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 1 evaluation.

Rank	Family	Model	Avg. score	Δ rank	Notes
1	Qwen	Qwen3-8B	63.12	-	-
2	Mistral	Ministral-3-8B-Instruct-2512	61.86	-	-
3	IBM	granite-4.0-h-tiny	47.21	-	-
4	Meta	Llama-3.1-8B-Instruct	44.86	-	-
5	AllenAI	Olmo-3-7B-Instruct	41.25	-	-

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

Per-Benchmark Breakdown

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

Benchmark	Ministral-3-8B-Instruct-2512	Qwen3-8B	Olmo-3-7B-Instruct	Llama-3.1-8B-Instruct	granite-4.0-h-tiny
MMLU-Pro (2)	61.86	63.12	41.25	44.86	47.21
?	[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]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
?	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Benchmark 1: MMLU-Pro

MMLU-Pro (2) 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` (1)). 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] TIGER-AI-Lab/MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. https://github.com/TIGER-AI-Lab/MMLU-Pro, 2024.

[2] 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.