MODULE 06

어떤 모델이 "좋다"고 말할 근거

모델을 도입하려면 "써보니 괜찮더라"보다 나은 근거가 필요합니다. 이 모듈에서는 같은 문제지로 여러 모델을 채점하는 법, 그리고 무엇을 재야 하는지를 다룹니다.

01 인상이 아니라 측정

"써보니 괜찮던데"는 데모 몇 번을 해본 인상일 뿐입니다. 도입 근거로 쓰려면 같은 문제지로 여러 모델을 채점한 표가 있어야 합니다.

그 문제지가 **골든셋**입니다. 입력과 정답(모범 답)의 쌍을 모아둔 것으로, 50~100건 정도면 의미 있는 비교를 시작할 수 있습니다.

골든셋이 있으면 다음을 전부 같은 문제지로 확인할 수 있습니다.

- ✔ **모델 비교** — 후보 모델들의 성능을 같은 기준으로 나란히 비교합니다.
- ✔ **양자화 품질 확인** — 모듈 3에서 다룬 양자화가 품질을 얼마나 깎았는지 측정합니다.
- ✔ **프롬프트 개선의 전후 비교** — 프롬프트를 고치기 전과 후의 점수를 비교합니다.

02 민감 데이터 없이 평가하기 — 합성 데이터

평가하려면 실제 업무 데이터가 필요합니다. 그런데 실데이터에는 개인정보가 들어 있어서 그대로 쓸 수 없습니다.

해법은 실데이터의 분포(어떤 유형이 몇 %씩 섞여 있는지의 비율 구성), 즉 유형 비율·길이·난이도를 먼저 분석하고, 그 분포를 따르는 가짜 데이터를 LLM으로 생성하는 것입니다. 이것이 **합성 데이터**입니다.

실사례 — 합성 골든셋 (리더의 사전 실증, 2026 상반기)
연구회 리더가 올해 상반기에 진행한 사전 실증에서는, 전년도 민원 통계에서 유형 분포를 뽑아내고 그 분포를 그대로 따르는 가상 민원 100건을 생성해 골든셋을 만들었습니다. 개인정보는 0건이지만, 실제 업무의 난이도는 그대로 반영되어 있습니다. 이 방법을 9월에 우리 공통 태스크에 그대로 적용해 볼 겁니다.

03 무엇을 재나 — 네 가지 지표

평가라고 다 같은 지표를 재는 것이 아닙니다. 태스크마다 중요한 지표가 다릅니다.

지표	뜻	이런 태스크에서 중요
정답률(Accuracy)	정답과 일치하는 비율	분류·추출
근거성(Groundedness)	주어진 근거에만 기반해 답했나	RAG·요약
지연(Latency)	요청부터 응답까지 시간	대화형, 실시간 보조
처리량(Throughput)	단위 시간당 처리 건수	야간 배치, 대량 처리

야간 배치 처리에는 지연이 덜 중요하고, 실시간 보조 업무에는 지연이 결정적입니다. 지표를 하나만 보고 "이 모델이 좋다"고 말할 수 없는 이유가 여기에 있습니다.

04 환각 — 그럴듯한 거짓

모델은 모듈 2에서 본 것처럼 다음 토큰을 예측하는 프로그램입니다. 그래서 모르는 내용도 그럴듯하게 지어낼 수 있습니다. 이것을 **환각(hallucination)**이라고 부릅니다.

- **RAG로 근거를 준다** — 지어낼 필요를 애초에 줄입니다.
- **출력과 근거를 대조 검사한다** — 답이 근거 문서에 실제로 있는 내용인지 확인합니다.
- **기계 검증 가능한 부분은 게이트로 잡는다** — 숫자나 형식처럼 결정적으로 확인 가능한 부분은 게이트

에서 걸러냅니다.

05 채점을 누가 하나 — 기계, 사람, 그리고 LLM 심판

채점에도 우선순위가 있습니다.

- ① **기계 채점** — 정답 일치, 정규식(010-####-#### 같은 글자 모양 규칙을 기계적으로 검사하는 방법입니다 — 처음 보는 말이면 워밍업에서 먼저 익히세요), 형식 검사처럼 결정적으로 확인 가능한 것부터 채점합니다.
- ② **LLM-as-judge** — 더 강한 모델에게 채점 기준표를 주고 채점을 맡깁니다. 요약 품질처럼 기계 채점이 어려운 항목에 씁니다.
- ③ **사람 검수** — 최종 확인 차원에서 일부를 샘플링(전체 중 일부만 무작위로 뽑아 확인)해 사람이 직접 봅니다.

LLM 심판을 쓸 때는 주의할 점이 있습니다. 반출이 허용된 데이터만 외부의 강력한 모델로 보내야 하고, 심판이 내린 결과도 다시 샘플링해서 사람이 재확인해야 합니다.

실측 교훈 (리더의 사전 실증)
같은 민원 분류 사전 실증에서 품질 1위는 thinking 모델이었습니다. 하지만 건당 약 18초(일반 모델의 6배)에, 한 건은 65분까지 폭주하는 꼬리 리스크까지 있어 실무에는 쓸 수 없다는 판정을 받았습니다. 평가의 결론은 "가장 좋은 모델"이 아니라, "업무 요구(품질 기준과 시간 기준)를 만족하는 가장 가벼운 모델"을 고르는 것이어야 합니다.

🔗 **우리 연구회에선**
9월에는 공통 태스크 2개의 골든셋을 함께 만들고, 우리 서버 모델과 후보 모델들을 같은 문제지로 채점해서 '모델 선정 사유서'를 씁니다. 이 모듈은 그 실습을 위한 예습입니다.

↗ **심화 학습**

평가 개념을 실습으로: **Hugging Face LLM Course** · 영어(일부 한국어 번역), 무료.
평가 사고방식의 뿌리(선택 심화): **Stanford CS229 — Machine Learning (Andrew Ng, 공개 강의)** · 영어(자막 지원). LLM 특화는 아니지만 "측정으로 판단한다"는 사고의 원류입니다.

CHECK

이해했는지 확인해 봅시다

틀려도 괜찮습니다 — 오답을 고르면 왜 아닌지 설명이 나오고, 정답을 찾을 때까지 다시 고를 수 있습니다. 점수는 첫 시도 기준입니다.

문제 1 / 4

Q1. 골든셋(Golden Set)이란?

① 모델이 학습할 때 사용한 원본 데이터셋

② 입력과 정답 쌍으로 이뤄진, 채점 기준이 되는 평가용 문제지

③ 모델이 절대 답하면 안 되는 금칙 사례 목록

④ 실제 사용자들의 질문 로그를 그대로 모아둔 것

← 이전
모듈 5 · RAG와 에이전트

다음 →