판단 모델 도입 전 체크리스트 10

1 **선택지** — 답의 선택지를 겹치지 않게, 빠짐없이 적었다. "기타 · 해당 없음"도 넣었다.

근거 · TypeSafe 문서: 선택지 전체 목록을 주고 other · none of the above를 추가하라 (/primitives/choice)

2 **기준 문장** — 선택지 · 단계마다 판단 기준을 한 문장으로 적었다. "정도"가 아니라 "상황"으로.

근거 · 문서: Describe situations, not degrees (/primitives/score)

3 **코드 먼저** — 계산 · 날짜 비교 · 개수 세기 · 조회는 판단 모델에 묻지 않고 코드로 처리했다.

근거 · Jev 1.13 약점 문서: 계산 · 날짜 비교 · 개수 세기 (/model-jaggedness/jev-1.13)

4 **검증 데이터** — 정답을 아는 우리 업무 샘플 50~100건을 모았다.

근거 · 문서: 보수적인 기준값에서 시작해 우리 데이터로 시험하고 조정하라 (/confidence)

5 **구간별 정확도** — 확신도 구간마다 실제 정확도를 쟀다. 확신도 0.9를 "90% 맞음"으로 읽지 않는다.

근거 · OpenRouter 실측: 확신도는 순서는 잘 매기지만 보정된 확률은 아님

6 **기준값** — 틀렸을 때 비용을 보고 자동 처리 기준값을 정했다. 되돌리기 어려운 일일수록 높게.

근거 · 문서: Thresholds scale with risk (/confidence)

7 **넘길 곳** — 기준값 미만일 때 넘길 곳(LLM 재검토 · 사람)을 정했다.

근거 · Decision Stack 규칙: 확신이 낮으면 한 층 위로

8 **버전 기록** — 응답의 모델 이름을 버전까지 기록하고, 버전이 바뀌면 검증을 다시 한다.

근거 · 세미나 Lab: 버전이 바뀌자 18개 중 4개 예제에서 판정이 바뀜

9 **한도** — 요청 크기가 한도 안에 드는지 확인했다 — 요청당 64k 토큰, 판단 재료 + 가장 긴 질문 32k.

근거 · TypeSafe 문서 (/models)

10 **표현** — "환각 없음 · 200배" 같은 벤더 표현을 보고서에 그대로 옮기지 않는다.

근거 · 「환각 없음」은 형식 밖 답이 없다는 뜻 · 193.6배는 제조사 자체 시험의 최대치