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.

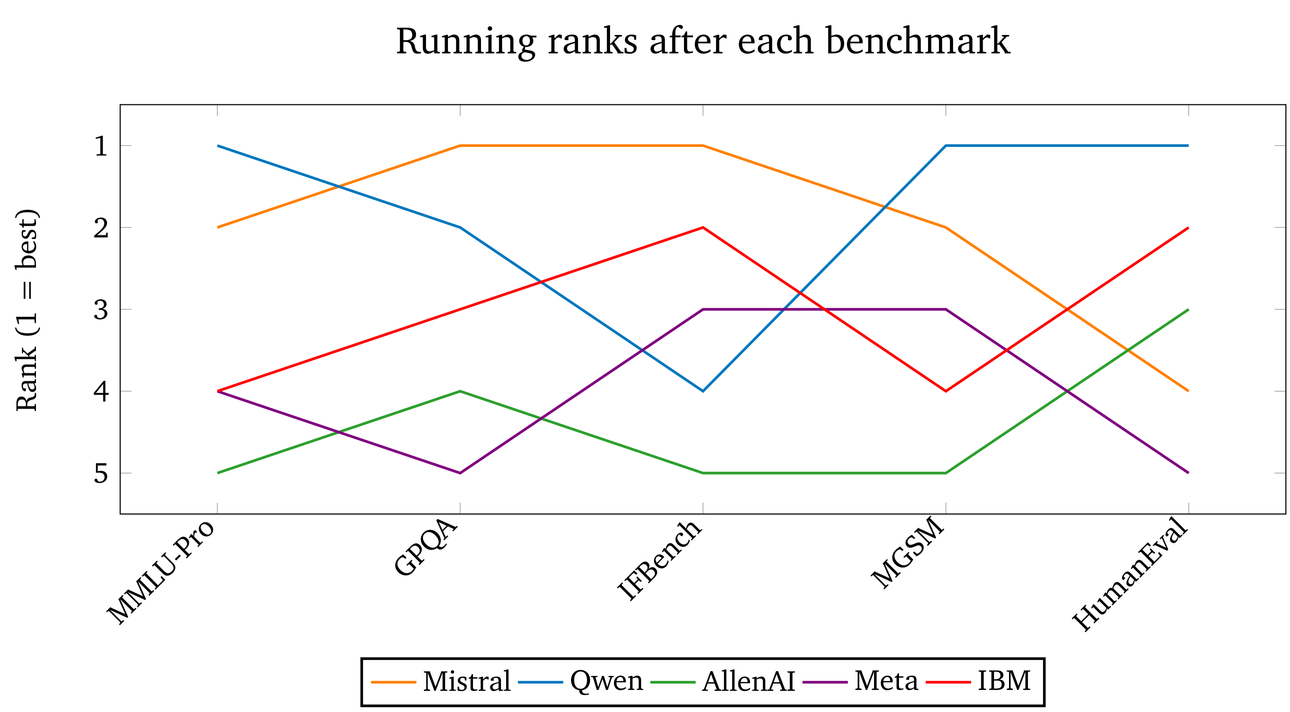
Entrant models

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



Rank	Family	Model	Avg. score	Δ rank	Notes
1	Qwen	Qwen3-8B	45.74	-	-
2	IBM	granite-4.0-h-tiny	42.51	↑ 2	-
3	AllenAI	Olmo-3-7B-Instruct	40.75	↑ 2	-
4	Mistral	Ministral-3-8B-Instruct-2512	40.38	↓ 2	-
5	Meta	Llama-3.1-8B-Instruct	38.18	↓ 2	-

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 (11)	61.86	63.12	41.25	44.86	47.21
GPQA-Diamond (9)	38.48	16.57	32.22	28.43	29.04
IFBench(8)	27.41	26.12	23.47	34.15	32.99
MGSM(10)	37.60	65.45	33.05	43.24	35.09
HumanEval(6)	36.59	57.44	73.78	40.24	68.26
$\boxed{?}$	—	—	—	—	—
$\boxed{?}$	—	—	—	—	—
$\boxed{?}$	—	—	—	—	—
$\boxed{?}$	—	—	—	—	—
$\boxed{?}$	—	—	—	—	—
$\boxed{?}$	—	—	—	—	—

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

Benchmark 5: HumanEval

The HumanEval benchmark (6) assesses code generation capabilities by making models complete Python function implementations given a function signature and docstring. It consists of 164 handwritten programming problems covering fundamental algorithms, data structures, string manipulation, and mathematical operations. We use the EvalPlus framework(7), which extends the original benchmark with comprehensive test suites containing approximately 80 times more test cases per problem to rigorously evaluate both functional correctness against edge cases. HumanEval is a valuable benchmark as its compact, specification-based challenges effectively measure how well a model can convert natural language descriptions into working code. It offers clear insights into its reasoning capabilities and ability to generalize when generating programs.

Experimental setup. We evaluate all 164 problems from HumanEval, using the EvalPlus library (2), which extends the original benchmark with comprehensive test suites. For pass@k evaluation, we generate k=10 samples per problem with temperature=0.2, yielding 1,640 total samples. Results are presented as pass@10 metrics. Note that the models are given a maximum output length of 2048 tokens.

We make the following observations:

- AllenAI's Olmo-3-7B-Instruct and IBM's granite-4.0-h-tiny performed significantly better than the others in this first coding benchmark.
- Qwen3-8B finished middle-of-the-pack, while Ministral-3-8B-Instruct-2512 and Llama-3.1-8B-Instruct performed the worst.

Benchmark 4: Multilingual Grade School Math Benchmark (MGSM)

The Multilingual Grade School Math (MGSM) benchmark (10) assesses mathematical reasoning abilities across eleven languages. It is built from 250 elementary-level math word problems drawn from the English GSM8K dataset (5), which were professionally translated into Bengali, Chinese, French, German, Japanese, Russian, Spanish, Swahili, Telugu, and Thai. The tasks involve multi-step arithmetic reasoning that a middle-schooler should be able to solve.

Experimental setup. We evaluate all 250 problems in each of the eleven languages, yielding 2,750 total instances, and score performance using exact match on the final extracted numerical answer after normalizing for commas and trailing decimal zeros. We use the evaluation scripts from the `openai/simple-evals` repository (3), adapted for local model inference using vLLM. Prompts are formatted in the target language with language-specific instructions that request reasoning followed by a numeric answer in a specified format. Models are sampled with greedy decoding (temperature=0.0) and a maximum generation length of 2,048 tokens. Reported results correspond to the mean accuracy across the 11 languages.

We make the following observation:

- Qwen3-8B performed significantly better than the others, scoring around 65%, while the others only scored 33-44%.

Benchmark 3: IF Bench

The Instruction Following benchmark (IFBench) from AllenAI assesses instruction-following across diverse tasks like counting, formatting, and text manipulation (8). 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 (9) and using the OpenAI simple-eval library (3). 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 (11) 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` (4)). 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] evalplus/evalplus: Rigourous evaluation of LLM-synthesized code. https://github.com/evalplus/evalplus/tree/master.

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

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

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[6] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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

[9] 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.

[10] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

[11] 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.