A completed collection.
One qualified comparison.

Public Banking77 benchmark. Plain prompt, no retrieval. Adaptive configuration comparison using reused cases; no customer assessment or rollout approval.

2650	2649	1
Model evaluations attempted	Resolved outcomes	Unresolved outcomes

ADAPTIVE RECOVERY

1 required retry completed; 2,651 total HTTP attempts for 2,650 scheduled model evaluations across 530 shared cases. First-pass failures, retry results and both charges remain in the evidence.

REPORT STATUS

Collection closed with unresolved evidence. The original all-five-complete gate did not pass.

CUSTOMER / WORKFLOW

Public Banking77 example, banking-support intent classification. Original plain prompt, no retrieval; no customer data or customer acceptance policy.

CURRENT MODEL

o4-mini / OpenAI, freshly measured in this segmented collection.

PREFERRED AMONG FULLY EVIDENCED ADAPTIVE CONFIGURATIONS

GLM-5.3-Flash

This recommendation excludes configurations lacking full evidence. It does not establish the best model among all five. Unresolved outputs and unknown charges remain unknown, and excluded models remain unranked.

Weights remain correctness 50%, end-to-end agreement 20%, actual cost 20% and p95 cumulative HTTP time 10%. The original per-arm evidence requirements and incumbent error floor remain unchanged. Preference scores are not statistical confidence.

COLLECTION WINDOW

2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC

Known outcomes. Visible gaps.

500 primary cases per model. Confirmed errors exclude unresolved outcomes.

Configuration	Correct labels	Confirmed errors	Unresolved	Preference score
o4-mini	74.8% 374/500	126/500	0	72.40
DeepSeek V4 Flash	Unavailable 374/500 confirmed 1 unresolved	125/500	1	Unranked
Qwen3.7-Flash	75.4% 377/500	123/500	0	76.65
Nemotron 3.5 Lightning	76.0% 380/500	120/500	0	Unranked
GLM-5.3-Flash	77.8% 389/500	111/500	0	82.06

O4-MINI

Fully evidenced under the original arm-level rules; no more primary errors than the incumbent.

DEEPSEEK V4 FLASH

unresolved outputs; incomplete all attempt pricing; Unresolved or missing full-run evidence; Incomplete full-run charges

QWEN3.7-FLASH

Fully evidenced under the original arm-level rules; no more primary errors than the incumbent.

NEMOTRON 3.5 LIGHTNING

incomplete all attempt pricing; Incomplete full-run charges

GLM-5.3-FLASH

Fully evidenced under the original arm-level rules; no more primary errors than the incumbent.

Where an outcome is unresolved, full-cohort accuracy is unavailable. Confirmed correct and error counts are observations; the unresolved count is separate. No timeout is silently converted into a wrong reference label.

Agreement with explicit denominators.

Conditional agreement excludes unusable and unresolved pairs. It is not the scoring input.

Configuration	End-to-end consistency	Valid-answer agreement	Valid-output rate
o4-mini	100.0% 500/500	100.0% 500/500 both-valid	100.0% 500/500
DeepSeek V4 Flash	Unavailable 437/500 confirmed 1 unresolved	87.8% 437/498 both-valid	Unavailable 498/500 confirmed 1 unresolved
Qwen3.7-Flash	84.4% 422/500	84.6% 422/499 both-valid	99.8% 499/500
Nemotron 3.5 Lightning	83.8% 419/500	84.3% 419/497 both-valid	99.4% 497/500
GLM-5.3-Flash	84.4% 422/500	84.4% 422/500 both-valid	100.0% 500/500

Unresolved configuration	Accuracy bounds	Agreement bounds	Valid-output bounds
DeepSeek V4 Flash	74.8% to 75.0%	87.4% to 87.6%	99.6% to 99.8%

These arithmetic bounds cover the possible outcomes of unresolved cases; they are not confidence intervals. A bound is not substituted into a weighted score. The CSV retains unrounded endpoints.

HOW TO READ AGREEMENT

End-to-end consistency divides matching valid labels by all planned cases. Valid-answer agreement divides matches by observed pairs where both labels are accepted. Both may include shared wrong answers. Valid-output rate measures accepted labels, which can still be wrong.

Excluding failed or unknown answers can inflate conditional agreement. Read it with valid-output coverage and unresolved counts. An incomplete whole-cohort rate remains unavailable.

Retries and their outcomes.

One retry at twice the initial output limit, only after a returned, billed response reports that it reached the output limit. The model, provider, prompt and reasoning settings stay the same. Wrong labels, upstream errors and timeouts do not trigger retries. A second failure is retained; there is no third attempt.

500 PRIMARY CASES PER MODEL

Recovered valid answers can still be wrong. Correct recoveries are checked against the reference label.

Configuration	First valid	First cap / other / unknown	Retries / required	Recovered valid / correct	Retry failed / unresolved	HTTP calls
o4-mini	499/500	1 / 0 / 0	1/1	1 / 0	0 / 0 0 charge unknown	501
DeepSeek V4 Flash	498/500	0 / 1 / 1	0/0	0 / 0	0 / 0 0 charge unknown	500
Qwen3.7-Flash	499/500	0 / 1 / 0	0/0	0 / 0	0 / 0 0 charge unknown	500
Nemotron 3.5 Lightning	497/500	0 / 3 / 0	0/0	0 / 0	0 / 0 0 charge unknown	500
GLM-5.3-Flash	500/500	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	500

Separate challenge recovery.

30 cases per model, excluded from the primary score.

30 SEPARATE CHALLENGE CASES PER MODEL

Recovered valid answers can still be wrong. Correct recoveries are checked against the reference label.

Configuration	First valid	First cap / other / unknown	Retries / required	Recovered valid / correct	Retry failed / unresolved	HTTP calls
o4-mini	30/30	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	30
DeepSeek V4 Flash	30/30	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	30
Qwen3.7-Flash	30/30	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	30
Nemotron 3.5 Lightning	30/30	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	30
GLM-5.3-Flash	30/30	0 / 0 / 0	0/0	0 / 0	0 / 0 0 charge unknown	30

Known first-pass failures remain in this recovery record even when a retry returns a valid label. The final quality table uses one result per case under the declared policy. A pending or budget-refused required retry is incomplete evidence.

A known upstream rejection is a failed answer even when its bill is unknown. Unknown charges are shown separately from unresolved answers and prevent ranking.

Recovery means that a retry returned a usable or correct answer under the stated policy. It does not show that the higher output limit caused that recovery. Repeated calls can differ, and a retry may finish within the original limit.

Cost includes every attempt, including the failed first response. Time is reconstructed per case by adding its recorded HTTP-attempt durations. It excludes automatic cooldown waits and manual pauses between collection segments, so it is not observed end-to-end customer latency. Earlier segments used different client deadlines and timing instrumentation; this is not a controlled model-speed comparison.

Reported charges. Qualified timing.

All-attempt charges and cumulative HTTP time per case, excluding cooldowns and manual pauses.

Configuration	Known primary charges	Fully priced cases	USD / 1,000	p50 / p95 Cumulative HTTP time
o4-mini	$0.8806	500/500	$1.7612	2.07 s / 5.87 s 500/500 timed
DeepSeek V4 Flash	$0.0732	499/500	Unknown	5.65 s / 24.16 s 500/500 timed
Qwen3.7-Flash	$0.0367	500/500	$0.0735	4.47 s / 14.57 s 500/500 timed
Nemotron 3.5 Lightning	$0.1256	499/500	Unknown	3.69 s / 13.23 s 500/500 timed
GLM-5.3-Flash	$0.0141	500/500	$0.0282	1.32 s / 3.03 s 500/500 timed

FULL-RUN ACCOUNTING

Known reported subtotal: $1.1844958149. Unknown attempted charges: 2. Held uncertainty reserve: $0.006677. The reserve is a budget allowance, not a measured bill. The known subtotal is not a complete total or final invoice.

Per-1,000 cost stays unavailable for a phase with unknown attempted charges. No missing charge becomes zero. Complete-arm values normalize actual primary charges, not a measured scaled workload.

TIMING PROTOCOL

Cost includes every attempt, including the failed first response. Time is reconstructed per case by adding its recorded HTTP-attempt durations. It excludes automatic cooldown waits and manual pauses between collection segments, so it is not observed end-to-end customer latency. Earlier segments used different client deadlines and timing instrumentation; this is not a controlled model-speed comparison.

Percentiles use linear interpolation over each case's summed HTTP-attempt time. They include failed first attempts and exclude automatic cooldowns and manual intervention gaps. They do not measure uninterrupted customer latency.

Errors, unknowns and changed answers.

Primary cohort. Known failure categories remain separate from unresolved evidence.

Configuration	Wrong label	Invalid output	Provider / unusable	Unresolved
o4-mini	126	0	0	0
DeepSeek V4 Flash	124	1	0	1
Qwen3.7-Flash	122	1	0	0
Nemotron 3.5 Lightning	117	1	2	0
GLM-5.3-Flash	111	0	0	0

Configuration	Both correct	Corrections	Regressions	Shared wrong	Different wrong
o4-mini	374	0	0	126	0
DeepSeek V4 Flash	354	20	20	83	21
Qwen3.7-Flash	350	27	23	72	27
Nemotron 3.5 Lightning	352	28	21	67	29
GLM-5.3-Flash	354	35	20	68	23

OUTPUT-LIMIT DIAGNOSTICS

1 output-limit truncation(s) were reported across all arms and cohorts. These are returned unusable responses, not upstream rejections or an unknown transport outcome.

A client transport timeout does not reveal the model answer or its final charge. Invalid labels, billed unusable responses, explicit upstream rejections and unresolved transport outcomes are different categories.

Separate checks. Unchanged priorities.

30 challenge cases per model, excluded from the primary weighted score.

Configuration	Correct labels	Confirmed errors	Unresolved	End-to-end agreement
o4-mini	83.3% 25/30	5/30	0	100.0% 30/30
DeepSeek V4 Flash	83.3% 25/30	5/30	0	93.3% 28/30
Qwen3.7-Flash	80.0% 24/30	6/30	0	96.7% 29/30
Nemotron 3.5 Lightning	83.3% 25/30	5/30	0	93.3% 28/30
GLM-5.3-Flash	76.7% 23/30	7/30	0	90.0% 27/30

Profile	Quality / agreement / cost / time	Preferred fully evidenced 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

Complete outcomes, prices, timing and identity across all 530 cases remain necessary for an arm to rank, with no more primary errors than the incumbent. The incumbent must itself be fully evidenced. Co-winners are retained; unranked arms do not have their weights redistributed.

The same cases. A declared recovery policy.

One retry at twice the initial output limit, only after a returned, billed response reports that it reached the output limit. The model, provider, prompt and reasoning settings stay the same. Wrong labels, upstream errors and timeouts do not trigger retries. A second failure is retained; there is no third attempt.

Segment	Attempts	Client deadline	Window / prior pause
Original prefix	248	120 s	2026-10-13 01:45:24 UTC 2026-10-13 02:07:38 UTC No prior segment
First continuation	1091	300 s	2026-10-13 02:26:08 UTC 2026-10-13 04:01:17 UTC Prior pause 1110.32 s
Adaptive continuation	28	300 s	2026-10-13 04:21:57 UTC 2026-10-13 04:26:58 UTC Prior pause 1240.53 s
Collection continuation	1284	300 s	2026-10-13 04:49:13 UTC 2026-10-13 06:29:05 UTC Prior pause 1335.46 s

After a verified upstream rejection, collection paused before continuing with untouched cases. Rejected calls were not retried: they remain failed answers with unknown charges reserved. Automatic cooldowns and a per-model circuit limit changed collection operations, not the model settings, scoring weights or output-limit retry policy.

Exact model ID	Provider route	Reasoning	Output cap
openai/o4-mini	openai	Enabled	2048 / 4096 retry
deepseek/deepseek-v4-flash-0731	streamlake/fp8	Low effort	8192 / 16384 retry
qwen/qwen3.7-flash	alibaba/us-east-1	1,024 reasoning tokens	2048 / 4096 retry
nvidia/nemotron-3.5-lightning	deepinfra/bf16	Enabled	8192 / 16384 retry
z-ai/glm-5.3-flash	z-ai/fp8	Low effort	2048 / 4096 retry

Reasoning can consume the output allowance before a usable label is returned. Initial and retry caps are maximum token budgets, not the measured length of the visible answer. First-attempt evidence remains preserved.

A separate report. Traceable evidence.

The amended report preserves the earlier frozen records and their identities.

The adaptive policy was declared after operational failures and before inspecting label quality. This new configuration can be reported only after all first attempts and every required retry have durable results. A pending or budget-refused retry prevents publication.

Reused public Banking77 cases and unchanged publisher labels support descriptive comparison only. They are not fresh held-out confirmation, customer adjudication, production traffic or rollout approval. The October 10 archive and historical retrieved-Luna assessment remain separate.

PREPARED

2026-10-13 06:43:08 UTC

INDEPENDENTLY AUDITED COMBINED-SUMMARY SHA-256

31d2f1a4ddc87eb9a6682c2c69f51a679d0c51f91bac9279772aa51c3d209b1d

REPORTING AMENDMENT SHA-256

b940a96188f7adfa9affcc629f668e90be015faea136c126847b76ff1b726da9

ORIGINAL PLAN SHA-256

fc86a515c397671c38d6ed40083938075dcc30df692c771de7b660f66a25d193

UNTOUCHED-SUFFIX PLAN SHA-256

fe1274037d6c43988e0f1f8035ec130e731a9df694dd1cbe4a46a4d2d98955a1

ADAPTIVE PLAN / TERMINAL SHA-256

9233745dc9c1a76050b1d2922aac48dcc4454ab2831b20fc29dec5c11b4c91b5
b02ea1b381ba9ed0934c7c66857f42d4b76c9f3aa69f51ce44680b5c2b9e111e

COLLECTION PLAN / TERMINAL SHA-256

fa28afcfde818f2234f7d94fc946db34b2d7254acca6a8fc1b9ea7a829a12c33
3d7f93d43856670f0ca5fb77783384308fab20c25825da2c6e50de11235ce097

COLLECTION AMENDMENT SHA-256

6a6da5802d9cdab3a8f08c1e43499fae4436592bba7579394a660daf4cfb46f8

ORIGINAL TERMINAL / SUFFIX TERMINAL SHA-256

011822529936c0e244456e1abce733cac12bd44b70c9337c4e3ad83d75e26c91
78cf2288bb9462aa0fb0813c6ce49b3768832c81c271bf77af5e7a0af5a6e1b9

Companion CSV preserves unrounded metrics, unknown values, arithmetic bounds, configuration settings and source identity. Curated JSON contains presentation inputs; the manifest pins exact file bytes. Private paths, account identifiers and raw responses remain outside the public packet.

Matching and correct

Both models return the reference intent. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

DeepSeek V4 Flash / banking77:test:20

I did not get my card yet, is it lost?

Reference intent: card_arrival

o4-mini / incumbent	DeepSeek V4 Flash / candidate
Visible answer {"intent":"card_arrival"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_arrival Correct against the reference	Parsed intent: card_arrival Correct against the reference

Qwen3.7-Flash / banking77:test:2

I ordered a card but it has not arrived. Help please!

Reference intent: card_arrival

o4-mini / incumbent	Qwen3.7-Flash / candidate
Visible answer {"intent":"card_arrival"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_arrival Correct against the reference	Parsed intent: card_arrival Correct against the reference

Matching and correct

Both models return the reference intent. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

Nemotron 3.5 Lightning / banking77:test:2

I ordered a card but it has not arrived. Help please!

Reference intent: card_arrival

o4-mini / incumbent	Nemotron 3.5 Lightning / candidate
Visible answer {"intent":"card_arrival"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_arrival Correct against the reference	Parsed intent: card_arrival Correct against the reference

GLM-5.3-Flash / banking77:test:2

I ordered a card but it has not arrived. Help please!

Reference intent: card_arrival

o4-mini / incumbent	GLM-5.3-Flash / candidate
Visible answer {"intent":"card_arrival"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_arrival Correct against the reference	Parsed intent: card_arrival Correct against the reference

Matching and wrong

Both models return the same intent, but it differs from the reference. Matching is not necessarily correctness. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

DeepSeek V4 Flash / banking77:test:21

Status of the card I ordered.

Reference intent: card_arrival

o4-mini / incumbent	DeepSeek V4 Flash / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_delivery_estimate"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_delivery_estimate Incorrect against the reference

Qwen3.7-Flash / banking77:test:27

I'm still waiting for delivery of my new card, why is it taking so long?

Reference intent: card_arrival

o4-mini / incumbent	Qwen3.7-Flash / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_delivery_estimate"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_delivery_estimate Incorrect against the reference

Matching and wrong

Both models return the same intent, but it differs from the reference. Matching is not necessarily correctness. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

Nemotron 3.5 Lightning / banking77:test:27

I'm still waiting for delivery of my new card, why is it taking so long?

Reference intent: card_arrival

o4-mini / incumbent	Nemotron 3.5 Lightning / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_delivery_estimate"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_delivery_estimate Incorrect against the reference

GLM-5.3-Flash / banking77:test:21

Status of the card I ordered.

Reference intent: card_arrival

o4-mini / incumbent	GLM-5.3-Flash / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_delivery_estimate"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_delivery_estimate Incorrect against the reference

A corrected incumbent error

The candidate returns the reference intent; the incumbent does not. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

DeepSeek V4 Flash / banking77:test:43

I have received my card, can you help me put it in the app?

Reference intent: card_linking

o4-mini / incumbent	DeepSeek V4 Flash / candidate
Visible answer {"intent":"activate_my_card"}	Visible answer {"intent":"card_linking"}
Parsed intent: activate_my_card Incorrect against the reference	Parsed intent: card_linking Correct against the reference

Qwen3.7-Flash / banking77:test:21

Status of the card I ordered.

Reference intent: card_arrival

o4-mini / incumbent	Qwen3.7-Flash / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_arrival Correct against the reference

A corrected incumbent error

The candidate returns the reference intent; the incumbent does not. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

Nemotron 3.5 Lightning / banking77:test:21

Status of the card I ordered.

Reference intent: card_arrival

o4-mini / incumbent	Nemotron 3.5 Lightning / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_arrival Correct against the reference

GLM-5.3-Flash / banking77:test:27

I'm still waiting for delivery of my new card, why is it taking so long?

Reference intent: card_arrival

o4-mini / incumbent	GLM-5.3-Flash / candidate
Visible answer {"intent":"card_delivery_estimate"}	Visible answer {"intent":"card_arrival"}
Parsed intent: card_delivery_estimate Incorrect against the reference	Parsed intent: card_arrival Correct against the reference

A new error against the reference

The incumbent returns the reference intent; the candidate does not. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

DeepSeek V4 Flash / banking77:test:2

I ordered a card but it has not arrived. Help please!

Reference intent: card_arrival

o4-mini / incumbent	DeepSeek V4 Flash / candidate
Visible answer {"intent":"card_arrival"}	Visible answer {"intent":"card_delivery_estimate"}
Parsed intent: card_arrival Correct against the reference	Parsed intent: card_delivery_estimate Incorrect against the reference

Qwen3.7-Flash / banking77:test:243

Does your system support multiple currency.

Reference intent: fiat_currency_support

o4-mini / incumbent	Qwen3.7-Flash / candidate
Visible answer {"intent":"fiat_currency_support"}	Visible answer {"intent":"supported_cards_and_currencies"}
Parsed intent: fiat_currency_support Correct against the reference	Parsed intent: supported_cards_and_currencies Incorrect against the reference

A new error against the reference

The incumbent returns the reference intent; the candidate does not. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

Nemotron 3.5 Lightning / banking77:test:137

Why is your exchange rate for my card payment wrong

Reference intent: card_payment_wrong_exchange_rate

o4-mini / incumbent	Nemotron 3.5 Lightning / candidate
Visible answer {"intent":"card_payment_wrong_exchange_rate"}	Visible answer {"intent":"exchange_rate"}
Parsed intent: card_payment_wrong_exchange_rate Correct against the reference	Parsed intent: exchange_rate Incorrect against the reference

GLM-5.3-Flash / banking77:test:243

Does your system support multiple currency.

Reference intent: fiat_currency_support

o4-mini / incumbent	GLM-5.3-Flash / candidate
Visible answer {"intent":"fiat_currency_support"}	Visible answer {"intent":"supported_cards_and_currencies"}
Parsed intent: fiat_currency_support Correct against the reference	Parsed intent: supported_cards_and_currencies Incorrect against the reference

Different answers, both wrong

Both answers differ from the reference and from each other. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

DeepSeek V4 Flash / banking77:test:382

I broke my card

Reference intent: card_not_working

o4-mini / incumbent	DeepSeek V4 Flash / candidate
Visible answer {"intent":"getting_spare_card"}	Visible answer {"intent":"order_physical_card"}
Parsed intent: getting_spare_card Incorrect against the reference	Parsed intent: order_physical_card Incorrect against the reference

Qwen3.7-Flash / banking77:test:586

what is the charges for new card?

Reference intent: contactless_not_working

o4-mini / incumbent	Qwen3.7-Flash / candidate
Visible answer {"intent":"order_physical_card"}	Visible answer {"intent":"card_arrival"}
Parsed intent: order_physical_card Incorrect against the reference	Parsed intent: card_arrival Incorrect against the reference

Different answers, both wrong

Both answers differ from the reference and from each other. Selected primary-cohort examples; no frequency or prevalence claim.

Selection: lowest numeric case ID in each candidate/category. Matching means exact parsed intent-label agreement, not identical answer text. Invalid, provider-failed, unresolved and challenge outputs are excluded.

Nemotron 3.5 Lightning / banking77:test:255

How do I change currencies to euros?

Reference intent: fiat_currency_support

o4-mini / incumbent	Nemotron 3.5 Lightning / candidate
Visible answer {"intent":"exchange_via_app"}	Visible answer {"intent":"exchange_rate"}
Parsed intent: exchange_via_app Incorrect against the reference	Parsed intent: exchange_rate Incorrect against the reference

GLM-5.3-Flash / banking77:test:586

what is the charges for new card?

Reference intent: contactless_not_working

o4-mini / incumbent	GLM-5.3-Flash / candidate
Visible answer {"intent":"order_physical_card"}	Visible answer {"intent":"getting_spare_card"}
Parsed intent: order_physical_card Incorrect against the reference	Parsed intent: getting_spare_card Incorrect against the reference

Source: the audited primary comparison and unchanged Banking77 test labels. Companion case-examples.json preserves exact visible text and request, response and receipt fingerprints; case-examples.csv provides the same twenty examples. Model and route settings are listed in the collection section.