Migration decision

Public benchmark / 2026-10-13 06:29:05 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

GLM-5.3-Flash

Recommend for workflow-specific validation. Conditional. Preference score 82.06/100, not confidence.

389/500 correct versus 374/500 for o4-mini: 35 corrections, 20 regressions. Lowest known cost and shortest p95 among fully evidenced configurations; preferred under all four declared weight profiles.

z-ai/fp8; low effort; 2,048-token cap, one length-only retry at 4,096. Challenge: 23/30 correct versus 25/30. Customer validation is still required.

5 / OTHER MODELS EVALUATED

Model	Errors / 500	Consistency	USD / 1,000	p95 HTTP	Score
o4-mini	126	100.0%	$1.7612	5.87 s	72.40
DeepSeek V4 Flash	125 confirmed 1 unresolved	437/500 known Incomplete	Unknown	24.16 s	Unranked
Qwen3.7-Flash	123	84.4%	$0.0735	14.57 s	76.65
Nemotron 3.5 Lightning	120	83.8%	Unknown	13.23 s	Unranked
GLM-5.3-Flash	111	84.4%	$0.0282	3.03 s	82.06

500 primary cases per model. Consistency = identical valid intent labels / 500; matching can be wrong. Weights: correctness 50%, consistency 20%, cost 20%, p95 10%.

DeepSeek has one unresolved answer/charge; Nemotron one known failed answer with unknown charge. Both are unranked. Cost is normalized known charges; HTTP time sums all attempts, excluding cooldowns and manual pauses; not production latency.

Failures and next action

Observed examples and the condition for moving forward.

6 / CURRENT-MODEL ISSUES

o4-mini: 126/500 errors. GLM corrects 35; 91 remain wrong in both.

7 / RECOMMENDED-MODEL ISSUES

GLM: 111/500 errors, including 20 incumbent-correct regressions. Challenge: 23/30 correct versus 25/30.

MATCHING / BOTH CORRECT / CASE 2

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, GLM WRONG / CASE 243

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

Illustrative lowest-ID cases, not prevalence estimates. Answers and reference labels are unchanged. The appendix includes all twenty paired examples.

8 / OTHER DETAILS AND NEXT ACTION

Validate the exact GLM configuration on fresh, representative workflow cases. Triage regressions and agree quality, latency and cost thresholds before a pilot; retain o4-mini until those conditions pass.

Reused public cases support workflow validation, not production approval. Dataset, configurations, scoring, billing uncertainty and recovery details are in the separate appendix.

Technical evidence, methodology and all twenty examples