

Frontier Language Models Under a Verifiable Protocol: Accuracy, Cost and Speed on Expert-Level Questions

Suhas Bhairav

Independent researcher - suhasbhairav.com

Preprint - October 30, 2026

Abstract

I evaluate 31 frontier language models, grouped by a published list-price rule, on 80 expert-level ten-option multiple-choice questions from SuperGPQA and MMLU-Pro, using a verifiable methodology in which every answer is linked to a hashed record of its API request and response. Scores range from 26.6% to 85.0% (median 75.0%; chance is 10%). The highest-scoring model, anthropic/claude-opus-5.5, answered 68 of 80 correctly (85.0%, 95% Wilson interval 75.6% to 91.2%). A two-proportion z-test ($|z| \geq 1.96$) cannot separate 19 other models from the leader, so the top of the ranking is too close to call on this sample. Median billed cost was \$0.0088 per question, and the rank correlation between score and cost was 0.06. Per-question verdicts, proof hashes and generation IDs are published for every model.

Keywords: large language models; LLM evaluation; reproducibility; SuperGPQA; MMLU-Pro; Wilson interval; two-proportion z-test; McNemar test; frontier models; cost per question; output speed

1 Introduction

This paper reports results for 31 frontier language models under the methodology described in the companion paper, "How Every Number Is Produced: A Verifiable Methodology for Evaluating Large Language Models". A model is classed as frontier when its blended list price, $(3 \times \text{input} + 1 \times \text{output}) / 4$ in USD per million tokens from the OpenRouter catalogue at test time, is at least \$1.00; every other model is non-frontier. The companion results paper covers the 41 non-frontier models evaluated under the identical methodology.

I report what the measurements support and no more: a score with its 95% interval for every model, which differences from the leading model pass a significance test, what each model cost per question as billed, and how fast it streamed. I do not combine these into a single index.

2 Setup

Each model answered the same 80 ten-option multiple-choice questions (50 from SuperGPQA, hard subset; 30 from MMLU-Pro) through OpenRouter at temperature 0, reasoning effort "low" where supported and at most 4000 output tokens. Answers were graded by a deterministic extraction rule against the datasets' answer keys. Models were tested between October 25, 2026 and October 27, 2026. Chance accuracy is 10%.

No request in this group failed after retries. moonshotai/kimi-k2.7-code (1), qwen/qwen3.8-2.4t-a95b (1), cohere/command-a (1), mistralai/mistral-medium-3-5 (1) and nousresearch/hermes-4-405b (1) had answers excluded because the run manifest did not confirm their proof hash (methodology, Sections 7 and 9). Scores below are therefore out of the number of graded answers shown.

3 Accuracy

Table 1. Scores of the 31 frontier models, sorted by score. Model IDs link to the model's OpenRouter page. CI: 95% Wilson interval. No ans.: graded answers with no extractable answer line.

Model	Score	95% CI	Correct / graded	No ans.	Cost / question	Blended list \$/M
anthropic/claude-opus-5.5	85.0%	75.6-91.2	68 / 80	0	\$0.015	\$8.000
google/gemini-4-argon	83.8%	74.2-90.3	67 / 80	1	\$0.0062	\$4.000
google/gemini-3.8-flash	82.5%	72.7-89.3	66 / 80	0	\$0.0037	\$1.650
openai/gpt-5.3-codex	82.5%	72.7-89.3	66 / 80	0	\$0.010	\$4.813
anthropic/claude-fable-5.1	81.2%	71.3-88.3	65 / 80	1	\$0.016	\$20.000
x-ai/grok-4.7	80.0%	70.0-87.3	64 / 80	0	\$0.025	\$3.000
anthropic/claude-sonnet-5.5	78.8%	68.6-86.3	63 / 80	0	\$0.0054	\$4.000
meta/muse-wonder	78.8%	68.6-86.3	63 / 80	5	\$0.021	\$8.000
google/gemini-3.1-pro-preview	77.5%	67.2-85.3	62 / 80	0	\$0.017	\$4.500
meta/muse-spark-open	77.5%	67.2-85.3	62 / 80	1	\$0.0045	\$2.000
moonshotai/kimi-k3.1-pro	77.5%	67.2-85.3	62 / 80	5	\$0.010	\$4.000
openai/gpt-6-astra	76.2%	65.9-84.2	61 / 80	0	\$0.013	\$20.000
openai/gpt-6.1-sol	76.2%	65.9-84.2	61 / 80	0	\$0.0029	\$4.000
openai/gpt-6.1-terra-pro	76.2%	65.9-84.2	61 / 80	0	\$0.0050	\$2.000
openai/gpt-5.6-terra	75.0%	64.5-83.2	60 / 80	0	\$0.0063	\$4.500
qwen/qwen3.8-max-0902	75.0%	64.5-83.2	60 / 80	10	\$0.011	\$3.000
openai/gpt-6-sol	73.8%	63.2-82.1	59 / 80	0	\$0.0024	\$4.000
openai/gpt-6.1-terra	73.8%	63.2-82.1	59 / 80	0	\$0.0013	\$2.000
openai/gpt-5.4	72.5%	61.9-81.1	58 / 80	0	\$0.011	\$5.625
openai/o3	72.5%	61.9-81.1	58 / 80	3	\$0.0088	\$3.500
deepseek/deepseek-v4-pro-0813	71.2%	60.5-80.0	57 / 80	12	\$0.0036	\$1.980
thinkingmachines/inkling	66.2%	55.4-75.7	53 / 80	0	\$0.0012	\$1.762
moonshotai/kimi-k2.7-code	64.6%	53.6-74.2	51 / 79	19	\$0.0061	\$1.195
stepfun/step-5-preview	63.8%	52.8-73.4	51 / 80	17	\$0.0047	\$1.425
nvidia/nemotron-3-ultra-550b-a55b	60.0%	49.0-70.0	48 / 80	18	\$0.0057	\$1.050
qwen/qwen3.8-2.4t-a95b	58.2%	47.2-68.5	46 / 79	26	\$0.015	\$3.000
mistralai/mistral-large-4-0	50.0%	39.3-60.7	40 / 80	35	\$0.011	\$2.065
cohere/command-a	45.6%	35.0-56.5	36 / 79	2	\$0.0071	\$4.375
qwen/qwen3.5-397b-a17b	32.5%	23.2-43.4	26 / 80	50	\$0.013	\$1.275
mistralai/mistral-medium-3-5	29.1%	20.3-39.9	23 / 79	47	\$0.025	\$3.000
nousresearch/hermes-4-405b	26.6%	18.1-37.2	21 / 79	0	\$0.010	\$1.500

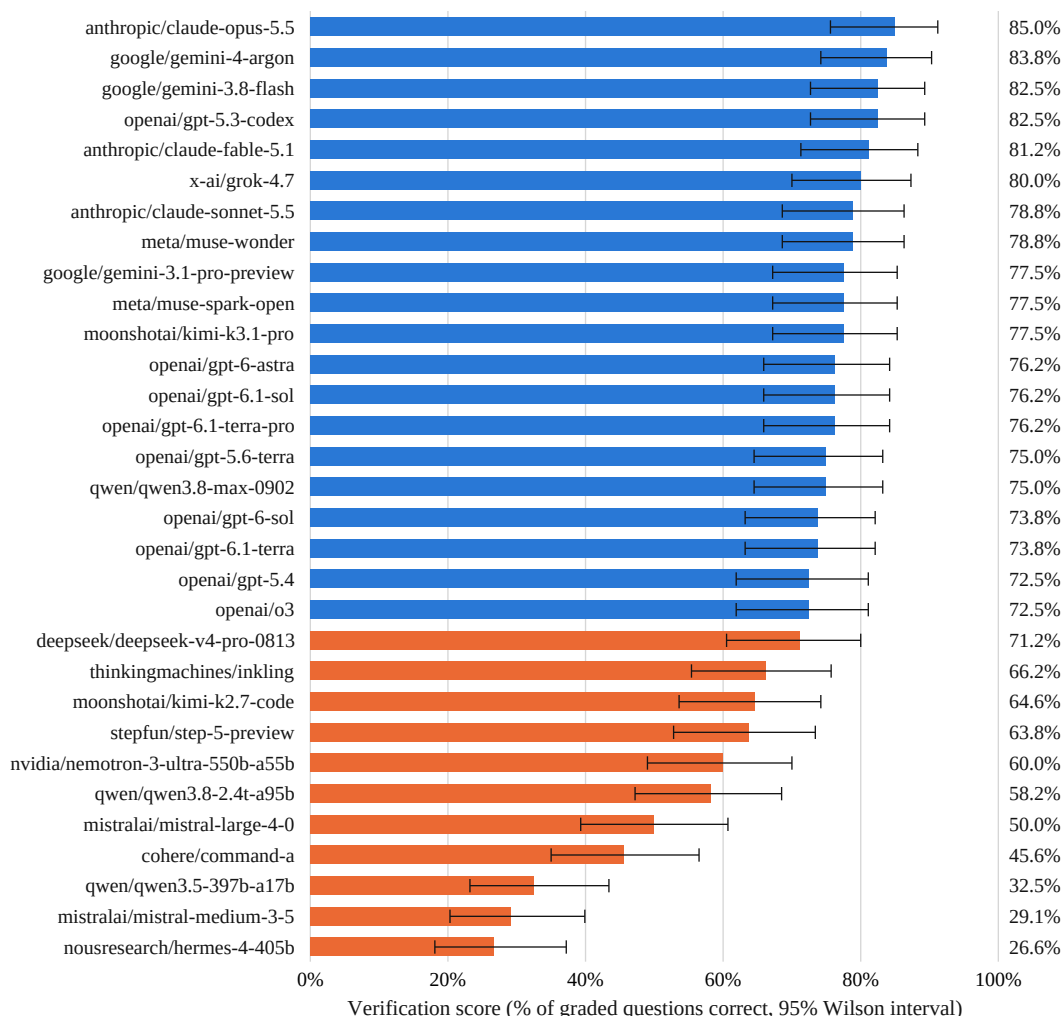


Figure 1. Score with 95% Wilson interval, frontier models. Blue: the leader and the models too close to call against it (two-proportion z-test, $|z| < 1.96$). Orange: the leader is significantly more accurate.

Scores range from 26.6% to 85.0%, with a median of 75.0%. The highest score is anthropic/claude-opus-5.5 with 68 of 80 (85.0%; 95% CI 75.6% to 91.2%). For comparison, the median of the non-frontier group is 48.8%.

4 Which differences are real

Table 2 tests the leader, anthropic/claude-opus-5.5, against every other model in the group with a two-proportion z-test on the two scores (pooled standard error). A gap counts only when $|z|$ is at least 1.96, the usual threshold for 95% confidence on a two-sided test. Of the 30 comparisons, 11 gaps are significant and 19 are too close to call. The models too close to call against the leader are google/gemini-4-argon, google/gemini-3.8-flash, openai/gpt-5.3-codex, anthropic/claude-fable-5.1, x-ai/grok-4.7, anthropic/claude-sonnet-5.5, meta/muse-wonder, google/gemini-3.1-pro-preview, meta/muse-spark-open, moonshotai/kimi-k3.1-pro, openai/gpt-6-astra, openai/gpt-6.1-sol, openai/gpt-6.1-terra-pro, openai/gpt-5.6-terra, qwen/qwen3.8-max-0902, openai/gpt-6-sol, openai/gpt-6.1-terra, openai/gpt-5.4 and openai/o3. On this suite their order relative to the leader is unresolved, not a ranking.

Table 2. Each model against anthropic/claude-opus-5.5 (85.0%). Gap: leader's score minus the model's, in points. Verdict: significant when $|z| \geq 1.96$.

Model	Score	Gap (pts)	z	Verdict	Only leader / only model right	McNemar p (Holm)
google/gemini-4-argon	83.8%	1.2	0.22	too close to call	4 / 3	1.000
google/gemini-3.8-flash	82.5%	2.5	0.43	too close to call	5 / 3	1.000
openai/gpt-5.3-codex	82.5%	2.5	0.43	too close to call	4 / 2	1.000
anthropic/claude-fable-5.1	81.2%	3.7	0.63	too close to call	5 / 2	1.000
x-ai/grok-4.7	80.0%	5.0	0.83	too close to call	7 / 3	1.000
anthropic/claude-sonnet-5.5	78.8%	6.3	1.03	too close to call	6 / 1	1.000
meta/muse-wonder	78.8%	6.3	1.03	too close to call	7 / 2	1.000
google/gemini-3.1-pro-preview	77.5%	7.5	1.22	too close to call	7 / 1	0.703
meta/muse-spark-open	77.5%	7.5	1.22	too close to call	9 / 3	1.000
moonshotai/kimi-k3.1-pro	77.5%	7.5	1.22	too close to call	7 / 1	0.703
openai/gpt-6-astra	76.2%	8.8	1.40	too close to call	7 / 0	0.234
openai/gpt-6.1-sol	76.2%	8.8	1.40	too close to call	7 / 0	0.234
openai/gpt-6.1-terra-pro	76.2%	8.8	1.40	too close to call	7 / 0	0.234
openai/gpt-5.6-terra	75.0%	10.0	1.58	too close to call	9 / 1	0.258
qwen/qwen3.8-max-0902	75.0%	10.0	1.58	too close to call	10 / 2	0.424
openai/gpt-6-sol	73.8%	11.2	1.76	too close to call	9 / 0	0.078
openai/gpt-6.1-terra	73.8%	11.2	1.76	too close to call	9 / 0	0.078
openai/gpt-5.4	72.5%	12.5	1.93	too close to call	11 / 1	0.114
openai/o3	72.5%	12.5	1.93	too close to call	12 / 2	0.216
deepseek/deepseek-v4-pro-0813	71.2%	13.7	2.10	leader more accurate	14 / 3	0.216
thinkingmachines/inkling	66.2%	18.8	2.76	leader more accurate	18 / 3	0.031
moonshotai/kimi-k2.7-code	64.6%	20.4	2.97	leader more accurate	16 / 0	<0.001
stepfun/step-5-preview	63.8%	21.3	3.08	leader more accurate	18 / 1	0.002
nvidia/nemotron-3-ultra-550b-a55b	60.0%	25.0	3.54	leader more accurate	21 / 1	<0.001
qwen/qwen3.8-2.4t-a95b	58.2%	26.8	3.75	leader more accurate	22 / 0	<0.001
mistralai/mistral-large-4-0	50.0%	35.0	4.73	leader more accurate	28 / 0	<0.001
cohere/command-a	45.6%	39.4	5.23	leader more accurate	34 / 3	<0.001
qwen/qwen3.5-397b-a17b	32.5%	52.5	6.74	leader more accurate	42 / 0	<0.001
mistralai/mistral-medium-3-5	29.1%	55.9	7.12	leader more accurate	45 / 0	<0.001
nousresearch/hermes-4-405b	26.6%	58.4	7.42	leader more accurate	46 / 0	<0.001

As a paired check, an exact McNemar test with Holm-Bonferroni correction (last two columns) agrees with the z-test verdict in 29 of 30 comparisons. It does not confirm the z-test's verdict for deepseek/deepseek-v4-pro-0813, so that gap should be read with extra caution. The terms are explained below and in the methodology paper.

Two-proportion z-test

Tests whether two success rates differ by more than chance would explain. The gap between the two scores is divided by its standard error (pooled from both models); $|z| \geq 1.96$ corresponds to 95% confidence, two-sided. See: https://en.wikipedia.org/wiki/Two-proportion_Z-test

McNemar's test

A test for paired yes/no outcomes. Because both models answered the same questions, it looks only at the questions where exactly one of them was right and asks whether that split could be a coin flip. It ignores questions both got right or both got wrong. See: https://en.wikipedia.org/wiki/McNemar%27s_test

Holm-Bonferroni correction

When one model is compared with many others, some comparisons look significant by luck alone.

Holm's method raises the bar step by step so that the chance of any false "significant" result across the whole set stays at 5%. See: https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method

5 Results by subject category

Categories hold between 6 and 30 questions each, so per-category results are shown as counts and are noisy; they indicate where models differ, not reliable sub-rankings.

Table 3. Best result per category within the group and the model(s) achieving it.

Category	Questions	Best	Share	Model(s)
Science	30	28/30	93.3%	anthropic/claude-opus-5.5, openai/gpt-5.3-codex
Engineering	15	13/15	86.7%	google/gemini-4-argon, anthropic/claude-fable-5.1, anthropic/claude-sonnet-5.5
Medicine	11	7/11	63.6%	7 models tied
Economics	9	8/9	88.9%	17 models tied
Law	6	6/6	100.0%	google/gemini-4-argon, google/gemini-3.8-flash
Humanities & social science	9	9/9	100.0%	meta/muse-wonder, qwen/qwen3.8-max-0902, nvidia/nemotron-3-ultra-550b-a55b

6 Cost and the cost-accuracy trade-off

Billed cost per question ranges from \$0.0012 (thinkingmachines/inkling) to \$0.025, with a median of \$0.0088. Spearman's rank correlation between score and cost per question is 0.06: within this group, paying more per question is only weakly related to a higher score.

6 models are on the cost-accuracy Pareto set, from cheapest to most accurate: thinkingmachines/inkling (66.2%, \$0.0012), openai/gpt-6.1-terra (73.8%, \$0.0013), openai/gpt-6.1-sol (76.2%, \$0.0029), google/gemini-3.8-flash (82.5%, \$0.0037), google/gemini-4-argon (83.8%, \$0.0062) and anthropic/claude-opus-5.5 (85.0%, \$0.015). Every other model in the group is matched or beaten on both score and cost by at least one of these. Because scores carry the uncertainty shown in Table 1, models just off the set may be equivalent in practice.

Under the 10% rule, thinkingmachines/inkling and openai/gpt-6.1-terra cost about the same per question as the cheapest model.

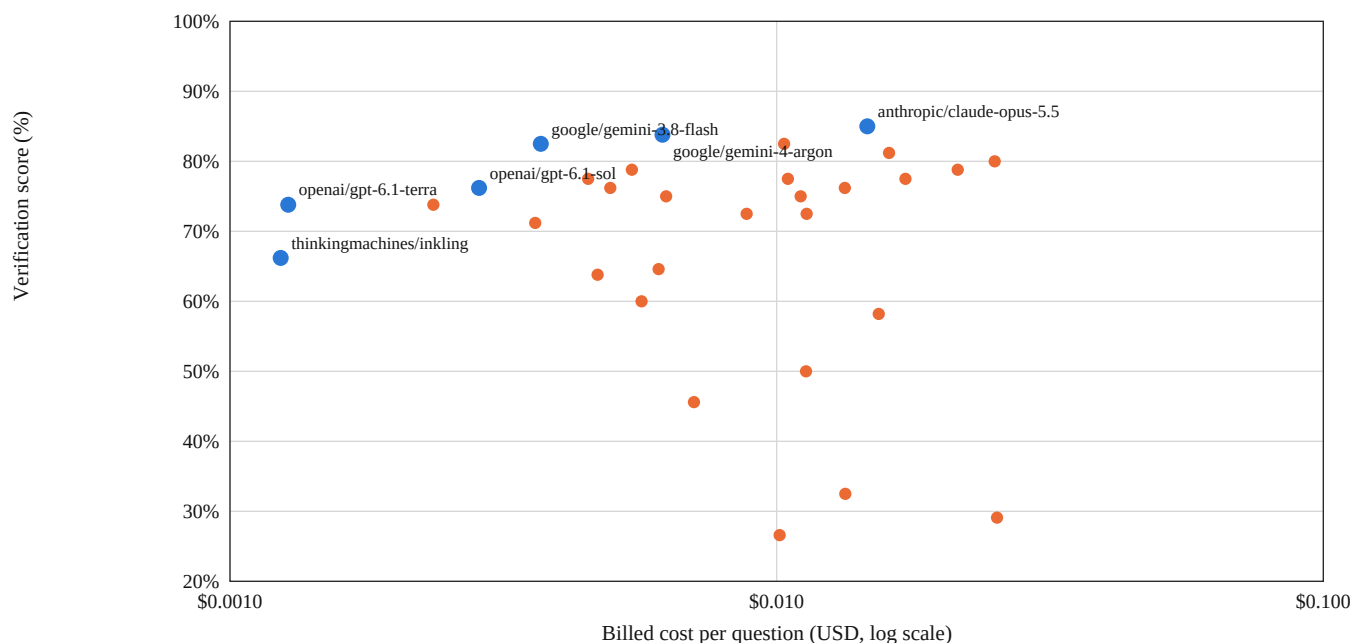


Figure 2. Score versus billed cost per question (log scale), frontier models. Blue, labelled points: the cost-accuracy Pareto set.

Spearman's rank correlation

Measures how consistently one quantity rises with another using only their rank order, from -1 to +1.

Here it asks whether more expensive models tend to score higher. See:

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Pareto set (cost-accuracy frontier)

The models for which no other model is both at least as accurate and at least as cheap. Any model off the set is beaten on both counts by one on it. See: https://en.wikipedia.org/wiki/Pareto_front

7 Speed and latency

Median output speed across the group is 94 tokens per second and median time to first token is 1.29 s. The highest median output speed is qwen/qwen3.5-397b-a17b at 240 tokens per second; the lowest median time to first token is nousresearch/hermes-4-405b at 0.35 s. Output speed is withheld for openai/gpt-6.1-terra-pro, whose runs delivered the answer in a single burst after a silent interval (methodology, Section 5); latency is still reported. Under the 10% rule, meta/muse-spark-open is similar in output speed to qwen/qwen3.5-397b-a17b, and moonshotai/kimi-k2.7-code is similar to nousresearch/hermes-4-405b in time to first token. Speed was measured from one client machine; absolute latencies depend on network distance to the host.

Table 4. Median streaming measurements (fixed prose prompt, medians over runs). TTFT includes reasoning tokens.

Model	Output tok/s	TTFT	End-to-end	Runs
qwen/qwen3.5-397b-a17b	240	0.65 s	3.53 s	3
meta/muse-spark-open	220	0.50 s	3.25 s	3
meta/muse-wonder	175	16.06 s	20.80 s	2
openai/o3	163	1.23 s	3.92 s	3
mistralai/mistral-medium-3-5	145	0.46 s	5.36 s	3

Model	Output tok/s	TTFT	End-to-end	Runs
google/gemini-3.8-flash	142	3.17 s	5.64 s	3
moonshotai/kimi-k2.7-code	136	0.36 s	5.26 s	3
nvidia/nemotron-3-ultra-550b-a55b	136	0.81 s	3.65 s	3
anthropic/claude-opus-5.5	129	2.34 s	7.58 s	3
anthropic/claude-sonnet-5.5	126	0.73 s	6.22 s	3
google/gemini-3.1-pro-preview	124	3.43 s	8.90 s	3
openai/gpt-6.1-terra	122	2.40 s	5.83 s	3
thinkingmachines/inkling	118	0.55 s	4.98 s	3
openai/gpt-5.6-terra	101	1.15 s	5.29 s	3
openai/gpt-5.3-codex	95	1.36 s	5.25 s	3
stepfun/step-5-preview	93	0.63 s	8.16 s	3
openai/gpt-6-sol	92	0.91 s	4.35 s	3
mistralai/mistral-large-4-0	83	1.05 s	9.49 s	3
x-ai/grok-4.7	77	1.88 s	8.00 s	3
anthropic/claude-fable-5.1	70	2.10 s	11.04 s	3
google/gemini-4-argon	69	3.29 s	9.88 s	3
deepseek/deepseek-v4-pro-0813	56	0.81 s	10.07 s	3
openai/gpt-5.4	55	1.72 s	8.73 s	3
openai/gpt-6-astra	49	1.29 s	8.92 s	3
qwen/qwen3.8-max-0902	46	1.16 s	12.22 s	3
qwen/qwen3.8-2.4t-a95b	44	3.14 s	14.79 s	3
openai/gpt-6.1-sol	41	2.10 s	10.81 s	3
nousresearch/hermes-4-405b	41	0.35 s	17.52 s	3
moonshotai/kimi-k3.1-pro	38	1.31 s	19.62 s	3
cohere/command-a	7	1.85 s	53.35 s	3
openai/gpt-6.1-terra-pro	withheld	8.54 s	9.02 s	3

8 Missing answers

9 models returned no extractable answer on at least 8 of the graded questions: qwen/qwen3.5-397b-a17b (50), mistralai/mistral-medium-3-5 (47), mistralai/mistral-large-4-0 (35), qwen/qwen3.8-2.4t-a95b (26), moonshotai/kimi-k2.7-code (19), nvidia/nemotron-3-ultra-550b-a55b (18), stepfun/step-5-preview (17), deepseek/deepseek-v4-pro-0813 (12) and qwen/qwen3.8-max-0902 (10). Across all models in both groups, 927 of the 939 missing answers ended with finish reason "length" and 932 had an empty answer text, which points to the 4000-token limit being spent on reasoning rather than to wrong answers. These models' scores measure performance under this budget; with a larger budget they may score higher. I do not adjust for this.

9 Discussion

The main result is a limit on what can be claimed. With 80 questions, 19 models in this group are too close to call against the leader, so a single ordered leaderboard would overstate what was measured. The data support coarser statements: which models sit clearly below the leader (Table 2), which models give the best accuracy for their billed cost (the Pareto set), and where low scores come from missing answers under a fixed token budget rather than from wrong answers.

Readers choosing a model for production work should weigh these results against task fit: the suite measures expert-level multiple-choice knowledge and reasoning, not coding, tool use or long-context work.

10 Limitations

- The group boundary is a list-price rule; it does not imply that every model above it is more capable than every model below it.
- 80 questions in one task format; intervals are roughly +/-10 points.
- One reasoning setting and a 4000-token limit for all models.
- OpenRouter routing may serve a model from different hosts; prices, hosts and model versions change after the test dates.
- Speed and latency come from a single client machine.

11 Data availability

Per-model results (llm-evaluation-models.csv) and per-question evidence with proof hashes and generation IDs (llm-evaluation-evidence.csv, llm-evaluation-evidence.json) are published at suhasbhairav.com/research under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>). Every table in this paper can be recomputed from those files.

References

- [1] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209-212, 1927. <https://doi.org/10.1080/01621459.1927.10502953>
- [2] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153-157, 1947. <https://doi.org/10.1007/BF02295996>
- [3] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65-70, 1979. <https://www.jstor.org/stable/4615733>
- [4] M-A-P Team et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. arXiv:2502.14739, 2025. <https://arxiv.org/abs/2502.14739>
- [5] Y. Wang et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. arXiv:2406.01574, 2024. <https://arxiv.org/abs/2406.01574>
- [6] OpenRouter. API reference: chat completions and generation records. <https://openrouter.ai/docs/api-reference/overview>
- [7] S. Bhairav. How every number is produced: a verifiable methodology for evaluating large language models. [suhasbhairav.com](https://suhasbhairav.com/research/verifiable-llm-evaluation-methodology), 2026. <https://suhasbhairav.com/research/verifiable-llm-evaluation-methodology>