Migration decision

Public benchmark / 2026-10-13 19:11:49 UTC / Illustrative customer

1 / Customer	2 / Identified workflow	3 / Current model
Banking77 Demo Bank	Support intent classification 77 labels; plain prompt	o4-mini

4 / RECOMMENDED MODEL

GPT-6 Luna

Conditional: workflow validation. Preference score 82.45/100, not confidence.

Default weights: correctness 50% / consistency 20% / cost 20% / p95 time 10%.

405/500 correct vs 374/500; 43 corrections, 12 regressions.

GPT-6 Luna leads GLM-5.3-Flash by 0.39 preference points; this does not establish statistical superiority. Gemini 3.8 Flash has the fewest primary errors: 82/500.

5 / OTHER MODELS EVALUATED / 500 PRIMARY CASES EACH

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

Claude Sonnet 5.5: 45 wrong labels + 109 invalid outputs. Claude Haiku 5.5: 104 wrong labels + 7 invalid outputs + 4 provider failures.

Failures and next action

Observed failures and the condition for moving forward.

6 / CURRENT-MODEL ISSUES

o4-mini: 126 wrong labels.

7 / RECOMMENDED-MODEL ISSUES

GPT-6 Luna: 95 wrong labels. 12 incumbent-correct regressions. Challenge: 28/30 correct vs 25/30.

Matching / both correct

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

Reference: card_arrival
o4-mini: card_arrival (correct)
GPT-6 Luna: card_arrival (correct)

Non-matching / incumbent correct

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

Reference: fiat_currency_support
o4-mini: fiat_currency_support (correct)
GPT-6 Luna: supported_cards_and_currencies (wrong)

8 / OTHER DETAILS AND NEXT ACTION

How priorities change the choice

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

Illustrative scenarios added after collection; default unchanged. Same eligibility and measurements. Scores compare models within a profile, not across profiles; they are not confidence.

Strict-format violations: Gemini 3.8 Flash 3/500; Gemini 3.5 Flash Lite 132/500. Intent scoring accepts valid fenced JSON. Validate the output contract.

Consistency includes shared wrong labels; the incumbent matches itself. Costs and HTTP time include retries; timing spans separate collection windows. ? = unresolved. Challenges are not ranked.

Next: agree priorities and acceptance thresholds, then validate on fresh workflow cases before a pilot.

Technical evidence, configurations and observed examples