기술보고서

은행 업무 50개에서 비교한 AI 모델 14개의 통과율

Flitto Agent Bench (Finance – Global BankOps)

발행 · Flitto Agent Data Researcher Team

2026년 11월 24일

데이터 버전 2026-11-23 · 평가 기준일 2026-11-23 · 문항 세트 26.11.23

초록

14개 모델이 가상 은행 업무 50개를 두 번씩 수행했습니다. 총 1,400회 중 425회가 모든 필수 조건을 충족했고, 975회는 실패했습니다. GPT-6.2 Astra가 50문항 중 42문항을 한 번 이상 통과해 가장 높은 통과율(84.0%)을 기록했습니다.

평가 과제와 채점 방법

에이전트는 가상 은행의 담당자가 되어 계약서와 원장, 내부 메일과 은행 간 SWIFT 전문을 대조합니다. 업무에 따라 담당 부서의 회신을 확인한 뒤 대금을 지급하거나 거래를 보류하고, 처리 결과를 원장이나 보고서에 남깁니다. 무역금융, 여신 관리, 자금 운영, 회계, 연금·신탁의 다섯 분야에 각각 10문항을 배정했습니다.

한 번의 실행이 통과하려면 해당 업무의 필수 조건을 모두 충족해야 합니다. 채점기는 거래 금액과 계정, 처리 후 원장 잔액을 확인합니다. 사전 승인이 필요하면 회신을 읽은 시점과 지급 시점도 대조합니다. 보고서와 메시지는 별도의 언어 모델이 근거 문장을 인용해 채점하고, 채점기는 그 문장이 실제 기록에 있는지 확인합니다.

리더보드에서는 유효한 두 실행 중 한 번이라도 통과한 문항 수를 전체 50문항으로 나눈 값(pass@2)으로 순위를 정합니다. 유효한 판정을 받지 못한 실행과 두 회차를 확정한 뒤의 추가 실행은 집계에서 제외합니다.

모델별 통과율

같은 모델이 같은 문항을 두 번 모두 통과했는지도 함께 봤습니다. 전체 700쌍 중 181쌍은 두 번 모두 통과했고, 63쌍은 한 번만 통과했습니다. 나머지 456쌍은 두 번 모두 실패했습니다.

표 1. 모델별 통과율과 통과 문항 수

모델	추론 강도	pass@2	한 번 이상 통과	두 번 모두 통과
GPT-6.2 Astra	max	84.0%	42/50	35/50
GPT-6 Astra	max	74.0%	37/50	34/50
Claude Fable 5.2	max	62.0%	31/50	25/50
Gemini 4 Carbon	high	54.0%	27/50	23/50
Claude Opus 5.5	max	40.0%	20/50	13/50
Gemini 4 Argon	high	30.0%	15/50	12/50
Qwen4 Max	xhigh	28.0%	14/50	9/50
Kimi K3.1 Pro	max	24.0%	12/50	7/50
Grok 4.7	xhigh	22.0%	11/50	5/50
GLM-5.3 Prime	max	20.0%	10/50	7/50
Hunyuan 4	high	18.0%	9/50	3/50
Qwen3.8 Max Prime	xhigh	16.0%	8/50	4/50
Muse Spark 1.3	max	14.0%	7/50	4/50
Solar Pro 4	max	2.0%	1/50	0/50

업무 분야별 결과

분야별로 각 모델이 같은 문항을 수행한 두 실행을 한 쌍으로 묶었습니다. 예를 들어 한 모델이 무역금융 1번 문항을 두 번 모두 통과해도 무역금융의 통과 건수는 하나만 늘어납니다. 표 2의 비율은 해당 분야의 전체 쌍을 기준으로 계산했습니다.

표 2. 업무 분야별로 한 번 이상 통과한 비율

업무 분야	한 번 이상 통과 / 전체	통과율
무역금융	40/140	28.6%
여신 관리	38/140	27.1%
자금 운영	76/140	54.3%
회계	38/140	27.1%
연금·신탁	52/140	37.1%

회신 확인과 지급 승인

업무를 처리하기 전에 회신을 확인해야 했던 1,232회 중 856회에서 모델은 필요한 회신을 모두 확인하지 못했습니다. 사전 승인이 필요한 56회 중 실제 지급이 있었던 실행은 51회였고, 그중 2회는 승인 회신을 읽기 전에 대금을 지급했습니다. 나머지 5회에는 해당 거래의 지급 기록이 없었습니다.

두 지표는 각각 해당 절차가 필요한 실행을 대상으로 합니다. 한 실행이 두 지표에 모두 포함될 수 있습니다.

연구 자료

이 보고서는 현재 리더보드의 14개 모델을 다룹니다. 11월 23일 연구에서 발표한 8개 모델의 800회 평가와 문항 설계, 채점 검증 절차는 원 연구 보고서에서 확인할 수 있습니다.