Delphi Model Evaluation Report: General Reasoning in Lightweight LLMs

Gensyn AI Team

We track the performance of “lightweight” variants of open source LLMs published by popular frontier labs. Over three weeks we will run evaluations across 11 general AI reasoning benchmarks that have been curated to capture broad real-world reasoning ability.

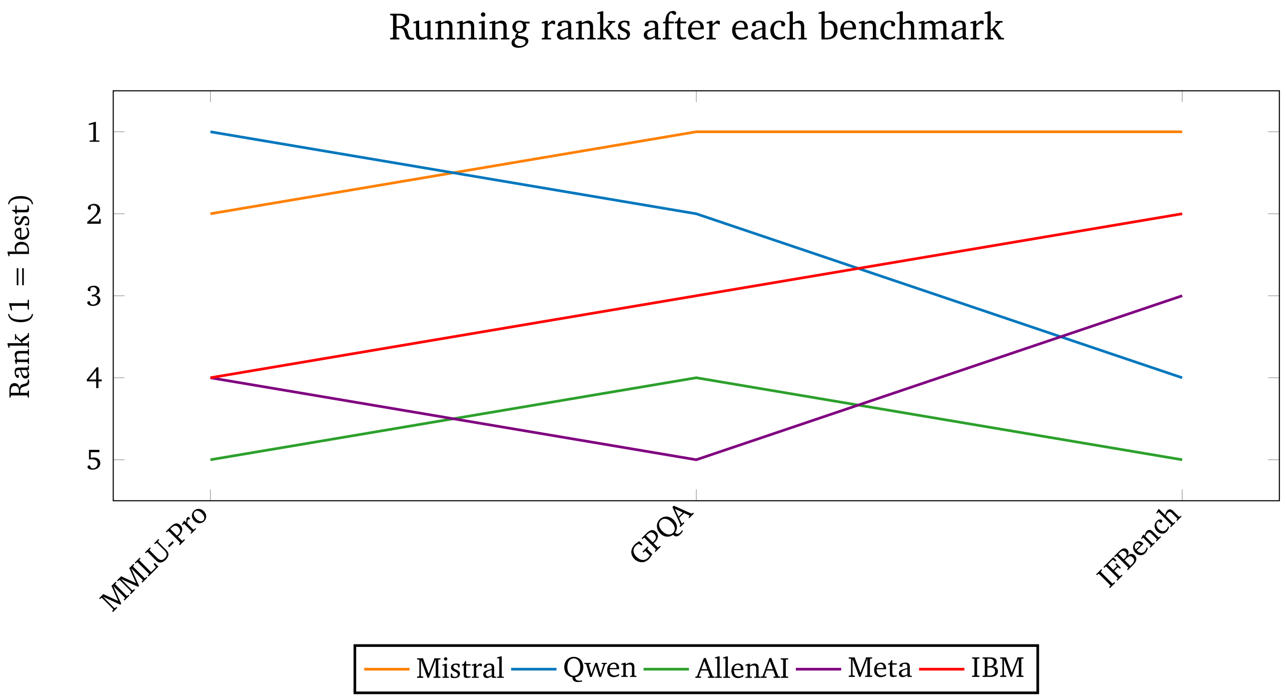
Entrant models

We evaluate the following models, as provided on HuggingFace, using the default datatype provided and versions current as of January 7, 2026.

- Ministral-3-8B-Instruct-2512 from Mistral.
- Qwen3-8B from Qwen.
- Olmo-3-7B-Instruct from AllenAI.
- Llama-3.1-8B-Instruct from Meta.
- granite-4.0-h-tiny from IBM.

Evaluation setup: We perform inference with temperature 0.0 on machines with NVIDIA H100 GPUs using vLLM, unless specified.

Average scores



The running average after 3 evaluations.

Rank	Family	Model	Avg. score	Δ rank	Notes
1	Mistral	Ministral-3-8B-Instruct-2512	42.58	-	-
2	IBM	granite-4.0-h-tiny	36.41	↑ 1	-
3	Meta	Llama-3.1-8B-Instruct	35.81	↑ 2	-
4	Qwen	Qwen3-8B	35.27	↓ 2	-
5	AllenAI	Olmo-3-7B-Instruct	32.31	↓ 1	-

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

Per-Benchmark Breakdown

Benchmark	Ministral-3-8B-Instruct-2512	Qwen3-8B	Olmo-3-7B-Instruct	Llama-3.1-8B-Instruct	granite-4.0-h-tiny
MMLU-Pro (6)	61.86	63.12	41.25	44.86	47.21
GPQA-Diamond (5)	38.48	16.57	32.22	28.43	29.04
IFBench(4)	27.41	26.12	23.47	34.15	32.99
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Rows with “?” as the benchmark will be revealed in the near future as we run those evaluations. Stay tuned for more.

Benchmark 3: IF Bench

The Instruction Following benchmark (IFBench) from AllenAI assesses instruction-following across diverse tasks like counting, formatting, and text manipulation (4). The models are evaluated on how well they follow constraints in their responses, such as “Include keyword beneath in the 23rd sentence, as the 34th word of that sentence,” “Use at least 3 different coordinating conjunctions in the response,” or “The second word in your response and the second to last word in your response should be the word vibrant.”

Experimental setup. We use the 294-question single-turn dataset with 5 repetitions per question, evaluated via pass@1 scoring (1). Responses are scored using the repository’s official evaluation code in loose mode, which tolerates formatting variations and extraneous text by testing multiple output formats (removing leading/trailing lines, asterisks, etc.). Results reflect average prompt-level accuracy across all questions and runs. In our experiments, models are sampled with temperature 0.7, top-p 0.95, and a 4096 token limit.

We make the following observations:

- Llama-3.1-8B-Instruct and granite-4.0-h-tiny have the two highest scores, making this the first benchmark where either of these models places above third.

- The two lowest scores are from Olmo-3-7B-Instruct and Qwen3-8B. Olmo-3-7B-Instruct was lowest in the first benchmark, and Qwen3-8B was lowest in the second.

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 (5) and using the OpenAI simple-eval library (2). 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:

• Qwen3-8B, the model with the best performance from the first benchmark, scored significantly lower than the other models.

• Ministral-3-8B-Instruct-2512 again performed strongly, while Olmo-3-7B-Instruct, Llama-3.1-8B-Instruct, and granite-4.0-h-tiny continued to achieve comparable scores.

Benchmark 1: MMLU-Pro

MMLU-Pro (6) 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` (3)). The models are allowed a maximum output length of 2048 tokens.

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

• Two models, Ministral-3-8B-Instruct-2512 and Qwen3-8B, significantly outperformed the other three. They scored between 61–64%, while the other three scored 41–48%.

References

[1] allenai/IFBench: A new, challenging benchmark for precise instruction following. https://github.com/allenai/IFBench.

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

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

[4] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025.

[5] 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.

[6] 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.