Five configurations.
One comparison.

Banking77 public benchmark. Plain intent prompt, no retrieval. Descriptive comparison, not a published customer assessment or rollout approval.

500 + 30	5	None
Primary + challenge cases	Tested configurations	Retrieval

WORKFLOW

Banking support intent classification

CURRENT MODEL

o4-mini / OpenAI route

PREFERRED ELIGIBLE CONFIGURATION

GLM-5.3-Flash

Default weights: correctness 50%, consistency 20%, observed cost 20%, p95 response time 10%. The score is a preference score, not statistical confidence.

COLLECTION WINDOW

2026-10-10 05:37:25 UTC to 2026-10-10 08:12:08 UTC

2,650 of 2,650 requests attempted. The 500 primary cases and 30 challenge cases are separate. The existing published assessment and historical retrieved-Luna results are unchanged.

Quality and consistency

500 primary cases. All rows retain their source denominators.

Configuration	Consistency	Correct labels	All errors	Score / 100
o4-mini	100.0% 500/500 valid matching labels	75.8% 379/500 correct	121/500	72.90
DeepSeek V4 Flash	30.6% 153/500 valid matching labels	25.6% 128/500 correct	372/500	Unranked
Qwen3.7-Flash	84.2% 421/500 valid matching labels	73.6% 368/500 correct	132/500	Unranked
Nemotron 3.5 Lightning	82.2% 411/500 valid matching labels	74.6% 373/500 correct	127/500	Unranked
GLM-5.3-Flash	86.0% 430/500 valid matching labels	79.6% 398/500 correct	102/500	83.27

O4-MINI, GLM-5.3-FLASH

Eligible under the declared requirements

DEEPSEEK V4 FLASH, QWEN3.7-FLASH

Phase charges are incomplete; Some attempt charges are unknown; More errors than the incumbent

NEMOTRON 3.5 LIGHTNING

More errors than the incumbent

Consistency includes shared wrong answers. Provider and output failures are distinct from incorrect reference labels. Unknown charges prevent ranking; they are never set to zero.

The operational trade-offs

Primary cohort. Latency is one complete HTTP attempt, not application response time.

Configuration	USD / 1,000	Known charges	p50	p95
o4-mini	$1.7566	500/500	2.26 s	5.92 s
DeepSeek V4 Flash	Unknown	177/500	0.16 s	15.72 s
Qwen3.7-Flash	Unknown	491/500	4.04 s	13.19 s
Nemotron 3.5 Lightning	$0.2020	500/500	3.22 s	5.28 s
GLM-5.3-Flash	$0.0274	500/500	1.31 s	3.08 s

ALL 530 CASES, ALL CONFIGURATIONS

Known reported subtotal $1.0938. Retained reserve $0.2685 for uncertain charges. The reserve is a budget allowance, not a measured bill.

An unknown attempted charge anywhere in an arm prevents its cost-based ranking. Per-1,000 values normalize recorded primary charges arithmetically. Subtotals exclude unknown charges and are not final invoice totals.

TIMER AND COLLECTION

Serial requests with rotating model order. Durations include failed attempts and full response reading. Pauses between collection segments are excluded. Percentiles use linear interpolation. Server load was not controlled.

Segment	First receipt window start	Last receipt	Prior pause
parent	2026-10-10 05:37:25 UTC	2026-10-10 05:40:56 UTC	Unavailable
continuation	2026-10-10 05:51:50 UTC	2026-10-10 05:57:08 UTC	653.77 s
second continuation	2026-10-10 06:01:16 UTC	2026-10-10 08:12:08 UTC	248.70 s

Where the outputs differ

Incumbent and candidate issues, on the same 500 primary cases.

Configuration	Wrong labels	Service / output	Unresolved / missing	Corrections	Regressions
o4-mini	121	0	0	0	0
DeepSeek V4 Flash	49	323	0	8	11
Qwen3.7-Flash	121	11	0	24	25
Nemotron 3.5 Lightning	118	9	0	24	23
GLM-5.3-Flash	102	0	0	33	14

Corrections and regressions use pairs with valid labels. Service/output counts provider errors, invalid outputs and wrong model identity. Missing and unresolved results are not described as observed wrong labels.

Configuration	Valid pairs	Both correct	Shared wrong label	Both wrong, different
o4-mini	500	379	121	0
DeepSeek V4 Flash	177	120	33	5
Qwen3.7-Flash	489	344	77	19
Nemotron 3.5 Lightning	491	349	62	33
GLM-5.3-Flash	500	365	65	23

Separate checks. Explicit priorities.

Challenge results do not contribute to the primary score.

Configuration	Consistency	Correct labels	All errors
o4-mini	100.0% 30/30 valid matching labels	83.3% 25/30 correct	5/30
DeepSeek V4 Flash	30.0% 9/30 valid matching labels	26.7% 8/30 correct	22/30
Qwen3.7-Flash	93.3% 28/30 valid matching labels	83.3% 25/30 correct	5/30
Nemotron 3.5 Lightning	90.0% 27/30 valid matching labels	80.0% 24/30 correct	6/30
GLM-5.3-Flash	86.7% 26/30 valid matching labels	80.0% 24/30 correct	6/30

Preference profile	Quality / agreement / cost / time	Preferred configuration
default	50% / 20% / 20% / 10%	GLM-5.3-Flash
quality	70% / 10% / 10% / 10%	GLM-5.3-Flash
economy	50% / 10% / 30% / 10%	GLM-5.3-Flash
latency	50% / 10% / 10% / 30%	GLM-5.3-Flash

Ties retain all co-winners. An unranked arm may have useful measurements, but missing evidence cannot be ignored to produce a winner.

Read the evidence in context

Public benchmark, plain prompt, no retrieval.

All configurations use the same 500 primary cases and original plain intent prompt, without retrieval. The 30 challenge cases are reported separately and do not contribute to weighted scores. These public cases were previously exposed; this is not fresh held-out confirmation.

Default weights are correctness 50%, valid label agreement 20%, observed cost 20%, and p95 response time 10%. Cost and time utilities are incumbent / (incumbent + candidate); equal measurements yield 0.5. Both-zero utility is 0.5. The displayed score is a preference score, not statistical confidence.

Eligibility requires complete outcomes, known attempt charges and timing across all 530 cases, verified identity, and no more primary errors than the incumbent. Missing evidence never redistributes weights. Co-winners are retained. The incumbent remains a retain-current option.

Consistency is matching valid intent labels divided by all planned primary cases. It can include shared wrong answers. Errors include wrong labels, invalid or unusable outputs, provider failures and unresolved or missing outcomes; those categories are shown separately.

Requests were serial with rotating model order, a 2,048-token output cap and no retries. p50/p95 use linear interpolation. Duration covers one full HTTP attempt including response reading and failed attempts; collection pauses are excluded. This is not production latency or a controlled server-load comparison.

Cost per 1,000 is arithmetic normalization, not a measured scaled workload. Unknown failed-attempt charges remain unknown and exclude the arm from ranking. Known subtotals are not complete invoice totals. Provider, route and reasoning settings are part of the configuration.

Publisher labels remain unchanged and are not expert customer adjudication. Historical Terra, Sol, Luna and retrieved-Luna results are separate configurations and are not pooled here. The existing published assessment, its policy and its revision are unchanged.

Exact settings. Traceable source.

A separate descriptive supplement; no assessment revision was changed.

Model ID	Provider route	Reasoning
openai/o4-mini	openai	Enabled
deepseek/deepseek-v4-flash-0731	deepinfra/fp8	Low effort
qwen/qwen3.7-flash	alibaba/us-east-1	1,024 reasoning tokens
nvidia/nemotron-3.5-lightning	deepinfra/bf16	Enabled
z-ai/glm-5.3-flash	z-ai/fp8	Low effort

PREPARED

2026-10-10 08:13:29 UTC

VERIFIED SOURCE SUMMARY SHA-256

7687c9fbec7063ae1835e8d44ff9611e51f3bc882d315325c917ca6bc3e54f5e

COMPANION FILES

comparison.html: browser preview. comparison.csv: exact unrounded metric values. comparison.json: curated presentation inputs. manifest.json: file digests. Keep this packet together.

pisama-benchmark-supplement/1. Source observations and scores are presented without rescoring or reweighting. Private filesystem paths, account IDs and raw provider errors are excluded.