기술보고서

은행 거래를 처리할 때 AI 에이전트는 필요한 사실을 먼저 확인하는가?

Flitto Agent Bench (Finance – Global BankOps)

발행 · 플리토 에이전트 데이터 리서처 팀

2026년 11월 23일

초록

Global BankOps는 AI에게 가상 은행의 업무를 맡겨 서로 다른 기록을 대조하고 거래를 처리하는 과정을 평가한다. 무역금융, 여신 관리, 자금 운영, 회계, 연금·신탁의 다섯 부문을 다룬다. 원장과 계약서, SWIFT 전문의 내용이 다르면 에이전트는 사정을 아는 부서에 문의한 뒤 처리 방향을 정한다. 채점할 때는 필요한 사실을 언제 확인했는지, 무엇을 기표하거나 발송했는지, 마지막 보고가 실제 처리 결과와 일치하는지 살핀다. 8개 모델에 50문항을 두 번씩 맡긴 결과, 800회 중 342회가 통과했다. 같은 모델이 같은 문항을 수행한 400쌍 중 40쌍(10.0%)은 두 시도의 판정이 달랐다. 실패한 458회 중 413회에서 근거를 확인하는 절차나 업무 처리 순서를 지키지 못했다. 최종 보고서와 함께 회신을 읽은 시점, 지급이나 기표를 실행한 시점을 살펴야 이런 실패를 찾을 수 있다.

1. 서론

은행에서 원장에 거래를 기록하거나 처리 결과를 통지하려면 여러 부서의 자료를 먼저 맞춰 봐야 할 때가 있다. 담당자는 어느 자료를 기준으로 삼을지 정하고 필요한 사실을 확인한 뒤 그에 따라 업무를 처리한다. 이를 평가하려면 최종 보고서와 함께 담당자가 어떤 순서로 일했고 그 결과 무엇이 달라졌는지 살펴야 한다.

각 문항은 모델에게 가상 은행의 한 직무와 업무 지시를 부여한다. 모델은 그 직무에 허용된 도구로 기록을 조회하고 업무를 처리한다. 평가 환경은 모델이 내린 결정과 처리 순서, 그 결과를 기록한다.

1.1 선행 연구와 연구 질문

FinQA는 금융 보고서를 읽고 추론하는 능력을 평가한다[1]. Bank-World는 가상 은행에서 에이전트의 업무 수행을 평가하며[2], τ-Bench는 에이전트가 업무 규정에 따라 행동한 뒤 데이터베이스에 남긴 결과를 확인한다[3]. Global BankOps에서는 서로 다른 기록을 대조한 뒤 업무를 처리했는지, 그때 남긴 원장 기록과 보고 내용이 일치하는지를 살핀다.

이 연구는 같은 은행 업무를 두 번 맡겼을 때 결과가 얼마나 달라지는지, 거래 전에 필요한 사실을 확인한 시점이 업무 통과 여부와 어떻게 맞물리는지 살폈다 [6].

2. 은행 업무의 범위와 문항 설계

2.1 다섯 부문의 업무

모델은 맡은 업무에 따라 약정서와 원장, 메시지를 읽고 권한이 있는 기록을 조회한다. 다른 부서에 문의하고 회신을 읽을 수도 있다. 지급, 기표, 보고, 거래 보류 등 모델이 내린 결정은 업무 기록에 남는다.

50문항 중 40개는 부다페스트 현지법인 업무이고, 10개는 서울 본점의 연금·신탁 업무다. 31개는 은행 내부에서 처리하며, 19개는 다른 금융기관과 SWIFT 전문을 주고받는다.

표 1. 다섯 부문의 업무와 사용하는 기록

부문	문항 수	담당 업무
무역금융	10	신용장과 은행 간 청구를 대조하고 지급·보류 여부를 알린다.
여신 관리	10	대출 약정과 거래 내역을 대조해 이자와 상환 조건을 확인한다.
자금·결제	10	계좌 명세와 거래 기록을 대사하고 외환·단기자금 결제를 처리한다.
회계	10	월말 발생액과 평가액을 기표하고 본점에 보고할 명세를 원장과 맞춘다.
연금·신탁	10	연금 수급 자격과 신탁 지시 권한을 확인한 뒤 지급하고 기록을 갱신한다.

2.2 자료가 서로 다를 때의 판단

자료가 오래됐거나 기록끼리 맞지 않으면 담당자는 해당 정보를 관리하는 부서에 사실을 확인할 수 있다. 일부 문항에서는 최신 전산 기록도 근거로 인정한다. 이때는 조치에 앞서 관련 기록을 조회해야 한다. 평가마다 모델 설정, 도구 사용 내역, 업무 처리 뒤의 상태와 채점 결과를 보관했다.

각 문항에는 업무 문서와 원장, SWIFT 전문을 비롯한 메시지가 있다. 직무별 열람 권한과 다른 부서가 알려 줄 수 있는 정보를 정하고, 지급이나 기표를 하면 거래 기록에 그 결과가 반영되도록 했다. 일부 문항에서는 내부 승인 절차를 단순화했다.

2.3 문항을 만든 근거

은행 실무자 인터뷰, 실무자가 제공한 자료, 공개된 업무 지침을 바탕으로 문항을 만들었다. 각 문항에는 해당 업무 절차를 어디에서 참고했는지 기록했다.

세부 절차는 Oracle FLEXCUBE 설명서[4]와 한국의 퇴직급여 규정 등을 참고했다. 문항에 등장하는 기관과 인물, 거래 금액은 가상이다.

문항을 고를 때는 사전 평가에서 관찰한 난이도도 참고했다.

3. 평가 절차

3.1 두 번의 실행과 채점

문항판 v1.0.0에서는 모델에게 실제 은행 업무처럼 지시하고 난수 시드는 0으로 고정했다. 한 차례 응답에서 도구는 한 번만 호출할 수 있다. 문항별 호출 한도는 40~90회이며 이 중 42문항은 60회다. 기본 제한 시간은 시도당 1,800초다. 표 B1에 제시한 모델별 추론 강도는 두 시도에서 같았다.

각 모델·문항 조합에서 채점이 끝난 첫 두 시도를 채택했다. 시작 시각이 같으면 실행 ID 순으로 정렬했다. 서비스 연결 오류 1회, 시간 초과 1회, 두 시도 뒤에 채점된 4회는 제외했다. 비교 대상은 50문항을 모두 마친 8개 모델 설정이다. 채점 근거의 인용을 바로잡거나 판정 이견을 검토해야 했던 경우에는 원래 업무 기록에서 확인한 판정을 썼다.

채점 방식 v6은 코드 검사와 언어 모델 판정을 함께 사용한다. 메시지나 보고서의 내용이 중요한 경우 Gemini 3.8 Flash가 서술형 기준에 따라 판정하고 근거가 된 구절을 인용한다. 이때 온도는 0, 추론 강도는 낮음으로 설정했다.

3.2 통과율의 계산

시작 시각을 기준으로 각 모델이 같은 문항을 처음 수행한 때와 두 번째 수행한 때를 구분했다. pass@1은 첫 번째에, pass@2는 둘 중 적어도 한 번, pass²는 두 번 모두 통과한 문항의 비율이다. 모델별 분모는 50문항이고 분야별 분모는 10문항이다.

세 지표 모두 같은 400개 모델·문항 조합을 쓴다. 각 모델이 모든 문항을 정확히 두 번씩 수행했기 때문이다. 부문별 통과율, 문항 난이도, 모델별 회신 확인 결과와 통계 검정은 첫 시도를 기준으로 삼았다. 실패 횟수, 비용, 도구 호출 횟수는 두 시도를 합쳐 800회 전체에서 집계했다.

4. 두 번의 시도에서 달라진 결과

확정된 50문항을 기준으로 결과를 집계했다. 이전에 검증한 문항에서 얻은 기록도 입력과 채점기가 그대로면 포함했다. 같은 기록을 재검토해 판정을 바로잡은 경우에는 검증한 판정을 적용하고 원래 실행 기록과 판정도 보관했다. 모델마다 문항당 두 번씩, 전체 800회 중 342회가 통과했고 458회가 실패했다.

표 2. 첫 시도, 두 번 중 한 번 이상, 두 번 모두 통과한 문항의 비율

모델	pass@1	pass@2	pass²	두 시도가 달랐던 문항
GPT-6.2 Astra	78.0%	84.0%	70.0%	7
GPT-6 Astra	74.0%	74.0%	68.0%	3
Fable 5.2	56.0%	62.0%	50.0%	6
Gemini 4 Carbon	48.0%	54.0%	46.0%	4
Opus 5.5	34.0%	40.0%	26.0%	7
Gemini 4 Argon	28.0%	30.0%	24.0%	3
Grok 4.7	16.0%	22.0%	10.0%	6
Qwen3.8 Max Prime	14.0%	16.0%	8.0%	4

표 2의 세 지표에서 모델 순위는 같았다. 모델·문항 조합 400개 중 40개는 첫 시도와 두 번째 시도의 통과 여부가 달랐다. 이런 문항은 GPT-6 Astra와 Gemini 4 Argon이 각 3개로 가장 적었고, GPT-6.2 Astra와 Opus 5.5가 각 7개로 가장 많았다. 두 번 중 한 번이라도 통과한 문항을 세면 GPT-6.2 Astra는 78.0%에서 84.0%로 올랐고 GPT-6 Astra는 74.0%로 같았다. 두 번 모두 통과한 문항만 세면 GPT-6.2 Astra는 70.0%, GPT-6 Astra는 68.0%, Grok 4.7은 10.0%, Qwen3.8 Max Prime은 8.0%였다.

GPT-6.2 Astra는 첫 시도에 39문항, 두 번 중 한 번 이상은 42문항을 통과했다. 하지만 두 번 모두 통과한 문항은 35개였다. 42문항을 한 번 이상 처리할 수 있었다는 결과와, 같은 업무를 반복해 같은 판정을 받은 결과는 구분할 필요가 있다.

5. 근거를 확인한 시점과 거래 처리

5.1 회신을 받았을 때와 읽었을 때

44문항은 지정된 부서나 거래 상대 은행에 문의하고 회신을 읽은 뒤에야 다음 업무를 처리할 수 있다. 이 중 4문항은 그런 확인 절차가 여러 개다. 다른 3문항은 조치 전에 최신 전산 기록을 조회하는 방법도 인정한다. 나머지 3문항에는 이런 확인 절차가 없다.

두 시도를 합쳐 실패한 458회 중 413회에서는 근거 확인이나 처리 순서에 문제가 있었다. 작성·저장한 결과물이 기준에 못 미친 경우는 351회, 지급·원장 처리 오류는 160회, 허용된 행동 범위를 벗어난 경우는 114회, 완료 보고 요건을 놓친 경우는 8회였다. 한 시도에서 여러 문제가 함께 나타날 수 있다. 필수 회신 확인을 하나라도 놓친 401회를 따로 보면, 247회에서는 지정된 상대에게 필요한 내용을 묻는 요청을 하나 이상 찾을 수 없었다. 이 중 29회는 적어도 한 건에 대해 문의할 상대에게 연락은 했고, 218회는 조건에 맞지 않았던 요청 중 어느 건에서도 맞는 상대에게 연락한 기록이 없었다. 나머지 154회 중 92회는 요청을 제대로 보냈지만 필요한 회신을 하나 이상 읽지 않았고, 62회는 일을 처리한 뒤에야 회신을 읽었다.

첫 시도에서 필수 회신을 모두 제때 읽은 155회 중 139회가 업무 전체도 통과했다. 회신을 읽는 것만으로 통과가 결정되지는 않았다. 다른 요건까지 충족하려면 그 내용을 실제 지급, 기표와 보고에 반영해야 했다.

5.2 지급 승인과 원장 기표

표 3. 원 연구 800회 중 절차별 적용 대상과 관찰 결과

점검한 업무	관찰된 실행 / 적용 대상
필요한 회신을 제때 확인하지 못함	401 / 704
승인 회신을 읽기 전에 대상 지급을 실행	0 / 31
승인 확인이 필요한 업무에서 대상 지급 기록이 없음	1 / 32
원장에 필요한 기표 또는 계정별 순변동 요건을 어김	103 / 560

표 3은 각 절차를 확인할 수 있었던 실행끼리 비교한다 [7]. 승인 절차가 적용된 32회 가운데 31회에 대상 지급이 기록됐고, 31회 모두 승인 회신을 읽은 뒤에 지급했다. 나머지 한 번에는 대상 지급 기록이 없어 지급 시점과 승인의 선후관계를 판단할 수 없다.

6. 결과를 해석할 때의 한계

50문항은 은행 업무 중 일부를 골라 구성했고, 일부 문항의 승인 절차와 정보 접근 방식은 단순화했다. 두 번의 시도만으로 업무를 언제나 일관되게 수행할지 추정하기는 어렵다. 통계 검정도 첫 시도 결과만을 대상으로 했다. 기록 재현으로 확인한 것은 같은 행동에 대한 가상 환경의 응답이다. 두 시도의 모델 응답이 달랐던 데에는 생성 과정의 변동, API 경로, 실행 환경, 캐시가 영향을 줬을 수 있

다. 모델별 추론 설정과 응답 한도도 달랐다. 표 A4에서는 한 모델의 응답을 대상으로 v5 채점기의 판정을 비교했다. 개발 중 일부 기준은 사람이 검토했지만, v6 채점기 전체의 판정이 독립적인 사람의 판정과 얼마나 일치하는지는 측정하지 않았다.

가상 은행에서는 요건에 맞게 문의하면 짧고 일정한 시간 뒤에 비용 없이 정확한 답이 온다. 현업에서는 회신이 늦거나 불완전할 수 있고 문의에도 비용이 들어, 이 평가에서는 문의하는 편이 실무보다 유리할 수 있다. 반대로 올바른 내용을 다른 말로 물어도 시스템이 표현을 인식하지 못하면 확인 기록이 남지 않을 수 있다. GPT-6 Astra는 연금·신탁 10문항을 두 번 모두 통과했고, 자금 운영 10문항 중 9개를 첫 시도에 통과했다. 8개 모델이 모두 첫 시도에 통과한 3문항 중 2개는 두 번째 시도에도 모두 통과해, 이 문항으로는 성능이 더 나아졌는지 가리기 어렵다.

7. 결론

Global BankOps는 은행의 다섯 부문에서 모델이 필요한 사실을 확인하고 그에 따라 업무를 처리한 뒤 결과를 정확히 보고하는지 평가했다. 실패한 458회 중 413회에서는 근거 확인이나 처리 순서를 놓쳤다. 문의가 필수인 문항에서 제때 확인한 횟수의 순위는 하위 두 모델을 제외하면 전체 통과율 순위와 같았다. 두 번 모두 통과한 문항만 세면 첫 시도 기준보다 모든 모델의 통과율이 2~8%포인트 낮아졌고 순위는 같았다. 최종 보고서만 읽어서는 자금을 움직이기 전에 필요한 사실을 확인했는지 알 수 없다. 문의와 회신, 실제 지급이 기록에 어떤 순서로 남았는지 함께 봐야 한다.

연구 윤리와 자료

은행 실무자들이 부서별로 어떤 기록을 관리하고 서로 다른 자료를 어떻게 대조하는지 설명했고, 이를 문항 설계에 반영했다. 등장하는 기관, 거래 상대, 계좌, 인물은 모두 가상이며 고객 기록은 사용하지 않았다. 절차 설계에 참고한 공개 자료는 출처를 밝혔다. 메시지 발송과 지급은 모두 가상 은행 안에서 이루어졌다.

저자는 벤치마크와 원고 작성에 AI 도구를 사용하고 결과를 검토했다.

참고문헌

1. FinQA: A Dataset of Numerical Reasoning over Financial Data (2021).
2. Bank-World: A Benchmark for Evaluating Agents on Banking Workflows (2026).
3. τ-Bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains (2024).
4. Oracle. FLEXCUBE Enterprise Limits and Collateral Management User Guide.
5. A Coefficient of Agreement for Nominal Scales (1960).
6. Global BankOps. 연구 원고, 2026년 11월 23일.
7. 플리토. 8개 모델의 회신 확인, 지급 승인과 원장 기표. 2026.

부록 A. 업무 환경과 채점 검증

800회 모두에서 에이전트는 파일 조회 도구를 통해 워크시트 내용을 한 번 이상 받아 보았다.

표 A1. 50문항의 구성

항목	개수
분야별 문항 · 무역금융 / 여신 / 자금 / 회계 / 연금·신탁	10 / 10 / 10 / 10 / 10
부다페스트 현지법인 / 서울 본점	40 / 10
다른 기관과 SWIFT를 주고받는 문항	19
공유 폴더의 문서	605개(중앙값 12.5, 범위 3~18)
49문항의 구조화된 통합문서 / 워크시트	146 / 289
시작 시점 내부 메일 / 수신 SWIFT	132 / 66
내부 메일 회신 담당 부서 / 회신 선택 규칙	143 / 403
채점 요건	585개(중앙값 11.5, 범위 3~22)
참조 풀이 / 잘못된 풀이 변형	50 / 744
실행별 도구 호출 한도	42문항은 60회, 8문항은 40~90회

각 문항에는 올바른 처리 절차와 일부 단계를 잘못 처리하는 예시가 있다. 모두 실제로 도구를 호출해 실행할 수 있다. 올바른 절차 50개와 오류를 넣은 절차 744개(문항당 중앙값 14개)를 재현해 예상대로 통과하거나 실패하는지 검사했다. 오류 사례는 어느 기준에서 실패해야 하는지도 지정할 수 있다. 여기에 올바른 절차에서 파생한 대조 사례 276개, 채점기 점검 사례, 이전에 발견한 문제가 다시 생기지 않는지 확인하는 사례를 더했다.

모든 문항에서 문서끼리 모순이 없는지, 필요한 사실을 기록에서 찾거나 담당 부서에 물어볼 수 있는지, 주어진 권한으로 업무를 마칠 수 있는지 확인했다. 평가 후에는 모델의 실제 처리 기록도 검토했다. 같은 거래를 다른 이름으로 부르거나 한 번에 여러 가지를 물었을 때 정당한 요청을 놓친 경우, 문항을 수정하고 다시 검증했다. 서로 비슷하지만 다른 답을 줘야 할 요청도 검사했다. 수정 전 문항에서 얻은 결과는 집계에서 제외했다.

표 A2. 유형별 채점 요건

유형	검사 범위	개수
지급·원장	이체, 지급, 원장 순변동, 거래 참조 정보	124
작성·저장 결과	보고서, 메일, SWIFT, 저장 기록과 파일, 통지와 자금 예측	168
근거 확인·순서	행동 전에 담당 부서에 문의해 회신을 읽거나 근거가 되는 서명 기록을 읽었는지	64
행동 제한	허용한 메시지, 이체, 지급 및 SWIFT 발송 범위	180
업무 완료	담당자가 완료를 보고했는지	49

표 A3. 누가 무엇을 채점하는가

요소	검사 방법
코드 검사	금액, 통화, 계좌, 수신자, 기산일과 상태 변화를 검증하고 근거 열람과 행동 순서를 비교한다.
LLM 판정	메시지, 보고서, 저장 기록이 서술 기준을 충족했는지 판정한다. 순서 확인에 필요한 메시지의 의미를 판별한다.
근거 검증	대상 기록과 인용문을 대조하고 지정 수치를 확인한다.
실행 판정	모든 필수 기준을 충족하면 통과한다. 필수 판정이 누락되거나 유효하지 않으면 미채점으로 남긴다.

실행 재현과 채점자 간 일치도

2026년 11월 12일부터 23일까지 800회의 기록을 수집했다. 400개 모델·문항 조합에서 두 시도의 업무 지시, 가상 은행의 초기 상태, 채점 기준이 각각 같은지 해시값으로 확인했다. 난수 시드, 지시 방식, 추론 강도와 도구 호출 한도도 같았다. 이어 모델을 다시 호출하지 않고 기록에 남은 행동 23,561건을 같은 입력으로 재현했다. 800회 모두에서 처음 제공한 정보와 각 도구의 응답, 업무를 마친 뒤의 상태가 원래 기록과 일치했다.

문항을 확정한 뒤 저장된 판정으로 1,214개 검증 사례를 다시 확인했다. 직접 작성한 업무 절차 794개, 대조 사례 276개, 그 밖의 채점기 점검·회귀 사례 144개가 모두 예상한 결과를 냈다. 개발 중에는 이전 문항판에서 Opus 5.5가 문항당 한 번씩 수행한 45건을 두 채점 모델에 맡겼다. 채점 방식 v5에서 Gemini 3.8 Flash와 Claude Opus 5.5(max)의 판정을 비교해 기준 문구와 인용 범위를 다듬었다. 보완 뒤 두 모델이 다르게 판정한 세부 항목은 하나였고, 그 차이는 해당 업무의 최종 통과 여부에 영향을 주지 않았다.

표 A4. 이전 45문항판에서 두 채점 모델의 판정 일치도

점검 시점	기준 일치	Cohen’s κ	실행 판정 일치	Opus 미채점
초기 점검	144/146 (98.6%)	0.962	40/40	5
표현 보완 뒤	151/152 (99.3%)	0.982	45/45	0

별도의 개발 표본에서는 검토자가 세부 기준 35건을 직접 판정했다. 두 채점 모델의 판단이 같았던 30건과 달랐던 5건을 골랐다. 사람의 판정과 일치한 건수는 Flash가 31건, Opus가 28건이었다. 이렇게 고른 표본으로는 전체 문항의 채점 정확도를 추정할 수 없다.

부록 B. 분야별 결과와 회신 확인

표 B1는 각 모델이 부문별 10문항을 첫 시도에 얼마나 통과했는지 보여 준다. 50문항 중 3개는 8개 모델이 모두 첫 시도에 통과했고, 8개는 한 모델도 통과하지 못했다. 모델 쌍 28개의 첫 시도 통과율을 비교한 결과, Holm 보정 후 15개 쌍에서 5% 수준의 유의한 차이가 있었다. GPT-6.2 Astra와 GPT-6 Astra 사이, 그리고 각 Astra와 Fable 5.2 사이의 차이는 유의하지 않았다. 두 Astra 모델은 각각 그 밖의 모델들과 유의한 차이를 보였다.

표 B1. 50문항 및 분야별 10문항의 첫 시도 통과율. 마지막 두 열은 두 시도 합계

모델	추론	전체	TF	LS	TO	AC	PT	실행	통과
GPT-6.2 Astra	max	78.0%	70.0%	70.0%	80.0%	80.0%	90.0%	100	77
GPT-6 Astra	max	74.0%	60.0%	50.0%	90.0%	70.0%	100.0%	100	71
Fable 5.2	max	56.0%	40.0%	50.0%	70.0%	50.0%	70.0%	100	56
Gemini 4 Carbon	high	48.0%	40.0%	40.0%	50.0%	40.0%	70.0%	100	50
Opus 5.5	max	34.0%	20.0%	20.0%	50.0%	30.0%	50.0%	100	33
Gemini 4 Argon	high	28.0%	20.0%	0.0%	50.0%	10.0%	60.0%	100	27
Grok 4.7	xhigh	16.0%	10.0%	10.0%	30.0%	10.0%	20.0%	100	16
Qwen3.8 Max Prime	xhigh	14.0%	20.0%	0.0%	50.0%	0.0%	0.0%	100	12

TF 무역금융 · LS 여신 관리 · TO 자금·결제 · AC 회계 · PT 연금·신탁.

표 B2. 분야별 업무와 실패 유형

분야(각 10문항)	업무 범위	두 시도 전체의 실패 실행에서 관찰한 사항
무역금융	무역 서류, 지급, 은행 간 통신	104회 중 92회에서 확인 또는 순서 요건 누락
여신 관리	대출 기록, 조건, 여신 관리 판단	112회 중 111회에서 확인 또는 순서 요건 누락
자금·결제	대사, 외환·단기자금 결제, 지급 해제	64회 중 확인·순서 50회, 지급·원장 37회
회계	월말 발생액·재평가, 명세와 본점 보고	107회 중 확인·순서 95회, 보고·명세·저장 86회
연금·신탁	자격, 계좌 지시, 기록 갱신	71회 중 70회에서 보고, 메시지 또는 저장 요건 누락

표 B3. 첫 시도에 통과한 모델 수에 따른 문항 분포

분야	0	1	2	3	4	5	6	7	8	전체
무역금융	2	1	2	2	1	1	0	0	1	10
여신 관리	3	1	0	3	1	2	0	0	0	10
자금·결제	1	1	1	0	1	1	2	1	2	10
회계	2	1	1	1	2	3	0	0	0	10
연금·신탁	0	0	1	2	1	3	2	1	0	10
전체	8	4	5	8	6	10	4	2	3	50

첫 시도 통과 모델 수로 난이도를 나눴다. 2개 이하가 통과한 어려운 문항은 17개, 3~4개가 통과한 중간 난이도 문항은 14개, 5개 이상이 통과한 쉬운 문항은 19개였다.

표 B4. 필수 회신이 있는 44문항의 첫 시도 결과와 나머지 6문항의 통과 수

모델	통과 / 44	제때 확인	늦게 열람	미열람	요건에 맞는 요청 없음	다른 문항 통과 / 6
GPT-6.2 Astra	34	35	1	1	7	5
GPT-6 Astra	32	33	6	3	2	5
Fable 5.2	24	26	7	2	9	4
Gemini 4 Carbon	19	20	3	6	15	5
Opus 5.5	12	14	4	9	17	5
Gemini 4 Argon	10	12	2	9	21	4
Grok 4.7	4	7	2	6	29	4
Qwen3.8 Max Prime	4	8	5	13	18	3

가운데 네 열은 모델마다 합계 44다. 여러 요건을 놓친 실행은 요청 누락, 회신 미열람, 뒤늦은 열람 순으로 한 번만 센다.

첫 시도에서 Grok 4.7은 29문항, Qwen3.8 Max Prime은 18문항에서 필요한 내용을 묻는 요청을 하나 이상 빠뜨렸다. Qwen3.8 Max Prime은 다른 13문항에서는 회신을 읽지 않았다. 문의가 필수가 아닌 6문항에서는 모든 모델이 3문항 이상을 통과했다. 문의가 필수인 문항에서 제때 회신을 확인한 횟수는 GPT-6.2 Astra가 35회, Grok 4.7이 7회, Qwen3.8 Max Prime이 8회였다. 하위 두 모델을 제외하면 제때 확인한 횟수의 순위가 전체 통과율 순위와 같았다. 제때 확인한 첫 시도 155회 중 139회가 해당 문항을 통과했다. 두 Astra 모델은 각각 제때 확인한 문항 중 하나만 실패했다.

부록 C. 모델 실행 비용과 평가 언급

모델 실행 비용과 도구 사용

그림 1은 모델별 100회의 시도를 기준으로 통과율과 도구 사용량을 비교한다. 세로축은 100회 중 통과한 비율로, 각 회차 통과율의 평균과 같다. 가로축은 종료 동작을 포함해 환경이 실제로 처리한 도구 호출 횟수를 100으로 나눈 값이다. 모델이 실행할 수 있는 동작을 보내지 않은 응답은 호출 횟수에서 제외했다.

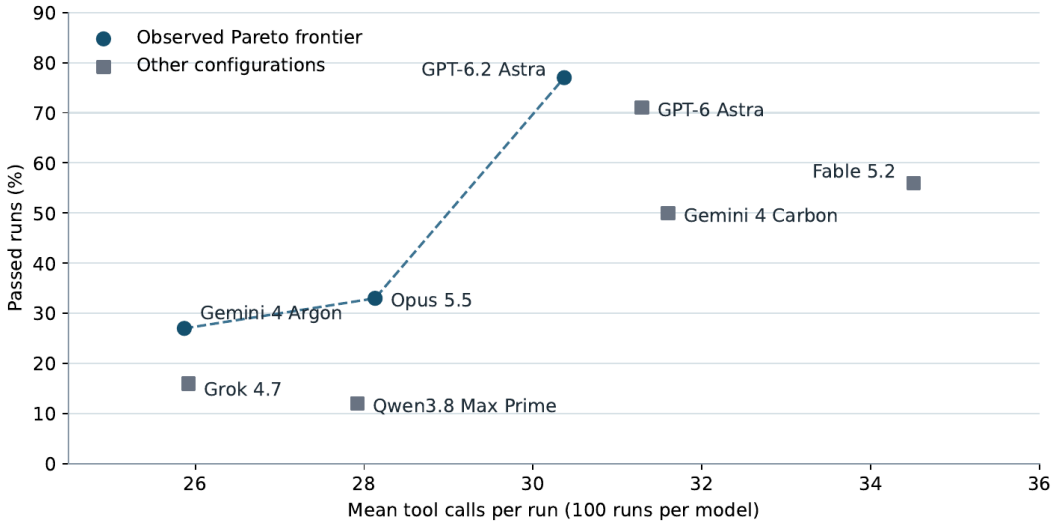

그림 1. 모델별 실행당 평균 도구 호출 수와 100회 실행의 통과율. 점선은 파레토 경계를 나타낸다.

그림의 점선은 도구 호출 수와 통과율을 함께 볼 때 다른 모델에 뒤지지 않은 세 모델을 잇는다. Gemini 4 Argon은 시도당 평균 25.87회 호출해 100회 중 27회를 통과했다. Opus 5.5는 28.13회 호출해 33회, GPT-6.2 Astra는 30.37회 호출해 77회를 통과했다. Grok 4.7은 평균 25.92회 호출해 16회를 통과했다.

표 C1. 모델별 100회 실행 비용(미국 달러)

모델	호출 경로(실행 수)	실행당 비용	합계
GPT-6.2 Astra	OpenAI direct (50), OpenRouter (50)	0.55	54.86
GPT-6 Astra	OpenRouter (96), OpenAI direct (4)	1.07	106.62
Fable 5.2	OpenRouter (62), Anthropic direct (38)	4.81	480.75
Gemini 4 Carbon	OpenRouter (100)	0.68	67.71
Opus 5.5	OpenRouter (94), Anthropic direct (6)	1.70	169.88
Gemini 4 Argon	OpenRouter (100)	0.41	40.87
Grok 4.7	OpenRouter (100)	0.41	40.71
Qwen3.8 Max Prime	OpenRouter (100)	0.71	70.82

표 C1의 비용은 평가 대상 모델이 800회 동안 생성한 응답을 기준으로 한다. 실행할 동작 없이 끝난 마지막 응답도 과금됐다면 포함했다. OpenRouter를 거친 702회의 청구액은 687.95달러였다. API를 직접 호출한 98회는 기록된 토큰 사용량에 공시 단가를 적용해 344.28달러로 추산했다. 합계는 1,032.23달러다. 같은 모델도 서비스 경로와 프롬프트 캐시 사용 여부에 따라 비용이 달라질 수 있다.

평가를 언급한 수행 기록

채점을 마친 뒤 800회의 도구 호출 기록 중 서비스가 제공한 추론 내용에서 채점, 출제자, 기대 답안, 평가를 받고 있다는 언급을 찾았다. 검색된 643건을 모두 읽고 자금 요청 심사처럼 은행 업무를 가리

킨 13건을 제외했다. 이 분류는 채점이나 시도 선택에 사용하지 않았다. 모델이 채점 방식을 이용하려 한 정황이 보이면 실제로 어떤 행동을 했고 그 행동이 어떤 판정을 받았는지 대조했다.

표 C2은 평가를 명시적으로 언급한 시도를 센 결과다. 별도의 71회에서는 평가를 직접 언급하지 않고 주어진 환경을 시뮬레이션이라고 표현했다. 모델에게 준 지시와 문항 자료에는 이 단어가 없었다.

표 C2. 평가를 언급한 실행 수와 추론 기록이 공개된 도구 호출 비율

모델	평가를 언급한 실행 / 100회	시뮬레이션만	추론 기록이 공개된 호출 비율
GPT-6.2 Astra	4	10	19.3%
GPT-6 Astra	0	0	13.8%
Fable 5.2	42	5	62.4%
Gemini 4 Carbon	10	17	51.3%
Opus 5.5	18	2	62.2%
Gemini 4 Argon	10	4	38.1%
Grok 4.7	9	15	63.7%
Qwen3.8 Max Prime	36	18	100.0%

검토한 기록에서는 모델이 채점기의 허점을 이용해 잘못 처리한 업무를 통과한 사례를 발견하지 못했다.

이 수치는 서비스가 추론 내용을 얼마나 공개하는지에 따라서도 달라진다. 도구를 호출한 응답 중 추론 기록을 볼 수 있었던 비율은 Qwen3.8 Max Prime이 100%, GPT-6 Astra가 13.8%였다. 문서 형식이 이런 언급을 유도했는지, 통과율에 어떤 영향을 줬는지는 검증하지 않았다.