Lean vs. Verbose: Three Western Models on Middle-School Competition Math

A difficulty-resolved, cost-matched head-to-head of GPT-5.5, Gemini-Pro and Llama-4-Maverick on AMC 8 / AJHSME

GenAware Model Lab · 22 June 2026

Executive summary

> **Governing thought.** At the top, **accuracy on grade-school competition math is solved — the differentiator is no longer whether a model is right, but how many tokens it burns to get there.** GPT-5.5 and Gemini-Pro both clear ≈99–100% on every difficulty tier, yet Gemini spends **2.5× the output tokens** to do it — which erases its cheaper sticker price and leaves the two at an identical cost per correct answer. One rung down, Llama-4-Maverick is **≈50× cheaper** but cracks on hard problems. The choice between these three is not a quality choice; it is a **verbosity-and-price** choice.

We ran **three Western models** through **432 graded attempts** of middle-school competition math (AMC 8 / AJHSME), and — the decisive design choice — resolved every result by **problem difficulty** (easy → medium → hard → stretch) rather than reporting one pooled average. Three findings:

1. **Capability is saturated at the top.** GPT-5.5 scored a **perfect 144/144** across all four tiers; Gemini-Pro **143/144** (one medium slip). Neither degrades as problems get harder. On accuracy alone, these two are interchangeable.
2. **Verbosity is the hidden price.** Gemini emits **955 tokens per question** to GPT-5.5's **374** — 2.5× more. Because cost is tokens × price, Gemini's lower per-token rate ($12/M out vs $30/M) is exactly cancelled: both land at **≈1.17 ¢ per correct answer**. The cheaper model is not cheaper in practice.
3. **The budget tier buys 50× cost at the cost of hard-problem reliability.** Llama-4-Maverick costs **0.023 ¢ per correct** — a different universe — but its accuracy falls **100 → 89 → 81 → 78%** from easy to stretch. Fine for easy volume; a liability on hard reasoning.

The positioning in one exhibit — accuracy against output verbosity (the axis that sets real cost):

Accuracy ↓ / Output per answer →	Lean (< 400 tok)	Verbose (> 900 tok)
Near-perfect (≥ 99%)	**GPT-5.5** — perfect and lean. The efficient frontier: cheapest path to a correct answer at this quality, and the friendliest to latency-sensitive, agentic, output-metered use.	**Gemini-Pro** — just as accurate, but pays 2.5× the tokens for it. Same ¢/correct as GPT despite a cheaper rate card. Right answer, dear delivery.
Cracks on hard (< 90%)	**Llama-4-Maverick** — lean and dirt-cheap (1/50th the cost), but unreliable past the easy tier. Volume workhorse, not a reasoner.	(empty — no model is both verbose and weak here)

Where to play.

- **Default for quality + efficiency: GPT-5.5** — perfect accuracy, the leanest output, tied-cheapest per correct answer. No reason to pay more tokens for the same result.
- **If you are already on Google's stack: Gemini-Pro** — equal accuracy and equal ¢/correct; you give up tokens and a little latency headroom, not quality.

- **High-volume, easy-skewed, latency-tolerant work: Llama-4-Maverick** — 50× cheaper per correct answer; just keep it away from hard/stretch reasoning, where it drops to ≈78–81%.

1. The buyer's question

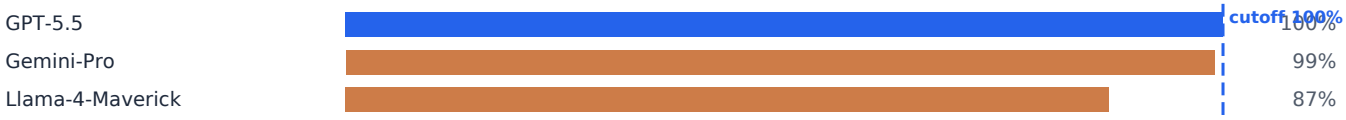
Three Western models sit at very different price points — Llama-4-Maverick at **$0.15 / $0.60** per million tokens, Gemini-Pro at **$2 / $12**, GPT-5.5 at **$5 / $30**. The sticker spread is 33× on output. For anyone choosing one, two questions follow: does the price buy accuracy, and does the cheaper-per-token option actually cost less in practice? Middle-school competition math is a clean probe — every question has one objectively correct multiple-choice answer, the problems span a real difficulty gradient, and no tool use or retrieval can paper over weak reasoning.

2. Method

- **Problems:** AMC 8 / AJHSME, text-only multiple choice, drawn from the GenAware Lab's curated pool, bucketed into four difficulty tiers (easy / medium / hard / stretch).
- **Design:** 12 rounds (3 per tier), 12 questions each. The same sampled question set is sent to all three models in a round, so every comparison is head-to-head on identical problems. **144 attempts per model, 432 total.**
- **One call per round per model.** All a round's questions go in a single request; a model that runs out of room and drops answers scores those **wrong** — a fair capability strike. Web search is **off**: pure reasoning, graded objectively against the known answer.
- **Cost:** live OpenRouter per-token pricing for GPT-5.5 and Llama. Gemini-Pro was run via the rolling google/gemini-pro-latest alias, which carries **no catalog price**, so its cost is **estimated** at gemini-3.1-pro-preview rates ($2/M in, $12/M out). See §7.

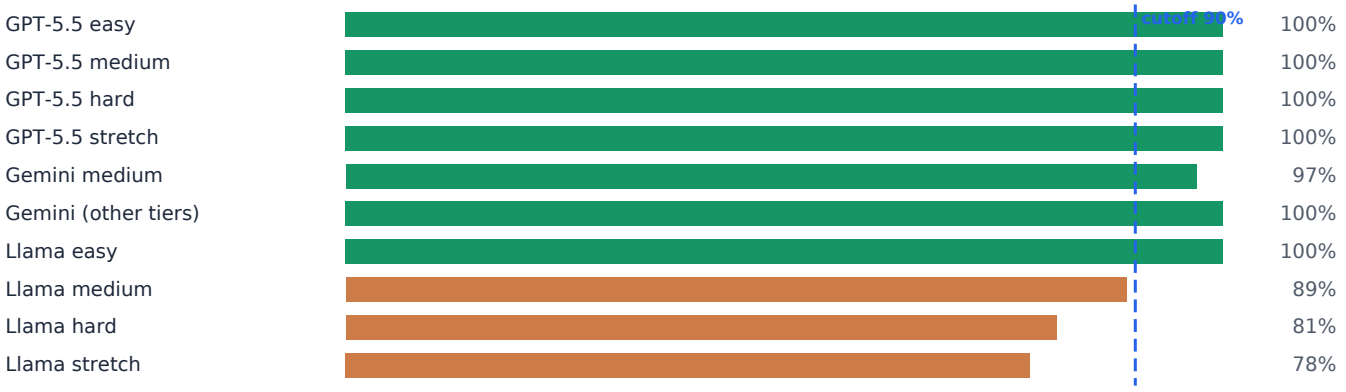
3. Accuracy — solved at the top, graded below

Overall accuracy (144 attempts each)



GPT-5.5 and Gemini-Pro are, for this task, a solved problem — 100% and 99%. The gap to Llama (87%) is real but pooled it understates the story, because the budget model's weakness is concentrated where it matters. Split by difficulty:

Accuracy by difficulty — the budget model cracks as problems harden



Model	Org	easy	medium	hard	stretch	overall
GPT-5.5	OpenAI	100 (36/36)	100 (36/36)	100 (36/36)	100 (36/36)	**100%**
Gemini-Pro	Google	100 (36/36)	97 (35/36)	100 (36/36)	100 (36/36)	**99%**

Model	Org	easy	medium	hard	stretch	overall
Llama-4-Maverick	Meta	100 (36/36)	89 (32/36)	81 (29/36)	78 (28/36)	**87%**

The two frontier models are **difficulty-inelastic** — GPT-5.5 doesn't miss a single problem at any tier; Gemini's only error is a lone medium question. Llama-4-Maverick shows the classic budget-model signature: perfect on easy, a clean monotonic decline to **78%** on stretch. If your workload is hard, the headline 87% is the wrong number — the relevant number is 78–81%.

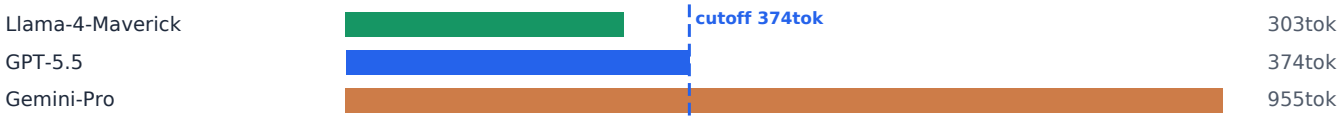
4. Latency — frontier pair tied, budget model slowest

Average latency per question (lower is better)

GPT-5.5 (9.3 s) and Gemini-Pro (9.2 s) are a dead heat; Llama is ≈65% slower (15.2 s) **despite emitting fewer tokens than Gemini** — its latency is throughput/overhead-bound, not length-bound. These per-tier latencies are noisy: one medium round ran during an OpenRouter congestion spike (≈8 min wall-clock vs ≈2–3 min for every other round), which inflates the medium tier for all three models above hard and stretch. Treat the **overall** latency ranking as solid and the per-tier latency shape as indicative only (see §7).

5. Verbosity — the hidden cost driver

Output tokens per question (Gemini is 2.5x GPT-5.5)



This is the chart that decides the money. GPT-5.5 and Llama are lean (374 and 303 tokens/question); Gemini-Pro is **2.5× more verbose at 955 tokens** — and on medium problems it climbs to **1,207**. Output tokens are the expensive half of every rate card, so verbosity is not a stylistic footnote — it is the cost.

6. Cost per correct answer — the cheaper rate card disappears

Cents per correct answer (effective cost, lower is better)

Here is the punchline. Gemini-Pro's output price ($12/M) is **2.5× cheaper** than GPT-5.5's ($30/M) — but Gemini spends **2.5× the tokens**, so the two arrive at a **statistically identical** cost per correct answer: **1.175 ¢ vs 1.173 ¢**. The cheaper rate card buys nothing once verbosity is priced in. You don't pay the sticker price; you pay the sticker price times verbosity.

Llama-4-Maverick is in a different economic universe: **0.023 ¢ per correct answer**, roughly **50× cheaper** than either frontier model — the reward for being lean and cheap-per-token, and the compensation for its hard-problem misses.

Model	¢ / correct	tokens / Q	accuracy	the trade
Llama-4-Maverick	**0.023**	303	87%	50× cheaper; cracks on hard

Model	¢ / correct	tokens / Q	accuracy	the trade
GPT-5.5	**1.173**	374	100%	perfect & leanest; tied-cheapest at the top
Gemini-Pro	**1.175**	955	99%	equal quality & cost, 2.5× the tokens

7. Methodology notes & caveats

- **Sample size.** 36 attempts per model per tier (144 per model). Accuracy at the top is stable; **latency carries real variance** and the per-tier latency shape is distorted by a single congested medium round (≈8 min vs ≈2–3 min typical). Read overall latency as solid, per-tier as indicative.
- **Gemini cost is estimated.** google/gemini-pro-latest is a rolling alias absent from the Lab's OpenRouter price catalog, so the Lab recorded its dollar cost as $0. We reconstructed it from measured token usage at **gemini-3.1-pro-preview** rates ($2/M in, $12/M out) — the closest concretely-priced Gemini Pro tier. Gemini's **accuracy, latency, and token counts are measured, not estimated**; only the ¢ figures are modelled. If the alias resolves to a different concrete version, scale Gemini's cost linearly with the true output price (its ¢/correct = 0.0979 ¢ per $/M of output price).
- **Grading.** Objective, against the known answer; dropped/omitted answers score wrong.
- **No tools.** Web search off; pure reasoning. Real-world accuracy with retrieval/tools would differ.
- **Scope.** Middle-school competition math (AMC 8 / AJHSME). Findings transfer to short, self-contained reasoning; they say nothing about long-context, coding, or agentic tasks.

8. Bottom line

For grade-school competition math, **the accuracy question is closed at the top and the contest is about delivery cost.** GPT-5.5 wins outright on efficiency — perfect accuracy, the leanest output, tied-cheapest per correct answer, and the lowest latency. Gemini-Pro matches it on the metric a buyer ultimately pays (¢/correct) but only by spending 2.5× the tokens, so it wins nothing it doesn't already own on Google's stack. Llama-4-Maverick is the volume play — 50× cheaper, genuinely usable on easy work, and genuinely risky past the easy tier. Pick on verbosity and price, not on accuracy.

GenAware Model Lab — 432 graded attempts, 22 June 2026. Data workbook, aggregate CSV and raw per-attempt CSV accompany this paper.