ChatGPT

LMSYS Chatbot ArenaにおけるGPT-5の順位とスコア

LMSYSのChatbot Arena（総合リーダーボード）では、**GPT-5**は最新モデルの中でトップクラスの評価を受けています。2025年8月時点のArena Overviewランキングでは、Google DeepMindの**Gemini 2.5 Pro**がEloレーティング1471で1位、**GPT-5**はそれに僅差の1462で2位につけています [1]。xAI（エックスAI）の**Grok-4**も1436で3位と健闘しています [2]。このElo値は数百万件に及ぶ匿名バトルでのユーザー投票データに基づき算出されたもので、GPT-5は発売直後に急速にスコアを伸ばし、先行モデルに迫る位置を占めています [1]。特にGPT-5は**コーディング部門**で突出しており、コード生成に関するArenaのエローレーティングで1500超え（1502）と他モデルを大きく上回る記録的スコアを達成しています [1]。これは、GPT-5が従来トップだったGemini 2.5 Proをコード生成タスクで75ポイント以上引き離したことを意味し、現時点で**「実世界のコーディングに最適なモデル」**との評価を受けています [1]。

各種ベンチマークにおけるGPT-5の性能と他モデル比較

OpenAIのGPT-5は、公開されているさまざまなベンチマークにおいても**最先端のスコア**を記録しています。以下に主要モデル（OpenAI GPT-5、Anthropic Claude Opus 4.1、Google DeepMind Gemini 2.5 Pro、xAI Grok-4）の評価指標を比較した表を示します。Claude Opus 4.1はAnthropic社のClaude 4世代の最新強化版、Gemini 2.5 ProはGoogle DeepMindの大規模言語モデル、Grok-4 HeavyはElon Musk率いるxAIの最新モデルです。

ベンチマーク	GPT-5	Claude Opus 4.1	Gemini 2.5 Pro	Grok-4 Heavy
コード生成 (SWE-Bench Verified) – 実問題でのプログラミング正解率	**74.9%** [3]	74.5% [3]	59.6% [4]	（未公表）
科学QA (GPQA Diamond) – 博士レベルの科学質問正答率	**89.4%** [5]	80.9% [5]	（非公開）	88.9% [5]
知識テスト (MMLU-Pro) – 学術・常識問題正答率	**87.3%** [6]	87.3% [7]	86.2% [8]	86.6% [2]
総合学力テスト (Humanity's Last Exam) – 大学レベル試験・ツール使用可	**42.0%** [9]	（未公表）	（未公表）	**44.4%** [9]

表：主要ベンチマークにおける各モデルのスコア比較。GPT-5 Pro（一部ツール使用や思考延長モード）は特に高度な科学問題やマルチステップ推論で高得点を記録。

上記のように、GPT-5は**総合的な知能指標**で他モデルを僅差で凌駕しています。例えば、PhDレベルの科学問題集である**GPQA Diamond**ではGPT-5 (Proモード)が**89.4%**と過去最高スコアを樹立し、Claude Opus 4.1の80.9%やGrok-4 Heavyの88.9%を上回りました [5]。プログラミング実務能力を測る**SWEベンチマーク**（GitHub由来の課題集）でもGPT-5は**74.9%**を初回実行で達成し、Claude 4.1（74.5%）とほぼ拮抗しつつ、Gemini 2.5 Pro（59.6%）には大差をつけています [3]。特にGPT-5のコーディング力は従来モデルを一段引き離し、「コード一発生成」の文脈では現状トップクラスです [3] [10]。

MMLU（大規模知識テスト）など伝統的な学習ベンチマークでもGPT-5は高い正答率を示しています。GPT-4世代が約86%前後だったMMLU精度をGPT-5は約**87〜89%**程度まで僅かに引き上げており、Claude 4.1やGemini 2.5とほぼ横並びの最高水準です [6]。一方で、**ARC**や**HellaSwag**といった常識推論ベンチマーク

や、**GSM8K**（算数文章題8K）・**HumanEval**（コーディングテスト）などの従来指標は、**2025年現在ほぼモデル間で飽和状態**に達しています [11]。GPT-5もこれらでトップクラスの成績を収めているものの、GPT-4からの向上幅は限定的です [11]。実際、OpenAIはGPT-5発表時に「従来ベンチマークは成熟しつつあり、今後は**MMMU**や**GPQA**のような新指標や、**Humanity's Last Exam**（厳しい学力テスト：GPT-5 Proで42% [9]）など**より難度の高い評価**に焦点が移っている」と述べています [11]。

総じて**GPT-5**は、**Claude Opus 4.1**（Anthropic）や**Gemini 2.5 Pro**（Google DeepMind）、**Grok-4**（xAI）といった最先端モデルと互角以上の戦いを繰り広げています。Chatbot Arenaのユーザー対話評価ではGeminiに次ぐ2位につけつつ、特定分野ではトップ（例えばコード生成）となるなど、モデルごとの強みの違いも明確です [1] [3]。Claude Opus 4.1は推論の「思考モード」を駆使することで知識問題でGPT-5に匹敵する成果を残し [7]、Gemini 2.5 Proはマルチモーダル（テキスト・画像・Web検索）統合能力で総合1位を獲得するなど [12]、各モデルが競い合う状況です。GPT-5はこれら最新モデルに対し**わずかながらリードを広げた分野（高度な科学QAや創造的文章生成など）**と、**互角もしくは追い上げられている分野（長文推論や特定専門領域など）**があり、現時点で「全領域で他を圧倒する決定的なAGI」には至らないものの、**総合知性では最強クラス**と言えるでしょう [13] [14]。今後も各社のモデル開発競争が激化する中、GPT-5がベンチマーク上で示した優位性がどこまで持続するか注目されています。

参考文献・出典：LMSYS Chatbot Arenaリーダーボード [1]、OpenLM/OpenAIによる各モデル評価値 [3] [5] [9] [6]、Stanford HAI 2025年AIインデックス報告 [11]、Anthropic (Claude 4.1発表) [10] ほか。

---

[1] [2] [6] [7] [8] Chatbot Arena + | OpenLM.ai
https://openlm.ai/chatbot-arena/

[3] [4] [5] [9] [13] [14] OpenAI's GPT-5 is here | TechCrunch
https://techcrunch.com/2025/08/07/openais-gpt-5-is-here/

[10] Claude Opus 4.1 \ Anthropic
https://www.anthropic.com/news/claude-opus-4-1

[11] Technical Performance | The 2025 AI Index Report | Stanford HAI
https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance

[12] Gemini 2.5 Pro: The Undisputed LMSYS Champion
https://learnprompting.org/blog/gemini_pro_update?srsltid=AfmBOopt7xeZ15B_-G-4wCVw6AY4d59oN-11Qn_ek7yjmofXMHbY12AW