

# Non-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 41 non-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 1.3% to 75.0% (median 48.8%; chance is 10%). The highest-scoring model, google/gemini-3.5-flash-lite, answered 60 of 80 correctly (75.0%, 95% Wilson interval 64.5% to 83.2%). A two-proportion z-test ( $|z| \geq 1.96$ ) cannot separate 12 other models from the leader, so the top of the ranking is too close to call on this sample. Median billed cost was \$0.00062 per question, and the rank correlation between score and cost was 0.18. 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; non-frontier models; cost per question; output speed*

## 1 Introduction

This paper reports results for 41 non-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 31 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%.

meta-llama/llama-4-maverick (2 host failures) had requests that failed after retries; these are excluded from the denominator. qwen/qwen3.7-plus (1), inclusionai/ling-3.1-flash (1), meta-llama/llama-4-scout (1), google/gemma-4-31b-it (1), google/gemma-4-26b-a4b-it (2) and rekaai/reka-flash-3 (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 41 non-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 |
|---------------------------------------|-------|-----------|------------------|---------|-----------------|-------------------|
| google/gemini-3.5-flash-lite          | 75.0% | 64.5-83.2 | 60 / 80          | 0       | \$0.0034        | \$0.850           |
| deepseek/deepseek-v4.1-flash-1017     | 73.8% | 63.2-82.1 | 59 / 80          | 12      | \$0.00091       | \$0.263           |
| openai/gpt-6-luna                     | 72.5% | 61.9-81.1 | 58 / 80          | 0       | \$0.00025       | \$0.200           |
| deepseek/deepseek-v4-flash-vision-exp | 71.2% | 60.5-80.0 | 57 / 80          | 2       | \$0.0011        | \$0.323           |
| anthropic/claude-haiku-5.5            | 68.8% | 57.9-77.8 | 55 / 80          | 1       | \$0.00048       | \$0.200           |
| xiaomi/mimo-v2.6-pro                  | 68.8% | 57.9-77.8 | 55 / 80          | 14      | \$0.0020        | \$0.544           |
| thinkingmachines/inkling-small        | 67.5% | 56.6-76.8 | 54 / 80          | 0       | \$0.00047       | \$0.637           |
| z-ai/glm-5.3                          | 67.5% | 56.6-76.8 | 54 / 80          | 4       | \$0.00093       | \$0.409           |
| z-ai/glm-5.3-flash                    | 67.5% | 56.6-76.8 | 54 / 80          | 2       | \$0.00015       | \$0.237           |
| minimax/minimax-m3.1-flash            | 66.2% | 55.4-75.7 | 53 / 80          | 9       | \$0.00087       | \$0.263           |
| qwen/qwen3.8-27b                      | 63.8% | 52.8-73.4 | 51 / 80          | 14      | \$0.0043        | \$0.775           |
| qwen/qwen3-coder-next                 | 62.5% | 51.5-72.3 | 50 / 80          | 0       | \$0.0017        | \$0.290           |
| xiaomi/mimo-v2.6-flash                | 62.5% | 51.5-72.3 | 50 / 80          | 18      | \$0.00056       | \$0.175           |
| qwen/qwen3.7-plus                     | 59.5% | 48.5-69.6 | 47 / 79          | 23      | \$0.0033        | \$0.560           |
| meta-llama/llama-4-maverick           | 59.0% | 47.9-69.2 | 46 / 78          | 0       | \$0.00059       | \$0.304           |
| openai/gpt-oss-120b                   | 58.8% | 47.8-68.9 | 47 / 80          | 0       | \$0.00013       | \$0.263           |
| tencent/hy3                           | 58.8% | 47.8-68.9 | 47 / 80          | 21      | \$0.0012        | \$0.231           |
| inception/mercury-2.5                 | 57.5% | 46.6-67.7 | 46 / 80          | 1       | \$0.00011       | \$0.068           |
| minimax/minimax-m3                    | 57.5% | 46.6-67.7 | 46 / 80          | 22      | \$0.0026        | \$0.525           |
| openai/gpt-oss-20b                    | 51.2% | 40.5-61.9 | 41 / 80          | 0       | \$0.00007       | \$0.036           |
| meituan/longcat-2.0                   | 48.8% | 38.1-59.5 | 39 / 80          | 34      | \$0.0032        | \$0.525           |
| upstage/solar-pro4                    | 45.0% | 34.6-55.9 | 36 / 80          | 39      | \$0.0018        | \$0.263           |
| nvidia/nemotron-3.5-lightning         | 43.8% | 33.4-54.7 | 35 / 80          | 29      | \$0.00049       | \$0.102           |
| meta-llama/llama-3.3-70b-instruct     | 42.5% | 32.3-53.4 | 34 / 80          | 0       | \$0.00053       | \$0.155           |
| mistralai/ministral-8b-2512           | 42.5% | 32.3-53.4 | 34 / 80          | 1       | \$0.00027       | \$0.150           |
| inclusionai/ling-3.1-flash            | 41.8% | 31.5-52.8 | 33 / 79          | 40      | \$0.00062       | \$0.113           |
| meta-llama/llama-4-scout              | 41.8% | 31.5-52.8 | 33 / 79          | 0       | \$0.00037       | \$0.150           |
| microsoft/phi-4                       | 41.2% | 31.1-52.2 | 33 / 80          | 0       | \$0.00012       | \$0.087           |
| upstage/solar-mini4                   | 41.2% | 31.1-52.2 | 33 / 80          | 40      | \$0.00053       | \$0.087           |
| nvidia/nemotron-3.5-super-vl          | 40.0% | 30.0-51.0 | 32 / 80          | 43      | \$0.0013        | \$0.237           |
| qwen/qwen3.6-35b-a3b                  | 37.5% | 27.7-48.5 | 30 / 80          | 40      | \$0.0037        | \$0.313           |
| ibm-granite/granite-4.2-8b            | 36.2% | 26.6-47.2 | 29 / 80          | 0       | \$0.00008       | \$0.107           |
| mistralai/ministral-14b-2512          | 33.8% | 24.3-44.6 | 27 / 80          | 1       | \$0.00031       | \$0.200           |
| mistralai/mistral-small-2603          | 32.5% | 23.2-43.4 | 26 / 80          | 44      | \$0.0019        | \$0.263           |
| google/gemma-4-31b-it                 | 30.4% | 21.3-41.2 | 24 / 79          | 50      | \$0.0014        | \$0.250           |
| arcee-ai/trinity-large-thinking       | 28.8% | 20.0-39.5 | 23 / 80          | 50      | \$0.0025        | \$0.388           |
| mistralai/ministral-3b-2512           | 26.2% | 17.9-36.8 | 21 / 80          | 2       | \$0.00019       | \$0.100           |
| ibm-granite/granite-4.0-h-micro       | 25.0% | 16.8-35.5 | 20 / 80          | 1       | \$0.00005       | \$0.041           |
| amazon/nova-micro-v1                  | 22.5% | 14.7-32.8 | 18 / 80          | 0       | \$0.00009       | \$0.061           |
| google/gemma-4-26b-a4b-it             | 15.4% | 9.0-25.0  | 12 / 78          | 66      | \$0.0011        | \$0.142           |
| rekaai/reka-flash-3                   | 1.3%  | 0.2-6.8   | 1 / 79           | 66      | \$0.00081       | \$0.125           |

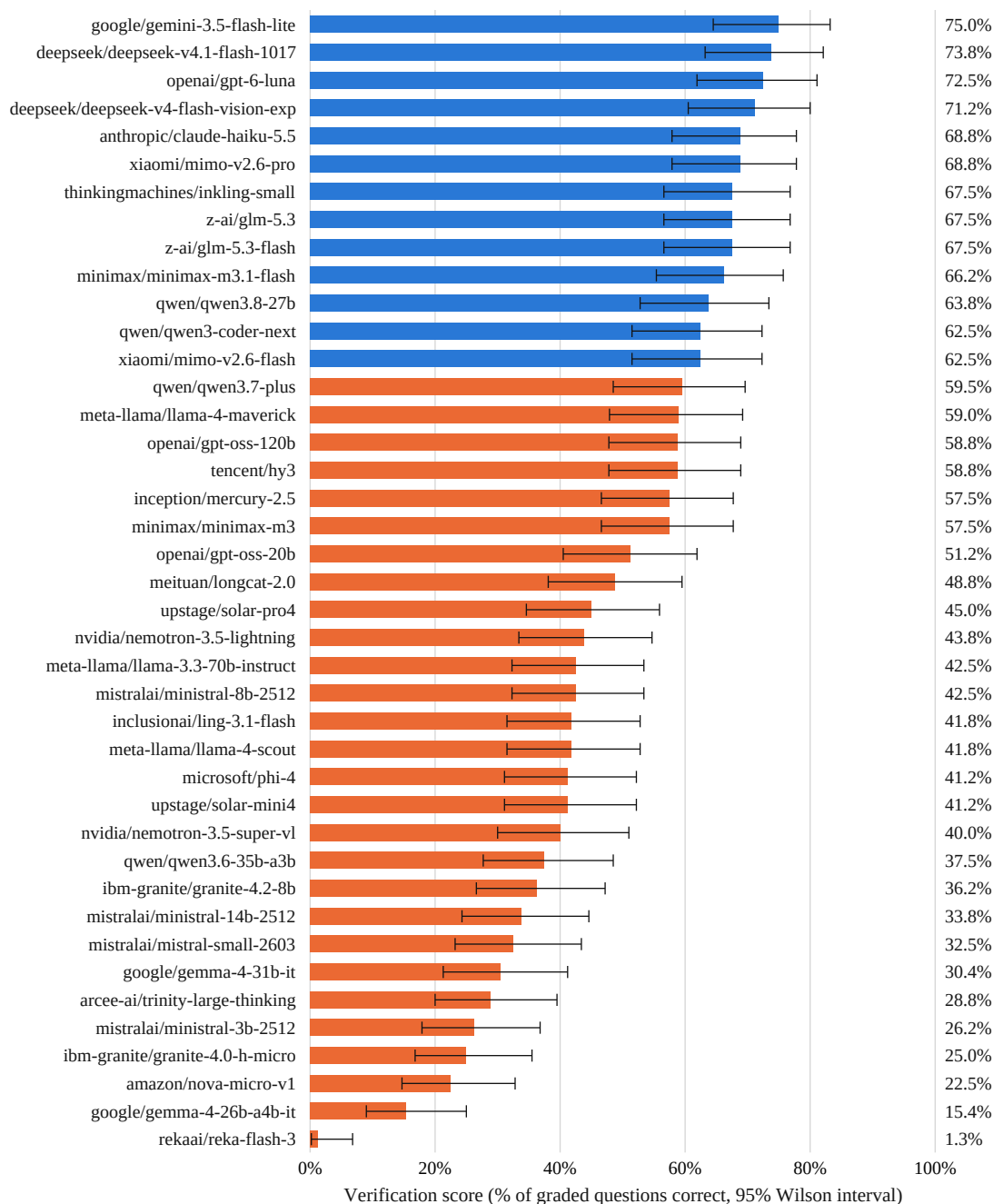


**Figure 1. Score with 95% Wilson interval, non-frontier models. Blue: the leader and the models too close to call against it (two-proportion z-test,  $|z| < 1.96$ ). Orange: the leader is significantly more accurate.**

Scores range from 1.3% to 75.0%, with a median of 48.8%. The highest score is google/gemini-3.5-flash-lite with 60 of 80 (75.0%; 95% CI 64.5% to 83.2%). For comparison, the median of the frontier group is 75.0%.

#### 4 Which differences are real

Table 2 tests the leader, google/gemini-3.5-flash-lite, 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 40 comparisons, 28 gaps are significant and 12

are too close to call. The models too close to call against the leader are deepseek/deepseek-v4.1-flash-1017, openai/gpt-6-luna, deepseek/deepseek-v4-flash-vision-exp, anthropic/claude-haiku-5.5, xiaomi/mimo-v2.6-pro, thinkingmachines/inkling-small, z-ai/glm-5.3, z-ai/glm-5.3-flash, minimax/minimax-m3.1-flash, qwen/qwen3.8-27b, qwen/qwen3-coder-next and xiaomi/mimo-v2.6-flash. On this suite their order relative to the leader is unresolved, not a ranking.

**Table 2. Each model against google/gemini-3.5-flash-lite (75.0%). 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) |
|---------------------------------------|-------|-----------|------|----------------------|--------------------------------|------------------|
| deepseek/deepseek-v4.1-flash-1017     | 73.8% | 1.2       | 0.18 | too close to call    | 8 / 7                          | 1.000            |
| openai/gpt-6-luna                     | 72.5% | 2.5       | 0.36 | too close to call    | 9 / 7                          | 1.000            |
| deepseek/deepseek-v4-flash-vision-exp | 71.2% | 3.7       | 0.54 | too close to call    | 9 / 6                          | 1.000            |
| anthropic/claude-haiku-5.5            | 68.8% | 6.3       | 0.88 | too close to call    | 13 / 8                         | 1.000            |
| xiaomi/mimo-v2.6-pro                  | 68.8% | 6.3       | 0.88 | too close to call    | 10 / 5                         | 1.000            |
| thinkingmachines/inkling-small        | 67.5% | 7.5       | 1.05 | too close to call    | 10 / 4                         | 1.000            |
| z-ai/glm-5.3                          | 67.5% | 7.5       | 1.05 | too close to call    | 9 / 3                          | 1.000            |
| z-ai/glm-5.3-flash                    | 67.5% | 7.5       | 1.05 | too close to call    | 9 / 3                          | 1.000            |
| minimax/minimax-m3.1-flash            | 66.2% | 8.8       | 1.21 | too close to call    | 13 / 6                         | 1.000            |
| qwen/qwen3.8-27b                      | 63.8% | 11.3      | 1.54 | too close to call    | 16 / 7                         | 0.931            |
| qwen/qwen3-coder-next                 | 62.5% | 12.5      | 1.71 | too close to call    | 12 / 2                         | 0.181            |
| xiaomi/mimo-v2.6-flash                | 62.5% | 12.5      | 1.71 | too close to call    | 13 / 3                         | 0.234            |
| qwen/qwen3.7-plus                     | 59.5% | 15.5      | 2.08 | leader more accurate | 17 / 5                         | 0.220            |
| meta-llama/llama-4-maverick           | 59.0% | 16.0      | 2.14 | leader more accurate | 15 / 3                         | 0.113            |
| openai/gpt-oss-120b                   | 58.8% | 16.2      | 2.18 | leader more accurate | 20 / 7                         | 0.230            |
| tencent/hy3                           | 58.8% | 16.2      | 2.18 | leader more accurate | 16 / 3                         | 0.078            |
| inception/mercury-2.5                 | 57.5% | 17.5      | 2.34 | leader more accurate | 18 / 4                         | 0.078            |
| minimax/minimax-m3                    | 57.5% | 17.5      | 2.34 | leader more accurate | 19 / 5                         | 0.106            |
| openai/gpt-oss-20b                    | 51.2% | 23.8      | 3.11 | leader more accurate | 23 / 4                         | 0.006            |
| meituan/longcat-2.0                   | 48.8% | 26.3      | 3.42 | leader more accurate | 24 / 3                         | 0.001            |
| upstage/solar-pro4                    | 45.0% | 30.0      | 3.87 | leader more accurate | 25 / 1                         | <0.001           |
| nvidia/nemotron-3.5-lightning         | 43.8% | 31.3      | 4.02 | leader more accurate | 28 / 3                         | <0.001           |
| meta-llama/llama-3.3-70b-instruct     | 42.5% | 32.5      | 4.18 | leader more accurate | 29 / 3                         | <0.001           |
| mistralai/ministral-8b-2512           | 42.5% | 32.5      | 4.18 | leader more accurate | 27 / 1                         | <0.001           |
| inclusionai/ling-3.1-flash            | 41.8% | 33.2      | 4.25 | leader more accurate | 28 / 2                         | <0.001           |
| meta-llama/llama-4-scout              | 41.8% | 33.2      | 4.25 | leader more accurate | 29 / 3                         | <0.001           |
| microsoft/phi-4                       | 41.2% | 33.8      | 4.33 | leader more accurate | 30 / 3                         | <0.001           |
| upstage/solar-mini4                   | 41.2% | 33.8      | 4.33 | leader more accurate | 30 / 3                         | <0.001           |
| nvidia/nemotron-3.5-super-vl          | 40.0% | 35.0      | 4.48 | leader more accurate | 29 / 1                         | <0.001           |
| qwen/qwen3.6-35b-a3b                  | 37.5% | 37.5      | 4.78 | leader more accurate | 34 / 4                         | <0.001           |
| ibm-granite/granite-4.2-8b            | 36.2% | 38.8      | 4.93 | leader more accurate | 34 / 3                         | <0.001           |
| mistralai/ministral-14b-2512          | 33.8% | 41.3      | 5.24 | leader more accurate | 36 / 3                         | <0.001           |
| mistralai/mistral-small-2603          | 32.5% | 42.5      | 5.39 | leader more accurate | 36 / 2                         | <0.001           |
| google/gemma-4-31b-it                 | 30.4% | 44.6      | 5.64 | leader more accurate | 36 / 1                         | <0.001           |
| arcee-ai/trinity-large-thinking       | 28.8% | 46.3      | 5.85 | leader more accurate | 38 / 1                         | <0.001           |
| mistralai/ministral-3b-2512           | 26.2% | 48.8      | 6.17 | leader more accurate | 40 / 1                         | <0.001           |
| ibm-granite/granite-4.0-h-micro       | 25.0% | 50.0      | 6.32 | leader more accurate | 42 / 2                         | <0.001           |
| amazon/nova-micro-v1                  | 22.5% | 52.5      | 6.64 | leader more accurate | 44 / 2                         | <0.001           |
| google/gemma-4-26b-a4b-it             | 15.4% | 59.6      | 7.52 | leader more accurate | 47 / 0                         | <0.001           |
| rekaai/reka-flash-3                   | 1.3%  | 73.7      | 9.56 | leader more accurate | 59 / 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 34 of 40 comparisons. It does not confirm the z-test's verdict for qwen/qwen3.7-plus, meta-llama/llama-4-maverick, openai/gpt-oss-120b, tencent/hy3, inception/mercury-2.5 and minimax/minimax-m3, so those gaps 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](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](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](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        | 26/30 | 86.7%  | deepseek/deepseek-v4.1-flash-1017                                                                             |
| Engineering                 | 15        | 10/15 | 66.7%  | google/gemini-3.5-flash-lite, deepseek/deepseek-v4-flash-vision-exp, anthropic/claude-haiku-5.5, z-ai/glm-5.3 |
| Medicine                    | 11        | 7/11  | 63.6%  | google/gemini-3.5-flash-lite                                                                                  |
| Economics                   | 9         | 8/9   | 88.9%  | 6 models tied                                                                                                 |
| Law                         | 6         | 4/6   | 66.7%  | 7 models tied                                                                                                 |
| Humanities & social science | 9         | 9/9   | 100.0% | 5 models tied                                                                                                 |

## 6 Cost and the cost-accuracy trade-off

Billed cost per question ranges from \$0.00005 (ibm-granite/granite-4.0-h-micro) to \$0.0043, with a median of \$0.00062. Spearman's rank correlation between score and cost per question is 0.18: within this group, paying more per question is only weakly related to a higher score.

8 models are on the cost-accuracy Pareto set, from cheapest to most accurate: ibm-granite/granite-4.0-h-micro (25.0%, \$0.00005), openai/gpt-oss-20b (51.2%, \$0.00007), inception/mercury-2.5 (57.5%, \$0.00011), openai/gpt-oss-120b (58.8%, \$0.00013), z-ai/glm-5.3-flash (67.5%, \$0.00015), openai/gpt-6-luna (72.5%, \$0.00025), deepseek/deepseek-v4.1-flash-1017 (73.8%, \$0.00091) and google/gemini-3.5-flash-lite (75.0%, \$0.0034). 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, no other model comes within 10% of ibm-granite/granite-4.0-h-micro's cost per question.

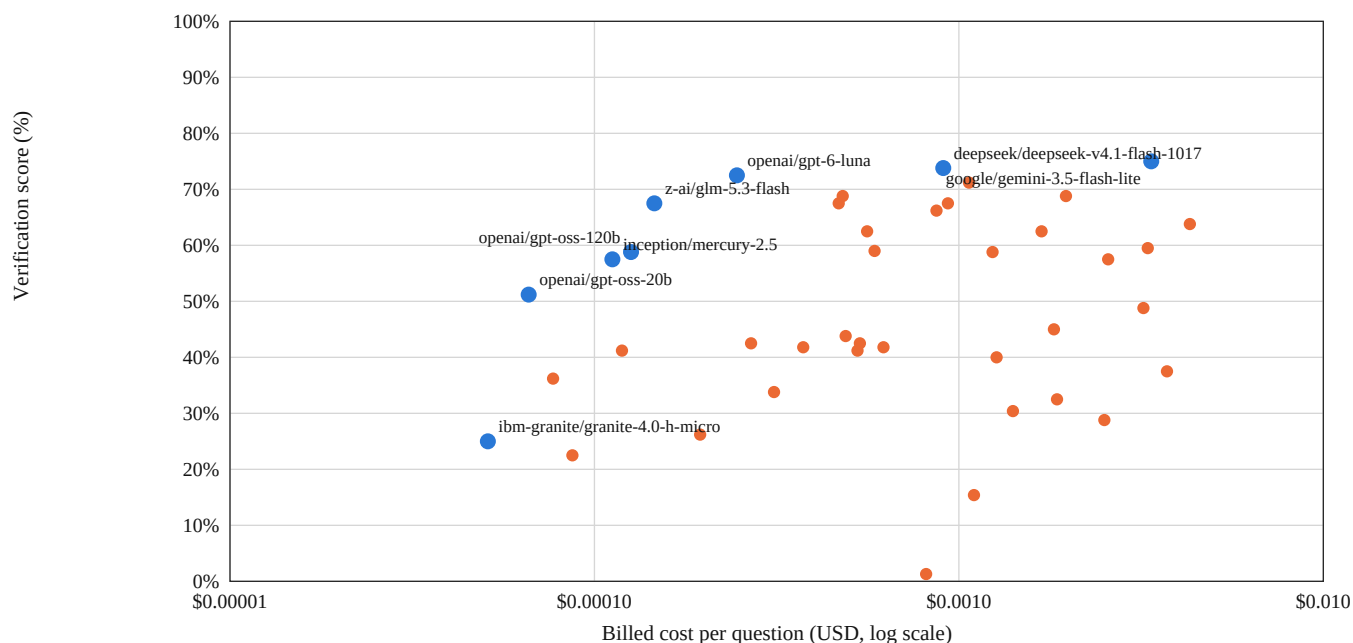


**Figure 2. Score versus billed cost per question (log scale), non-frontier models. Blue, labelled points: the cost-accuracy Pareto set.**

### 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](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](https://en.wikipedia.org/wiki/Pareto_front)

## 7 Speed and latency

Median output speed across the group is 103 tokens per second and median time to first token is 0.65 s. The highest median output speed is nvidia/nemotron-3.5-super-vl at 715 tokens per second; the lowest median time to first token is mistralai/ministral-3b-2512 at 0.30 s. Under the 10% rule, no other model comes within 10% of nvidia/nemotron-3.5-super-vl's output speed, and mistralai/ministral-8b-2512 is similar to mistralai/ministral-3b-2512 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 |
|-------------------------------------------------|--------------|--------|------------|------|
| <a href="#">nvidia/nemotron-3.5-super-vl</a>    | 715          | 0.65 s | 1.53 s     | 3    |
| <a href="#">google/gemini-3.5-flash-lite</a>    | 409          | 1.07 s | 2.77 s     | 3    |
| <a href="#">arcee-ai/trinity-large-thinking</a> | 329          | 0.57 s | 2.73 s     | 3    |
| <a href="#">mistralai/ministral-3b-2512</a>     | 252          | 0.30 s | 1.61 s     | 3    |
| <a href="#">thinkingmachines/inkling-small</a>  | 233          | 0.51 s | 3.59 s     | 3    |
| <a href="#">anthropic/claude-haiku-5.5</a>      | 192          | 0.68 s | 3.92 s     | 3    |
| <a href="#">amazon/nova-micro-v1</a>            | 180          | 0.47 s | 2.11 s     | 3    |

| Model                                 | Output tok/s | TTFT   | End-to-end | Runs |
|---------------------------------------|--------------|--------|------------|------|
| google/gemma-4-26b-a4b-it             | 172          | 0.68 s | 4.84 s     | 3    |
| mistralai/mistral-small-2603          | 170          | 0.48 s | 4.51 s     | 3    |
| deepseek/deepseek-v4-flash-vision-exp | 170          | 0.74 s | 3.18 s     | 3    |
| inception/mercury-2.5                 | 157          | 2.25 s | 6.50 s     | 3    |
| z-ai/glm-5.3                          | 154          | 0.76 s | 3.35 s     | 3    |
| openai/gpt-6-luna                     | 137          | 1.20 s | 4.26 s     | 3    |
| qwen/qwen3.6-35b-a3b                  | 135          | 0.42 s | 5.62 s     | 3    |
| minimax/minimax-m3.1-flash            | 131          | 0.53 s | 4.17 s     | 3    |
| mistralai/ministral-8b-2512           | 124          | 0.32 s | 3.50 s     | 3    |
| minimax/minimax-m3                    | 121          | 0.52 s | 4.23 s     | 3    |
| nvidia/nemotron-3.5-lightning         | 116          | 0.50 s | 6.52 s     | 3    |
| upstage/solar-mini4                   | 107          | 0.69 s | 7.43 s     | 3    |
| qwen/qwen3-coder-next                 | 106          | 0.42 s | 4.19 s     | 3    |
| microsoft/phi-4                       | 103          | 0.47 s | 4.88 s     | 3    |
| openai/gpt-oss-20b                    | 103          | 0.69 s | 4.03 s     | 3    |
| rekaai/reka-flash-3                   | 101          | 0.48 s | 7.38 s     | 3    |
| tencent/hy3                           | 96           | 1.70 s | 9.02 s     | 3    |
| upstage/solar-pro4                    | 91           | 0.61 s | 8.32 s     | 3    |
| z-ai/glm-5.3-flash                    | 83           | 0.50 s | 7.39 s     | 3    |
| openai/gpt-oss-120b                   | 82           | 0.53 s | 5.43 s     | 3    |
| xiaomi/mimo-v2.6-flash                | 77           | 0.41 s | 5.38 s     | 3    |
| meituan/longcat-2.0                   | 67           | 1.36 s | 9.99 s     | 3    |
| qwen/qwen3.7-plus                     | 62           | 1.14 s | 12.35 s    | 3    |
| ibm-granite/granite-4.2-8b            | 60           | 0.63 s | 7.34 s     | 3    |
| google/gemma-4-31b-it                 | 58           | 0.64 s | 12.62 s    | 3    |
| xiaomi/mimo-v2.6-pro                  | 58           | 0.67 s | 7.92 s     | 3    |
| deepseek/deepseek-v4.1-flash-1017     | 57           | 0.88 s | 7.31 s     | 3    |
| qwen/qwen3.8-27b                      | 49           | 0.53 s | 10.59 s    | 3    |
| meta-llama/llama-3.3-70b-instruct     | 42           | 0.68 s | 8.82 s     | 3    |
| ibm-granite/granite-4.0-h-micro       | 38           | 0.70 s | 8.89 s     | 3    |
| mistralai/ministral-14b-2512          | 32           | 0.65 s | 12.74 s    | 3    |
| meta-llama/llama-4-scout              | 15           | 1.16 s | 25.62 s    | 3    |
| meta-llama/llama-4-maverick           | 13           | 0.67 s | 27.93 s    | 3    |
| inclusionai/ling-3.1-flash            | 7            | 1.06 s | 81.35 s    | 3    |

## 8 Missing answers

20 models returned no extractable answer on at least 8 of the graded questions: google/gemma-4-26b-a4b-it (66), rekaai/reka-flash-3 (66), google/gemma-4-31b-it (50), arcee-ai/trinity-large-thinking (50), mistralai/mistral-small-2603 (44), nvidia/nemotron-3.5-super-vl (43), inclusionai/ling-3.1-flash (40), upstage/solar-mini4 (40), qwen/qwen3.6-35b-a3b (40), upstage/solar-pro4 (39), meituan/longcat-2.0 (34), nvidia/nemotron-3.5-lightning (29), qwen/qwen3.7-plus (23), minimax/minimax-m3 (22), tencent/hy3 (21), xiaomi/mimo-v2.6-flash (18), xiaomi/mimo-v2.6-pro (14), qwen/qwen3.8-27b (14), deepseek/deepseek-v4.1-flash-1017 (12) and minimax/minimax-m3.1-flash (9). 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, 12 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](https://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), 2026. <https://suhasbhairav.com/research/verifiable-llm-evaluation-methodology>