Evidence appendix

12 configurations. 500 primary and 30 separate challenge cases each.

RECOMMENDATION

GPT-6 Luna

The report uses the same frozen public inputs, prompt and publisher references across configurations. 5 earlier model records are retained unchanged; 7 additional models were collected later. No incumbent requests were repeated for this extension.

The operational prefix produced five strict-format violations. Its original failed readiness gate remains in the evidence. A separately declared continuation collected only the untouched suffix. All prefix answers and charges are retained; no prompt tuning or output-based retries were introduced.

The unchanged native parser accepts fenced JSON and records it as a separate formatting violation. Prose around JSON is invalid. Intent accuracy and strict response formatting are distinct; the parser was not relaxed for these models.

Configurations and collection dates 2

Primary measurements and exclusions 5

Agreement and output coverage 6

Challenge cases and preference profiles 9

Recovery and timing 12

Recorded matching and regression examples 14

Collection history and rejected requests 25

Source identity 27

Declared configurations

Provider settings describe tested configurations, not an equal compute budget.

Model	Route / reasoning	First / retry cap	Collection / origin
o4-mini	openai Enabled	2,048 / 4,096	2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC Retained
DeepSeek V4 Flash	streamlake/fp8 Low effort	8,192 / 16,384	2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC Retained
Qwen3.7-Flash	alibaba/us-east-1 1,024 reasoning tokens	2,048 / 4,096	2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC Retained
Nemotron 3.5 Lightning	deepinfra/bf16 Enabled	8,192 / 16,384	2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC Retained
GLM-5.3-Flash	z-ai/fp8 Low effort	2,048 / 4,096	2026-10-13 01:45:24 UTC to 2026-10-13 06:29:05 UTC Retained
Gemini 3.5 Flash Lite	google-ai-studio Low effort	2,048 / 4,096	2026-10-13 16:22:28 UTC to 2026-10-13 19:11:44 UTC New collection
Gemini 3.8 Flash	google-ai-studio Low effort	2,048 / 4,096	2026-10-13 16:22:29 UTC to 2026-10-13 19:11:46 UTC New collection
GPT-6 Luna	openai Low effort	2,048 / 4,096	2026-10-13 16:22:31 UTC to 2026-10-13 19:11:47 UTC New collection
GPT-6.1 Sol	openai Low effort	2,048 / 4,096	2026-10-13 16:22:33 UTC to 2026-10-13 19:11:49 UTC New collection
Claude Haiku 5.5	anthropic Low effort	2,048 / 4,096	2026-10-13 16:22:36 UTC to 2026-10-13 19:11:37 UTC New collection

Configurations / continued

Model	Route / reasoning	First / retry cap	Collection / origin
Claude Sonnet 5.5	anthropic Low effort	2,048 / 4,096	2026-10-13 16:22:37 UTC to 2026-10-13 19:11:39 UTC New collection
Grok 4.7	xai Low effort	2,048 / 4,096	2026-10-13 16:22:40 UTC to 2026-10-13 19:11:43 UTC New collection

Routes are pinned without automatic fallback. A larger output allowance is used only after a verified length-limit finish, once per case. All attempt charges and HTTP time are retained. Low-effort settings across providers do not imply equal reasoning compute.

Response format diagnostics

Newly collected models only. Intent-label scoring remains unchanged.

Model / cohort	Strict-format violations	Valid fenced labels
Gemini 3.5 Flash Lite / primary	132	132
Gemini 3.5 Flash Lite / challenge	14	14
Gemini 3.8 Flash / primary	3	3
Gemini 3.8 Flash / challenge	0	0
GPT-6 Luna / primary	0	0
GPT-6 Luna / challenge	0	0
GPT-6.1 Sol / primary	0	0
GPT-6.1 Sol / challenge	0	0
Claude Haiku 5.5 / primary	7	0
Claude Haiku 5.5 / challenge	0	0
Claude Sonnet 5.5 / primary	109	0
Claude Sonnet 5.5 / challenge	5	0
Grok 4.7 / primary	0	0
Grok 4.7 / challenge	0	0

Fenced JSON can supply a valid intent under the frozen parser while violating the strict one-key JSON contract. Prose around JSON is invalid. These diagnostic counts are not added to the intent-error counts. Historical formatting counts are not inferred as zero.

Primary comparison

500 cases per configuration. Unknown outcomes and charges are not zero.

Model	Errors	Consistency	USD / 1,000	p50 / p95	Score / status
o4-mini	126	100.0%	$1.7612	2.07 s / 5.87 s	72.40
DeepSeek V4 Flash	125 + 1?	437/500 known	Unknown	5.65 s / 24.16 s	Charge gap
Qwen3.7-Flash	123	84.4%	$0.0735	4.47 s / 14.57 s	76.65
Nemotron 3.5 Lightning	120	83.8%	Unknown	3.69 s / 13.23 s	Charge gap
GLM-5.3-Flash	111	84.4%	$0.0282	1.32 s / 3.03 s	82.06
Gemini 3.5 Flash Lite	117	85.0%	$0.2319	0.65 s / 0.97 s	81.56
Gemini 3.8 Flash	82	84.0%	$1.0833	2.09 s / 4.38 s	76.71
GPT-6 Luna	95	84.8%	$0.0842	2.52 s / 4.08 s	82.45
GPT-6.1 Sol	86	83.4%	$1.3176	2.07 s / 4.53 s	75.17
Claude Haiku 5.5	115	81.8%	Unknown	0.98 s / 2.83 s	Charge gap
Claude Sonnet 5.5	154	73.6%	$2.2768	1.68 s / 3.72 s	More errors
Grok 4.7	98	88.0%	$3.2315	2.81 s / 9.59 s	68.65

Consistency: matching valid intent labels divided by 500, including shared errors. Errors exclude unresolved outcomes; the main table marks these with ?. Cost is normalized from complete recorded charges. p50/p95 are HTTP seconds per logical case, summing attempts and excluding scheduling waits.

All 12 compared model configurations; primary and challenge cases. Known reported subtotal: $5.5974860649; retained uncertainty reserve: $0.013845. A reserve is not an observed charge. Subtotals exclude unknown charges and are not a final invoice.

DEEPSEEK V4 FLASH / UNRANKED

unresolved outputs; incomplete all attempt pricing; full run unresolved or missing; full run incomplete all attempt pricing

NEMOTRON 3.5 LIGHTNING / UNRANKED

incomplete all attempt pricing; full run incomplete all attempt pricing

CLAUDE HAIKU 5.5 / UNRANKED

incomplete all attempt pricing; full run incomplete all attempt pricing

CLAUDE SONNET 5.5 / UNRANKED

error rate worse than incumbent

Primary answer coverage

500 cases per model. Conditional agreement is read alongside output coverage.

Model	End-to-end consistency	Valid-answer agreement	Valid-output rate
o4-mini	100.0% 500/500	100.0% 500/500 valid pairs	100.0% 500/500
DeepSeek V4 Flash	Unavailable 437/500 confirmed; 1 unresolved	87.8% 437/498 valid pairs	Unavailable 498/500 confirmed; 1 unresolved
Qwen3.7-Flash	84.4% 422/500	84.6% 422/499 valid pairs	99.8% 499/500
Nemotron 3.5 Lightning	83.8% 419/500	84.3% 419/497 valid pairs	99.4% 497/500
GLM-5.3-Flash	84.4% 422/500	84.4% 422/500 valid pairs	100.0% 500/500
Gemini 3.5 Flash Lite	85.0% 425/500	85.0% 425/500 valid pairs	100.0% 500/500
Gemini 3.8 Flash	84.0% 420/500	84.0% 420/500 valid pairs	100.0% 500/500
GPT-6 Luna	84.8% 424/500	84.8% 424/500 valid pairs	100.0% 500/500
GPT-6.1 Sol	83.4% 417/500	83.4% 417/500 valid pairs	100.0% 500/500
Claude Haiku 5.5	81.8% 409/500	83.6% 409/489 valid pairs	97.8% 489/500
Claude Sonnet 5.5	73.6% 368/500	94.1% 368/391 valid pairs	78.2% 391/500
Grok 4.7	88.0% 440/500	88.0% 440/500 valid pairs	100.0% 500/500

Valid-answer agreement excludes invalid, failed and unresolved pairs. A high conditional rate can conceal poor output coverage. End-to-end consistency remains the scored metric; shared wrong labels can still match. Unresolved outcomes are unknown, not confirmed failures.

Challenge answer coverage

30 cases per model. Conditional agreement is read alongside output coverage.

Model	End-to-end consistency	Valid-answer agreement	Valid-output rate
o4-mini	100.0% 30/30	100.0% 30/30 valid pairs	100.0% 30/30
DeepSeek V4 Flash	93.3% 28/30	93.3% 28/30 valid pairs	100.0% 30/30
Qwen3.7-Flash	96.7% 29/30	96.7% 29/30 valid pairs	100.0% 30/30
Nemotron 3.5 Lightning	93.3% 28/30	93.3% 28/30 valid pairs	100.0% 30/30
GLM-5.3-Flash	90.0% 27/30	90.0% 27/30 valid pairs	100.0% 30/30
Gemini 3.5 Flash Lite	93.3% 28/30	93.3% 28/30 valid pairs	100.0% 30/30
Gemini 3.8 Flash	93.3% 28/30	93.3% 28/30 valid pairs	100.0% 30/30
GPT-6 Luna	90.0% 27/30	90.0% 27/30 valid pairs	100.0% 30/30
GPT-6.1 Sol	86.7% 26/30	86.7% 26/30 valid pairs	100.0% 30/30
Claude Haiku 5.5	83.3% 25/30	83.3% 25/30 valid pairs	100.0% 30/30
Claude Sonnet 5.5	83.3% 25/30	100.0% 25/25 valid pairs	83.3% 25/30
Grok 4.7	100.0% 30/30	100.0% 30/30 valid pairs	100.0% 30/30

Valid-answer agreement excludes invalid, failed and unresolved pairs. A high conditional rate can conceal poor output coverage. End-to-end consistency remains the scored metric; shared wrong labels can still match. Unresolved outcomes are unknown, not confirmed failures.

What changed relative to the incumbent

Only pairs with valid labels contribute to these five categories.

Candidate	Both correct	Corrections	Regressions	Same wrong	Different wrong
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
Gemini 3.5 Flash Lite	355	28	19	70	28
Gemini 3.8 Flash	370	48	4	50	28
GPT-6 Luna	362	43	12	62	21
GPT-6.1 Sol	365	49	9	52	25
Claude Haiku 5.5	347	38	22	62	20
Claude Sonnet 5.5	330	16	2	38	5
Grok 4.7	368	34	6	72	20

Correction: candidate correct, incumbent wrong. Regression: incumbent correct, candidate wrong. Missing, invalid, failed and unresolved outputs stay in aggregate results but cannot form a pair of valid labels.

Separate challenge results

30 cases per configuration; excluded from the primary weighted score.

Model	Correct	Confirmed errors	Unresolved
o4-mini	25/30	5	0
DeepSeek V4 Flash	25/30	5	0
Qwen3.7-Flash	24/30	6	0
Nemotron 3.5 Lightning	25/30	5	0
GLM-5.3-Flash	23/30	7	0
Gemini 3.5 Flash Lite	27/30	3	0
Gemini 3.8 Flash	27/30	3	0
GPT-6 Luna	28/30	2	0
GPT-6.1 Sol	27/30	3	0
Claude Haiku 5.5	24/30	6	0
Claude Sonnet 5.5	23/30	7	0
Grok 4.7	25/30	5	0

Profile	Accuracy / agreement / cost / p95	Preferred eligible models
default	50% / 20% / 20% / 10%	GPT-6 Luna
quality	70% / 10% / 10% / 10%	GPT-6 Luna
economy	50% / 10% / 30% / 10%	GPT-6 Luna
latency	50% / 10% / 10% / 30%	Gemini 3.5 Flash Lite

Cost and timing utility = incumbent / (incumbent + candidate); equal values score 0.5, including both zero. Missing evidence never redistributes weights. Eligibility requires complete outcomes, charges, timing and identity across all 530 cases, with no more primary errors than the incumbent. Ties and retaining the incumbent remain available.

How priorities change the choice

Illustrative post-collection analysis. The recorded default and original profiles above are unchanged.

Priority	Correctness / consistency / cost / p95 time	Preferred model / score	Vs default
Default	50% / 20% / 20% / 10%	GPT-6 Luna / 82.45	Baseline
Correctness	80% / 10% / 5% / 5%	Gemini 3.8 Flash / 81.24	Changes
Consistency	30% / 50% / 10% / 10%	Gemini 3.5 Flash Lite / 82.90	Changes
Cost	30% / 10% / 50% / 10%	GLM-5.3-Flash / 87.59	Changes
Latency	30% / 10% / 10% / 50%	Gemini 3.5 Flash Lite / 83.23	Changes

Weights total 100%. Correctness is agreement with the reference label; consistency is agreement with the incumbent, including shared wrong labels. Cost and p95 time use the same incumbent-relative utilities as the recorded profiles. Score = 100 times the weighted sum of these four components.

The four added scenarios emphasize a business priority; they were not prospective acceptance policies. Every eligible model, including the incumbent, remains in the comparison. Missing evidence is never filled or reweighted. Scores and margins are preference points within one profile, not confidence or statistical superiority. These five examples do not establish stability over all possible weights.

Keeping the incumbent is an available choice and gives self-agreement of 100%. A consistency-led preference can preserve existing errors. The remaining weights still matter: the highest-consistency or lowest-error model need not lead a mixed profile.

Gemini 3.8 Flash: 3/500 strict-format violations; intent scoring still accepts valid fenced JSON. Validate the output contract before use.

Gemini 3.5 Flash Lite: 132/500 strict-format violations; intent scoring still accepts valid fenced JSON. Validate the output contract before use.

Priority	Runner-up	Lead (points)
Default	GLM-5.3-Flash	0.39
Correctness	GPT-6 Luna	0.24
Consistency	o4-mini	0.46
Cost	GPT-6 Luna	1.19
Latency	GLM-5.3-Flash	8.62

Eligible models by priority

Read each column independently. Exclusions and measured components are unchanged.

Model	Default	Correctness	Consistency	Cost	Latency
o4-mini	72.40	74.84	82.44	62.44	62.44
Qwen3.7-Flash	76.65	75.00	77.29	81.93	55.02
GLM-5.3-Flash	82.06	78.90	81.98	87.59	74.60
Gemini 3.5 Flash Lite	81.56	78.49	82.90	84.25	83.23
Gemini 3.8 Flash	76.71	81.24	79.00	70.17	68.31
GPT-6 Luna	82.45	81.00	82.15	86.40	71.84
GPT-6.1 Sol	75.17	80.26	77.91	67.43	67.14
Grok 4.7	68.65	76.78	75.45	54.36	55.44

Default weights 50/20/20/10; correctness 80/10/5/5; consistency 30/50/10/10; cost 30/10/50/10; latency 30/10/10/50. Order: correctness / consistency / cost / p95 time. The four original recorded profiles remain in the preceding challenge section.

Primary recovery

Only verified output-limit failures may receive one larger retry.

Model	Firsts	Cap failures	Retries required / sent	Recovered valid / correct	HTTP attempts
o4-mini	500/500	1	1 / 1	1 / 0	501
DeepSeek V4 Flash	500/500	0	0 / 0	0 / 0	500
Qwen3.7-Flash	500/500	0	0 / 0	0 / 0	500
Nemotron 3.5 Lightning	500/500	0	0 / 0	0 / 0	500
GLM-5.3-Flash	500/500	0	0 / 0	0 / 0	500
Gemini 3.5 Flash Lite	500/500	0	0 / 0	0 / 0	500
Gemini 3.8 Flash	500/500	0	0 / 0	0 / 0	500
GPT-6 Luna	500/500	0	0 / 0	0 / 0	500
GPT-6.1 Sol	500/500	0	0 / 0	0 / 0	500
Claude Haiku 5.5	500/500	0	0 / 0	0 / 0	500
Claude Sonnet 5.5	500/500	0	0 / 0	0 / 0	500
Grok 4.7	500/500	0	0 / 0	0 / 0	500

Recorded wrong or invalid outputs are not retried. A successful retry supplies the final intent; its original attempt remains in cost and timing. Unknown charges prevent ranking until resolved. Different collection windows make HTTP timing descriptive, not a causal production latency claim.

Challenge recovery

Only verified output-limit failures may receive one larger retry.

Model	Firsts	Cap failures	Retries required / sent	Recovered valid / correct	HTTP attempts
o4-mini	30/30	0	0 / 0	0 / 0	30
DeepSeek V4 Flash	30/30	0	0 / 0	0 / 0	30
Qwen3.7-Flash	30/30	0	0 / 0	0 / 0	30
Nemotron 3.5 Lightning	30/30	0	0 / 0	0 / 0	30
GLM-5.3-Flash	30/30	0	0 / 0	0 / 0	30
Gemini 3.5 Flash Lite	30/30	0	0 / 0	0 / 0	30
Gemini 3.8 Flash	30/30	0	0 / 0	0 / 0	30
GPT-6 Luna	30/30	0	0 / 0	0 / 0	30
GPT-6.1 Sol	30/30	0	0 / 0	0 / 0	30
Claude Haiku 5.5	30/30	0	0 / 0	0 / 0	30
Claude Sonnet 5.5	30/30	0	0 / 0	0 / 0	30
Grok 4.7	30/30	0	0 / 0	0 / 0	30

Recorded wrong or invalid outputs are not retried. A successful retry supplies the final intent; its original attempt remains in cost and timing. Unknown charges prevent ranking until resolved. Different collection windows make HTTP timing descriptive, not a causal production latency claim.

DeepSeek V4 Flash

Actual visible outputs. Lowest numeric case ID in each observed category.

498/500 valid pairs; 2 invalid, failed or unresolved pairs excluded. Both correct 354; regressions 20; corrections 20; same wrong 83; different wrong 21.

MATCHING / BOTH CORRECT

banking77:test:20: "I did not get my card yet, is it lost?"

Reference: card_arrival

o4-mini	DeepSeek V4 Flash
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	DeepSeek V4 Flash
{"intent":"card_arrival"} Correct	{"intent":"card_delivery_estimate"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Qwen3.7-Flash

Actual visible outputs. Lowest numeric case ID in each observed category.

499/500 valid pairs; 1 invalid, failed or unresolved pairs excluded. Both correct 350; regressions 23; corrections 27; same wrong 72; different wrong 27.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Qwen3.7-Flash
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:243: "Does your system support multiple currency."

Reference: fiat_currency_support

o4-mini	Qwen3.7-Flash
{"intent":"fiat_currency_support"} Correct	{"intent":"supported_cards_and_currencies"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Nemotron 3.5 Lightning

Actual visible outputs. Lowest numeric case ID in each observed category.

497/500 valid pairs; 3 invalid, failed or unresolved pairs excluded. Both correct 352; regressions 21; corrections 28; same wrong 67; different wrong 29.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Nemotron 3.5 Lightning
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:137: "Why is your exchange rate for my card payment wrong"

Reference: card_payment_wrong_exchange_rate

o4-mini	Nemotron 3.5 Lightning
{"intent":"card_payment_wrong_exchange_rate"} Correct	{"intent":"exchange_rate"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

GLM-5.3-Flash

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 354; regressions 20; corrections 35; same wrong 68; different wrong 23.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	GLM-5.3-Flash
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:243: "Does your system support multiple currency."

Reference: fiat_currency_support

o4-mini	GLM-5.3-Flash
{"intent":"fiat_currency_support"} Correct	{"intent":"supported_cards_and_currencies"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Gemini 3.5 Flash Lite

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 355; regressions 19; corrections 28; same wrong 70; different wrong 28.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Gemini 3.5 Flash Lite
{"intent":"card_arrival"} Correct	```json {"intent":"card_arrival"} ``` Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:560: "Are contactless payments enabled on my new card?"

Reference: contactless_not_working

o4-mini	Gemini 3.5 Flash Lite
{"intent":"contactless_not_working"} Correct	{"intent":"activate_my_card"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Gemini 3.8 Flash

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 370; regressions 4; corrections 48; same wrong 50; different wrong 28.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Gemini 3.8 Flash
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:901: "Which forms of currency are accepted?"

Reference: supported_cards_and_currencies

o4-mini	Gemini 3.8 Flash
{"intent":"supported_cards_and_currencies"} Correct	{"intent":"fiat_currency_support"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

GPT-6 Luna

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 362; regressions 12; corrections 43; same wrong 62; different wrong 21.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	GPT-6 Luna
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:274: "How man currencies can I hold?"

Reference: fiat_currency_support

o4-mini	GPT-6 Luna
{"intent":"fiat_currency_support"} Correct	{"intent":"supported_cards_and_currencies"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

GPT-6.1 Sol

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 365; regressions 9; corrections 49; same wrong 52; different wrong 25.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	GPT-6.1 Sol
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:696: "I want to reverse my transaction from earlier"

Reference: cancel_transfer

o4-mini	GPT-6.1 Sol
{"intent":"cancel_transfer"} Correct	{"intent":"request_refund"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Claude Haiku 5.5

Actual visible outputs. Lowest numeric case ID in each observed category.

489/500 valid pairs; 11 invalid, failed or unresolved pairs excluded. Both correct 347; regressions 22; corrections 38; same wrong 62; different wrong 20.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Claude Haiku 5.5
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:543: "Why did I get blocked?"

Reference: pin_blocked

o4-mini	Claude Haiku 5.5
{"intent":"pin_blocked"} Correct	{"intent":"card_not_working"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Claude Sonnet 5.5

Actual visible outputs. Lowest numeric case ID in each observed category.

391/500 valid pairs; 109 invalid, failed or unresolved pairs excluded. Both correct 330; regressions 2; corrections 16; same wrong 38; different wrong 5.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Claude Sonnet 5.5
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:406: "How are currencies exchanged?"

Reference: exchange_via_app

o4-mini	Claude Sonnet 5.5
{"intent":"exchange_via_app"} Correct	{"intent":"exchange_rate"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Grok 4.7

Actual visible outputs. Lowest numeric case ID in each observed category.

500/500 valid pairs; 0 invalid, failed or unresolved pairs excluded. Both correct 368; regressions 6; corrections 34; same wrong 72; different wrong 20.

MATCHING / BOTH CORRECT

banking77:test:2: "I ordered a card but it has not arrived. Help please!"

Reference: card_arrival

o4-mini	Grok 4.7
{"intent":"card_arrival"} Correct	{"intent":"card_arrival"} Correct

NON-MATCHING / INCUMBENT CORRECT, CANDIDATE WRONG

banking77:test:596: "How do I enable contactless pay?"

Reference: contactless_not_working

o4-mini	Grok 4.7
{"intent":"contactless_not_working"} Correct	{"intent":"apple_pay_or_google_pay"} Wrong

All available category examples and their source receipts are available in case-examples.json and case-examples.csv. Hidden reasoning is not displayed. Examples illustrate outcomes; counts describe their frequency.

Preserved attempts and billing

Original receipts are retained across all three collection segments.

Segment / status	Initial / retry attempts	Known charges / reserved
Operational prefix readiness failed	35 / 0	$0.04103170 $0
Stopped continuation stopped	1,740 / 0	$2.08393090 $0.001792
Final continuation completed with billing unknown	1,935 / 0	$2.28802765 $0.005376

A rejected request produced no answer. Its charge remains unknown and fully reserved. It counts as a failure; an incomplete charge prevents cost-based ranking. Rejected requests are not replayed. Cooldowns and operator holds are excluded from HTTP latency.

COLLECTION WINDOW

2026-10-13T16:22:28.628360+00:00 to 2026-10-13T16:23:47.795105+00:00

PLAN / JOURNAL SHA-256

7d754c636ae5406ccd0ed86dee815bcc7b1e112167c68dedc9df0d05ae8368f4
9a6a559f8bfe814bacb318498cfc7b81a5c6b6bec80436ef7dfb9feb6743c85d

PREFIX REVIEW SHA-256

5b6889ddf760f78012cf73f374ecc96e88823651fec1733a6b1ad98fb7580ddf

COLLECTION WINDOW

2026-10-13T16:35:11.311412+00:00 to 2026-10-13T17:40:22.287501+00:00

PLAN / JOURNAL SHA-256

e1cbd00495afb9c5aa74e2b024c39d0034eb1c70130dbdc55109b1d79661075b
a2069c72f60b76f91dd1badf8b46e6900cd2522823e7b3b1fe75766efef65ad8

STOPPED RESULT SHA-256

2a19b2b06747ff660401ac72a54cb257ed4194c6391eb8c36b9ed84db2f2ec20

COLLECTION WINDOW

2026-10-13T18:02:59.300269+00:00 to 2026-10-13T19:11:49.317240+00:00

PLAN / JOURNAL SHA-256

d0e53b2ccbc823fb7a960bf5cf31e12c3c95b2cfc7465691725175e81e7e649c
e2e0b650a36824e09db0d4544ad45d7dc8c42149aea7ec89906baf3421e68a1b

Collection history / continued

COMPLETED RESULT SHA-256

9629be29651ded2dc859fbbc6efb2ca0dd65e9a79d91831066da94286cb20563

HTTP 429 / NO ANSWER

Claude Haiku 5.5
banking77:test:2602; charge unknown; $0.001792 reserved.

RECEIPT / RAW RESPONSE SHA-256

3e779bd93881982832879d4175c760b5de39ab51ea75225b781857d073480ce5
659d11b962d28a9e94f6a85b3e36706522f07c7e3ade1100f0996b8baf95b1a4

HTTP 429 / NO ANSWER

Claude Haiku 5.5
banking77:test:2606; charge unknown; $0.001792 reserved.

RECEIPT / RAW RESPONSE SHA-256

e80f6b7da176ea0b1806f0a325e8653c74bf164164bf5fe9f8e9ba3d2bee5c77
f2f5403d6211249b3d22d3c459e0df158979d608ba6979f8a664c2bf8659c2ad

HTTP 429 / NO ANSWER

Claude Haiku 5.5
banking77:test:2608; charge unknown; $0.001792 reserved.

RECEIPT / RAW RESPONSE SHA-256

8c3796ec2ee0bdd2947fc8ef542631fa4b8533f504c1444823feea9e591144f5
c937e2ba8a5d6d4a29bfc6e0d75951b0547f828048b344c85836eee640156405

HTTP 429 / NO ANSWER

Claude Haiku 5.5
banking77:test:2612; charge unknown; $0.001792 reserved.

RECEIPT / RAW RESPONSE SHA-256

0661b90d0edb9af0033279c028c530a3d7fb72294c001737ce4f798f437e0cc8
33f9c90ff70fcf71e001931dcd9da3a862fba9594098d73d785dd3f57946e16c

CONTINUATION AMENDMENT SHA-256

7b9e5e5933782f87c18368a75c52d341e778b5a1ab27ba630a159d4483007014

Sources and limits

Traceable measurements with explicit evidence boundaries.

SUMMARY SHA-256

3b41e33295c60f5b80807797c9446aed1ff644c022af72442246dbcb770f91ca

RETAINED FIVE-MODEL SUMMARY

31d2f1a4ddc87eb9a6682c2c69f51a679d0c51f91bac9279772aa51c3d209b1d

EXTENSION PLAN

7d754c636ae5406ccd0ed86dee815bcc7b1e112167c68dedc9df0d05ae8368f4

UNTOUCHED CONTINUATION PLAN

e1cbd00495afb9c5aa74e2b024c39d0034eb1c70130dbdc55109b1d79661075b

READINESS AMENDMENT

c2e3f6e758920525a56cc1e22882a2aab8b562bda1e39195d7cc2e9c57e94b19

REJECTION CONTINUATION PLAN

d0e53b2ccbc823fb7a960bf5cf31e12c3c95b2cfc7465691725175e81e7e649c

REJECTION AMENDMENT

7b9e5e5933782f87c18368a75c52d341e778b5a1ab27ba630a159d4483007014

FINAL RESULT

9629be29651ded2dc859fbbc6efb2ca0dd65e9a79d91831066da94286cb20563

INDEPENDENT AUDIT

ee9496dae06a559b90191b577fab02107ed411979801b78523d4479175206cdb

AUDITED INPUT MANIFEST

9d5b9237701042b222b2ebcccb5754225f7baefaa22e09d828df36343d449b17

Dataset: Banking77, test split. Revision sha256:d12d6e3bc4c3103966ae786dc435913c0c563dfa328f5a3646d0e62cfeeb474d.

https://github.com/PolyAI-LDN/task-specific-datasets/tree/master/banking_data

Dataset license: https://github.com/PolyAI-LDN/task-specific-datasets/blob/master/LICENSE

Banking77 Demo Bank is illustrative. Public cases were reused and reference labels were not customer-adjudicated. Preference scores and arithmetic unknown-outcome ranges are not statistical confidence or rollout approval. Validate on a fresh representative customer workflow before migration.