음.
- Retrieval 평가 및 최적화 지원 (Retrieval GT 기반의 정확한 Retrieval 평가 지원)
- 평가 데이터셋 생성에서 한국, 영어, 일본어 지원
- 평가 리더보드 지원


※ 평가 프레임워크 정리
회 원 하
Dashboard
Observability
LLM-as-a-judge
전통적인 NLP Metrics
기타 평가 메트릭
기능.
메트릭.