UNDERDOG SCORECARD

Don't trust the hype. Compare the facts.

JAN 2025 DATA v1



Created by Mr. Peace – AI character & AI voice

RESULT: DeepSeek-R1 (challenger) vs OpenAI o1 (incumbent)

Metric	DeepSeek-R1	OpenAI o1
Output price / 1M tokens lower wins	$2.19 WINNER	$60
AIME 2024 (math, pass@1) higher wins	79.8 WINNER	79.2
GPQA Diamond (science, pass@1) higher wins	71.5	75.7 WINNER

Shaded = benchmark scores

Scores self-reported by DeepSeek, not independently verified

CONDITIONS & SOURCES

Models: DeepSeek-R1 (released Jan 20, 2025) vs OpenAI o1-2024-12-17 ("o1-1217").
Prices: list API price per 1M output tokens as of Jan 2025 (R1 input $0.55 cache miss; o1 input $15).
Scores: DeepSeek-R1 technical report, Table (both models' numbers reported by DeepSeek).
Sources: deepseek.com/en/news/deepseek-r1 · arxiv.org/abs/2501.12948
web.archive.org/web/20250125082819/https://openai.com/api/pricing/ (OpenAI pricing, archived Jan 25, 2025)
Prices and models change. Check today's pages before deciding. Names are trademarks of their owners. Not affiliated.

YOUR TURN: BLANK SCORECARD

Pick any two tools. Fill in only what you can source.

	CHALLENGER	INCUMBENT
Model		
Price		
Performance		
Speed		
Best at		
Sources		

WHICH ONE WOULD YOU CHOOSE? Why?

UNDERDOG SCORECARD · EP2

Don't trust the hype. Compare the facts.

OCT 2026 PREPRINT v2



Created by Mr. Peace – AI character & AI voice

RESULT: Qwen3-8B, trained (small) vs Llama-4-Maverick, prompted (big)

Metric	Qwen3-8B	Maverick
Size (parameters) not a score	8B	402B total / 17B active
ToolBench G3 tool selection exact match, 2,124 instructions	41.5% WINNER	30.6%
TaskBench (different benchmark) success rate	72.8%	80.6% WINNER

Preprint (not peer-reviewed). Scores reported by the authors, not independently verified.

THE CATCH (from the paper)

- Qwen3-8B was trained for this task (SFT, then RL). Maverick was only prompted.
- Maverick's 30.6% is its best of four prompting setups (28.4, 8.6, 30.6, 10.1).
- Maverick was also the "teacher" used to create the small model's training examples.
- Other benchmarks in the paper: G2 25.8 vs 20.8, Gorilla 29.9 vs 16.0, CRAFT 69.0 vs 47.9 (8B vs best Maverick).

So: the small model won on tool selection after training. It is not shown to be smarter or cheaper overall.

SOURCE

Zhou, Magesh, Kwon, Kuriksha, Jiang, Zanette, Jain (Meta; Carnegie Mellon University).
"Training Small Language Models for Transferable Tool Selection", arXiv:2610.31483, Oct 22, 2026.
arxiv.org/abs/2610.31483 · Table 1 (ToolBench G3), Table 2 (TaskBench etc.), Appendix (Maverick 402B/17B active).
Names are trademarks of their owners. Not affiliated.

DID THE SMALL MODEL REALLY WIN? Or does it depend on how you train and test?