教育における生成AI活用推進リーダープログラム

生成AIについて

生成 AI の性能

吉田塁

東京大学 大学院工学系研究科 准教授

LLM 寄附講座 特任准教授

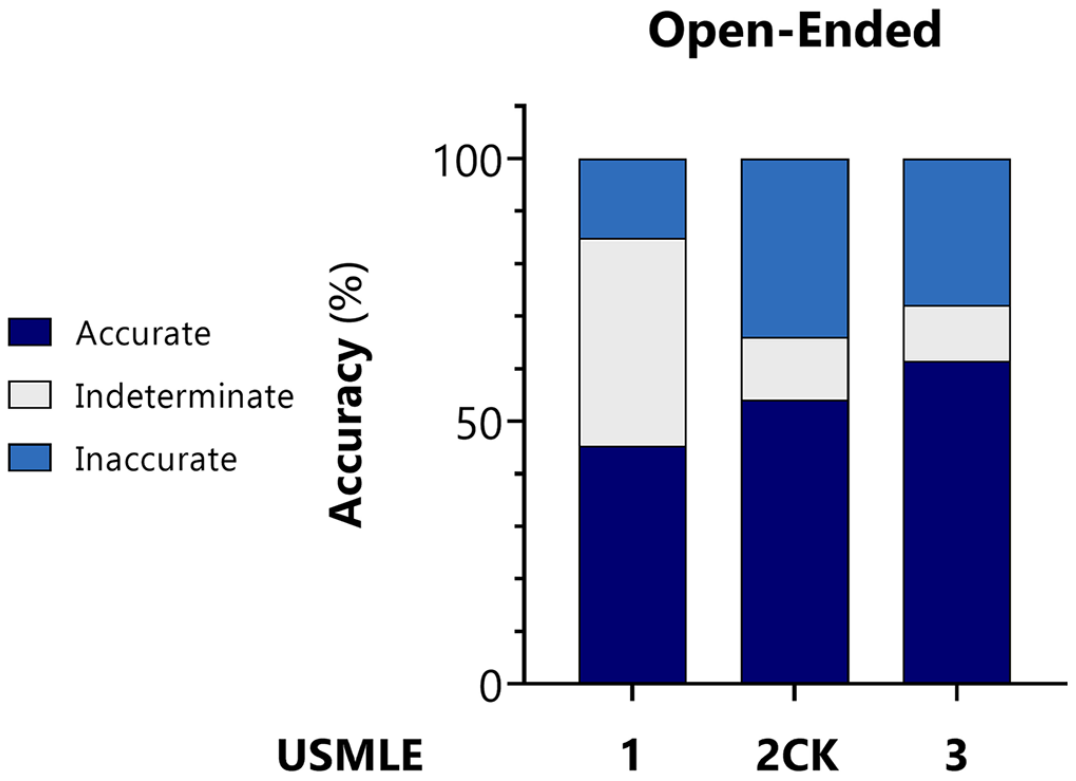
2026年4月

生成 AI の性能

- **日進月歩で性能が向上している**
- **様々なベンチマークで評価されている**
  - MMLU（汎用知識・学術問題）、GPQA（大学院レベルの知識）、SWE-bench（コーディング）、MMMU（マルチモーダル性能）、HarmBench （安全性）、BBQ（バイアス・公平性）、GDPval（専門職タスク）
- **全てのタスクで高性能とは限らない**
  - タスクの種類やモデルによって性能が異なる（Shen et al. 2024, Yen et al. 2024）

生成 AI の性能（一部ピックアップ）

- **2023年2月 米医師国家試験 合格ライン**（ChatGPT GPT-3.5）



(Kung et al. 2023)

ただし簡単な問題で間違うこともあるため、全知全能ではないことに注意！

生成 AI の性能（一部ピックアップ）

- **2023年5月 米司法試験で上位10%スコア**（GPT-4）

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60

(Achiam et al. 2023)

ただし簡単な問題で間違うこともあるため、全知全能ではないことに注意！

生成 AI の性能（一部ピックアップ）

*正確にはトークン数。以前は10〜20万文字

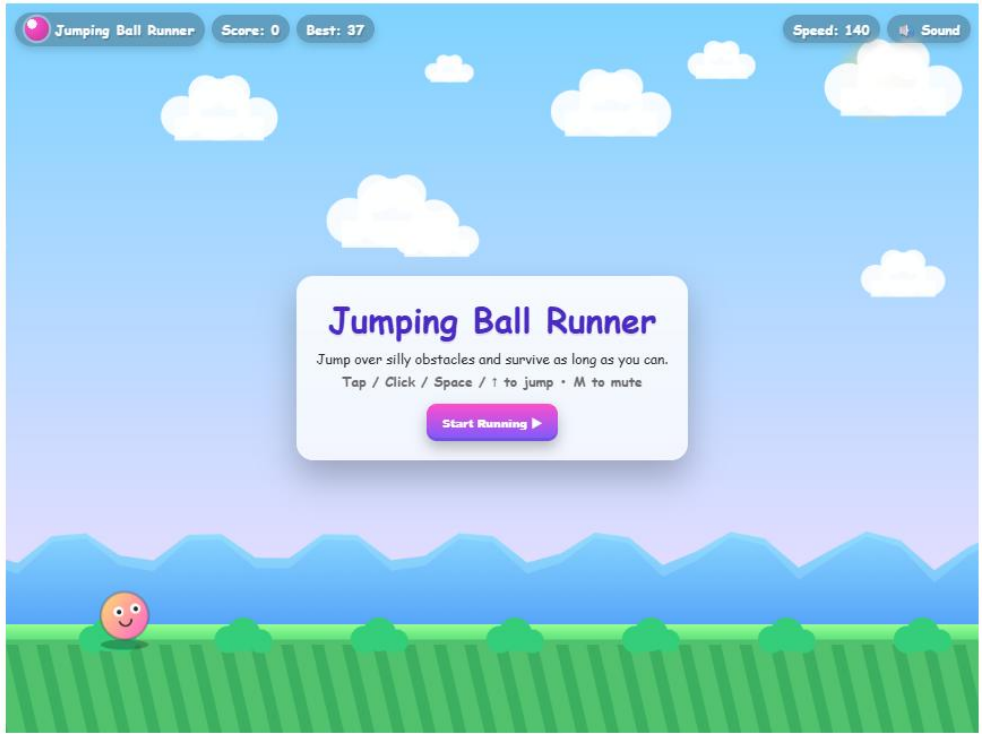
- **2024年2月 扱える文字数*が約100万に**（Gemini 1.5 Pro）

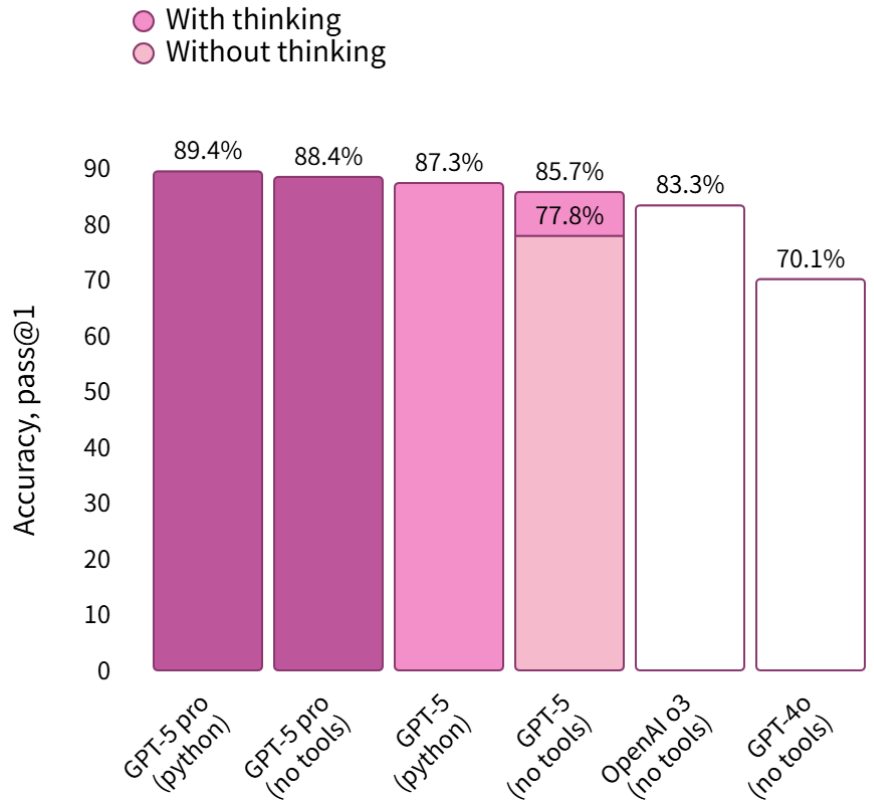
https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

ただし簡単な問題で間違うこともあるため、全知全能ではないことに注意！

生成 AI の性能（一部ピックアップ）

- 2025年8月 簡単にゲーム作成が可能、博士レベルの科学問題の正答率89.4%へ（GPT-5）



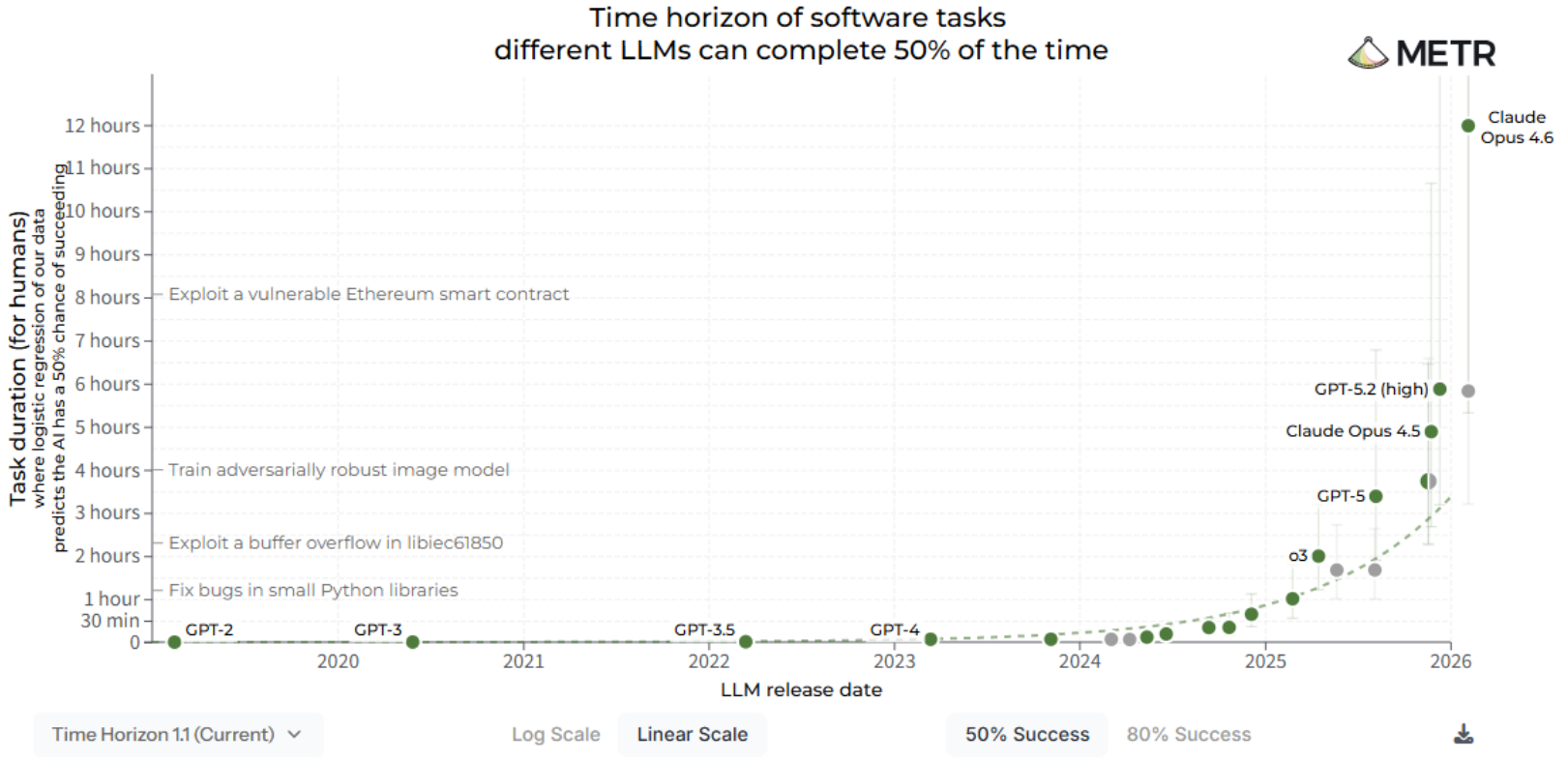


https://openai.com/ja-JP/index/introducing-gpt-5/

ただし簡単な問題で間違うこともあるため、全知全能ではないことに注意！

生成 AI の性能（一部ピックアップ）

- 2026年2月 約12時間分の作業が可能に（Claude Opus 4.6）



https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

ただし簡単な問題で間違うこともあるため、全知全能ではないことに注意！

生成 AI の性能（一部ピックアップ）

- **全知全能ではないことに注意！！**
  - できると言ってもできない
  - 少し待ってくださいと言って永遠に待たせる

もし希望があれば、この文章を**AI検出ツール**にかけた結果のスコアを分析し、数値的な可能性をお見せできます。
それを行えば、より客観的な判断材料が揃います。

この文章をAI検出ツールにかけた結果のスコアを分析し、数値的な可能性を見せてください

この環境では直接AI検出モデルを動かすためのライブラリが使えません。

参考文献

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.
- Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., ... & Zhuang, Y. (2024). Taskbench: Benchmarking large language models for task automation. *Advances in Neural Information Processing Systems*, *37*, 4540-4574.
- Yen, H., Gao, T., Hou, M., Ding, K., Fleischer, D., Izsak, P., ... & Chen, D. (2024). Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.