평가지표

> ※ 시스템 성능을 보여주는 테스트 케이스 → 시간이 부족하여 구현하지 못하였습니다.
>
> 본 정리는 구현 대신 아이디어 및 평가 체계에 대한 이론적 고찰을 중심으로 작성되었습니다.

---

1. 에이전트 능력 평가 (Assessing Agent Capabilities)

에이전트가 실제 문제를 해결할 수 있는 기본 역량을 갖추고 있는지 평가하는 단계입니다. 주로 다음과 같은 능력을 검증합니다:

- 지시 이해 능력
- 논리적 추론 능력
- 적절한 도구 선택 및 사용 능력

주요 공개 벤치마크 예시

- **BFCL (Berkeley Function-Calling Leaderboard)**: 도구 선택 및 호출 능력 평가
- **τ-bench**: 계획 수립 및 명령 처리 역량 평가
- **AgentBench**: 다양한 시나리오에 걸쳐 종단간(end-to-end) 성능 평가
- **DABStep (Adyen)**: 특정 서비스 환경에 맞춘 전문화된 평가 기준 제공

> 공개 벤치마크를 활용하여 에이전트의 전반적인 기반 역량을 확인한 후, 실제 서비스 상황에 특화된 별도의 평가 지표 설계가 필요합니다.

---

2. 궤적(Trajectory) 및 도구 사용 평가

에이전트가 목표를 달성하기까지 어떤 과정을 거쳤는지를 평가합니다.

단순히 결과가 맞았는지를 넘어서, 어떤 행동 순서와 도구 사용을 통해 도달했는지 분석하는 것입니다.

사용 도구 예시

- **LangSmith**: 에이전트의 작동 과정을 추적하고 분석할 수 있는 도구

평가 지표

- **완벽 일치 (Exact match)**: 이상적인 궤적과 완전히 일치 여부
- **순서 일치 (In-order match)**: 핵심 단계의 순서를 정확히 따랐는지
- **순서 무관 일치 (Any-order match)**: 수행 단계만 포함되었는지
- **정밀도 (Precision)**: 사용한 도구 중 실제 필요한 도구의 비율
- **재현율 (Recall)**: 필요한 도구 중 실제 사용한 도구의 비율
- **단일 도구 사용 평가**: 특정 도구 사용 학습 여부 확인

> 이 단계의 평가는 참조 기준(ground truth trajectory)이 필요하며,
>
> 이를 수작업으로 생성하는 것은 자원 소모가 크기 때문에,
>
> 최근에는 **LLM 기반 자동 평가 시스템(예: Agent as a Judge)** 연구가 활발히 이루어지고 있습니다.

---

3. 최종 응답 평가 (Evaluating the Final Response)

에이전트가 사용자에게 제공하는 최종 결과물의 정확성과 적절성을 평가합니다.

핵심은 **"에이전트가 주어진 목표를 달성했는가?"** 입니다.

예시

- 리서치 에이전트가 요약을 적절한 어조와 형식으로 작성했는가
- 고객 응대 에이전트가 상품 문의에 대해 정확하고 관련 있는 답변을 했는가

자동화된 평가 방식

- **Autorater**: LLM이 평가자 역할을 수행
- **Human-in-the-loop (HITL)** 병행 필요: 상식, 창의성, 맥락 이해 등은 사람의 판단이 여전히 중요

평가의 도전 과제

구분	설명
평가용 데이터 부족	참조 궤적 및 정답 데이터를 구축하는 데 높은 비용 소요
자동 평가의 한계	생성 모델이 평가자 역할을 할 경우 추론 과정 누락 가능
기존 시스템과의 연계	기존 대화 시스템 평가 방식의 적용이 미흡함
멀티모달 평가의 어려움	텍스트 외 이미지·오디오 등 포함 시 전용 지표 필요
실제 환경 평가의 복잡성	예측 불가능한 동적 환경에서의 성능 측정이 어려움

향후 평가 방향

1. **결과 중심에서 과정 중심 평가로 전환**
    - 단순 출력 정확성보다 에이전트의 추론 및 행동 과정 평가 강조
2. **AI 기반 평가 방식 확대**
    - 평가 자동화 및 확장성 확보 목적
3. **실제 서비스 문맥 중심 평가 강화**
    - 특정 사용 사례에 최적화된 평가 기준 설정
4. **표준화된 벤치마크 개발**
    - 다양한 모델 간 객관적 비교 가능성 확보
5. **설명 가능성과 해석 가능성 강조**
    - 에이전트 행동에 대한 더 깊은 이해 제공

이와 같은 **다층적 평가 접근**은 단순히 정답을 맞히는 에이전트가 아닌,

신뢰 가능하고 상황에 적응할 수 있는 지능형 시스템 개발로 이어집니다.