YU-GI-AI HARNESS BENCHMARK

Benchmarking frontier models on Yu-Gi-Oh deck building

Seven frontier LLMs on a real card pool, graded on strategist creativity, tactician deck quality, and real dollar cost. Vol. 1 of the Yu-Gi-AI Harness Benchmark.

April 16, 2026 · Yu-Gi-AI

Every LLM benchmark lives or dies on the question it asks. The Yu-Gi-AI question is concrete: which frontier model is actually best at helping a real Yu-Gi-Oh player build a real deck?

That turns out to be two questions. Strategist deck-building is the premium work — non-obvious bridges, clever packages, suggestions a strong player finds sharp. Tactician deck-building is grounded repair — identifying what is structurally wrong with a list and reconstructing the missing engine. Same harness, same card pool, very different jobs. The models that win one mode are usually not the models that win the other.

This is Vol. 1 of the Yu-Gi-AI Harness Benchmark. Seven frontier models, two suites, nine deck-building cases, real OpenRouter dollars.

PREMIUM STRATEGIST

GPT-5.4

91 / 100

Highest creative-track score, averaging $0.11 per case. The sharpest non-obvious deck-builder moments across the matrix.

CHEAP TACTICIAN

Gemini 3 Flash Preview

5 / 5

Perfect deck-quality pass rate at $0.02 per case and 13s latency — a 22× cost delta under Opus 4.6.

QUALITY BENCHMARK

Claude Opus 4.6

93 composite

Strongest aggregate output in the matrix, at $0.53 per case and 93s latency. Too expensive to run as a default.

How the harness works

The Yu-Gi-AI harness feeds each model a complete main deck, extra deck, and side deck, along with a deck-context summary and the current Yu-Gi-Oh format state. The model reasons about the deck using a controlled card-lookup tool and proposes concrete changes.

Two suites ran in this pass:

- `creative-upgrade` — four decks, each paired with a prompt to find an impressive package the deck is missing. Responses are scored 0–100 on originality, shell coherence, mechanical defensibility, and likelihood of impressing a strong player.
- `deck-quality` — five reconstruction cases. A known package is removed from a tournament list and the model is asked to restore the missing structure. Graded pass/fail per case.

Every run logs the OpenRouter dollar cost and end-to-end latency billed by the provider, not an estimate per million tokens.

REVIEWER DISCLOSURE

The creative track was scored by GPT-5.4 acting as a rubric-graded reviewer on each model's transcripts, and then audited by Claude Opus 4.7 as a second pass against the same transcripts. No human scored the creative track in this run. GPT-5.4's own responses were therefore reviewed by a sibling model; Claude Opus 4.7's own responses were reviewed partly by itself. Treat the absolute scores as directional rather than authoritative: they are reliable for relative ordering on this specific rubric, this specific deck set, and this specific run, and are not a substitute for a tournament-grade human reviewer.

WHY TWO TRACKS

A model that writes a thoughtful essay about a Crystron/K9 bridge can still miss that a Sky Striker list is running the wrong engine spells. Creative strategy and structural repair load different skills, so the benchmark scores them separately instead of averaging them into a single universal number.

The scorecard

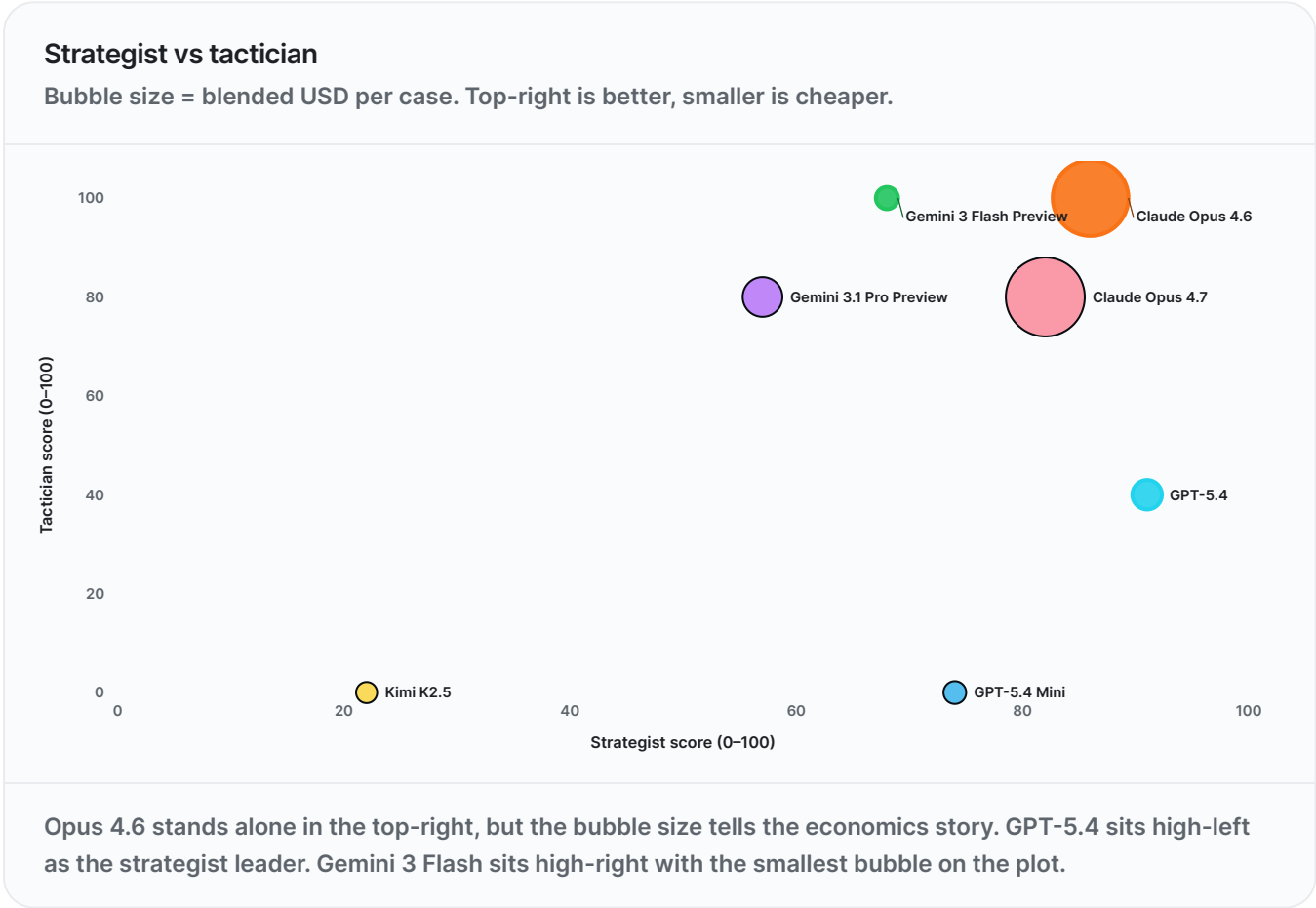
Scorecard

Strategist = 0-100 creative quality (GPT-5.4 as reviewer, Claude Opus 4.7 as auditor) · Tactician = deck-quality pass rate × 100 · Blended $ = mean of creative + deck-quality cost per case

MODEL	STRATEGIST	TACTICIAN	BLENDED $	LATENCY	VERDICT
Claude Opus 4.6 Anthropic	86	100	$0.53	93s	Benchmark Strongest quality benchmark
Gemini 3 Flash Preview Google	68	100	$0.024	14s	Cheap tactician Best cheap tactician
Claude Opus 4.7 Anthropic	82	80	$0.56	79s	Benchmark Strong challenger
Gemini 3.1 Pro Preview Google	57	80	$0.19	112s	Not selected Too unstable at the price
GPT-5.4 OpenAI	91	40	$0.093	46s	Premium strategist Best premium strategist
GPT-5.4 Mini OpenAI	74	0	$0.029	11s	Not selected Collapsed on structure
Kimi K2.5 Moonshot	22	0	$0.013	303s	Not selected Operationally not viable

The split is clearer than a single ranking would suggest. Opus 4.6 tops the balanced composite, but it costs 22× more per case than Gemini 3 Flash and 6× more than GPT-5.4 — a gap the transcripts do not justify on quality. GPT-5.4 leads the strategist track by nine points. Gemini 3 Flash ties Opus 4.6 on tactician. GPT-5.4 Mini and Kimi K2.5 both scored zero on deck-quality, via different failure modes.

Strategist vs tactician



Opus 4.6 stands alone in the top-right, but the bubble size tells the economics story. GPT-5.4 sits high-left as the strategist leader. Gemini 3 Flash sits high-right with the smallest bubble on the plot.

The interesting pattern is the diagonal. GPT-5.4 is 23 points stronger than Gemini 3 Flash on strategist and 60 points weaker on tactician. Consolidating both jobs into one model loses value on whichever side is underweighted.

THE KIMI OUTLIER

Kimi K2.5 averaged 303 seconds per case — over five minutes — and passed 1 of 9 cases across both suites. It is also the cheapest model in the matrix at $0.013 per case. The cost advantage over Gemini 3 Flash is $0.011, which does not come close to closing the reliability gap.

Cost versus what you actually get

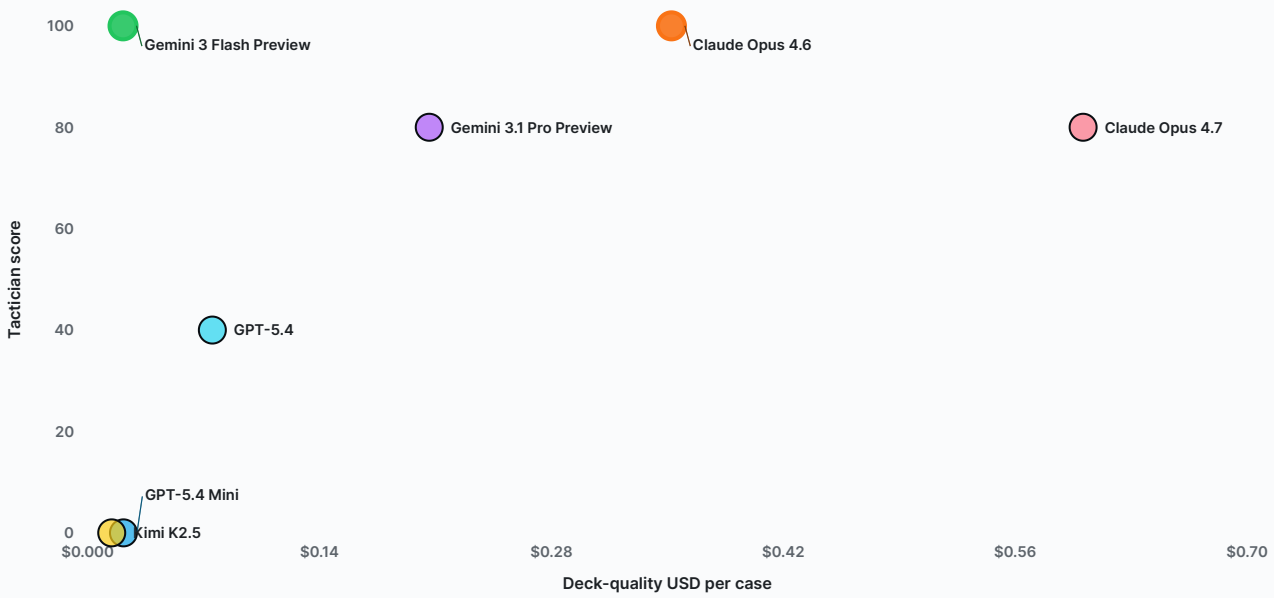
Strategist score vs real dollar cost

Creative-track average USD per case on X. Model-reviewed strategist score on Y.

The pareto frontier runs from Gemini 3 Flash at the bottom-left to GPT-5.4 at the middle, to Opus 4.6 at the top-right. Opus 4.7 sits below 4.6 at a slightly lower cost — strictly dominated for strategist work in this run.

Tactician score vs real dollar cost

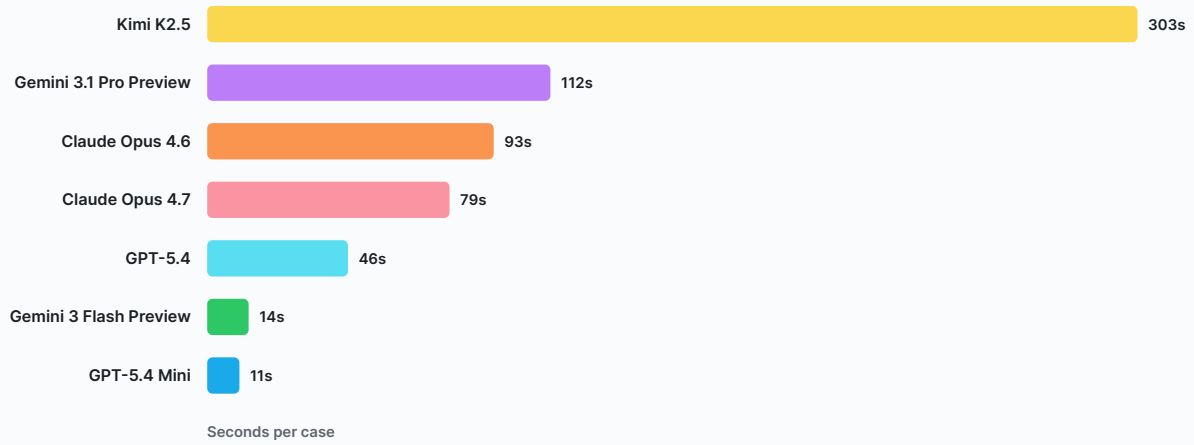
Deck-quality USD per case on X. Pass-rate × 100 on Y.



For structural deck repair, cheap wins outright. Gemini 3 Flash is a tenth the price of Opus 4.6 and ties it on pass rate.

Latency

Blended average duration per case, across both tracks



Model dossiers

OPENAI

Best premium strategist

GPT-5.4

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
91	40	$0.093	46s

GPT-5.4 set the top of the strategist track and no other model in the matrix surpassed it. Four of four on creative pass rate, the most consistent source of suggestions that read like a deck-builder thought specifically about the shell in front of them, and roughly a tenth of Opus 4.6's cost per case.
A representative moment from the Crystron K9 case:

> I'd seriously test a tiny P.U.N.K. bridge, not as a standalone engine, but as a discard-conversion and body-generation package for Crystron/K9. Foxy Tune discarding Crystron is real value, not a cost. Pitching Smiger, Tristaros, or even setup pieces you can recur later is much better here than in random decks.
>
> GPT-5.4 · creative-hybrid-combo-crystron-k9

That is the difference between surfacing generic staples and identifying a specific hand-coherence problem that a small package solves. The weakness is the stricter deck-quality track, where GPT-5.4 passed 40%. On literal reconstruction it sometimes answered at too high a level or fixed the wrong layer of the list. A strong strategist, a weaker literal repair technician.

GOOGLE

Best cheap tactician

Gemini 3 Flash Preview

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
68	100	$0.024	14s

Five of five on deck-quality, at $0.021 per case and 13 seconds of end-to-end latency. That ratio is the single biggest argument for a split policy rather than a single-flagship routing rule.

The creative suggestions are more obvious than GPT-5.4's — closer to "the correct ratio and engine fix" than to "a sharp bridge a strong player would not have seen." For structural deck audits, cheap automated review flows, and missing-package diagnosis, Gemini 3 Flash is the right engine. It is not the model that makes the product feel premium, and the routing policy does not ask it to be.

ANTHROPIC

Strongest quality benchmark

Claude Opus 4.6

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
86	100	$0.53	93s

Opus 4.6 produced the strongest aggregate output in the matrix. Four of four on creative, five of five on deck-quality, thoughtful long-form reasoning, and genuinely sharp single-card picks. At $0.71 per creative case, it is also the most expensive model tested.

Opus 4.6 is clearly good. The question is whether it is enough better than GPT-5.4 to justify six times the cost per case and roughly twice the latency. In this run the answer is no. It belongs in the toolkit as a benchmark reference and an optional second-opinion escalation, not as a default.

ANTHROPIC

Strong challenger

Claude Opus 4.7

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
82	80	$0.56	79s

Opus 4.7 was the strongest candidate to displace GPT-5.4. It is often clever, materially cheaper than 4.6 on creative, and it scored four of five on deck-quality — strong by any reasonable standard.

Strong is not the bar. Displacing the split policy would require beating GPT-5.4 and Gemini 3 Flash at their respective jobs by enough to justify consolidating. On the product-defining test — would a strong player find this genuinely impressive — Opus 4.7 was less

consistent than GPT-5.4. On structural repair it lost to Gemini 3 Flash at a fraction of the price. A close challenger, not a replacement.

OPENAI

Collapsed on structure

GPT-5.4 Mini

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
74	0	$0.029	11s

A mixed result. On creative, GPT-5.4 Mini is respectable: 74 strategist, 4/4 creative pass, $0.036 per case — close to Gemini 3 Flash on price. It was the only low-cost model that produced a handful of genuinely useful shell-specific bridges.
On deck-quality it scored 0/5 — zero passes across five reconstruction cases. The cheap-strategist slot requires a model that can also handle unglamorous structural work, which GPT-5.4 Mini did not clear in this run.

GOOGLE

Too unstable at the price

Gemini 3.1 Pro Preview

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
57	80	$0.19	112s

Real strategic ideas on its best runs. Also one no-final-response failure mid-suite, 112 seconds per blended case, and $0.19 average cost — a middle ground that is neither cheap enough for the tactician slot nor sharp enough for the strategist slot. Not selected for routing.

MOONSHOT

Operationally not viable

Kimi K2.5

STRATEGIST	TACTICIAN	BLENDED $	LATENCY
22	0	$0.013	303s

One decent Crystron-K9 idea, followed by operational collapse. Over five minutes per case on average, 1 of 9 aggregate case passes, and a failure surface that included both missed final responses and outright structural errors. Not a viable option for this product in its current configuration.

Recommendation

SPLIT MODEL POLICY

Premium deck-building, creative package discovery, and "impress me" prompts route to GPT-5.4. Cheap reconstruction, missing-package diagnosis, and automated deck audits route to Gemini 3 Flash Preview. Claude Opus 4.6 stays wired up as an optional second-opinion escalation, not a default. GPT-5.4 Mini, Gemini 3.1 Pro Preview, and Kimi K2.5 are not selected for production routing in this pass.

If the product were forced to pick a single model for everything, GPT-5.4 remains the better single-model bet. The premium creative mode is the product differentiator; giving it up for marginal gains on structural repair would trade away the capability users pay for.

Caveats

One snapshot, not a final answer.

- Creative-track scoring was produced by GPT-5.4 as reviewer and Claude Opus 4.7 as second-pass auditor, not a human. Two self-reference risks follow directly from that: GPT-5.4 scoring its own responses, and Opus 4.7 auditing its own. A different reviewer configuration could shift individual scores by several points and in edge cases could reorder adjacent models. The headline policy (GPT-5.4 as strategist, Gemini 3 Flash as tactician) is robust to that variance; the 91 / 86 / 82 spread between the top three is not.
- Banquet FTK was on the original creative-track shortlist but was not available in the current source deck set and was dropped from this pass.

- **Pricing for the shortlisted finalists (GPT-5.4, Gemini 3 Flash, Opus 4.6) still needs a dedicated multi-turn pricing benchmark before credit packs are repriced. The per-case numbers here are sufficient for choosing models, not yet sufficient for final product pricing.**
- **OpenRouter prices and provider model IDs change over time. These numbers are accurate as of the run timestamps recorded in the artifacts.**

Reproducibility

All inputs are frozen. Run order:

1. **Creative matrix — `2026-04-16T14:50:01Z` (six-model) and `2026-04-16T18:54:38Z` (Opus 4.7)**
2. **Deck-quality matrix — `2026-04-16T17:42:21Z` (six-model) and `2026-04-16T18:59:59Z` (Opus 4.7)**
3. **Synthesis — `packages/evals/src/assistant/phase31-synthesis.ts` emits the typed summary JSON**

The data module backing this article lives at `content/blog/phase31-openrouter-model-selection/data.ts` and matches `artifacts/evals/assistant-matrix/phase31-synthesis/phase31-model-selection-summary.json` entry for entry.

A printable PDF of this article is available from the `Download PDF` action at the top of the page.