

VANTAGE RUNTIMEAI · ABOUT

Market Reality

Model churn, pricing variance, and why manual gates cannot keep up

One gateway already holds **422 model routes** — **7.8× end-2024** — with input prices spanning **15,000×** and session costs **~17,000×**. On the same path: **~549** by YE '26 and **~1041** by YE '27. The market sells speed and rate cards; enterprises need control. RuntimeAI is the continuous ship / still-trust gate.

OpenRouter live catalog · as of 23 August 2026

01 Snapshot

422

Live model routes

15,000×

Input \$/M spread

20,000×

Output \$/M spread

~17k×

Same-session cost spread

7.8×

Catalog vs end-2024

202

New routes · 2026 YTD

The problem. OpenRouter's catalog is **422 model options** — most paid (395), plus 61 cheaper offline "batch" variants and 22 free ones. In 2026 alone it added **202 new options** — more than the entire catalog held at end of 2024. Nobody re-clears that surface with a quarterly spreadsheet and an LLM judge.

02 Catalog growth & forecast

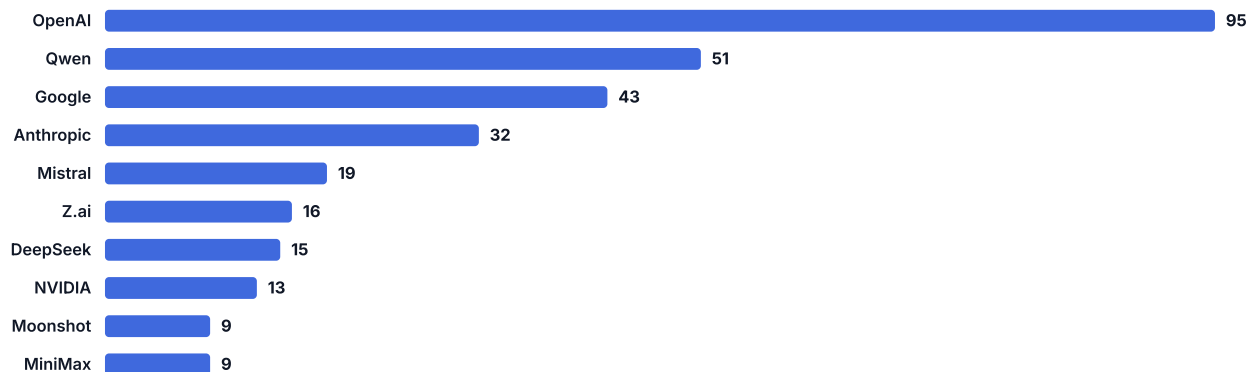
Observed catalog size (Jan 2025 – Aug 2026) with an anchored quadratic forecast through YE 2027. Extrapolation — not a commitment; silent defaults and delists are not modeled.

Forecast milestones

HORIZON	CATALOG	ADDS / MO	VS TODAY	BASIS
Aug 2026 (actual)	422	~32/mo	—	Actual
Dec 2026 · YE	549	~40/mo	+127 (1.3×	Quadratic forecast
Mar 2027 · Q1	657	~47/mo	+235 (1.6×	Quadratic forecast
Jun 2027 · Q2	774	~55/mo	+352 (1.8×	Quadratic forecast
Sep 2027 · Q3	903	~63/mo	+481 (2.1×	Quadratic forecast
Dec 2027 · YE	1,041	~72/mo	+619 (2.5×	Quadratic forecast

Where the 422 routes come from

Top providers by route count — one gateway, 53 organizations.



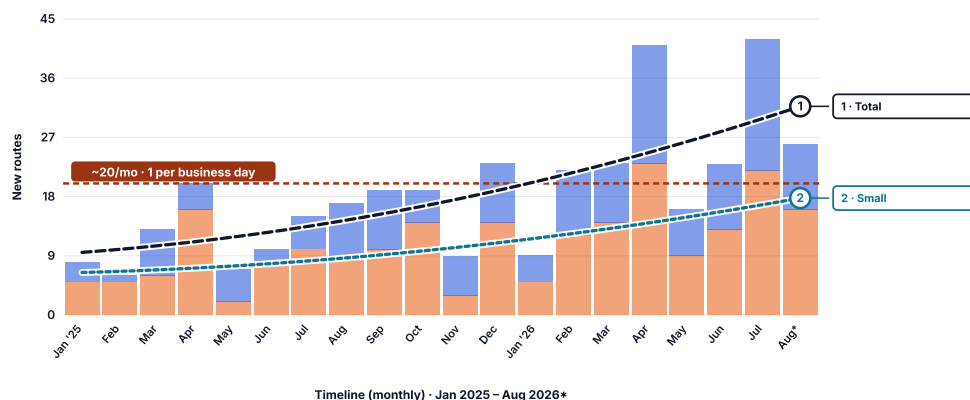
Catalog trajectory



Monthly adds (Jan 2025 – Aug 2026*)

New OpenRouter routes per month. Dashed curves are quadratic trends.

■ Small / cheap ■ Mid + frontier -- 1 · Total routes — quadratic trend - - 2 · Small / cheap — quadratic trend



03 Pricing is all over the place

Live OpenRouter rates (334 paid + 22 free; excludes offline batch). Input spans \$0.01 → \$150/M (15,000×). Output spans \$0.03 → \$600/M (20,000×).

42×

Input p90 / p10

75×

Output p90 / p10

51×

Session cost p90 / p10

488×

Context max / min

Price & metric chaos

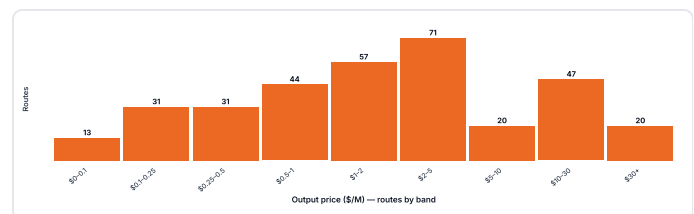
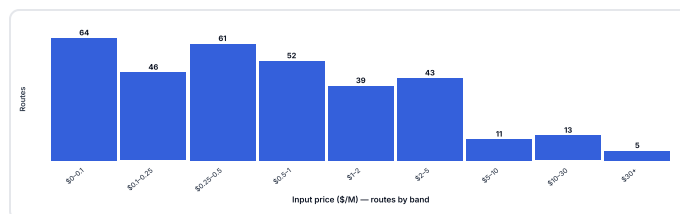
METRIC	SPREAD
Input \$/M	\$0.01 → \$150 · median \$0.44 · 15,000×
Output \$/M	\$0.03 → \$600 · median \$1.78 · 20,000×
Session cost	Cheapest → dearest · ~17,000×
Context length	~4K → 2M · 488×
Free vs paid	22 free · 334 paid

Even inside one lab

PROVIDER	INPUT \$/M	SPREAD
OpenAI	\$0.03–\$150/M in	5,000×
Anthropic	\$0.25–\$30/M in	120×
Qwen	\$0.03–\$2/M in	67×
Google	\$0.05–\$2/M in	40×
DeepSeek	\$0.04–\$1.12/M in	28×

Price sheets are not a gate. A 15,000× input spread means “pick the cheaper model” is undefined without a fixed scenario, pass rate, and effective USD.

Price distribution



04 Control is the empty seat

What teams do today

APPROACH	WHY IT BREAKS
Inspect	Traces, LLM judges — fuel, not a ship decision
Quarterly bake-off	Stale before the next Flash bump (<30 days)
Price sheet	Picks among 422 routes; never proves your workflow still passes
Router alone	Chooses a slug; does not clear still-trust after the swap

Evidence of chaos

PATTERN	EVIDENCE
Catalog mass	54 → 422 routes
Gemini Flash	3.5 → 3.6 → 3.7 in ~12 weeks
Anthropic Opus	4.6 → 4.7 → 4.8 → Opus 5 in ~5.5 months
July 2026	42 new OpenRouter routes in 31 days
Silent defaults	MAI into Copilot — no new public eval target

RuntimeAI: continuous attested **ship / still-trust** on partner-owned paths — fixed scenarios, deterministic rubric, pass rate + USD per completed run. Not FinOps, not a router, not an LLM judge average.

Control vs chaos

	WITHOUT CONTROL	WITH SHIP / STILL-TRUST GATE
Swap candidates	422 live routes	Same paths · pinned rubric · before/after
Decision unit	\$/M · arena rank · judge score	Pass rate + USD on your scenario
Cadence	Weekly drops · silent defaults	Continuous clearance on every change
Failure mode	Green dashboard · wrong invoice	Hard-fail on failure modes you refuse to ship

Sources: OpenRouter live API · AI Release Tracker · RuntimeAI Industry signals · scorecard freeze (+88 catalog ids Jul→Aug). Updated 23 August 2026.