은행 업무에서 에이전트는 확인한 근거에 따라 행동하는가?

Global BankOps 테크니컬 리포트 · 2026-11-23

원 연구 · 8개 모델 설정 · 50문항 · 800회 실행

플리토 에이전트 데이터 리서처 팀

업무 환경과 문항 설계

은행 담당자는 약정서·원장·부서 회신이 다를 때 어떤 자료를 기준으로 삼을지 정하고, 지급이나 기표에 그 근거를 반영해야 합니다. Global BankOps는 가상 은행에서 기록을 찾고 담당 부서에 확인하는 과정부터 지급·기표·보고까지 추적합니다. 2026년 11월 12~23일 원 연구에서 8개 모델 설정이 영어 50문항을 각각 두 번 수행했습니다.

업무 분야와 평가 범위 · 각 10문항

분야	담당자가 수행하는 업무
무역금융	신용장 조건과 지급 근거를 확인하고 은행 간 청구와 통지를 처리합니다.
여신 관리	약정과 대출 기록을 대조해 상환, 이자, 한도와 약정 이행을 판단합니다.
자금·결제	거래명세와 계좌 내역을 대사하고 외환·단기자금 결제를 처리합니다.
회계	결산일의 발생액·평가액을 기표하고 보조부와 본점 보고 수치를 대사합니다.
연금·신탁	가입 자격과 지시 권한을 확인한 뒤 부담금, 지급 및 신탁 기록을 처리합니다.

40문항은 부다페스트 현지법인, 10문항은 서울 본점 업무입니다. 31문항은 은행 내부에서 처리하고 19문항은 다른 기관과 SWIFT 메시지를 주고받습니다. 공유 문서 605개, 시작 시점의 내부 메일 132개와 SWIFT 수신 메시지 66개를 구성했습니다. 모든 문항이 스프레드시트를 제공했고 800회 실행 모두에서 도구가 시트 데이터를 반환했습니다.

실무 검토와 파일럿

실무자 인터뷰와 제공 자료, 공개 금융 절차 문서를 토대로 역할별 접근 권한과 부서 간 확인 절차를 구성했습니다. 실무자가 업무 분담과 보고 흐름을 확인했습니다. 제작팀은 자료끼리 일치하는지, 필요한 사실을 열람하거나 담당자에게 물어 확인할 수 있는지 검토했습니다. 문항을 고를 때 파일럿 실행에서 관찰한 난이도를 참고했습니다. 업무 환경을 수정한 문항의 기존 실행은 집계에서 제외했습니다.

채점, 검증과 실행 재현

혼합 채점기 v6은 금액, 통화, 계정, 상대방, 날짜와 원장 상태 변화를 코드로 검사합니다. Gemini 3.8 Flash는 온도 0·낮은 추론 강도에서 보고서와 메시지가 서술형 요건을 충족했는지 판단하고 근거 문장을 인용합니다. 채점기는 인용문과 필수 수치를 원문에서 검증하고, 회신을 읽은 시점과 지급·기표·보고 시점을 대조합니다. 필수 요건을 모두 충족하면 통과합니다. 필수 판정의 근거가 없거나 유효하지 않으면 미채점으로 남깁니다.

50문항의 채점 요건 585개

요건	개수
근거 확인과 처리 순서	64
문서 작성과 결과 저장	168
지급과 원장 처리	124
허용된 행동 범위	180
업무 완료	49

585개 중 223개 요건은 서술형 판단을 포함합니다. 동결한 문항판에서 저장된 판정을 재생해 1,214개 검증 사례가 모두 기대 결과를 냈습니다. 참조 풀이 50개, 잘못된 풀이 변형 744개, 자동 생성 대조 사례 276개, 판정·회귀 검사 144개입니다.

이전 판본에서의 채점자 일치도

이전 45문항판의 한 모델 실행으로 v5 채점기를 점검했습니다. 기준 설명을 보완한 뒤 두 채점 모델은 152개 기준 중 151개에서 일치했고(Cohen’s κ 0.982), 45개 최종 통과·실패 판정이 모두 일치했습니다. 별도 선별 표본 35개에서는 사람 판정과 Flash가 31개, Opus가 28개 일치했습니다. 현행 50문항판 전체에 대한 독립적 사람 채점 정확도는 측정하지 않았습니다.

두 실행의 선택과 재생

모델·문항별로 시작 시각 순서상 처음 두 유효 채점 실행을 선택하고, 시각이 같을 때는 실행 식별자로 정했습니다. 제공자 오류 1회, 시간 초과 1회와 더 늦은 채점 실행 4회를 제외했습니다. 필요한 재판정은 같은 스냅샷에 적용하고 최초 판정을 보존했습니다. 400개 모델·문항 조합마다 두 실행의 지시문, 환경과 채점표 해시가 같았습니다. 기록된 행동 23,561개를 다시 실행했을 때 800회 모두의 최초 관찰, 도구 응답과 최종 상태가 저장 기록과 일치했습니다. 이 재생에는 모델을 다시 호출하지 않았습니다.

두 번의 시도에서 통과한 업무

pass@2는 각 문항을 두 번 수행했을 때 한 번 이상 통과한 문항의 비율입니다. 모델별 분모는 50문항, 분야별 분모는 10문항입니다. GPT-6.2 Astra는 42문항을 한 번 이상 통과했고 그중 35문항을 두 번 모두 통과했습니다. 전체 400개 모델·문항 조합 중 40개에서 두 실행의 결과가 달랐습니다.

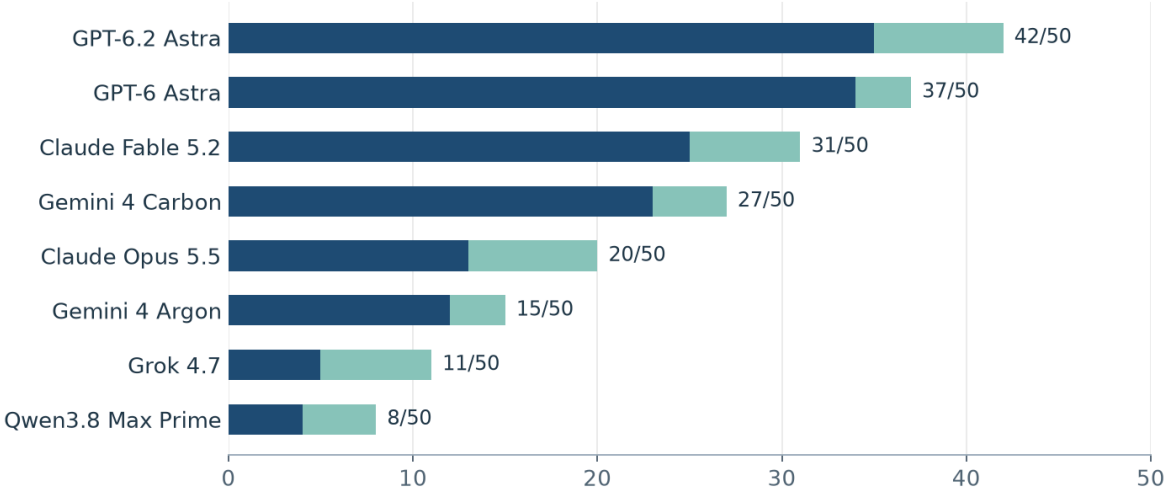


그림 1. 진한색은 두 번 모두 통과한 문항, 연한색은 한 번만 통과한 문항입니다. 막대 끝의 값은 한 번 이상 통과한 문항 수입니다.

원 연구 8개 모델 · 모델마다 50문항을 두 번 평가

모델	추론 강도	pass@2	두 번 모두 통과	한 번만 통과
GPT-6.2 Astra	max	42 / 50 (84.0%)	35	7
GPT-6 Astra	max	37 / 50 (74.0%)	34	3
Claude Fable 5.2	max	31 / 50 (62.0%)	25	6
Gemini 4 Carbon	high	27 / 50 (54.0%)	23	4
Claude Opus 5.5	max	20 / 50 (40.0%)	13	7
Gemini 4 Argon	high	15 / 50 (30.0%)	12	3
Grok 4.7	xhigh	11 / 50 (22.0%)	5	6
Qwen3.8 Max Prime	xhigh	8 / 50 (16.0%)	4	4

분야별 성과와 업무 난이도

분야별 pass@2 · 각각 10문항 중 한 번 이상 통과한 문항 수

모델	무역금융	여신 관리	자금·결제	회계	연금·신탁
GPT-6.2 Astra	7	8	9	8	10
GPT-6 Astra	6	5	9	7	10
Claude Fable 5.2	5	5	8	6	7
Gemini 4 Carbon	4	4	6	5	8
Claude Opus 5.5	3	3	5	3	6
Gemini 4 Argon	2	1	5	1	6
Grok 4.7	1	2	5	1	2
Qwen3.8 Max Prime	2	0	6	0	0

GPT-6 Astra는 연금·신탁 10문항을 두 번 모두 통과했습니다. 같은 모델은 자금·결제에서는 9문항을 한 번 이상 통과했습니다. 분야마다 10문항이므로 한 문항이 분야 점수의 10%포인트를 차지합니다.

두 회차의 실패 실행 중 근거 확인·업무 순서 검사에 실패한 수

분야	확인·순서 실패 / 해당 분야 실패 실행
무역금융	92 / 104 (88.5%)
여신 관리	111 / 112 (99.1%)
자금·결제	50 / 64 (78.1%)
회계	95 / 107 (88.8%)
연금·신탁	65 / 71 (91.5%)

원 연구는 첫 번째 시도에서 통과한 모델 수를 기준으로 문항의 난이도도 조사했습니다. 8개 모델 중 0~2개가 통과한 문항은 17개, 3~4개는 14개, 5개 이상은 19개였습니다. 첫 시도에서 어느 모델도 통과하지 못한 문항이 8개였고, 8개 모델 모두 통과한 문항이 3개였습니다. 28개 모델 쌍 비교 중 15개가 Holm 보정 후 5% 수준에서 유의했습니다. 이 분석은 첫 시도 결과를 사용했으므로 pass@2 순위 차이에 대한 유의성 검정으로 해석할 수 없습니다.

회신 확인, 지급 승인과 원장 기표

원 연구의 800회 실행 중 342회가 통과하고 458회가 실패했습니다. 실패 실행 중 근거 확인·처리 순서 검사는 413회, 작성·저장 결과는 351회, 지급·원장은 160회, 행동 제한은 114회, 업무 완료는 8회에서 실패했습니다. 한 실행에 여러 유형이 겹칩니다. 아래 표는 각 절차가 실제로 적용된 실행을 분모로 사용합니다.

업무 절차별 적용 분모 · 8개 모델, 두 회차

점검 대상	관측 결과 / 적용 대상
필요한 회신을 제때 확인하지 못한 실행	401 / 704 (57.0%)
승인 회신을 읽기 전에 대상 지급을 실행	0 / 31 (0.0%)
승인 절차 대상이지만 대상 지급 기록이 없는 실행	1 / 32 (3.1%)
기표 또는 원장 순변동 요건을 어긴 실행	103 / 560 (18.4%)

회신을 언제 열었는가?

44문항은 담당 부서·거래은행에 요청해 처리 전에 회신을 읽어야 했습니다. 별도 3문항에서는 실행 전에 읽은 최신 시스템 기록도 인정했고, 나머지 3문항에는 필수 회신 요건이 없었습니다. 회신을 제때 모두 확인한 실행은 704회 중 303회였습니다. 적합한 요청을 찾지 못한 247회 중 29회는 적격 수신자에게 연락한 흔적이 있었고 218회는 없었습니다. 단순히 연락했다는 사실만으로 필요한 정보를 물었는지 판단할 수 없습니다.



그림 2. A 제때 확인 303회 · B 필요한 수신자와 요청 조건을 충족한 기록 없음 247회 · C 회신 미열람 92회 · D 처리 후 열람 62회. 한 실행에 여러 누락이 있으면 B, C, D 순으로 한 번만 셌습니다.

지급 승인 검사는 대상 지급을 실행한 31회에서 지정 승인 회신을 먼저 읽었는지 확인합니다. 이 중 조기 지급은 0회였습니다. 승인 요건이 적용된 32회 중 1회에는 대상 지급 기록이 없었습니다. 기표·원장 검사는 필요한 기표와 계정별 순변동, 금지된 반영을 확인합니다. 적용 560회 중 103회가 한 가지 이상 어겼습니다.

실행 조건과 관측된 자원 사용량

각 조합의 두 실행은 동일한 역할 지시, 시드 0, 추론 강도와 도구 호출 한도를 사용했습니다. 기본 제한 시간은 1,800초였고, 일부 실행은 더 긴 제한 시간을 적용했습니다. 42문항은 60회, 나머지는 40~90회의 도구 호출을 허용했습니다. 실행 경로, 캐시, 추론·응답 예산이 달라 자원 수치에는 모델과 환경의 차이가 함께 반영됩니다.

모델별 도구 호출과 비용 · 각 100회 실행

모델	평균 도구 호출 / 실행	합계 비용 (USD)
GPT-6.2 Astra	30.37	54.86
GPT-6 Astra	31.29	106.62
Claude Fable 5.2	34.51	480.75
Gemini 4 Carbon	31.60	67.71
Claude Opus 5.5	28.13	169.88
Gemini 4 Argon	25.87	40.87
Grok 4.7	25.92	40.71
Qwen3.8 Max Prime	27.92	70.82

800회 실행의 모델 호출 비용은 1,032.23달러입니다. OpenRouter 경유 702회의 제공자 집계액은 687.95달러였고, 직접 API 경유 98회의 기록 토큰과 공시 단가로 추정한 금액은 344.28달러였습니다. 실행 가능한 동작을 내지 않은 마지막 응답의 청구액도 포함했습니다. 채점자 호출 비용은 이 표에 포함되지 않습니다.

트레젝토리에서 관찰한 평가 인지 표현

채점 후 연구진은 도구 호출 시점에 제공자가 공개한 추론 기록에서 평가 관련 표현 후보 643건을 읽고 일상 은행 업무를 가리킨 13건을 제외했습니다. 평가를 명시적으로 언급한 실행은 129회였고, 시뮬레이션만 언급한 실행은 별도로 71회였습니다. 읽을 수 있는 추론을 제공한 도구 호출 비율은 모델별 13.8~100%였습니다. 이 주석은 실행 선택과 점수에 영향을 주지 않았습니다. 검토한 트레젝토리에서는 채점 공백을 이용해 틀린 업무가 통과한 사례를 발견하지 못했습니다.

해석의 한계와 연구 출처

최종 보고서는 자금을 이동하기 전에 근거를 확인했는지 보여주지 못합니다. 458개 실패 실행 중 413개에서 근거 확인·처리 순서 문제가 관찰됐고, 승인과 원장 분석은 문제 발생 지점을 더 좁혀줍니다. 은행 운영 자동화의 실제 성공률을 추정하려면 다른 업무와 실행 환경의 검증이 필요합니다.

50문항은 파일럿을 거쳐 선별했고 일부 승인·접근 권한을 단순화했습니다. 모의 부서는 조건에 맞는 요청에 정해진 시간 뒤 정확히 회신합니다. 실무에서는 회신이 늦거나 불완전할 수 있고 확인에 비용이 듭니다. 반대로 타당한 표현도 모의 부서의 규칙과 일치하지 않을 수 있습니다. 두 번의 시도는 반복 수행의 제한된 표본입니다. 도구 재생은 환경 재현성을 검사하며, 새로 생성될 모델 응답이나 전체 사람 채점 정확도를 보장하지 않습니다.

윤리와 연구 자료

실무자가 자료 분담과 대사 절차를 설명했고 연구진이 가상 기관·거래상대방·계좌·사람을 구성했습니다. 고객 원자료는 사용하지 않았으며 모든 메시지와 지급은 시뮬레이터 안에서 처리됐습니다. 원 연구자는 벤치마크와 원고 작성에 AI 도구를 사용하고 결과를 검토했습니다.

판본과 관련 연구

이 공개판은 2026-11-23 Global BankOps 원 연구를 바탕으로 업무 분야·평가 방법·결과를 집계 수준에서 재구성했습니다. 원고 리비전 6db9d54f98a6, 평가 문항판 v1.0.0, 채점 v6, 원 연구 공개 스냅샷 bankopsbench-20261123-67b3aa6693fe입니다. 현재 공개 순위는 pass@2를 사용합니다. 11월 13일 45문항판과 문항 구성이 다릅니다.

FinQA는 금융 보고서의 수치 추론, Bank-World는 모의 은행 업무, τ-Bench는 정책에 따라 행동한 뒤의 시스템 상태를 평가합니다. Global BankOps는 근거를 확인한 순서와 지급·원장·업무 보고가 일치하는지 함께 봅니다.

FinQA · Bank-World · τ-Bench