Delphi Model Evaluation Report: General Reasoning in Middleweight LLMs

Gensyn AI Team

We track performance of "middleweight" 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 December 7, 2025.

- Qwen3-30B-A3B-Instruct-2507 from Qwen.
- gpt-oss-20b from OpenAI.
- gemma-3-27b-it from Google.
- GLM-4-32B-0414 from Zai-Org.
- Falcon-H1-34B-Instruct from TII-UAE.

Benchmark 1: MMLU-Pro

MMLU-Pro (1) 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 observed the following from the performance of the Delphi entrant models.

- Qwen3-30B-A3B-Instruct-2507 performed the best, scoring nearly 75%.
- The first place model scored less than 10% ahead of the fourth place model.
- The last place model, gpt-oss-20b, scored noticeably lower than the other models.

Experimental setup: We evaluate the models following the guidelines from the benchmark authors (on `TIGER-Lab/MMLU-Pro repository` (2)). We perform inference with a maximum output length of 2048 tokens and temperature 0.0 on machines with NVIDIA H100 GPUs using vLLM.

Overall Scores

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

Rank	Family	Model	Avg. score	Δ vs prev.	Notes
1	Qwen	Qwen3-30B-A3B-Instruct-2507	74.76	±0.0	TBD
2	TIIUAE	Falcon-H1-34B-Instruct	72.27	±0.0	TBD
3	Zai-Org	GLM-4-32B-0414	67.52	±0.0	TBD
4	Google	gemma-3-27b-it	65.19	±0.0	TBD
5	OpenAI	gpt-oss-20b	47.88	±0.0	TBD

Per-Benchmark Breakdown

Thus far we have revealed 1 out of 11 evaluations

Benchmark	Qwen3-30B-A3B-Instruct-2507	gpt-oss-20b	gemma-3-27b-it	GLM-4-32B-0414	Falcon-H1-34B-Instruct
MMLU-Pro (1)	74.76	47.88	65.19	67.52	72.27
?	[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]

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

References

[1] 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:2406.01574, 2004

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