Flitto Agent Bench

Finance – US Consumer Credit · Technical report

Flitto Agent Data Researcher Team | Data version 2026-11-18 · Results as of 2026-11-13

Report edition 26.11.18 · Task-set version 26.11.11 · Analysis / report revised 2026-11-18

Abstract

Agents doing financial work must reconcile records, apply relevant rules and submit the requested deliverables within a fixed budget. Flitto Agent Bench (Finance – US Consumer Credit) measures agent judgment, task execution and weaknesses through tasks that reproduce professional rules and workflows. Its first public track covers US consumer-credit and card back-office work. Development responses from five models informed the selection of 99 tasks across eight categories.

The tasks and data are synthetic. Some generators were calibrated to the schema and aggregate statistics of real financial data, and the judgment criteria come from public US regulations and practitioner material. No real customer records are included.

The main metric is completion within budget: every required value and submission rule must be satisfied. Mean pass rates across two evaluations of the same 99 tasks were 79.3% for Opus, 53.5% for Gemini, 28.8% for GPT, 27.8% for Qwen and 16.2% for Kimi. The share of tasks passed twice ranged from 10.1% to 75.8%; agreement between runs ranged from 78.8% to 92.9%. This report presents those five models and separate comparisons for Fable and GLM. Fable and GLM were also evaluated twice on the same 99 tasks, with mean pass rates of 85.9% and 27.3%, respectively.

1. Purpose and contribution

Financial work requires deciding which records, reporting periods, populations and policies apply before doing arithmetic. Missing required fields or failing to submit within budget also constitutes incomplete work. Flitto Agent Bench gives agents a data room containing business records and policy documents, then evaluates the full workflow under explicit submission rules.

The study makes three contributions. First, it translates US public regulations and practitioner judgment into synthetic consumer-credit and card tasks. Second, passing requires agents to use tools to explore records and submit every required result. Third, it evaluates five models twice on the same 99 tasks and reports mean completion alongside repeated success and failure.

1.1 Related evaluations

Financial agent evaluations emphasize different tasks. DABstep combines financial data and documents in multi-step analysis. Finch focuses on spreadsheet-centered finance and accounting. The banking domain of τ-Knowledge combines knowledge retrieval and tool use for customer support. APEX-Agents and GDPval assess professional deliverables, including financial work. Flitto Agent Bench focuses on the requested output and the completion criteria for US consumer-credit tasks.

Table 1. Tasks and grading in related evaluations

Study	Main tasks	Deliverables and grading
DABstep	Multi-step analysis of financial data and documents	Automatic checks of factoid answers
Finch	Spreadsheet-centered finance and accounting	File creation, editing and questions; human and automated evaluation
τ-Knowledge	Retrieval and tool use for banking customer support	Verifiable, policy-compliant state changes
APEX-Agents	Cross-application work in investment banking, consulting and law	Required criteria for responses and generated files
GDPval	Economically valuable work across occupations	Expert blind comparisons of deliverable quality
Flitto Agent Bench	Record reconciliation and policy application in US consumer credit	Correct required values, valid submission and completion within budget

The focus here is deciding the requested scope from a fixed collection of source material and submitting a valid final answer within budget.

2. Data and task design

2.1 Synthetic data and practitioner knowledge

The tasks and data are synthetic. Some generators were calibrated to the schema and aggregate statistics of real financial data, and the judgment criteria come from public US regulations and practitioner material. No real customer records are included.

Task design also draws on practitioner interviews, congressional hearings, webinars and training videos. Practitioner knowledge means the ways experts interpret material, order their decisions and handle exceptions. Sources include an interview on underwriting strategy and a US House hearing on credit-reporting accuracy. These informed task backgrounds and requirements.

Each task contains source records, policy and procedure documents, a work request, an output specification, grading criteria and an execution budget. Every task is assigned one primary category.

Table 2. Task composition by category

Task category	Tasks	Base designs
Personal finance	13	8
Payments & billing	13	10
Fraud & disputes	12	9
Credit files	11	7
Underwriting	13	11
Churn & retention	13	9
Collections	12	8
Portfolio	12	12
Total	99	74

The 99 tasks derive from 74 base designs and use 63 data-room packages. Some tasks share a design or a data room.

2.2 Selection using development responses

Development responses from Opus, Gemini, GPT, Qwen and Kimi informed the selection of 99 tasks from 311 candidates. Selection also considered category coverage. Appendix B reports development pass counts.

Selecting tasks that challenge models is also used in HLE and ANLI. This benchmark targets the interpretation, policy application, calculation and submission required by professional work.

2.3 Answer validation and review

During construction, finance experts reviewed the work logic and agreed with the answer determinations. External rules were checked against US primary sources. References to data-room documents were checked against the supplied documents. Calculation paths and source checks are maintained as supporting evidence.

Answers follow the scope of the request, data units and periods, and the priority of policies and records. 98 tasks have one reference answer, and one task accepts two definitionally valid answers; answers include numbers and strings.

A reviewer blinded to model names checked unclassified failures for alternative interpretations. Saved execution records and grading results were also independently reconciled to reproduce aggregation and statistical analysis. Each reference answer is linked to its solution, policy evidence and design sources.

3. Evaluation method

3.1 Models and common conditions

The main evaluation covers Claude Opus 5.5 (Opus), Gemini 4 Argon (Gemini), GPT-6 Astra (GPT), Qwen3.8 Max 0902 (Qwen) and Kimi K3.1 Pro (Kimi).

All models received the same 99 tasks, data rooms, output specifications and grading criteria. Each run allows one attempt per task, with a budget of 60 turns and 3,600 seconds. All seven models were evaluated twice; the primary score is the mean of those two runs.

The first valid execution matching the input conditions was graded. Incorrect answers, missing submissions and budget exhaustion count as failures. One rerun was allowed for predefined infrastructure errors. Only executions whose task, prompt and data-room hashes and grader version match the frozen set were graded.

Evaluation dates (start–end)

Model	Run 1	Run 2
Opus	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Gemini	2026-11-05–2026-11-11	2026-11-09–2026-11-11
GPT	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Qwen	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Kimi	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Fable	2026-11-11	2026-11-13
GLM	2026-11-08–2026-11-11	2026-11-13

3.2 Grading and metrics

Agents must submit an answer file containing every required key. A code verifier checks the file, its values and normal process completion. No LLM judge is used. A task passes only when all requirements are satisfied within budget. A correct file followed by forced termination at a turn or time limit fails. Partial answers receive no task-level credit. Missing files, required keys or valid formatting also fail.

Numeric fields accept numbers or numerically equivalent strings. Counts require integers; amounts and ratios use task-specific tolerances. Strings containing commas or currency symbols, and Boolean values, fall outside the numeric specification. Text fields follow each task's rules. Every required field must meet the reference criteria; conflicting values for the same field fail.

- Primary pass rate: the mean of two single-attempt task-completion rates, calculated as (run-1 passes + run-2 passes) / (2 × 99) × 100.
- Repeated completion (pass^2): the share of tasks passed in both runs.
- Agreement: the share of tasks with the same pass or fail outcome in both runs.

3.3 Statistical analysis

Appendix A gives 95% percentile bootstrap intervals for run-1 pass rates, based on 10,000 item-level resamples with seed 20261104. These intervals concern the first run on the selected 99 tasks. Appendix C uses Wilson intervals for category-level run-1 results (z=1.96, no continuity correction).

Model comparisons use an exact McNemar test on paired task outcomes, with Holm correction across all ten model pairs within each run at a significance level of 0.05. Tiers retain boundaries between adjacent models in score order that are significant in both runs. Appendix A lists the adjacent pairs and shared boundaries.

4. Results

4.1 Mean pass rates across two runs

Table 3. Two-run mean and tiers on the same 99 tasks

Model	Run 1 /99	Run 2 /99	Mean pass rate %	Tier
Claude Opus 5.5	81	76	79.3	1
Gemini 4 Argon	50	56	53.5	2
GPT-6 Astra	29	28	28.8	3
Qwen3.8 Max 0902	26	29	27.8	3
Kimi K3.1 Pro	14	18	16.2	3

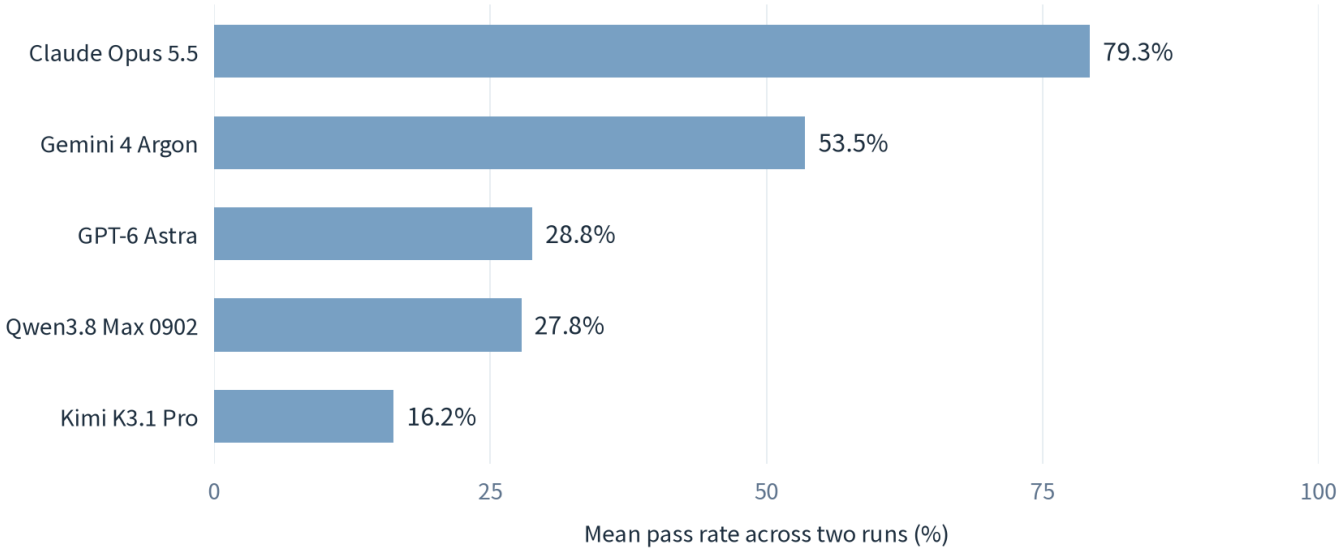


Figure 1. Mean pass rate across two evaluations of the same 99 tasks.

In run 1, all five models failed 12 tasks and passed 7 tasks. Outcomes differed across models on the other 80 tasks. Between one and three models passed 74 tasks.

Budget exhaustion ended 19 Gemini executions and 2 Qwen executions in run 1. Eighteen of those Gemini executions produced no answer file. One produced a file but did not satisfy normal completion within budget.

4.2 Repeated completion and tiers

The following table summarizes two evaluations of the same tasks.

Table 4. Repeated completion and outcome agreement

Model	Run 1 %	Run 2 %	Mean %	pass^2 %	Agreement %
Opus	81.8	76.8	79.3	75.8	92.9
Gemini	50.5	56.6	53.5	44.4	81.8
GPT	29.3	28.3	28.8	19.2	80.8
Qwen	26.3	29.3	27.8	17.2	78.8
Kimi	14.1	18.2	16.2	10.1	87.9

GPT scored above Qwen in run 1; Qwen scored above GPT in run 2. Qwen passed 26 tasks in run 1 and 29 in run 2, including 17 tasks passed twice (17.2%). Nine outcomes changed from pass to fail and twelve from fail to pass. Kimi's 87.9% agreement includes tasks failed in both runs. Exact McNemar p-values for changes between runs ranged from 0.125 to 1.000 across models.

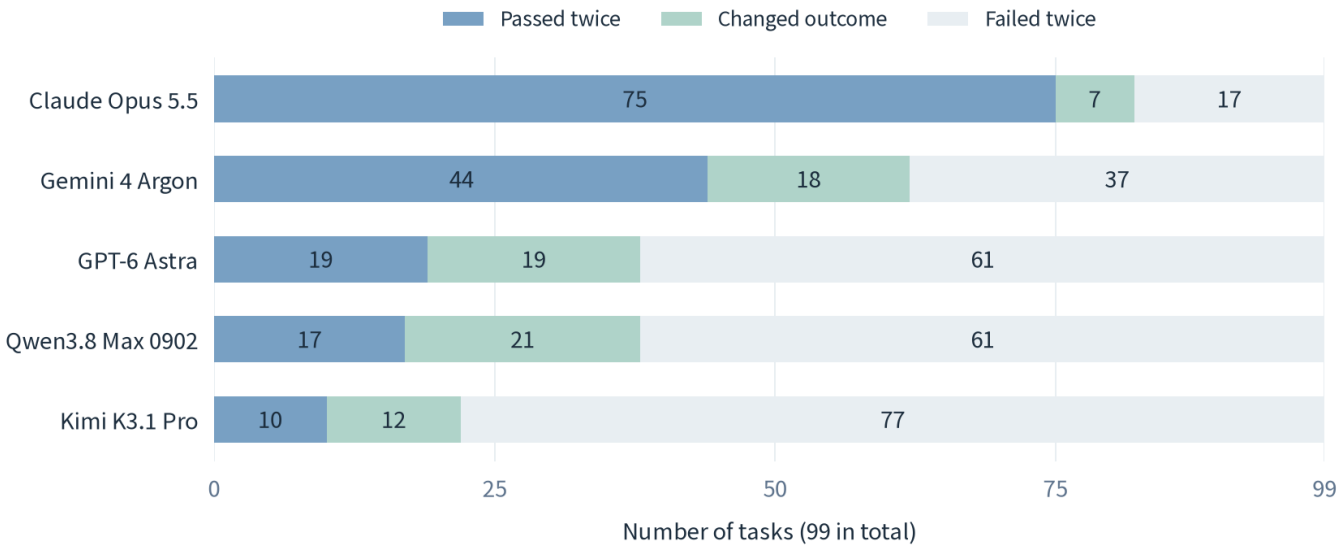


Figure 2. Repeat outcomes on the same 99 tasks: passed twice, changed outcome and failed twice.

The shared tiers are {Opus} > {Gemini} > {GPT, Qwen, Kimi}. GPT and Qwen did not differ significantly in run 1 (Holm p=0.728). The GPT–Kimi and Qwen–Kimi differences were significant in run 1 but not run 2, placing GPT, Qwen and Kimi in the same third tier.

4.3 Execution-record review

Saved execution records and grading results were reconciled to reproduce aggregation and statistical analysis. The record audit observed no access to reference answers, graders, the benchmark repository or other executions' material. The same checks covered all 495 executions in run 2.

A classifier blinded to model identity and correctness labeled evaluation-awareness expressions using a common rule. Such expressions appeared in 9 Opus, 1 Gemini and 1 Qwen executions in run 1, and 3 Opus and 1 Qwen executions in run 2. In a sensitivity analysis treating passes with these expressions as failures, run-1 pass counts became 77 for Opus and 49 for Gemini; run-2 counts became 75 for Opus and 28 for Qwen. Qwen and GPT were tied at 28 in that run-2 analysis.

4.4 Separate comparisons of models not used in selection

Claude Fable 5.2 and GLM-5.3 Prime were each evaluated twice on the same 99 tasks, output specifications, 60-turn and 3,600-second budget, and grading criteria.

Table 5. Two-run results for separate comparison models

Model	Run 1 passed /99	Run 2 passed /99	Mean pass rate %	Min–max %
Claude Fable 5.2	85	85	85.9	85.9–85.9
GLM-5.3 Prime	29	25	27.3	25.3–29.3

In run 1, Fable completed 98 executions normally and reached the 60-turn limit once. Its run-1 item-bootstrap 95% interval was 78.8–91.9%. In paired comparisons with the main models' first runs, Fable alone passed 7 tasks against Opus, while Opus alone passed 3 ($p=0.344$). Differences against Gemini, GPT, Qwen and Kimi each had Holm $p<0.001$.

In run 1, GLM scored below Opus and Gemini and above Kimi (Holm $p<0.001$, 0.002 and 0.008, respectively). Differences against GPT and Qwen were not significant (Holm $p=1.000$ each). Holm correction was applied separately to each comparison model's five pairs with the main models.

4.5 Failure types: submitted answers and execution status

This analysis covers 1,386 executions: seven models, the same 99 tasks and two runs. Of these, 631 passed and 755 failed. It pools 583 failures from the five primary models and 172 from the separate Fable/GLM comparisons for diagnostic purposes. Different models and runs count as separate executions even on the same task. This denominator differs from the live leaderboard's pass@2, which counts tasks passed at least once.

Data interpretation and cross-checking patterns appeared in 392 of 755 failed executions (51.9%). These answers matched predefined error patterns associated with unhandled data anomalies or relying on a field despite other evidence taking precedence. Population and balance definition patterns appeared in 158 failures; abstention on answerable values appeared in 65.

Patterns differ by model. Population and balance definitions were most common for Claude Opus 5.5, at 24 of 41 failures (58.5%), and Claude Fable 5.2, at 22 of 28 failures (78.6%). Rates across all evaluated executions are 24/198 (12.1%) and 22/198 (11.1%), respectively. The web diagnostics switch between failed-execution and all-execution denominators.

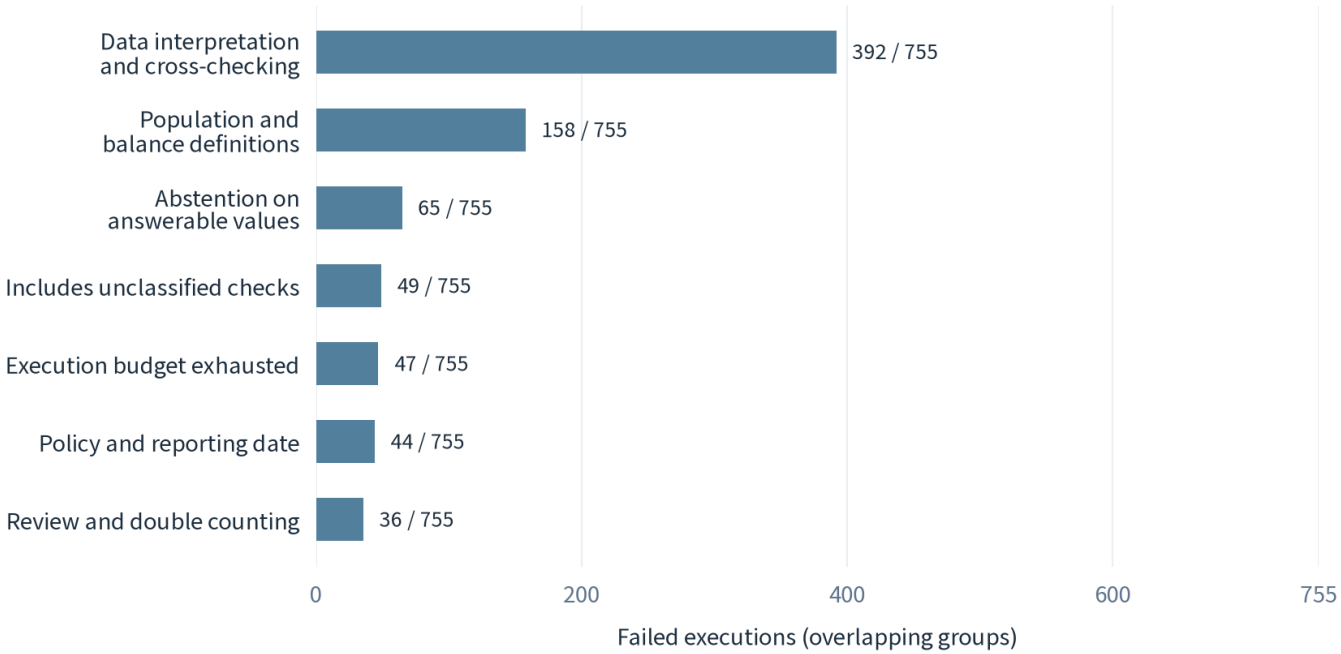


Figure 3. Diagnostic groups across both runs. The denominator is 755 failed executions; an execution can appear in several groups.

Table 6. Failed executions by diagnostic group and evaluation cohort (overlapping groups)

Diagnostic group	5 primary models	2 separate models	All
Data interpretation and cross-checking	304	88	392
Population and balance definitions	109	49	158
Abstention on answerable values	61	4	65
Includes unclassified checks	36	13	49
Execution budget exhausted	41	6	47
Policy and reporting date	32	12	44
Review and double counting	24	12	36
Failed executions (denominator)	583	172	755

Data interpretation and cross-checking groups patterns involving anomalies and precedence between fields or evidence. Population and balance definitions concerns the included records and balance definitions; policy and reporting date concerns applicable rules and reporting periods. Review and double counting covers draft review, reconciliation and duplicate aggregation. Abstention on answerable values means withholding a required numeric answer that the data supports. An execution counts once within each group but can appear in several groups, so row counts exceed the denominator when added.

Among the failures, 708 executions submitted an answer and finished normally but failed required checks. The other 47 exhausted the execution budget: 42 left no answer file and five left a file. Of those five, three passed every answer check but failed the requirement to finish within budget. Of 49 executions with unclassified checks, 46 are entirely unclassified and three also contain identified patterns.

The analysis compares saved answers and termination states with the frozen checks and registered error patterns. All 1,386 pass/fail outcomes match this report's aggregates; the diagnostics do not change scores. Pattern matches do not establish which records an agent actually read or skipped, or the order of its decisions. Establishing an individual behavioral cause requires a separate trajectory review. Individual tasks, specific error-pattern conditions, reference answers and execution records remain private.

5. Discussion

Flitto Agent Bench implements US consumer-credit and card judgment and submission workflows in synthetic data rooms. Some generators were calibrated to the schema and aggregate statistics of real financial data. Public regulations and selected practitioner knowledge inform task conditions. Evaluation checks whether agents submit all required deliverables within budget.

Failure diagnostics compare submitted answers with execution records to identify cases for human review. Reproduction records connect selected tasks, input and configuration identifiers, end states, grading decisions, aggregation and statistical procedures. Publication of aggregate results follows the disclosure scope for private tasks, answers and customer material.

6. Conclusion

Flitto Agent Bench evaluates the full process from interpreting business records to applying policy, calculating and submitting valid results. Mean pass rates across two evaluations of the same 99 tasks ranged from 16.2% to 85.9% for all seven models. Repeat outcomes and paired comparisons describe performance differences on this task set.

Flitto separately supplies AI training data.

7. Resource metrics on the public site

The public site reports tool calls, input tokens, output tokens and execution time. Usage bars aggregate first-run task records. Averages divide by the number of tasks with measurements; totals require records for every task. The Pareto chart compares models with complete scores and usage records for the same 99 tasks under the selected aggregation. Its dashed line connects points for which no other selected model uses no more resources and scores at least as highly, with a strict improvement in one of the two.

Appendix A. Intervals and paired comparisons

Table A1. Run-1 pass rates and item-bootstrap 95% intervals (%)

Percentile intervals use 10,000 item-level resamples with seed 20261104.

Model	Run-1 pass rate %	95% interval
Opus	81.8	73.7 - 88.9
Gemini	50.5	40.4 - 60.6
GPT	29.3	20.2 - 38.4
Qwen	26.3	18.2 - 35.4
Kimi	14.1	8.1 - 21.2

Table A2. Run-1 exact McNemar tests with Holm correction

The two middle columns count tasks passed by only the named member of each pair.

Model pair	First only	Second only	Exact p	Holm p
Opus-Gemini	33	2	<0.001	<0.001
Opus-GPT	54	2	<0.001	<0.001
Opus-Qwen	59	4	<0.001	<0.001
Opus-Kimi	67	0	<0.001	<0.001
Gemini-GPT	26	5	<0.001	<0.001
Gemini-Qwen	30	6	<0.001	<0.001
Gemini-Kimi	39	3	<0.001	<0.001
GPT-Qwen	18	15	0.728	0.728
GPT-Kimi	19	4	0.003	0.008
Qwen-Kimi	17	5	0.017	0.034

Table A3. Adjacent pairs in each run

Holm correction covers all ten pairs of the five models in each run. Adjacent pairs follow each run's score order; shared tier boundaries appear separately below.

Run	Adjacent pair	Holm p	Significant in run
1	Opus-Gemini	<0.001	Yes
1	Gemini-GPT	<0.001	Yes

Run	Adjacent pair	Holm p	Significant in run
1	GPT-Qwen	0.728	No
1	Qwen-Kimi	0.034	Yes
2	Opus-Gemini	0.002	Yes
2	Gemini-Qwen	<0.001	Yes
2	Qwen-GPT	1.000	No
2	GPT-Kimi	0.174	No

Table A4. Tier boundaries shared by both runs

Shared tier boundary	Run-1 adjacent pair	Run-2 adjacent pair
{Opus} / {Gemini, GPT, Qwen, Kimi}	Opus-Gemini	Opus-Gemini
{Opus, Gemini} / {GPT, Qwen, Kimi}	Gemini-GPT	Gemini-Qwen

Appendix B. Development results

Table B1. Development passes / graded tasks by selection status

Development pass-rate order was Gemini, Opus, Qwen, GPT and Kimi in all three sets.

Item set	Opus	Gemini	GPT	Qwen	Kimi
Selected 99	43/99	50/99	18/99	35/99	10/99
Excluded 212	150/212	156/212	111/212	140/212	101/212
All 311 candidates	193/311	206/311	129/311	175/311	111/311

Appendix C. Run-1 outcome distribution and task categories

Table C1. Tasks grouped by the number of models passing in run 1

Models passing	Tasks	Opus	Gemini	GPT	Qwen	Kimi
0	12	0	0	0	0	0
1	30	25	1	1	3	0
2-3	44	43	36	16	12	4
4-5	13	13	13	12	11	10

Table C2. Run-1 category passes / tasks [Wilson 95% interval, %]

Task category	Opus	Gemini	GPT	Qwen	Kimi
Personal finance	5/13 [17.7, 64.5]	7/13 [29.1, 76.8]	4/13 [12.7, 57.6]	6/13 [23.2, 70.9]	1/13 [1.4, 33.3]
Payments & billing	10/13 [49.7, 91.8]	7/13 [29.1, 76.8]	4/13 [12.7, 57.6]	4/13 [12.7, 57.6]	3/13 [8.2, 50.3]
Fraud & disputes	11/12 [64.6, 98.5]	6/12 [25.4, 74.6]	2/12 [4.7, 44.8]	1/12 [1.5, 35.4]	1/12 [1.5, 35.4]
Credit files	10/11 [62.3, 98.4]	6/11 [28.0, 78.7]	2/11 [5.1, 47.7]	5/11 [21.3, 72.0]	3/11 [9.7, 56.6]
Underwriting	13/13 [77.2, 100.0]	9/13 [42.4, 87.3]	6/13 [23.2, 70.9]	1/13 [1.4, 33.3]	0/13 [0.0, 22.8]
Churn & retention	13/13 [77.2, 100.0]	5/13 [17.7, 64.5]	6/13 [23.2, 70.9]	4/13 [12.7, 57.6]	2/13 [4.3, 42.2]
Collections	10/12 [55.2, 95.3]	5/12 [19.3, 68.0]	1/12 [1.5, 35.4]	3/12 [8.9, 53.2]	0/12 [0.0, 24.3]
Portfolio	9/12 [46.8, 91.1]	5/12 [19.3, 68.0]	4/12 [13.8, 60.9]	2/12 [4.7, 44.8]	4/12 [13.8, 60.9]

References

- DABstep: Data Agent Benchmark for Multi-step Reasoning.
- Finch: Benchmarking Finance & Accounting across Spreadsheet-Centric Enterprise Workflows.
- τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge.
- APEX-Agents.
- GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks.
- Humanity's Last Exam.
- Adversarial NLI: A New Benchmark for Natural Language Understanding.
- GDS Link: Credit Risk Expert Discusses the Art of Underwriting Strategy.
- U.S. House Committee on Financial Services: Consumer Credit Reporting: Assessing Accuracy and Compliance.

Task-set fingerprint (SHA-256): 72b60695c53f6e395464f2b37c383430d7aec7a770f9267b5f245bac5c469636