Do agents verify the evidence before acting in banking operations?

Global BankOps Technical Report · 2026-11-23

Original study · 8 configurations · 50 tasks · 800 runs

Flitto Agent Data Researcher Team

Work environment and task design

Bank officers resolve discrepancies across agreements, ledgers and departmental replies before releasing funds, posting entries or reporting a decision. Global BankOps records how an officer obtains evidence, acts on it and reports the outcome in a simulated bank. From 12 to 23 November 2026, eight model configurations attempted 50 English tasks twice each.

Work covered by division · ten tasks each

Division	Officer responsibilities
Trade finance	Check letter-of-credit terms and payment authority, then handle interbank claims and notices.
Loan servicing	Reconcile agreements and loan records to handle repayments, interest, limits and covenants.
Treasury operations	Reconcile statements and accounts, then settle FX and money-market transactions.
Accounting	Post accruals and valuation adjustments and reconcile schedules with head-office returns.
Pensions and trusts	Verify eligibility and the authority to instruct before allocating contributions, paying benefits or updating trust records.

Forty tasks take place at the Budapest subsidiary and ten at Seoul head office. Thirty-one stay within the bank and 19 involve SWIFT exchanges with another institution. The environments contain 605 shared documents, 132 initial mailbox messages and 66 initial incoming SWIFT messages. All tasks supply spreadsheets. The file tool returned worksheet data in all 800 runs.

Practitioner review and pilot selection

Practitioner interviews, practitioner-supplied materials and published procedures informed record access and cross-team verification. Practitioners checked selected responsibilities and reporting workflows. The team reviewed whether records agreed and whether officers could find needed facts or ask their owners. Difficulty observed in pilot runs informed task selection. After an environment change, runs on the earlier task were excluded.

Grading, validation and replay

Hybrid scoring v6 checks amounts, currencies, accounts, recipients, dates and ledger changes in code. Gemini 3.8 Flash applies written criteria at temperature zero and low reasoning effort, citing eligible messages or reports. The evaluator verifies citations and required figures and compares evidence-read and action steps. A run passes when it meets every required criterion. A missing or invalid required judgment leaves it unscored.

585 rubric checks across 50 tasks

Check group	Count
Verification and action order	64
Written or saved outputs	168
Payments and ledger entries	124
Action restrictions	180
Completion	49

Of the 585 checks, 223 include a written requirement judged by the model. Using cached judgments, the frozen release reproduced all 1,214 expected validation outcomes. These comprised 50 reference solutions, 744 incorrect variants, 276 generated controls and 144 additional fixtures and regression cases.

Judge agreement on the earlier edition

A development audit used v5 grading and one solver’s 45 saved runs on the earlier edition. After clarification, the two judges agreed on 151 of 152 criteria (Cohen’s κ 0.982) and all 45 run verdicts. On a separate selected sample of 35 human judgments, Flash matched 31 and Opus matched 28. The study did not measure independent human agreement across the full 50-task v6 release.

Trial selection and replay

For each model and task, the study selected the first two eligible scored runs by start time, breaking ties by run identifier. It excluded one provider error, one timeout and four later scored runs. Certified reassessments preserved original snapshots and grades. Both trials shared task, world and rubric hashes in all 400 model–task pairs. Replaying 23,561 recorded actions reproduced all initial observations, tool responses and final states across 800 runs without calling the models again.

Task completion over two trials

Pass@2 is the share of tasks passed at least once in two trials. The denominator is 50 tasks per configuration, or ten per division. GPT-6.2 Astra passed 42 tasks at least once and 35 on both trials. Across all 400 model–task pairs, the two outcomes differed on 40.

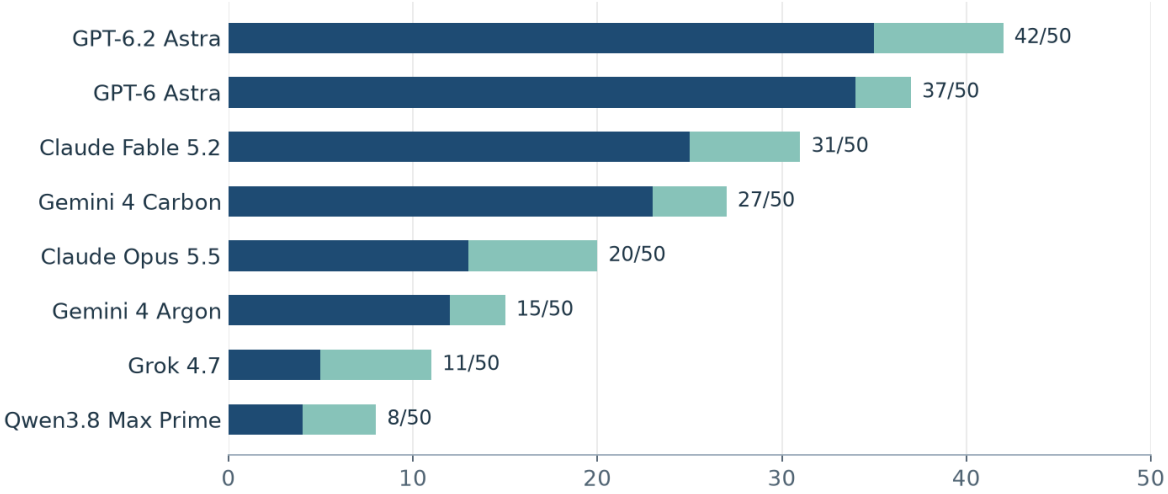

Figure 1. The dark segment counts tasks passed twice. The light segment counts tasks passed once. Each end label gives the total passed at least once.

Eight original configurations · 50 tasks and two trials each

Configuration	Effort	pass@2	Passed twice	Passed once
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

Results by division and task difficulty

Pass@2 by division · tasks passed at least once out of ten

Configuration	Trade finance	Loan servicing	Treasury operations	Accounting	Pensions and trusts
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 Astra passed all ten pension and trust tasks on both trials. It passed nine treasury tasks at least once. With ten tasks in each division, one task changes the division score by ten percentage points.

Failed runs with a verification or action-order error, both trials

Division	Verification or order errors / failed runs in division
Trade finance	92 / 104 (88.5%)
Loan servicing	111 / 112 (99.1%)
Treasury operations	50 / 64 (78.1%)
Accounting	95 / 107 (88.8%)
Pensions and trusts	65 / 71 (91.5%)

The original study also counted how many configurations passed each task on its first trial. Seventeen tasks were passed by zero to two configurations, 14 by three to four and 19 by five or more. Eight tasks were passed by none and three by all eight. Of 28 first-trial pairwise comparisons, 15 were significant at the 5% level after Holm correction. That test concerns first trials and does not establish significance for pass@2 rankings.

Confirmations, payment authority and ledger postings

Of the 800 original runs, 342 passed and 458 failed. Among the failures, 413 missed verification or action order, 351 missed written or saved outputs, 160 missed payment or ledger checks, 114 violated action restrictions and eight missed completion. These groups overlap. The table below uses the runs to which each banking control applied.

Applicable denominators by control · eight models, both trials

Banking control	Observed / applicable
Runs missing at least one timely confirmation	401 / 704 (57.0%)
Target payments released before reading the approval reply	0 / 31 (0.0%)
Runs under an approval requirement with no target payment recorded	1 / 32 (3.1%)
Runs failing a posting or net-ledger check	103 / 560 (18.4%)

When did the officer read the reply?

Forty-four tasks required a request to the designated team or correspondent and a reply read before acting. Three other tasks also accepted a current system record, and three had no mandatory reply check. All required replies were read in time in 303 of 704 applicable runs. Of the 247 runs with an unmatched request, 29 recorded contact with an eligible recipient for at least one unmatched check and 218 did not. Contact alone does not establish that the officer requested the governing fact.

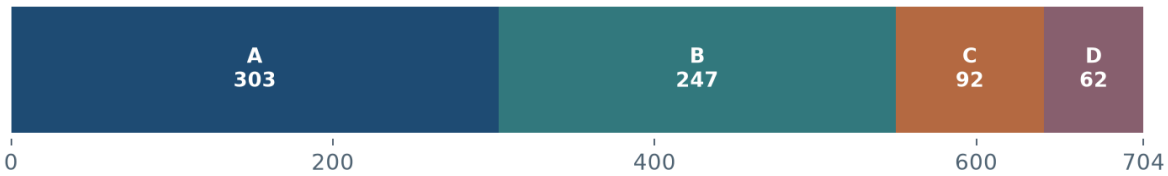

Figure 2. A, all replies in time (303). B, no matching request for at least one required confirmation (247). C, reply never opened (92). D, reply opened after acting (62). Each failed run counts once, prioritizing B, then C, then D.

The payment-authority check asks whether the officer read the designated approval before releasing a target payment. None of the 31 releases preceded approval. In one of 32 applicable runs, no target payment was recorded. Ledger checks examine required postings, net account movements and prohibited entries. At least one failed in 103 of 560 applicable runs.

Execution settings and observed resource use

Within each model–task pair, both trials used the same workplace framing, seed zero, reasoning effort and tool-call limit. The default timeout was 1,800 seconds, with longer limits for some runs. Forty-two tasks allowed 60 calls and the others allowed 40 to 90. Provider routing, caches and reasoning and response budgets varied across the study.

Tool calls and solver costs · 100 scored runs per configuration

Configuration	Mean calls / run	Total cost (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

Total solver cost for the 800 runs was US$1,032.23. OpenRouter reported US$687.95 for 702 runs. For 98 direct API runs, recorded token usage and list prices gave an estimated US$344.28. Costs include billed terminal responses with no executable action. The table excludes judge costs.

Evaluation references in recorded reasoning

After scoring, reviewers read 643 candidate mentions in provider-visible reasoning on tool-calling turns and set aside 13 about ordinary bank work. A total of 129 runs explicitly mentioned evaluation. A further 71 mentioned only simulation. Readable reasoning covered 13.8% to 100% of tool-calling turns, depending on the model. These annotations played no role in selection or scoring. The reviewed traces showed no incorrect work passing through exploitation of a verifier gap.

Limitations and study provenance

A final report alone cannot show whether an officer checked the governing evidence before moving funds. Verification or action-order errors appeared in 413 of 458 failed runs. Payment and ledger checks identify the operation affected. Estimating production performance would require additional workflows and environments.

The 50 pilot-selected tasks simplify some approvals and access controls. Simulated teams answer matching requests accurately after a fixed delay. In practice, requests can cost time and produce incomplete information. Conversely, a valid paraphrase can miss a simulated reply rule. Two trials provide limited evidence about repeatability. Replay establishes consistency for recorded actions. It cannot reproduce fresh model generations or establish full-suite agreement with human graders.

Ethics and research materials

Practitioners described how teams divide records and reconcile information. Institutions, counterparties, accounts and people in the benchmark are fictional. The materials contain no customer records, and every message and payment remained inside the simulator. The study author used AI tools in preparing the benchmark and manuscript and reviewed the resulting work.

Edition and related work

This public edition presents the aggregate methods and results of the 23 November 2026 Global BankOps study. Manuscript revision 6db9d54f98a6, task release v1.0.0, scoring v6, original public snapshot bankopsbench-20261123-67b3aa6693fe. The public ranking uses pass@2. The task mix differs from the 45-task edition of 13 November.

FinQA studies numerical reasoning over financial reports, Bank-World studies simulated banking workflows, and τ-Bench checks the resulting system state under a domain policy. Global BankOps examines the order of evidence checks alongside payment, ledger and reporting outcomes.

FinQA · Bank-World · τ-Bench