Technical report

How 14 AI models performed on 50 banking tasks

Flitto Agent Bench (Finance – Global BankOps)

Published by Flitto Agent Data Researcher Team

24 November 2026

Data version 2026-11-23 · Evaluation date 2026-11-23 · Task-set version 26.11.23

Abstract

14 models each attempted 50 simulated banking tasks twice. Of 1,400 scored trials, 425 passed every requirement and 975 failed. GPT-6.2 Astra passed 42 of 50 tasks at least once, leading the leaderboard at 84.0%.

Tasks and grading

The agent works as a bank officer, consulting contracts, ledger entries, internal mail and interbank SWIFT messages. Depending on the task, the officer must obtain confirmation, release or hold a payment, and record the outcome in the ledger or a report. Each of the five divisions contributes ten tasks: Trade Finance, Loan Servicing, Treasury Operations, Accounting, Pension & Trust.

A trial passes only when the agent meets every required criterion for that task. The evaluator checks amounts, accounts and resulting ledger balances. When a payment requires prior approval, it compares the reply-read time with the payment time. A separate language model cites passages from reports and messages to grade written responses. The evaluator checks those citations against the records.

The leaderboard ranks models by pass@2, the share of the 50 tasks passed in at least one of the two eligible trials. Runs without a valid grade and additional runs beyond the selected pair are excluded.

Results by model

The last column shows how often a model succeeded on both attempts. Of the 700 model–task pairs, 181 passed twice, 63 passed once and 456 failed twice.

Table 1. Pass rates and task counts by model

Model	Effort	pass@2	Passed at least once	Passed twice
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

Results by division

Each model contributes one trial pair per task in each division. A model that passes the same trade finance task twice adds one to the division's count of pairs passed. Each percentage in Table 2 uses all model–task pairs in that division as its denominator.

Table 2. Tasks passed at least once by division

Division	Passed at least once / total	Pass rate
Trade Finance	40/140	28.6%
Loan Servicing	38/140	27.1%
Treasury Operations	76/140	54.3%
Accounting	38/140	27.1%
Pension & Trust	52/140	37.1%

Reply checks and payment approvals

In 856 of 1,232 trials requiring prior confirmation, the agent did not read all required replies before acting. Of 56 trials requiring payment approval, 51 recorded the payment. In 2 of those, the agent paid before reading the approval reply. The other 5 trials contained no matching payment.

Each measure uses the trials to which the relevant control applied. A trial can contribute to both measures.

Source study

This edition covers the 14 models on the current leaderboard. The 23 November study documents the original eight-model, 800-trial evaluation, task design and validation of the grading procedure.