Global BankOps

Flitto Agent Bench · Finance / Global · Technical report

Flitto Agent Data Researcher Team

Data version 2026-11-13 · Results as of 2026-11-13

Task-set version 1.0.0 · Analysis / report revised 2026-11-14

Abstract

Global BankOps evaluates whether agents reconcile evidence, act on it, and accurately report the outcome in a synthetic bank environment. Six model configurations cover 45 tasks across four task categories. Using one eligible execution per model and task, the 270-run evaluation yields task-averaged pass rates from 22.2% to 82.2%. Of 137 failed executions, 123 missed an evidence-verification or action-order requirement.

1. Purpose and research context

Banking work depends on both the transaction performed and the evidence available before it. Global BankOps evaluates the path from finding evidence to payment, posting and reporting, including the order and effects of those actions.

FinQA [1] evaluates numerical reasoning over financial documents, while FinanceBench [2] covers financial questions with supporting evidence. ToolSandbox [3] evaluates intermediate and final milestones in a stateful tool environment. τ-bench [4] evaluates policy-guided tool and user interactions and resulting states. Global BankOps connects evidence timing, financial state changes and report consistency in bank back-office workflows. Scores from different tasks and execution systems are not directly comparable.

This edition uses the same 270 selected executions as the public leaderboard. It describes task categories, methods and aggregate results without including private evaluation tasks or answers.

2. Task construction and work environment

The 45 assignments were constructed from practitioner interviews, banking reference materials and public documents. Interviews informed record ownership and when officers reconcile information across systems. Oracle FLEXCUBE documentation and South Korean retirement-benefit rules informed the represented procedures. Documents and transactions are synthetic. Some internal approvals and information-access steps are simplified.

Table 1. Task categories and scope

Task category	Tasks	Work covered
Trade Finance	7	Reconcile trade documents and departmental confirmations before sending messages and processing payments.
Loan Servicing	14	Apply loan agreements and transaction records to repayments, interest, facility limits and servicing reports.
Treasury & Settlement	14	Settle and reconcile deposits, foreign exchange and money-market transactions. This task category also includes month-end accruals and foreign-currency revaluation.
Pension & Trust	10	Six pension tasks and four trust tasks cover allocation, investment and payouts after checking the relevant rules and authority.

Roles, records and state changes

Agents receive an officer role and an assignment, then use tools to inspect agreements, ledgers, mail, shared documents and SWIFT messages. Access to records and information from other teams depends on the assigned role. Payments, postings, reports and decisions to hold a transaction change the recorded environment. Each execution preserves its configuration, tool trajectory, final state and grade.

The setting is a Budapest subsidiary for 33 tasks and Seoul head-office or domestic branches for 12. Twenty-seven tasks stay within the bank; 18 involve SWIFT exchanges with another institution.

Accepted paths to evidence

Stale or conflicting records may require confirmation from the responsible team. Where the task permits a current system record as equivalent evidence, reading that record before acting is also accepted. Departmental enquiries and replies are not the only valid path for every task.

3. Grading accuracy and procedural alignment

Evaluation covers both accuracy and procedural alignment. Correct amounts may still fail if an agent skips a required check or takes a hard-to-reverse action before confirmation.

Table 2. Required evaluation dimensions

Dimension	What is checked
Financial accuracy	Code checks amounts, currencies, accounts, beneficiaries, value dates, balances and payment counts.
Business decisions and communication	Assess reports, instructions, decisions to defer a transaction and saved records with source quotations.
Evidence validity	Verify that cited text exists in the recorded evidence and contains the required financial figures.
Reconciliation and timing	Use event order to verify required enquiries and replies, or permitted current system records, before the relevant action.
Completion and scope	Check required outputs, task completion and the absence of unauthorized financial actions.

Code checks

Inspect recorded amounts, currencies, accounts, beneficiaries, value dates and state changes. Compare evidence and action steps to check required ordering.

Semantic judgment

Gemini 3.8 Flash assesses whether eligible messages, reports and saved records satisfy written requirements. It also identifies the relevant action when timing depends on a message’s meaning.

Evidence validation

Code verifies that quoted text exists in records allowed by the criterion. It also checks required financial figures where the rubric specifies them.

The evaluator uses hybrid scoring v6. Gemini 3.8 Flash judges meaning at temperature 0 with low reasoning effort. Every required criterion must pass for the execution to pass. A missing verdict or invalid required evidence leaves the execution unscored, rather than failed.

4. Execution protocol and aggregation

Item	Shared conditions
Evaluation set	45 English tasks · Four task categories
Setting	Synthetic bank offices · 33 tasks at a Budapest overseas subsidiary · 12 at Seoul head-office or domestic-branch settings
Scope	27 internal-bank tasks · 18 involving SWIFT exchanges with other banks or entities
Execution conditions	Workplace framing · Seed 0 · Isolated environment · 1,800-second agent timeout
Tool budget	At most 60 calls per task, one per turn. The agent is not told the limit or remaining count.
Evaluation	Hybrid scoring v6 · Code checks and Gemini 3.8 Flash semantic judgment

Select one execution per model and task

Eligible candidates match the current task inputs and validated grading provenance, have a scored verdict, and are not explicitly excluded. A validated reassessment of an unchanged task preserves the original execution and verdict. Candidates are ordered by time, then execution identifier, without using pass or fail.

pass@1 = passed selected runs / evaluated tasks

Give equal weight to the single selected result for each task. The model denominator is 45; category denominators follow their task counts. Scores, duration and token use come from the same selected executions.

Exclusions and additional executions

Eligibility does not mean identical settings. Reasoning settings and response budgets vary, including the budget change below.

Treatment and reason	Models and executions	Settings change	Aggregation
Infrastructure retry: Unscored provider error	Claude Fable 5.2: 1	Same settings	Original: 0 included · replacement: 1 included
Rerun with a changed response budget: Response-limit execution explicitly excluded from eligibility	Gemini 4 Carbon: 1	64,000 → 128,000 output tokens	Original: 0 included · replacement: 1 included
Additional eligible executions not selected: Earlier eligible executions already exist; all additional executions passed	GPT-6 Astra: 2; Claude Opus 5.5: 2	Not applicable · no replacement	All 4 unselected · these are scored executions

Selected 270 of 274 eligible scored executions. Treatment counts describe distinct states and must not be added to infer the total number of executions.

A before/after comparison under a matched current grading version and execution protocol is unavailable. Success on a new execution does not isolate the effect of a larger response budget.

Reasoning effort is listed with each model in the results. Provider routes and response budgets differ; max, high and xhigh are not treated as equal computation.

5. Performance and tool use

Table 3. Task-averaged pass rates (%) on the same 270 selected executions

Model · effort	pass@1	Trade Finance	Loan Servicing	Treasury & Settlement	Pension & Trust	Passed / runs
GPT-6 Astra (max)	82.2	71.4	64.3	92.9	100.0	37/45
Claude Fable 5.2 (max)	62.2	28.6	64.3	71.4	70.0	28/45
Gemini 4 Carbon (high)	55.6	28.6	57.1	57.1	70.0	25/45
Claude Opus 5.5 (max)	44.4	14.3	42.9	57.1	50.0	20/45
Grok 4.7 (xhigh)	28.9	0.0	35.7	42.9	20.0	13/45
Qwen3.8 Max Prime (xhigh)	22.2	14.3	21.4	42.9	0.0	10/45

Each model covers 45 tasks. Category denominators are 7 for Trade Finance, 14 for Loan Servicing, 14 for Treasury & Settlement and 10 for Pension & Trust. The overall score is not an unweighted mean of the four category scores.

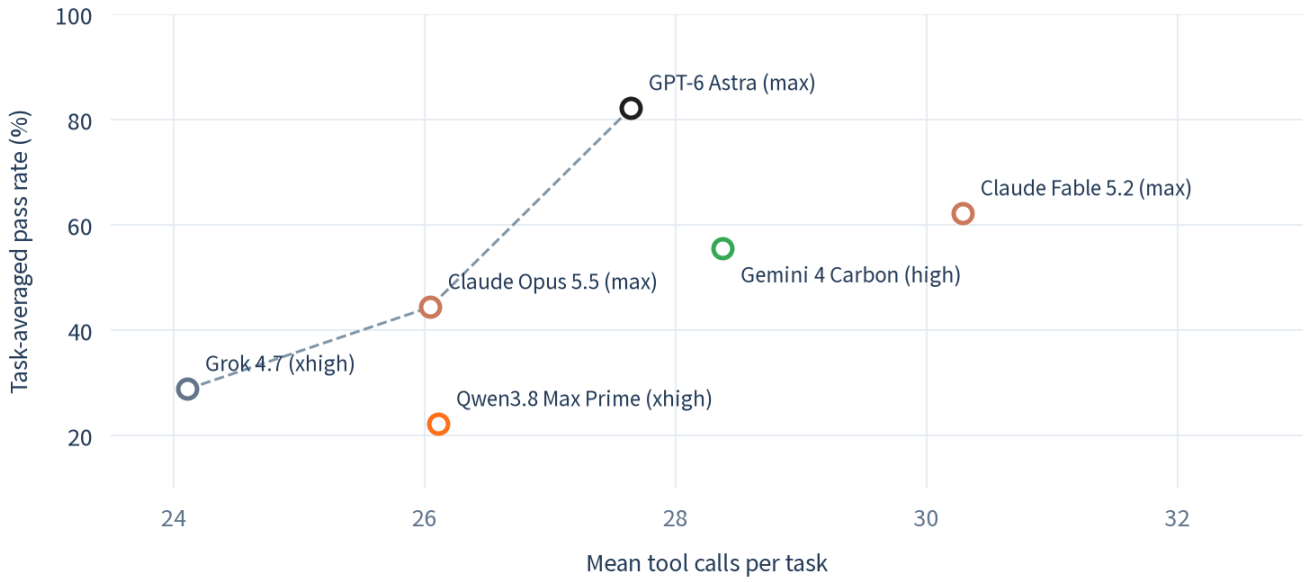


Figure 1. Task-averaged pass rate against mean tool calls per task. The dashed line marks the observed frontier.

Tool calls count actions processed by the environment, including the finish action. For each model, total calls over the 45 selected executions are divided by 45. Both axes use the same executions. Calls are distinct from tokens and latency. A configuration is on the observed frontier when no other of these six configurations has a higher pass rate with no more calls, or fewer calls with no lower pass rate.

GPT-6 Astra (max): 27.6 · Claude Fable 5.2 (max): 30.3 · Gemini 4 Carbon (high): 28.4 · Claude Opus 5.5 (max): 26.0 · Grok 4.7 (xhigh): 24.1 · Qwen3.8 Max Prime (xhigh): 26.1

6. Failed checks and category findings

Failed checks in the saved grades were grouped into four categories. Evidence verification and action order appeared most often. Requirements for reports, messages or saved records were also missed in 107 executions.

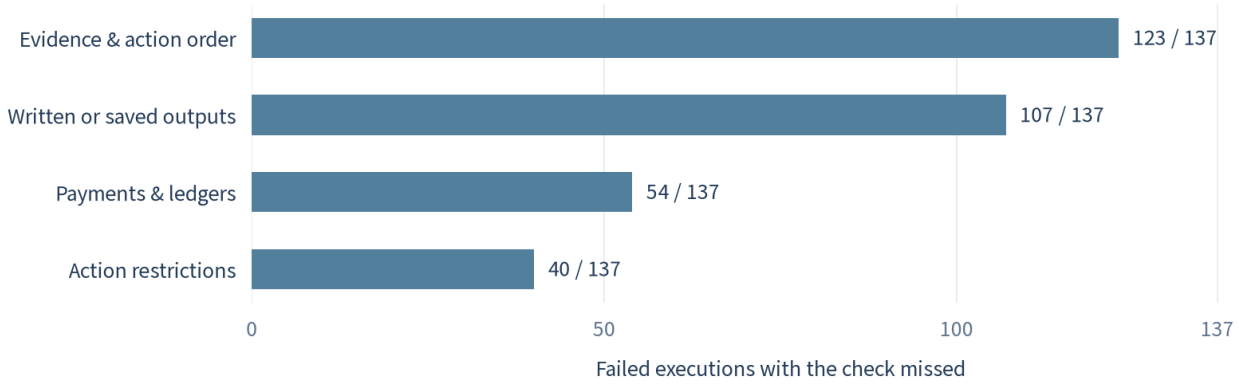


Figure 2. Failed checks among 137 failed executions. An execution can appear in several groups.

Check group	Executions	Definition
Evidence & action order	123 / 137	Required confirmation or record reading was absent, or an action occurred at or before the evidence.
Written or saved outputs	107 / 137	Reports, messages or saved records missed required content or fields.
Payments & ledgers	54 / 137	Payments, transfers or ledger states did not meet the required financial conditions.
Action restrictions	40 / 137	Actions exceeded the permitted scope of payments, transfers or messages.

Of the 123 evidence or ordering failures, 104 lacked a qualifying evidence event. In the remaining 19, the action occurred at or before that event. These are recorded check outcomes, not independent diagnoses of financial knowledge or reasoning failures.

Executions with applicable checks

An applicable execution has at least one recorded verdict in the group. This denominator differs from failed executions above and from the number of individual criterion checks.

Check group	Failed / applicable executions	Failure rate when applicable
Evidence & action order	123 / 246	50.0%
Written or saved outputs	107 / 258	41.5%
Payments & ledgers	54 / 234	23.1%
Action restrictions	40 / 270	14.8%

Table 4. Failed executions by category · overlapping groups

Task category	All failures	Evidence & action order	Written or saved outputs	Payments & ledgers	Action restrictions
Trade Finance	31	25	21	6	9
Loan Servicing	44	43	31	18	1
Treasury & Settlement	33	27	26	21	8
Pension & Trust	29	28	29	9	22

An execution can fail more than one group. Bars count executions with a failed check among the 137 failures; they are not mutually exclusive shares of root causes.

7. Evaluator validation and judge agreement

Validation separates reproduction of expected environment and evaluator outcomes from agreement between two model judges. These checks answer different questions.

Current v6: expected outcome reproduction

Using cached judgments, the frozen task pack reproduced all 1,276 expected outcomes: 744 authored workflows, 353 controls and 179 additional fixtures and regression cases. This does not represent 1,276 new solver executions or judge accuracy against human assessment.

Case group	Reproduced / expected
Authored workflows	744 / 744
Controls	353 / 353
Additional fixtures and regressions	179 / 179
Total	1,276 / 1,276

Earlier v5: two judges on a development sample

A separate development audit used earlier scoring v5 on 45 saved executions generated by Opus 5.5. It compared Gemini 3.8 Flash with Opus 5.5 (max). Rubric wording and citation scope were clarified using the same sample. Afterwards, criterion agreement was 151/152 and execution-outcome agreement was 45/45. Denominators include only criteria or executions scored by both judges.

Table 5. Judge agreement on the same 45 saved executions

Stage	Criterion agreement	Cohen’s κ	Run agreement	Unscored by Opus
Initial audit	144/146 (98.6%)	0.962	40/40	5
After clarification	151/152 (99.3%)	0.982	45/45	0

Cohen’s κ adjusts criterion agreement for chance agreement under the observed label marginals. The remaining criterion disagreement after clarification did not change the execution outcome.

The same development sample informed the clarification; it was not held out. Agreement under the earlier v5 evaluator is not human-validated accuracy for current v6. The current evaluation has no independent human estimate of judge accuracy.

8. Interpretation and limitations

The findings support examining evidence retrieval, permitted action order and outcome reporting together. A failed check alone does not establish a model's general financial ability or a causal explanation for its failure.

- Scope: 45 authored tasks in a synthetic bank, with simplified approval and access steps. Results do not measure live operational performance or a representative success rate across all banking work.
- Execution: tasks, environment, grading criteria, time limits and tool budgets are shared. Provider routes, reasoning settings and response budgets vary. Equally named effort settings do not imply equal computation across providers.
- Repetition: public pass@1 uses one eligible execution per model and task. It does not provide repeated-success reliability, confidence intervals or precise task-level success probabilities.
- Judging: code checks still depend on authored criteria and their implementation. Judge agreement differs from independent answer validation. Human-validated accuracy of current v6 has not been measured.
- Access: task categories, methods and aggregate results are public. Individual evaluation tasks, hidden conditions, answers, criterion-level solutions and complete execution records remain private.

References

[1] Chen et al. (2021). FinQA: A Dataset of Numerical Reasoning over Financial Data.

[2] Islam et al. (2023). FinanceBench: A New Benchmark for Financial Question Answering.

[3] Lu et al. (2024). ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities.

[4] Yao et al. (2024). τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.