

名前	MMLU	MMLU-Pro	MMMLU (名前)
Claude Opus 4.5	~91%	-	90.8%
Claude Sonnet 4.5	~89%	-	89.1%
GPT-5	91.4%	87.0%	-
GPT-5.2	~89.6%	-	-
Gemini 3 Pro	91.8%	-	91.8%
Grok 4	86.6%	87%	-

Model	GPQA Diamond	ARC-AGI-2
Claude Opus 4.5	87.0%	37.6%
Claude Sonnet 4.5	83.4%	13.6%
GPT-5	85.7%-88.4%	~17.6%
GPT-5.1	88.1%	17.6%
GPT-5.2 Thinking	92.4%	52.9%
GPT-5.2 Pro	93.2%	54.2%
Gemini 3 Pro	91.9%	31.1%
Gemini 3 Deep Think	93.8%	45.1%
Grok 4	87.5%-88%	15.9%

Humanity's Last Exam
~30% (with search)
-
~26.7%
26.7%
-
-
37.5%
41.0%
44.4%-50% (Heavy)

LLM	AIME 2025 (Problems)	AIME 2025 (Short Problems)	MATH-500
Claude Opus 4.5	~95%	100%	-
Claude Sonnet 4.5	87%	100%	-
GPT-5	94.6%	100%+	-
GPT-5.2	100%	100%	-
Gemini 3 Pro	95%	100%	-
Grok 4 Heavy	94%-100%	100%	96%

Model	SWE-bench Verified	LiveCodeBench	Terminal-Bench
Claude Opus 4.5	80.9%	-	59.3%
Claude Sonnet 4.5	77.2%	-	50.0%
GPT-5	74.9%	-	-
GPT-5.2	80%+	-	-
Gemini 3 Pro	76.2%	2,439 Elo	54.2%
Grok 4	72-75%	79.4%	-

Model	MMMU	MMMU-Pro	Video-MMMU
Claude Opus 4.5	80.7%	-	-
Claude Sonnet 4.5	77.8%	-	-
GPT-5	84.2%	-	-
GPT-5.1	85.4%	-	80.4%
Gemini 3 Pro	~85%	81.0%	87.6%
Grok 4	76.5%	-	-

名前	リリース日	ベンチマークスコア	特徴
Claude Opus 4.5	2025年11月	MMMLU 90.8%	ARC-AGI-2 37.6%
Claude Sonnet 4.5	2025年9月	MMMLU 89.1%	GPQA 83.4%
GPT-5	2025年8月	MMLU 91.4%	GPQA 88.4%
GPT-5.2	2025年12月	MMLU ~89.6%	ARC-AGI-2 52.9%★
Gemini 3 Pro	2025年11月	MMLU 91.8%★	GPQA 91.9%
Grok 4	2025年7月	MMLU 86.6%	HLE 44.4%★

□□□□	□□□□	□□□
AIME 100%	SWE-bench 80.9%	MMMU 80.7%
AIME 100%	SWE-bench 77.2%	MMMU 77.8%
AIME 94.6%	SWE-bench 74.9%	MMMU 84.2%
AIME 100%★	SWE-bench 80%+	-
AIME 100%	SWE-bench 76.2%	MMMU 81%★
AIME 94-100%	SWE-bench 72-75%	MMMU 76.5%

AIの最新動向

本記事では、最新のAI動向について、Artificial Analysis、Vellum、OpenAI、Google、Anthropic、xAIの最新動向について、2025年12月現在の状況を紹介します。

主要なAIモデル

- MMLU/MMLU-Pro: 総合的な知識と推論能力を評価するベンチマークで、57+のスコアを達成。
- GPQA Diamond: 高度な専門知識を評価するベンチマークで、70%以上のスコアを達成。
- ARC-AGI-2: 高度な推論能力を評価するベンチマークで、80%以上のスコアを達成。
- Humanity's Last Exam (HLE): 2500+の専門知識を評価するベンチマークで、80%以上のスコアを達成。
- AIME: 高度な数学能力を評価するベンチマークで、80%以上のスコアを達成。
- SWE-bench: ソフトウェア開発能力を評価するベンチマークで、GitHubのデータに基づいて評価。
- MMMU: 医学知識を評価するベンチマークで、80%以上のスコアを達成。

注目すべきニュース

- ★ OpenAIの最新発表: GPT-5.2の発表と、新しいモデルの紹介。
- - Googleの最新発表: Geminiの最新バージョンの発表。
- ~ Anthropicの最新発表: Claudeの最新バージョンの発表。

今後の展望

- AIの応用範囲の拡大: 医療、教育、ビジネスなどでの応用が期待される。
- GPT-5.2の発表: 2025年12月11日に発表される。
- Grok 4.1の発表: 2025年12月11日に発表される。