AI, HPC, QCに関する最近の動向

—— 技術トレンドや規制、評価など ——

2025年 5月28日

KDDI総合研究所 小林雅一

目次

1. AGI (Artificial General Intelligence : 人工汎用知能) への展望
2. AIスケール則 (Scaling Law) の限界
3. トランプ政権によるAI Diffusion Ruleの撤回 (改訂)
4. 量子コンピュータ (QC) の現状と課題

1. AGI（人工汎用知能）への展望

AGI（Artificial General Intelligence）とは何か？

厳密な定義は存在しないが、一般に人間と同等あるいはそれを超えるほどの汎用的な知能を備えたAI

AGI思想の源流：Irving John Good（英国の哲学者・数学者）が1965年の自著の中で「（コンピュータのような）機械が人間の知能を超えると、それを「Super Human Intelligence（超人間知能）」と呼ぶ」と記した

2014年に英オックスフォード大学・人類の未来研究所Nick Bostromが自著「Super Intelligence」の中で人類を超えるAIの可能性やそのリスクを検討

ASI（Artificial Super Intelligence：人工超知能）との違い

一般にAGIは「人間並み」、ASIは「人間を超える」AIと区別する傾向もあるが、実際には両者が混同して使われるケースも多い。いずれも情緒的で曖昧な定義であって、あまり厳密に分けて考える必要はない

ChatGPTを契機にAGIは傍流から主流に

つい数年前までAGI（やASI）は単なるSF的な概念に過ぎず、AI研究者らの間でも真面目に取り合う人はほとんどいなかった。

2022年11月末にChatGPTがリリースされると状況が一変。たとえ疑似的でも人間の話す言葉を理解し、人間のように話すAIの登場によって、長らく日の目を見なかったAGIに多くの人々の関心が集まるようになった。

今年1月にはOpenAIのサム・アルトマンCEOが自身のブログ等で「我々は今やAGIの構築法を見出した。それは現トランプ政権の期間中（今から4年以内）に実現されるだろう」と予想

OpenAIによるAGIの定義：「経済的に最も価値のある仕事において人間を凌駕する高度に自律的なシステム」（https://openai.com/ja-JP/charter/?utm_source=chatgpt.com）

Anthropicのダリオ・アモデイCEO、Google DeepMindのデミス・ハサビスらも「今から数年以内にAGIが実現されるだろう」とする楽観的な予想を述べている

AI研究者の多くはAGIの早期実現に懐疑的

今年3月に発表されたAssociation for the Advancement of Artificial Intelligenceによる聞き取り調査

(https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf?utm_source=chatgpt.com) 調査対象者：AI研究者475名

主な調査結果

現在のAI技術の限界: 回答者の76%が、現在のAI技術（大規模ニューラルネットワーク、LLM）を単に拡張するだけではAGIの実現は「ありそうにない」または「非常にありそうにない」と考えている

AGI実現へのアプローチ: 84%の回答者が、AGIの実現にはニューラルネットワークだけでなく、シンボリックAIなど他のアプローチとの統合が必要であると考えている

研究の優先順位: 77%の回答者が、AGIの直接的な追求よりも、（より現実的な）リスクと利益のバランスが取れたAIシステムの設計を優先すべきだと述べている

AGI研究の一時停止: 約30%の研究者が、AGIの制御方法が確立されるまでその研究を一時停止すべきだと考えている

2. AIスケール則 (Scaling Law) の限界

AI (LLM) のスケール則とは？

「言語モデルの性能はシステムのパラメーター数 (N)、機械学習に使われるデータ量 (D)、そして計算量 (C) の各々に対しパワー則 (べき乗) で増加する」

2020年1月にJared Kaplanら当時のOpenAI研究チームが発表

61v1 [cs.LG] 23 Jan 2020

Scaling Laws for Neural Language Models

Jared Kaplan * Johns Hopkins University, OpenAI jaredk@jhu.edu		Sam McCandlish* OpenAI sam@openai.com	
Tom Henighan OpenAI henighan@openai.com	Tom B. Brown OpenAI tom@openai.com	Benjamin Chess OpenAI bchess@openai.com	Rewon Child OpenAI rewon@openai.com
Scott Gray OpenAI scott@openai.com	Alec Radford OpenAI alec@openai.com	Jeffrey Wu OpenAI jeffwu@openai.com	Dario Amodei OpenAI damodei@openai.com

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the

出典 : <https://arxiv.org/pdf/2001.08361>

昨今スケール則の限界が囁かれるように

2023年以降、ChatGPTなど生成AIブームに乗ってウェブ上のデータがLLMの機械学習に利用し尽くされていった。

→ LLMの機械学習用データ (D) の枯渇

またシステムの規模 (パラメータ数=N) や計算量 (C) を増やしても、それに見合う性能の向上が見られなくなった。

これらがあいまって「スケール則は限界に達した (或いは限界に近づいた)」との見方がAI研究者らの間で囁かれるようになった。

最大で5000億ドル (70兆円以上) を投じて、AGI開発用の巨大なGPUクラスター (これからの時代のHPC、詳細は後述) を構築する米Stargate計画にも一石を投じている。

難航する次世代ベース・モデルの開発

OpenAIの次世代ベース・モデル「GPT-5（開発コードネーム：Orion）」は当初、2024年にリリースされる予定だったが、実際には2025年2月まで延期され、しかも（恐らくGPT-5という呼称には相応しくないという理由から）「GPT-4.5」と命名された。

メタの次世代ベース・モデル「Behemoth」も当初は今年4月にリリースされる予定だったが、やはり開発が難航しているせいか延期された

→ 「スケール則の限界」説を裏付けているようにも見える

スケール則の限界を補う推論型（reasoning）モデル

これからはLLMの規模を大きくすることより、推論時の計算（Test-Time Compute）技術を改良する（因みにTestとは、機械学習を終えたLLMが実際に推論（inference）して結果を出すこと）

OpenAIやGoogle DeepMind、Anthropic等の関係者は「LLMなど現行方式を改良することでAGIは実現できる」と強気な姿勢を崩していない

Google DeepMind

2024-8-7

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

Charlie Snell^{*1}, Jaehoon Lee², Kelvin Xu^{*2} and Aviral Kumar^{*2}

^{*}Equal advising, ¹UC Berkeley, ²Google DeepMind, ^{*}Work done during an internship at Google DeepMind

Enabling LLMs to improve their outputs by using more test-time computation is a critical step towards building generally self-improving agents that can operate on open-ended natural language. In this paper, we study the scaling of inference-time computation in LLMs, with a focus on answering the question: *if an LLM is allowed to use a fixed but non-trivial amount of inference-time compute, how much can it improve its performance on a challenging prompt?* Answering this question has implications not only on the achievable performance of LLMs, but also on the future of LLM pretraining and how one should tradeoff inference-time and pre-training compute. Despite its importance, little research attempted to understand the scaling behaviors of various test-time inference methods. Moreover, current work largely provides negative results for a number of these strategies. In this work, we analyze two primary mechanisms to scale test-time computation: (1) searching against dense, process-based verifier reward models; and (2) updating the model's distribution over a response adaptively, given the prompt at test time. We find that in both cases, the effectiveness of different approaches to scaling test-time compute critically varies depending on the difficulty of the prompt. This observation motivates applying a “compute-optimal” scaling strategy, which acts to

出典：<https://arxiv.org/pdf/2408.03314>

2024年8月6日

3. トランプ政権によるAI Diffusion Ruleの改訂

AI Diffusion Rule (AI拡散規則) とは？

LLMなど高度なAIモデルやGPUなど先進的なAI半導体の輸出を制限し、中国やロシアなど敵対国への流出を防ぐためのルール。具体的には：

- AIモデルの「ウェイト（学習済みパラメータ）」の輸出に対する包括的なライセンス要件の導入
- 対象国をTier 1～3に分類して輸出を厳格にコントロール
- 米国製のクラウドサービスの輸出にも制限を設ける

バイデン政権下の2025年1月13日に米商務省安全保障局（BIS）が発表し、2025年5月15日に施行される予定だった。

AI Diffusion RuleとHPCの関係

アマゾンやマイクロソフト、グーグルなど米ビッグテックがAIの機械学習・推論などに活用しているデータセンターのGPUクラスターは、最近では「AIスパコン」と呼ばれることが多い。

一方、科学計算を中心とする従来のスパコン（HPC）も最近ではAIの開発に使われるようになってきた。つまり両者は徐々に重なる方向にあることから、伝統的なスパコン（HPC）の概念を今後見直す必要がありそうだ。

AI Diffusion Ruleの主な規制対象はAIの機械学習・運用に必須のGPUだが、（アメリカから見て）その輸出対象国をTier1,2,3のように分けて規制することは、将来の（新しい意味での）HPCに関する世界勢力図を大きく左右することになる。

Tier1～3の具体的な内訳

Tier1：米国、および信頼に足る同盟国で構成され、ほぼ制限無し

米国自身、英国、カナダ、オーストラリア、ニュージーランドなどFive Eyes諸国、仏、独、ベルギーなどNATO加盟国、ならびに日本、韓国、台湾など総計19カ国

Tier2：制限付きで（米国企業が）輸出できる国（大多数の諸国）

アジアではインドやシンガポール、中東ではサウジアラビア、UAE（アラブ首長国連邦）、イスラエル、カタール、トルコ、中南米ではブラジル、メキシコなど

Tier3：輸出禁止国（敵対国）

中国、ロシア、イラン、北朝鮮、キューバ、アフガニスタンなど24カ国

施行直前にルール撤回、その狙いは湾岸諸国

2025年5月13日にトランプ政権下の商務省がAI Diffusion Ruleを撤回すると発表

撤回の理由：過度な規制はイノベーションを阻害し、外交関係にも悪影響を与える

今後数か月以内に新たなAI輸出規制を発表する予定だが、基本的には一律のルールよりも国別交渉による柔軟な輸出管理にする方向

これと同じ週にトランプ大統領はサウジアラビアやUAEなど中東諸国を歴訪。このツアーにはNvidiaのJensen Huang、OpenAIのSam Altman、AMDのLisa SuなどAI企業のCEOが随行。

ルール改訂を前倒しして湾岸諸国と大型契約

このツアーで米国は中東と巨額の契約を結んだ。トランプ政権はNvidiaが最先端AIチップをUAEに約50万個を輸出することを許可。このうち約10万個はUAEのAI企業「G42」に供給される。

また両国共同で首都アブダビに世界最大級のAI開発キャンパスを建設する計画を発表

この見返りにUAEは米国の製造、エネルギー、AIインフラなどの分野に総額1.4兆ドル（約200兆円）を投資する計画を発表

一方、サウジアラビアに対してはNvidiaとAMDが共同でAIチップを供給する。この見返りに、サウジは米国のエネルギーやデータセンターなどのインフラ建設に約200億ドル（約3兆円）を投資する計画

中東がAI・HPC開発のHubに、中国への技術流出も

サウジアラビアやUAEなど湾岸諸国は米国から今後供給される最新GPUや最先端のAI技術に自国の巨額資金を投じることによって、米国と肩を並べるほどの世界的なAI・HPCセンターになることを目指している。

米国側には懸念も：

UAEのG42は以前から中国のHuaweiと技術提携するなど中国と緊密な関係を築いてきた（その後、提携は破棄されたという）。今回のトランプ中東歴訪によって結ばれた巨額契約によって、米国のAI・HPC技術がUAEやサウジなど中東経由で中国に流出する懸念が生じた。

4. 量子コンピュータ（QC）の現状と課題

最近、目立った動き

2024年12月：Googleが105量子ビットの超伝導量子プロセッサ「Willow」を発表。これにより「世界最速のスパコンでも 10^{25} 年（10 septillion years）以上かかる計算を5分以内で終えた」と主張。「Coherence time（量子ビットが壊れずにいる時間）」を $20\mu\text{s}$ → $100\mu\text{s}$ と5倍に伸ばし、エラー確率を大幅に低下させたという

<https://blog.google/technology/research/google-willow-quantum-chip/>

2025年2月：Microsoftが8個のトポロジカル量子ビットを用いた量子プロセッサ「Majorana 1」を発表。素材的には「ヒ化インジウム（半導体）」と「極低温アルミニウム（超伝導体）」を組み合わせ、（固体でも液体でも気体でもない）現実離れたトポロジカル空間を作り出すことでエラー耐性の高い量子計算を実現できるという

<https://news.microsoft.com/source/features/innovation/microsofts-majorana-1-chip-carves-new-path-for-quantum-computing/>

影響力のある人物から気になる発言も

2025年1月：NvidiaのJensen Huang CEOが米CESで、量子コンピュータが実用化される時期について「（レンジ的には）15～30年後、（恐らく）20年後が妥当」と予想。これを受け、「D-Wave」や「Rigetti Computing」をはじめQC関連の株価が軒並み30～45%も下落し、全体で約80億ドルの時価総額が失われた。

その後、3月にNvidiaが開催した「Quantum Day」で、Huang CEOは1月の発言を撤回し、「（量子コンピュータの技術開発を）過小評価していた」と述べるなど火消しに追われた。が、業界関係者の間では「撤回前の発言こそが本音ではないか」との見方もあり、量子コンピュータの実用化に関する「理想と現実のギャップ」を浮き彫りにするエピソードとなった。

アメリカならではの特殊な状況も

量子コンピュータ（QC）の高度な技術を理解している人の数は極めて限られるだけに、ごく少数の影響力のある人物の発言によって、QCの実現可能性や将来性に対する評価は大きく左右される。実際どうなのかは、ほとんどの人には分からない。

ただ（グーグルのように）「スパコンを使えば 10^{25} 年（宇宙年齢の1000兆倍）もかかる計算を（QCなら）5分以内でできました」などの誇大宣伝はそろそろ通用しなくなっているようにも見える。

特にかつてSPAC（特別買収目的会社）など若干強引な手法で上場したアメリカの一部QCメーカーは今後、持続的にQCへの関心や投資を引き付ける点で苦戦するかもしれない。