# Delphi Model Evaluation Report: General Reasoning in Commercial Fast LLMs

**Gensyn AI Team**

We track the performance of the fast variants of commercial LLMs by popular frontier labs. Over three weeks, we will run evaluations across 11 general AI reasoning benchmarks curated to capture broad real-world reasoning ability. The score of each benchmark is normalized to a 0–100 scale, and a model's final score is the average across all of them, giving a single, clean metric of practical capability.

## Entrant models

We evaluate the following models, using the default API accesses provided and versions current as of February 3, 2026.

1. Claude Haiku 4.5 from Anthropic.
    - Cost: $1/input MTok, $5/output MTok.
2. Gemini 3 Flash Preview from Google.
    - Cost: $0.50/input MTok, $3/output MTok.
3. GPT 5 Mini from OpenAI.
    - Cost: $0.25/input MTok, $2/output MTok.
4. Grok 4.1 Fast Reasoning from xAI.
    - Cost: $0.20/input MTok, $0.50/output MTok.

**Model settings:** We use the default temperature setting (which is typically 1.0). Unless otherwise noted for a given benchmark, we also used default settings for thinking level (also referred to as thinking budget or effort, depending on the provider).

## Average scores

| Rank | Family | Model | Avg. score | Δ rank | Notes |
|---|---|---|---|---|---|
| 1 | xAI | Grok 4.1 Fast Reasoning | 84.71 | - | - |
| 2 | OpenAI | GPT 5 Mini | 79.46 | - | - |
| 3 | Anthropic | Claude Haiku 4.5 | 70.75 | - | - |
| 4 | Google | Gemini 3 Flash Preview | 59.46 | - | - |

*Avg. score* is the mean of normalized benchmark scores. Δ *rank* is the change in rank compared to the last evaluation.

| Benchmark | Claude Haiku 4.5 | Gemini 3 Flash Preview | GPT 5 Mini | Grok 4.1 Fast Reasoning |
|---|---|---|---|---|
| MMLU-Pro (4) | 78.86 | 52.50 | 81.65 | 85.08 |
| GPQA-Diamond (3) | 62.63 | 66.41 | 77.27 | 84.34 |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |
| ? | ▬ | ▬ | ▬ | ▬ |

## Per-Benchmark Breakdown

## Benchmark 2: GPQA Diamond

The GPQA (Graduate-Level Google-Proof Q&A) Diamond dataset consists of highly challenging multiple-choice problems in biology, physics, and chemistry, requiring scientific knowledge and reasoning abilities to answer. Each question is authored by subject-matter experts and is intentionally formulated to be difficult for non-experts to solve, even with access to online resources. The most difficult "Diamond" split includes 198 questions for which both expert annotators selected the correct answer, while most non-experts did not.

**Experimental setup.** We evaluate the models following the guidelines from the benchmark authors (3) and using the OpenAI simple-eval library (1). Every question is presented 10 times, each time with the answer choices in a different randomized order. The models are allowed a maximum output length of 2028 tokens.

We make the following observations:

- Grok 4.1 Fast Reasoning achieves the highest score again.
- Claude Haiku 4.5 has a good score in the first benchmark, but scored the lowest in this one.

## Benchmark 1: MMLU-Pro

MMLU-Pro (4) is an enhanced benchmark for evaluating language models, building on the original MMLU dataset with significantly harder, reasoning-intensive questions and 10 answer options instead of 4. It contains over 12,000 curated questions from academic sources spanning 14 fields, including Biology, Computer Science, Mathematics, Physics, and Law. Experiments demonstrate MMLU-Pro substantially increases difficult with model accuracies dropping 16–33% compared to standard MMLU. Notably, Chain-of-Thought prompting yields greater improvements on MMLU-Pro than on the original benchmark, indicating the dataset demands deeper, more structured reasoning.

We evaluate the models following the guidelines from the benchmark authors (on `TIGER-Lab/MMLU-Pro repository` (2)). The models are allowed a maximum output length of 2048 tokens.

We observed the following from the performance of the Delphi entrant models.

- Gemini 3 Flash Preview performed significantly worse than the other three, scoring around 52%, whereas the remaining models scored between 78% and 85%.

## References

[1] OpenAI simple-eval: a lightweight library for evaluating language models. https://github.com/openai/simple-evals/tree/main.

[2] TIGER-AI-Lab/MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. https://github.com/TIGER-AI-Lab/MMLU-Pro, 2024.

[3] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[4] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.