



# 最新版SwallowとAIの 最新動向

東京科学大学 石田 茂樹



石田茂樹

東京科学大学  
情報理工学院 情報工学系  
横田研究室 修士2年

研究テーマ:

大規模言語モデルの医療応用  
Mechanistic Interpretability

1

最新版Swallow

2

AIの最新動向

3

DeepSeekの登場

4

医療における推論能力の強化モデル

1

**最新版 Swallow**

2

AIの最新動向

3

DeepSeekの登場

4

医療における推論能力の強化モデル

東京科学大学の横田研究室と岡崎研究室と産業技術総合研究所で  
LLM(Swallow)を開発

## Swallow LLM

東京科学大学（旧・東京工業大学）情報理工学院の岡崎研究室と横田研究室を中心に、大規模言語モデルの研究・開発をしています。

↓ [VIEW ON HUGGINGFACE](#)



<https://swallow-llm.github.io/index.ja.html>

## メンバー



**岡崎 直観**  
東京科学大学 教授

全体の統括、事前学習コーパスチームのリーダー、ウェブ開発者



**藤井 一喜**  
東京科学大学 修士課程学生

大規模言語モデルの事前学習、指示チューニング



**前田 航希**  
東京科学大学 博士課程学生

大規模言語モデルの評価



**岡本 拓己**  
東京科学大学 修士課程学生

指示チューニング



**島田 比奈理**  
東京科学大学 修士課程学生

大規模言語モデルの評価



**横田 理央**  
東京科学大学 教授

学習チームのリーダー



**中村 泰士**  
東京科学大学 修士課程学生

大規模言語モデルの事前学習、指示チューニング、大規模言語モデルの評価



**服部 翔**  
東京科学大学 修士課程学生

事前学習コーパスの構築、大規模言語モデルの評価



**石田 茂樹**  
東京科学大学 修士課程学生

大規模言語モデルの評価



**齋藤 幸史郎**  
東京科学大学 学部生

大規模言語モデルの評価、トークン化の検討



**水木 栄**  
産総研 / 東京科学大学 非常勤研究員

指示チューニングのリーダー、評価チームのリーダー



**馬 尤咪**  
東京科学大学 助教

大規模言語モデルの評価



**大井 聖也**  
東京科学大学 修士課程学生

大規模言語モデルの評価



**塩谷 泰平**  
東京科学大学 修士課程学生

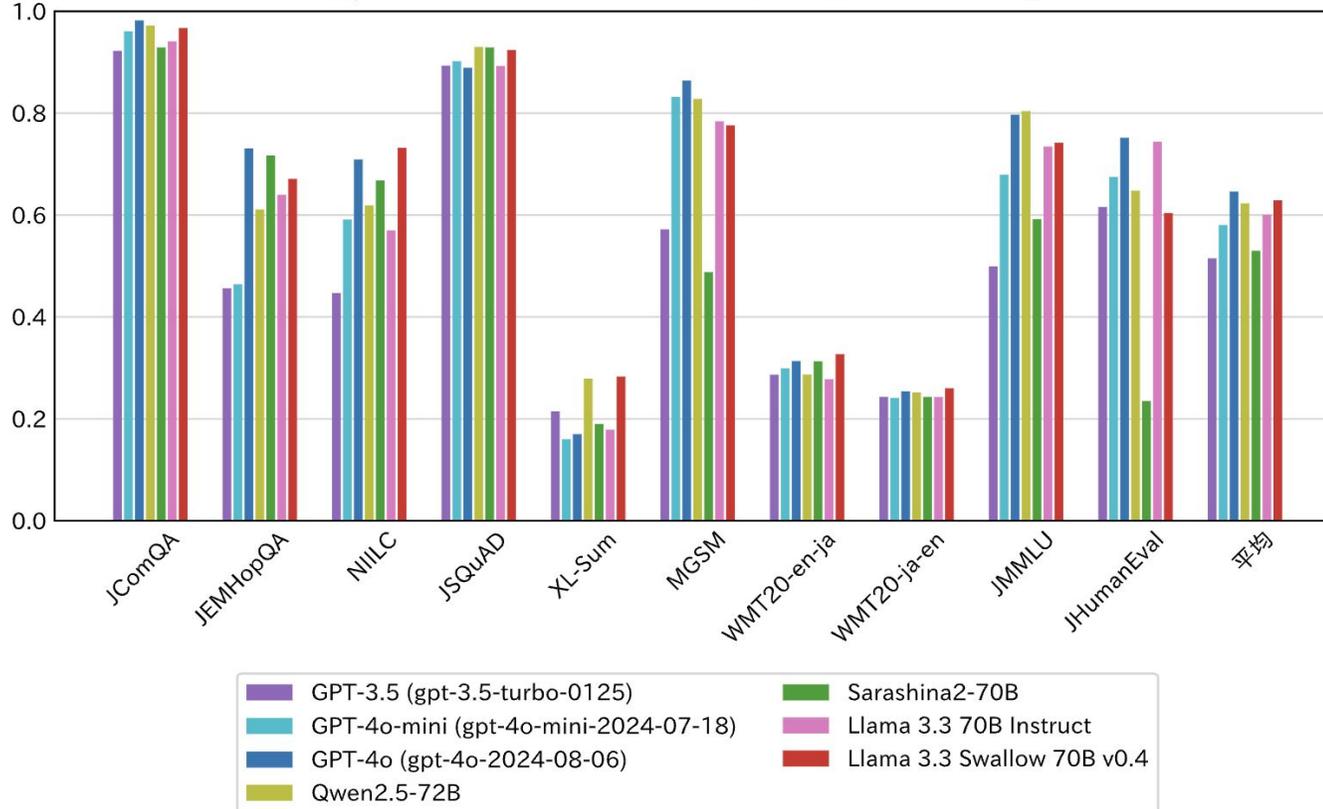
大規模言語モデルの評価



**高村 大也**  
産総研 AIRC チーム長

マネージャー

主要な大規模言語モデルの日本語理解・生成タスクでの性能

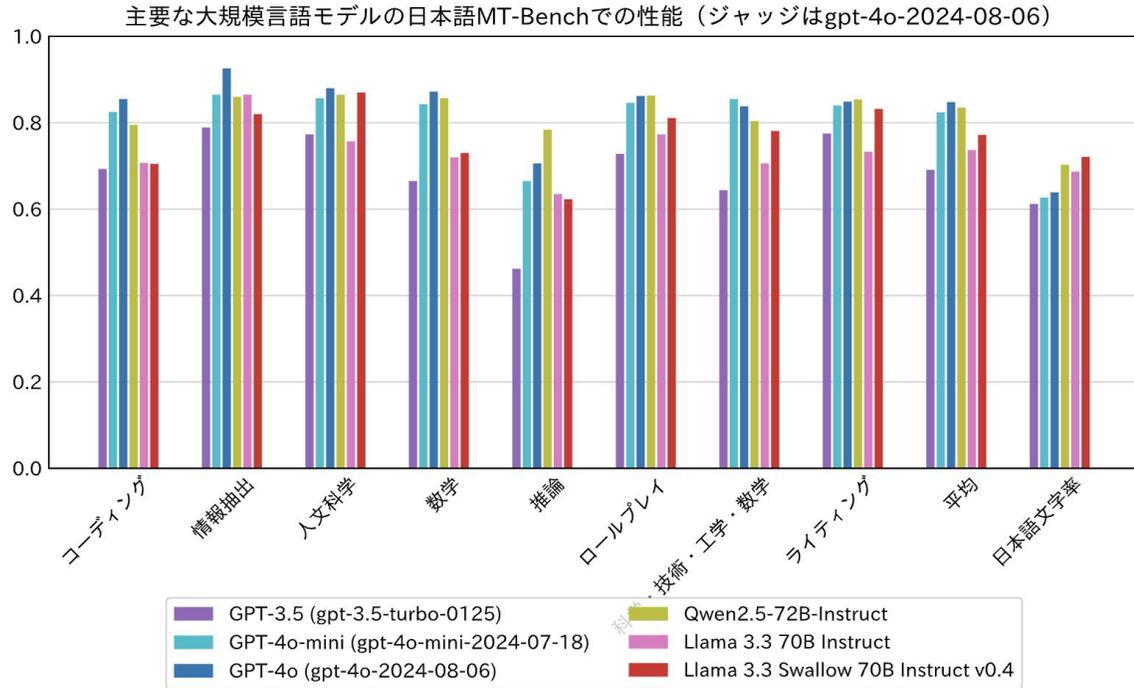


1. GPT-4o:0.646
2. Llama 3.3 Swallow 70B v0.4:0.629
3. Qwen2.5-72B:0.623

Llama-3.3-Swallow-70B-v0.4は日本語理解・生成タスクにおいて、Qwen2.5-72Bの平均スコアをわずかに上回り、同程度の能力をベンチマーク評価上示した。

## 70Bベースモデルの比較

<https://swallow-llm.github.io/llama3.3-swallow.ja.html>



Llama-3.3-70B-Instruct-v0.4はLlama-3.1-Swallow-70B-Instruct-v0.3よりもわずかに高いJapanese MT-Benchの性能を示しており、GPT-4oやQwen2.5-72B-Instructには及ばないものの高い性能を示す。

## 70B指示チューニングモデルの比較

※Sarashina2-70Bは指示チューニングモデルが無いため除外

Swallow開発の3つのキーポイント

- **大規模な日本語コーパスの開発**
- **継続事前学習**
- **指示チューニング**

※今回の発表では指示チューニングの紹介は割愛

初期状態のLLM(Llama)から大量の言語データを継続事前学習ベースモデルを適切な応答を生成できるように指示チューニング



1

### 継続事前学習

大量の言語データをLlamaに追加学習させることで基礎的な知識を獲得する

2

### 指示チューニング(SFT)

教師あり学習を通じて特定のタスクやドメインへの応用力を強化

継続事前学習は**大量のテキスト**を次単語予測で学習し、**基礎的な知識**を獲得



1

## 継続事前学習

大量の言語データをLlamaに追加学習させることで基礎的な知識を獲得する

2

## 指示チューニング(SFT)

教師あり学習を通じて特定のタスクやドメインへの応用力を強化

# 大規模な日本語コーパスの開発

日本語テキストで事前学習を行い、日本語に強いLLMを構築  
期待する効果

**学習データを高品質にすることにより、LLMの性能を向上**

## オープンソースLLMの日本語能力を高めた「Llama 3.1 Swallow」を公開

—英語力を維持しながら日本語の理解・生成・対話能力を強化した大規模言語モデル—

### ポイント

- 大規模言語モデルLlama 3.1の英語の能力を維持しながら、日本語の能力を強化
- Llama 3.1ライセンスにより、商用利用だけでなく他のモデルの改良にも利用可能
- 高度な日本語処理が求められる多くの場面で、生成AI技術の利活用を推進

[https://www.aist.go.jp/aist\\_j/press\\_release/pr2024/pr20241008\\_2/pr20241008\\_2.html](https://www.aist.go.jp/aist_j/press_release/pr2024/pr20241008_2/pr20241008_2.html)

Swallow corpus : ルールに基づくフィルタリングでWebデータによる事前学習の効果を向上

- (1) 繰り返しの多いウェブページ の除去
- (2) 品質のよい日本語の文章を含むウェブ ページの選別
- (3) 有害な表現を含むと思われるウェブページの除去
- (4) 分類器に基づくフィルタリング

Common Crawl から日本語  
テキストを抽出



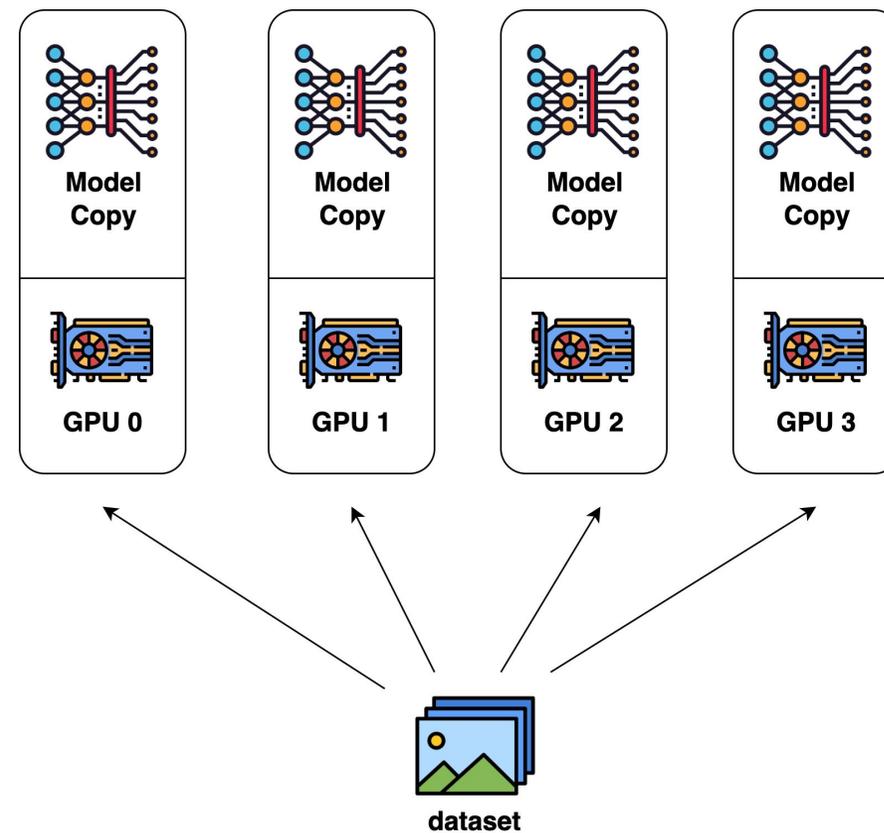
Common Crawl から日本語  
テキストを抽出

## 詳細

[https://speakerdeck.com/aya\\_se/data-centric-ai-swallow-corporus-56e2869a-f9bd-46cb-b030-1012235c37f7](https://speakerdeck.com/aya_se/data-centric-ai-swallow-corporus-56e2869a-f9bd-46cb-b030-1012235c37f7)

## Data Parallel

- データセットを GPU の個数に分割し、分割したデータセットを各 GPU に割り当て
- 割り当てられたデータセットをそれぞれ別々に学習
- 誤差逆伝播の後に、モデルの勾配を GPU 間で共有することで同期をとる



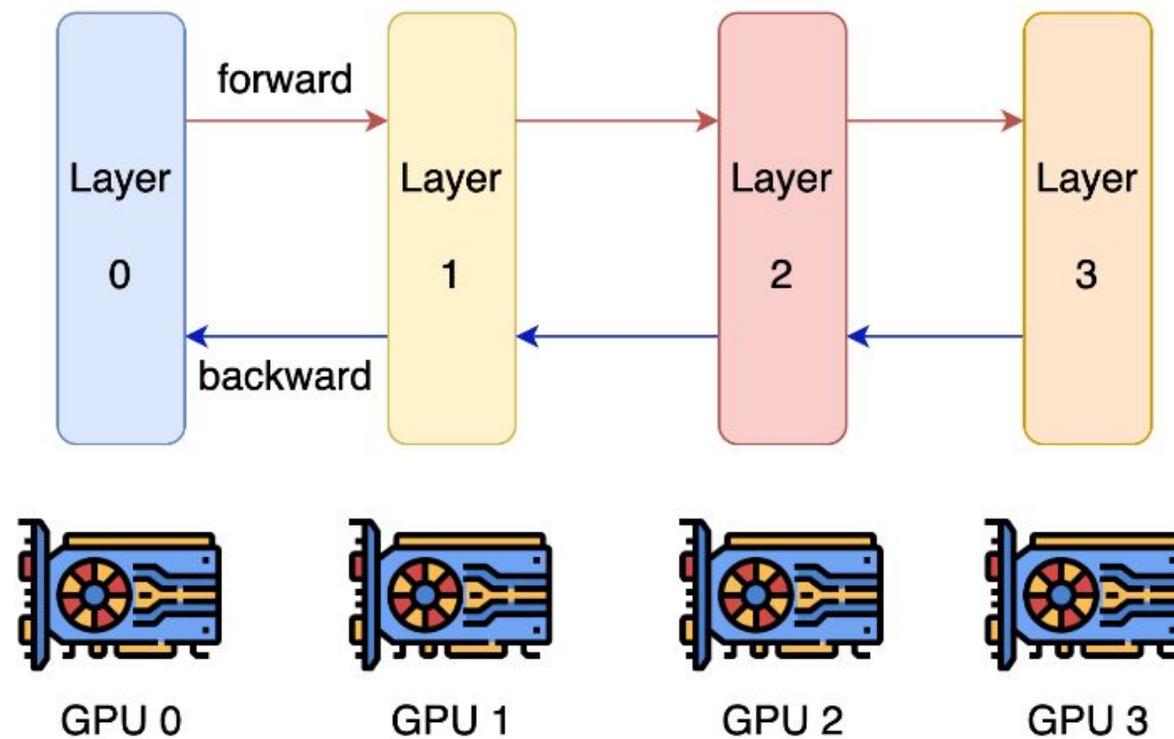
## 詳細解説

[https://zenn.dev/turing\\_motors/articles/0e6e2baf72ebbc](https://zenn.dev/turing_motors/articles/0e6e2baf72ebbc)

[https://colossalai.org/docs/concepts/paradigms\\_of\\_parallelism/](https://colossalai.org/docs/concepts/paradigms_of_parallelism/)

## Pipeline Parallel

- モデルを分割し、分割した Layer を別々の GPU が持つ
- 1つの GPU に乗りきらないサイズのモデルを学習を実現

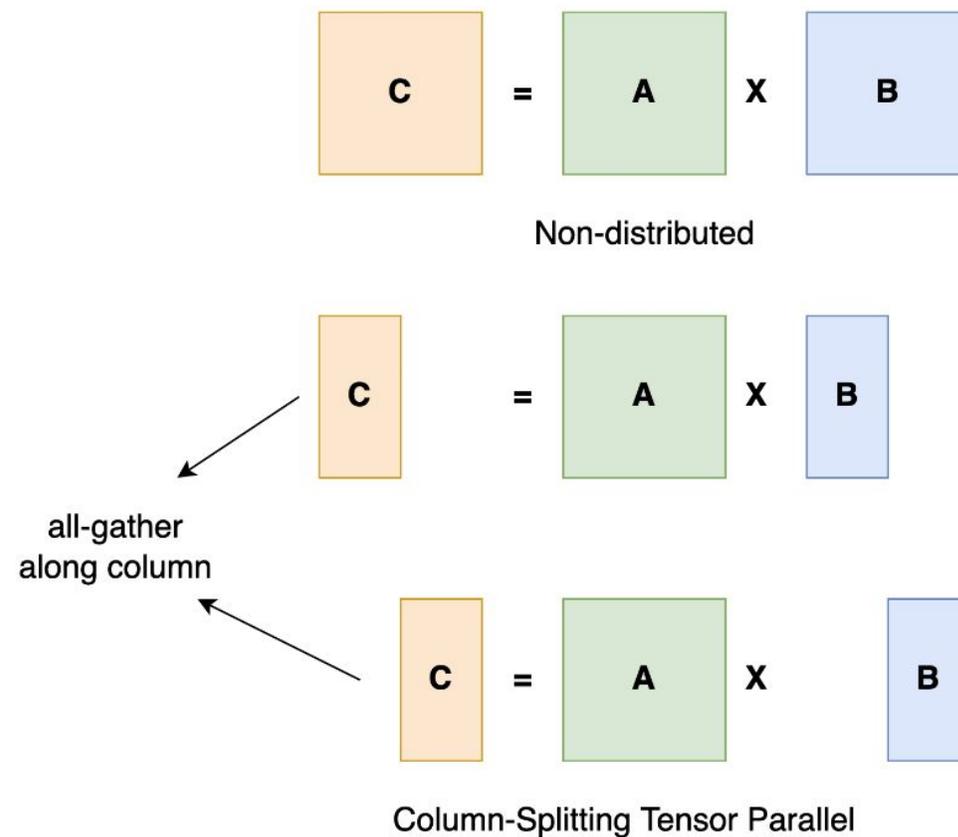


Pipeline parallel illustration

[https://colossalai.org/docs/concepts/paradigms\\_of\\_parallelism/#pipeline-parallel](https://colossalai.org/docs/concepts/paradigms_of_parallelism/#pipeline-parallel)

## Tensor Parallel

- テンソルを特定の次元に沿ってN個のチャンクに分割
- 各 GPU を分割された1/Nのテンソルだけ処理
- 各GPUを通信することで計算結果を集約



Tensor parallel illustration

[https://awsdocs-neuron.readthedocs-hosted.com/en/latest/libraries/neuronx-distributed/tensor\\_parallelism\\_overview.html](https://awsdocs-neuron.readthedocs-hosted.com/en/latest/libraries/neuronx-distributed/tensor_parallelism_overview.html)

1

最新版 Swallow

2

AIの最新動向

3

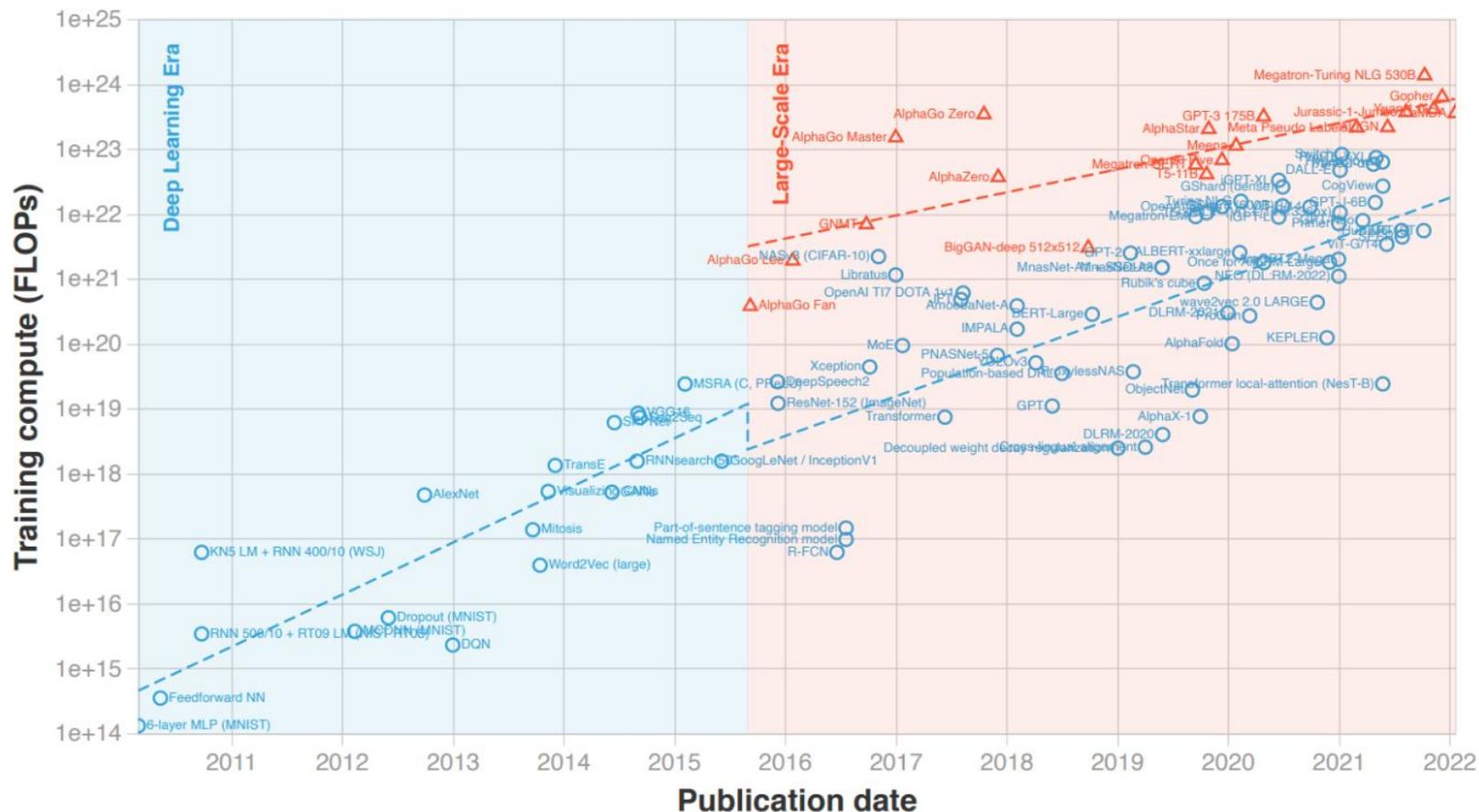
DeepSeekの登場

4

医療における推論能力の強化モデル

Training compute (FLOPs) of milestone Machine Learning systems over time

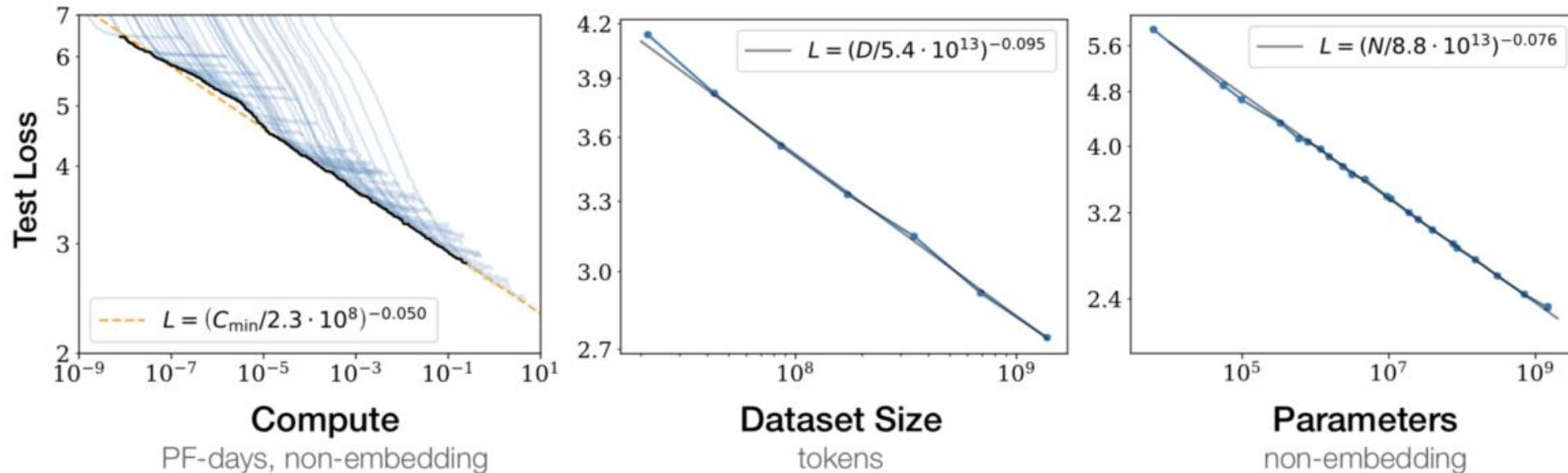
n = 102



学習に必要な計算量 (FLOPs)は以下のように年々膨大

<https://arxiv.org/abs/2202.05924>

- 大規模なモデルは小さいモデルと比較して性能が優れる
  - 性能を上げるには **モデルサイズ** と **学習トークン数** を一緒に大きくすると良い



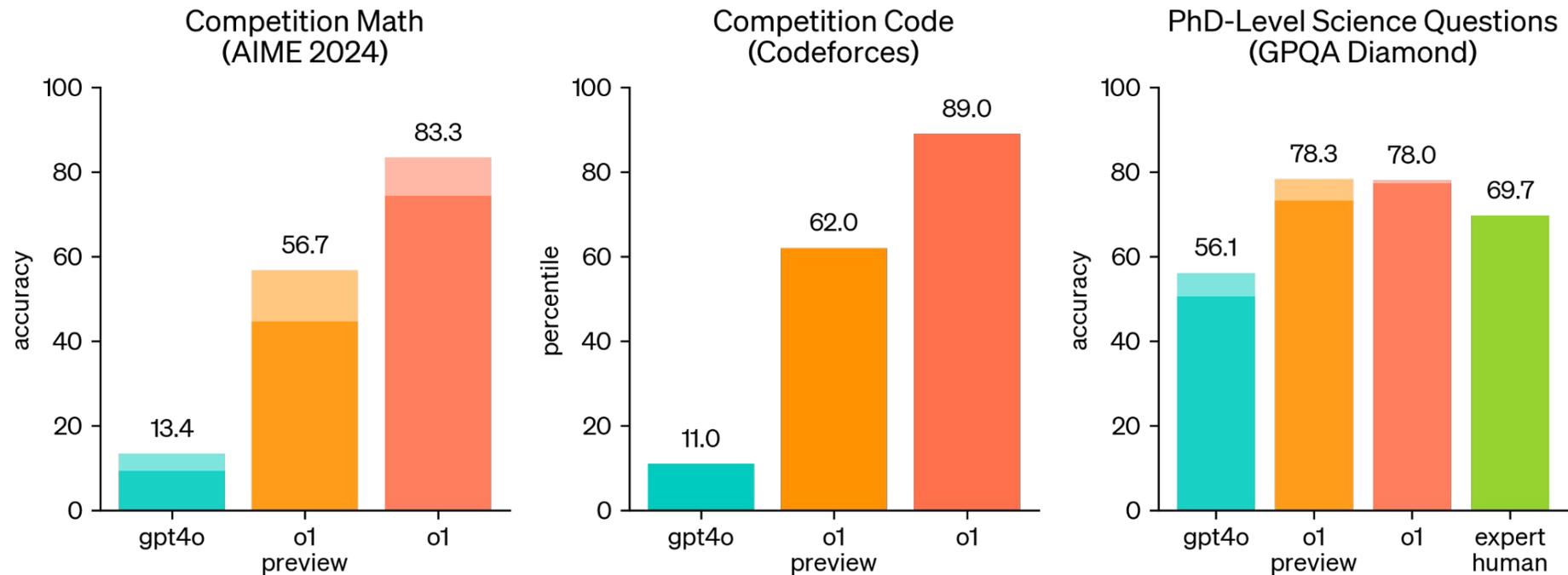
**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

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

<https://speakerdeck.com/iwiwi/stability-ai-japanniokeruda-gui-mo-yan-yu-moderunoyan-jju-kai-fa?slide=20>

- Test-Time Compute
  - (search/cot/sampling...etc)を用いてタスクを解く時に使う計算量
- Test-Time ComputeのScaling
  - (search/cot/sampling...etc)の計算量を増やすこと
- 主な特徴と目的
  - 推論時の計算強化:
    - 推論中に追加の計算を行う
    - 従来のAIモデルと比較して、より最適で精度の高い回答を導き出す

- chatGPT O1の登場
- 理数系の能力で、博士課程学生レベル
- 数学とコーディング能力: 数学オリンピック予選83%、Codeforces 89%



<https://openai.com/ja-JP/index/learning-to-reason-with-llms/>

## Test-Time Computeに使われるタスク

### MATH

USA Math Olympiad (AIME)

MATH [Hendrycks+, NeurIPS2021]

MathVista [Lu+, ICLR2024]

GSM8K [Cobbe+, 2021]

MiniF2F-MATH [Zheng+, ICLR2022]

FrontierMath [Glazer+, 2024]

### Code

Codeforces

SWE-bench [E Jimenez+, 2024]

LiveCodeBench [Jain+, 2024]

### その他？

GPQA [Rein+, COLM2024]

MMLU [Hendrycks+, ICLR2021]

ARC-AGI

1

最新版 Swallow

2

AIの最新動向

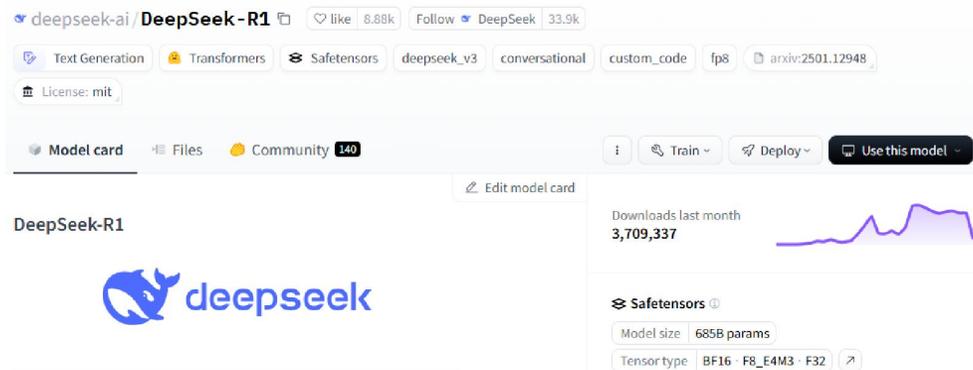
3

**DeepSeekの登場**

4

医療における推論能力の強化モデル

## DeepSeekショック



DeepSeek-R1をMITライセンスで公開 (2025年1月20日) [1]



NVIDIAの株価が一時17%下落 (2025年1月27日) [3]



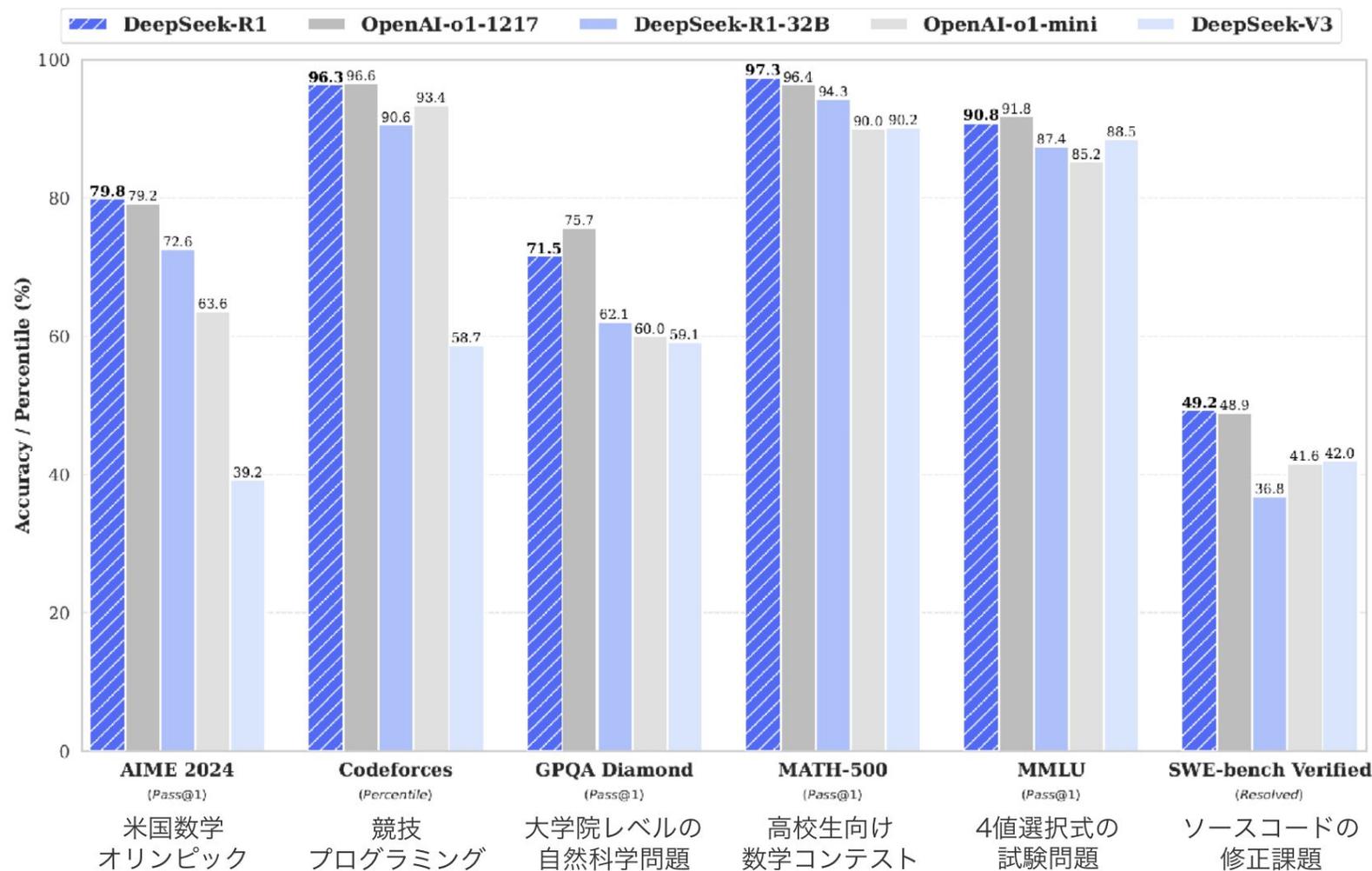
DeepSeekが米国のApp Store無料ランキングで1位に (2025年1月27日) [2]

- DeepSeek-R1はMITライセンスで無償公開されている(商用利用可)
  - DeepSeek-R1を手元にダウンロードして動かすのも、チューニングするの自由
  - モデルの再配布時に課される条件は著作権表示とMITライセンス全文の記載のみ

[1] <https://huggingface.co/deepseek-ai/DeepSeek-R1>

[2] ITmedia AI+. 中華AI「DeepSeek」、日本のApp Store無料アプリランキングでも1位に ChatGPT抑える (2025年1月28日)

[3] <https://www.nasdaq.com/market-activity/stocks/nvda>



Legend:

- OpenAI-o1-1217
- DeepSeek-R1

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](https://arxiv.org/abs/2501.12948). arXiv:2501.12948.

これは数学の大学入試問題です。人間が手計算で出来る範囲で、円周率は3.05より大きいことを証明してください。



DeepSeek R1

<think> じゃあ、円周率が3.05より大きいって証明する問題だよね。どうやってやるんだろう。そうだな、円に内接する正多角形を使う方法があったような気がする。アルキメデスが円周率を計算するのに、正多角形の周長で近似したって聞いたことがある。

まず、円の周りの長さは直径×円周率で、直径を1とすれば円周は $\pi$ になる。でも、内接する正多角形を使うと、その周長は円周より短くなるはず。だから、内接する正多角形の周長が3.05より大きくなるようなものを見つければ、 $\pi$ はそれより大きいから3.05より大きいと言えるよね。

例えば、正六角形だと、1辺の長さは半径と同じで、 $1/2$ 。正六角形の周長は $6 \times 1/2 = 3$ 。これは3.05より小さいからダメ。次に正十二角形を考える。内接する正十二角形の周長を計算したい。

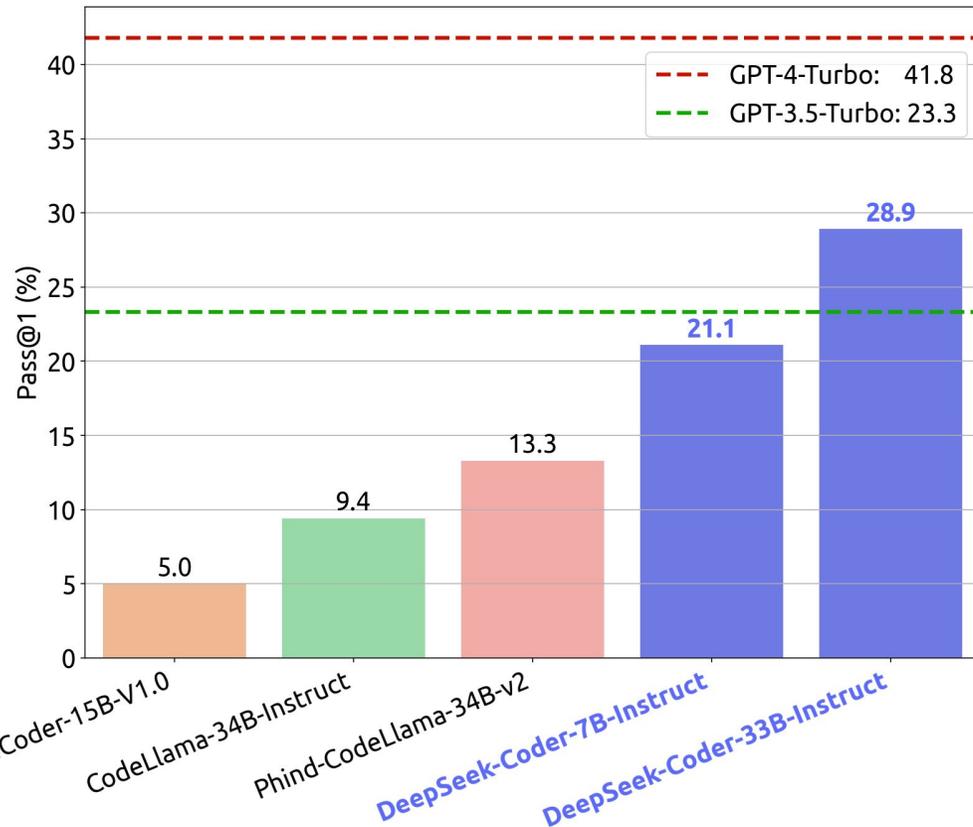
## コード生成と数学が思考を鍛える？

- LLMをプログラムコードとテキストで事前学習すると、推論力が高まる (Ma+ 2024)

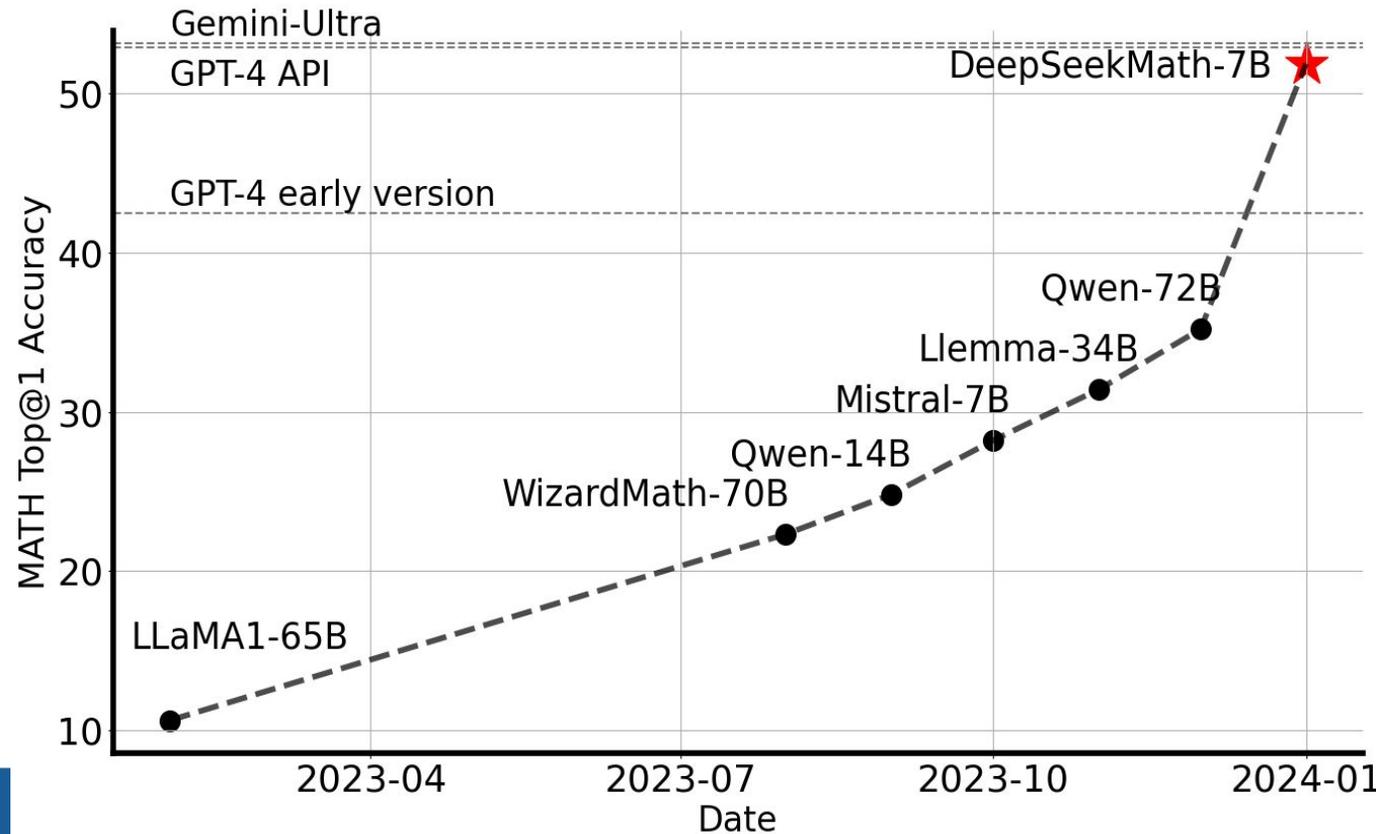
## コードや数学に特化した訓練データ構築

- コード: Githubのレポジトリからデータ構築, 数学: 高品質な数学関連ウェブサイトを収集

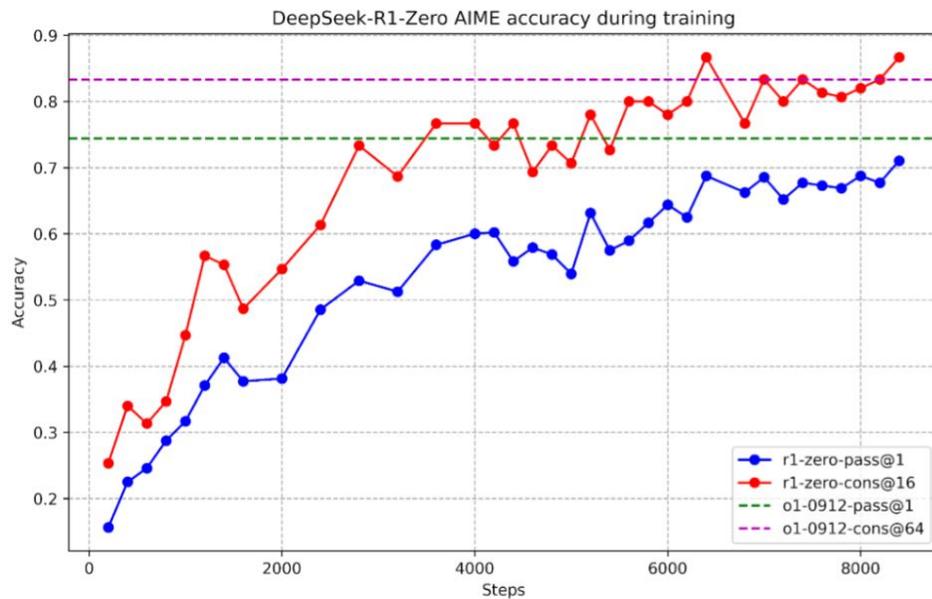
LeetCode Weekly Contest



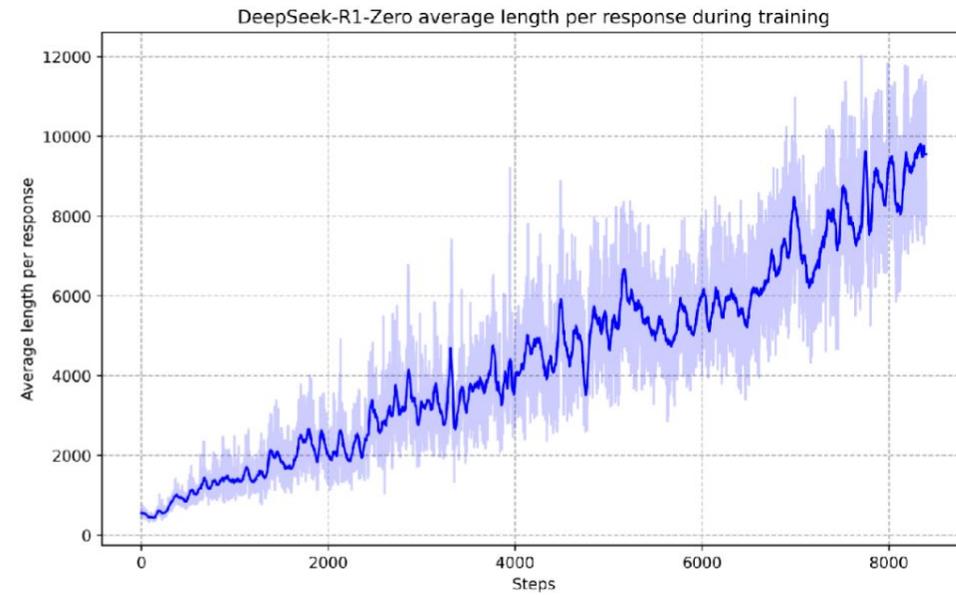
DeepSeek-Math (右) は数学でGemini UltraやGPT-4に迫る性能を示した



- 左図: 結果のフィードバックのみ(コード出力と数学の最終結果)で強化学習を行い、DeepSeek-R1はAIME(米国数学オリンピック)の正解率で学習が進むにつれて向上している。
- 右図: 学習が進むにつれて思考過程が長くなる傾向が見られ、DeepSeek-R1-Zeroは学習ステップ数の増加に伴い生成される思考過程の長さ(平均応答長)が増加している。(c.f., test time scaling)



横軸は学習の進み具合、縦軸は米国数学オリンピック(AIME)の正解率(DeepSeek-AI+, 2025)



横軸は学習の進み具合、縦軸は生成される思考過程の長さ(DeepSeek-AI+, 2025)

1

最新版 Swallow

2

AIの最新動向

3

DeepSeekの登場

4

医療における推論能力の強化モデル

## UC Santa CruzとAmazon Researchの研究 LLMが医療関連の質問に答える能力を向上させる新たなアプローチを提案

### 背景:

LLMは推論時に「考える時間」(生成するトークン数)を増やすことで性能が向上することが知られている(Test Time Scaling)

**しかし、数学などの分野とは異なり、知識表現や意思決定プロセスが複雑な医療分野での有効性は不明**

LLMに推論時により**長く考えさせる (Test Time Scaling)**ことで医療分野の質問応答性能を向上させるシンプルな手法「m1」を提案し、その有効性を示した

**しかし、単に長く考えさせるだけでは不十分で、根底にある医療知識の正確さが性能向上の鍵**であり、データやモデル自体の改善も必要であることを明らかにした

## データキュレーション

- 196K件の医療QAサンプル初期収集
- 難易度フィルタリング
- DeepSeek-R1による思考プロセス生成
- 医療ドメイン全体の多様性サンプリング

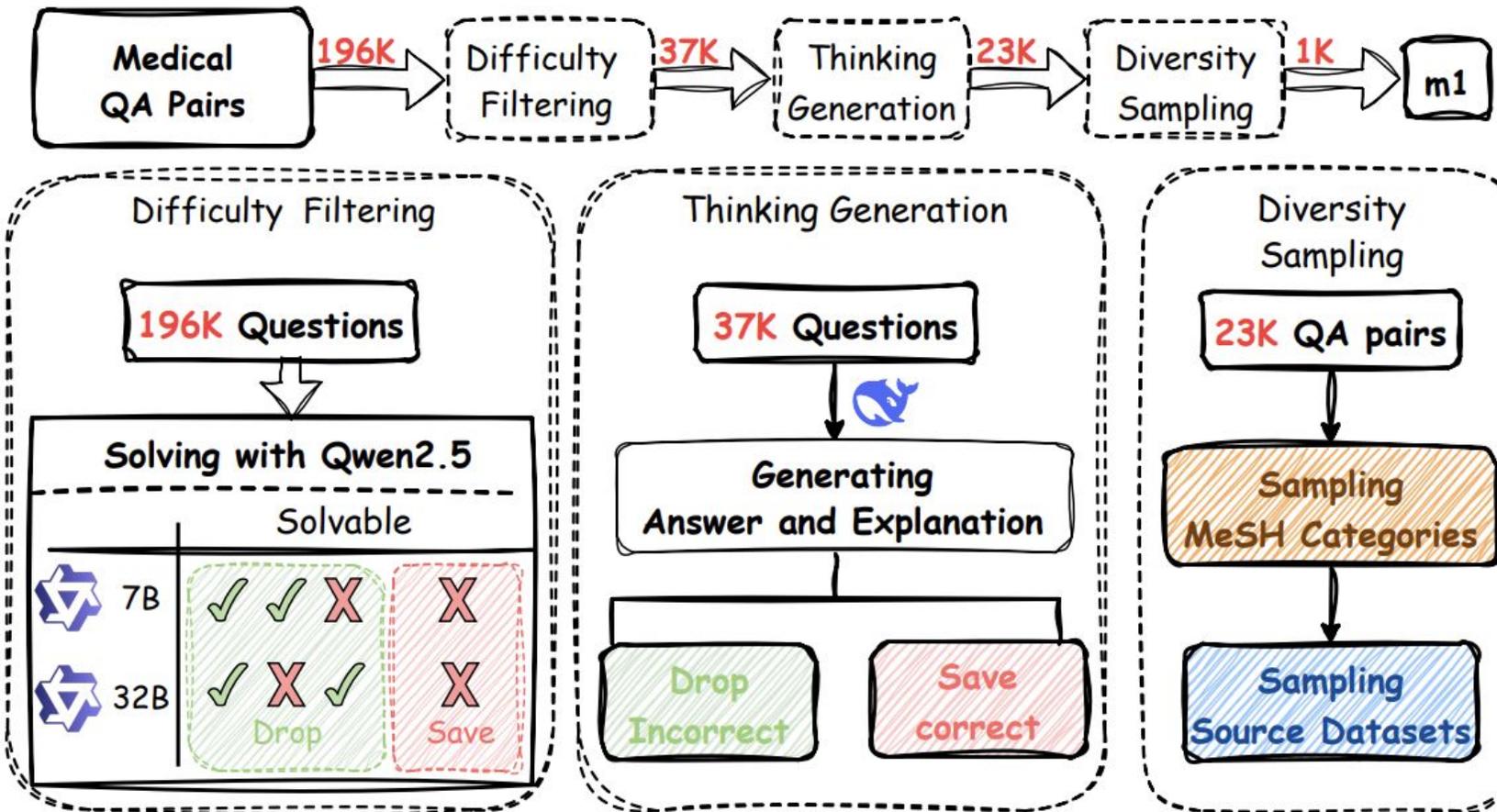
## 思考連鎖の推論形式

- 推論時の思考トークンを制御
- 推論トークンの拡張(128~8192トークン)

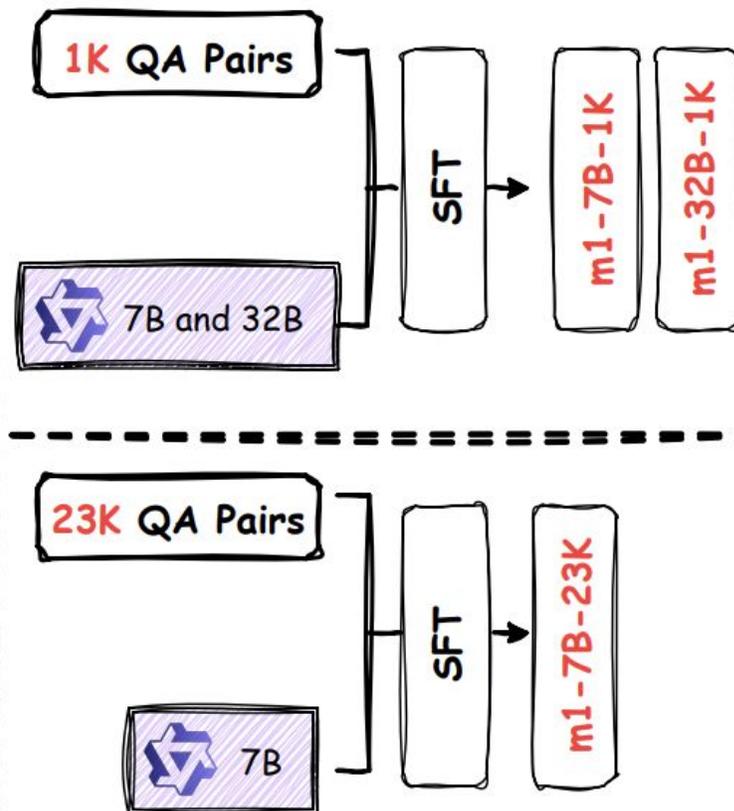
## 主な発見

- 医療用LLMにおける思考を強化したモデルにおける課題の特定
- データ品質が推論能力に影響
- 複雑な手法を上回る軽量アプローチ
- 1Kサンプルの32Bモデル ≈ 70B専門モデル

## Data Curation

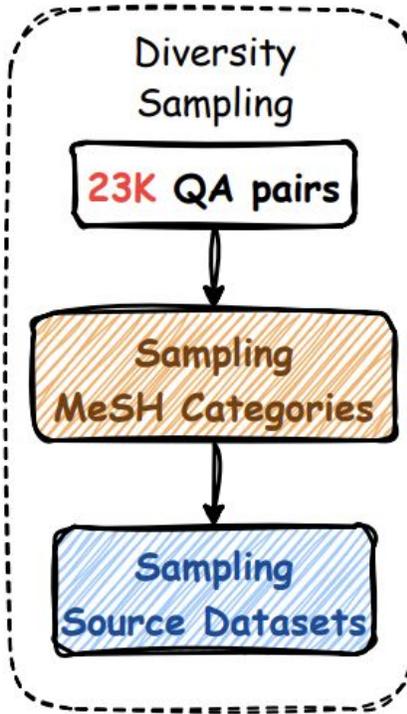
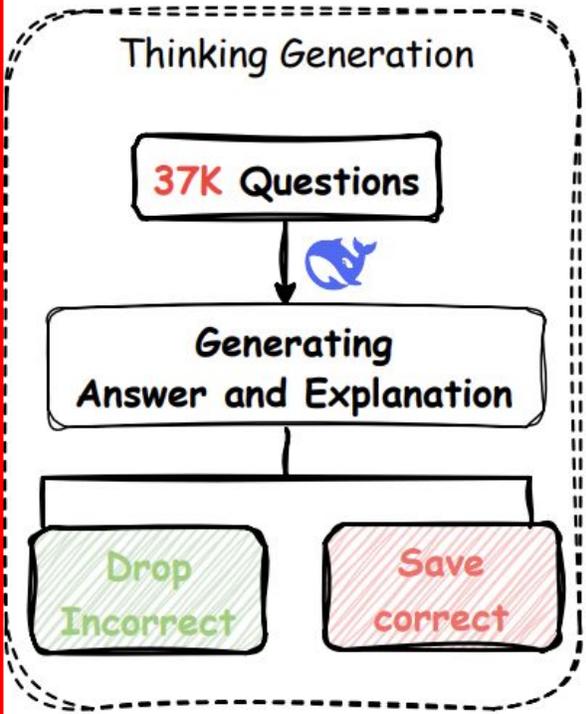
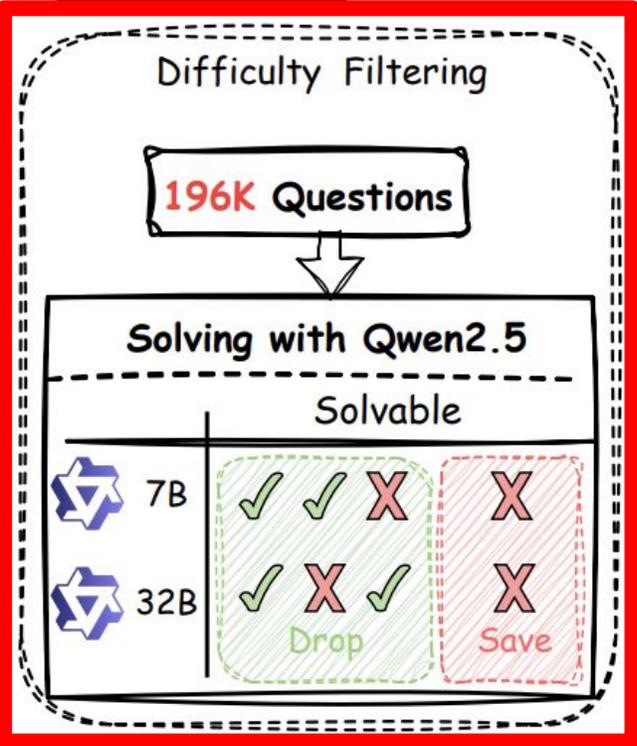
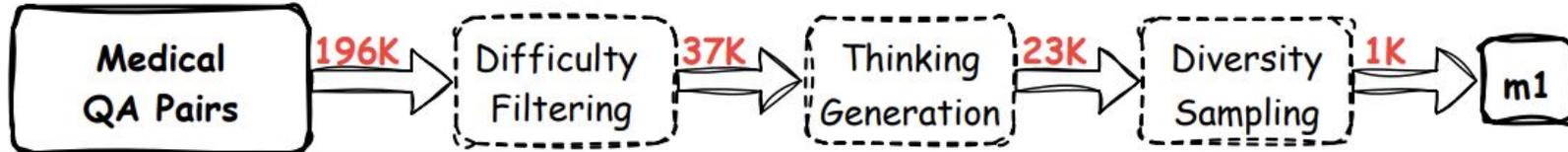


## Model Training

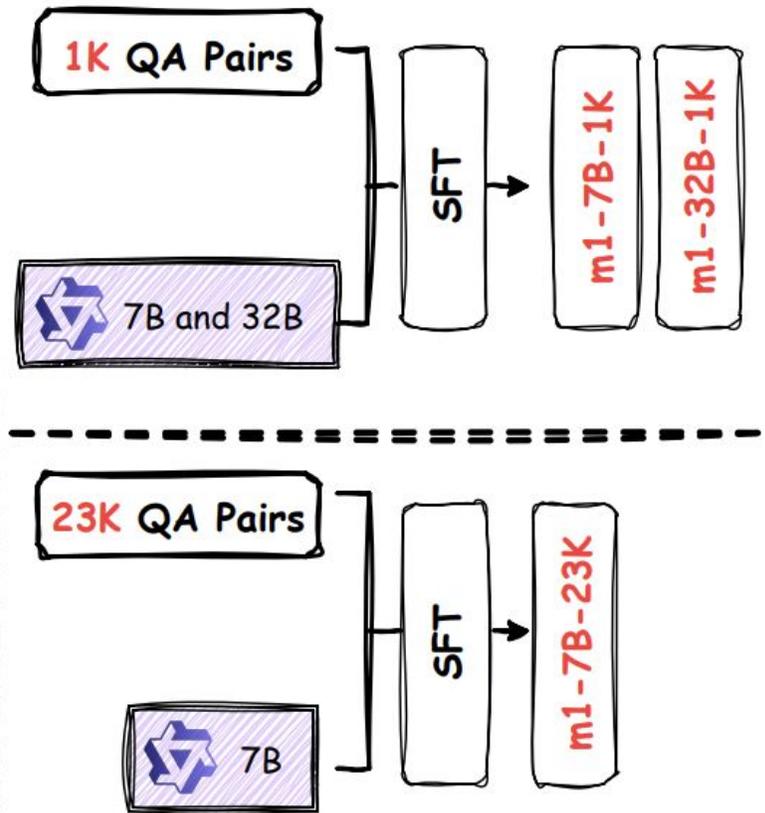


## 簡単すぎる問題を除外

### Data Curation

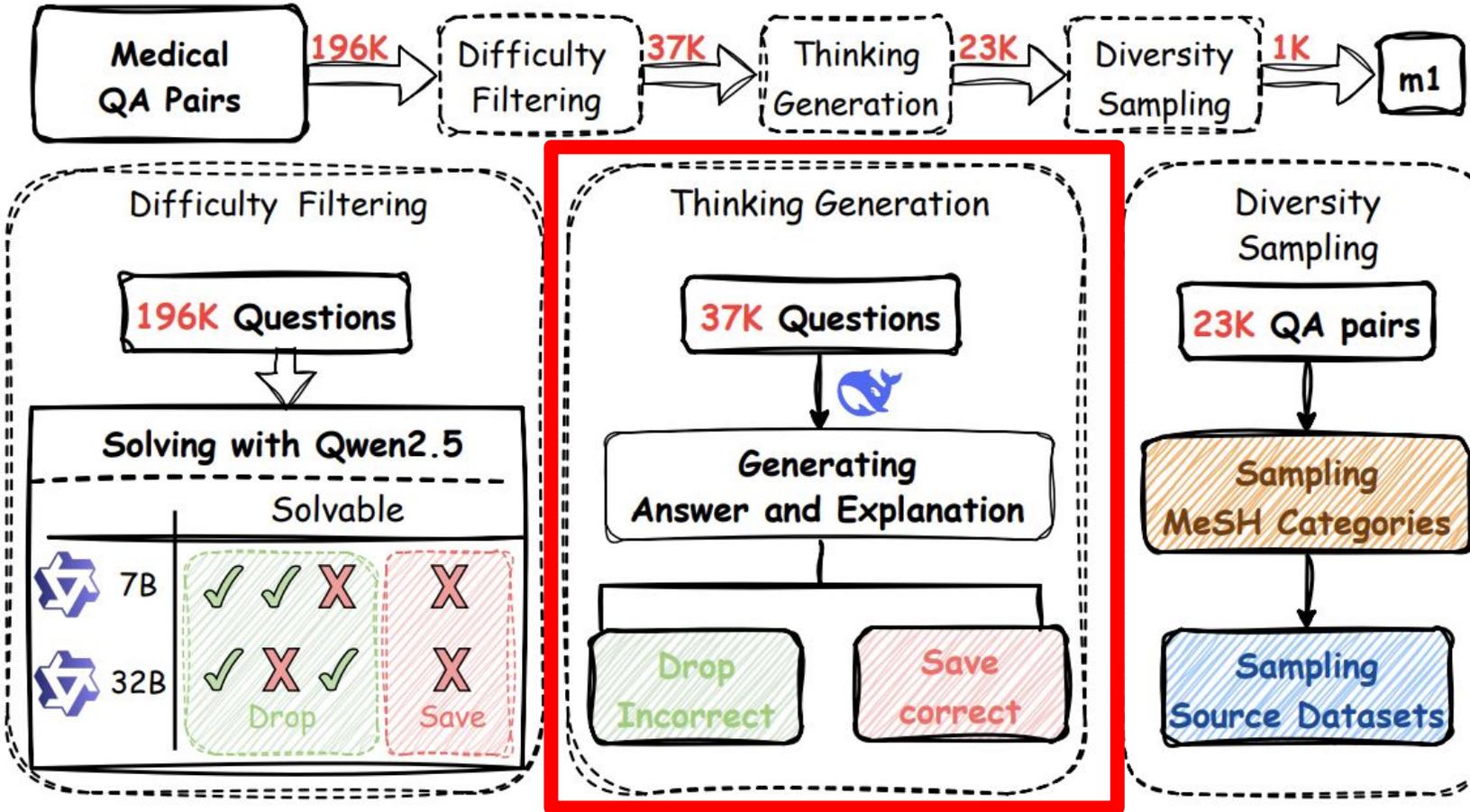


### Model Training

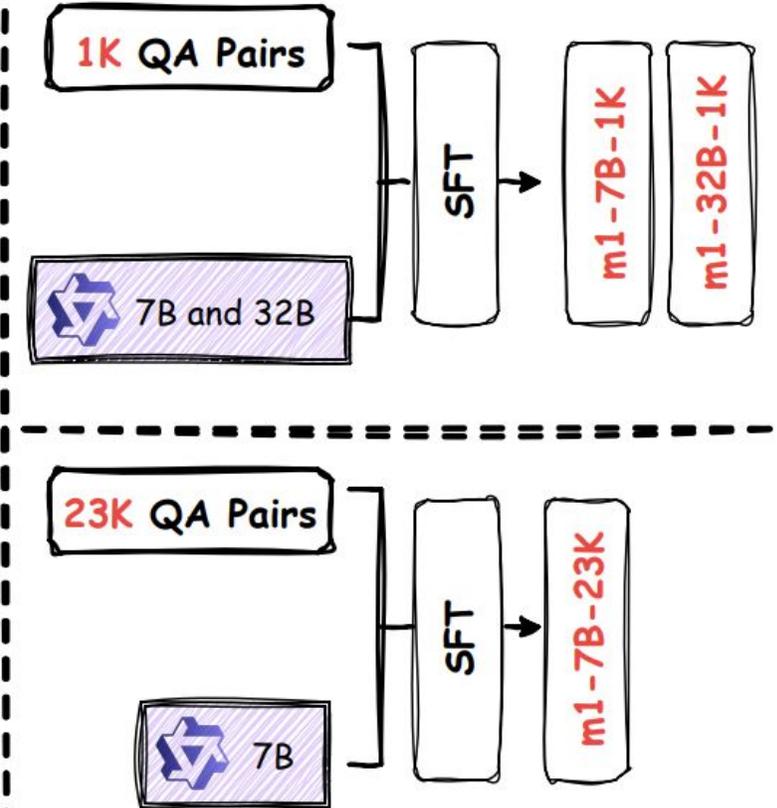


DeepSeek-R1により、難しい問題で LLMの思考プロセスを生成し、正解のみを採用

## Data Curation

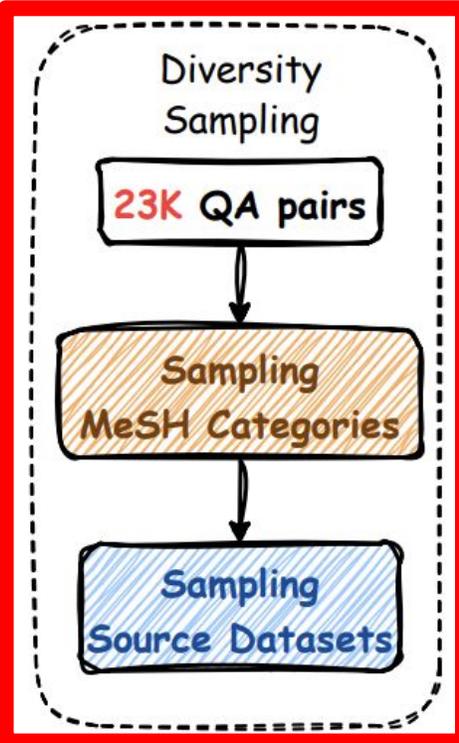
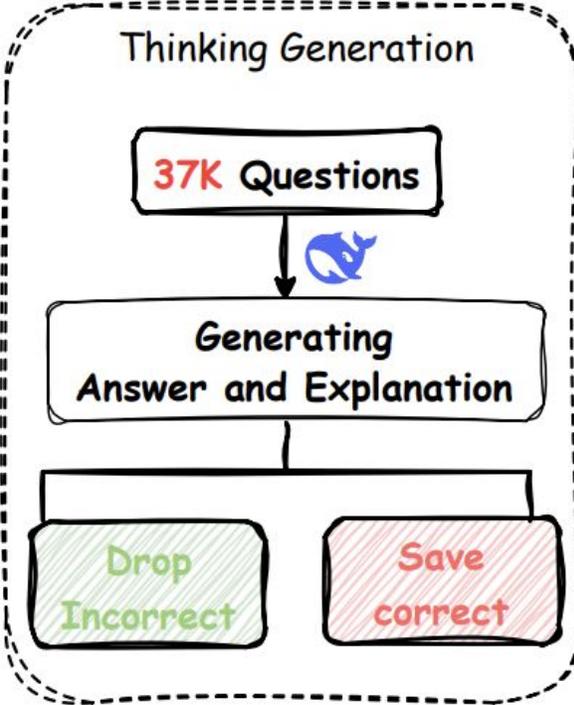
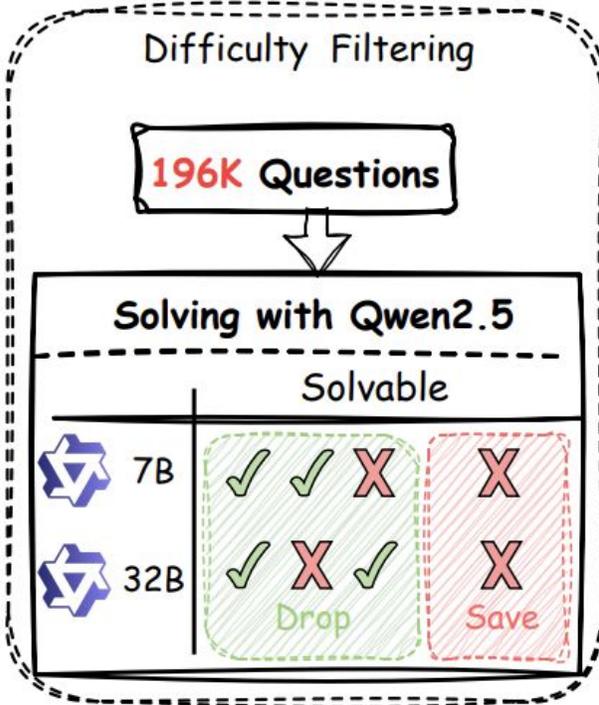
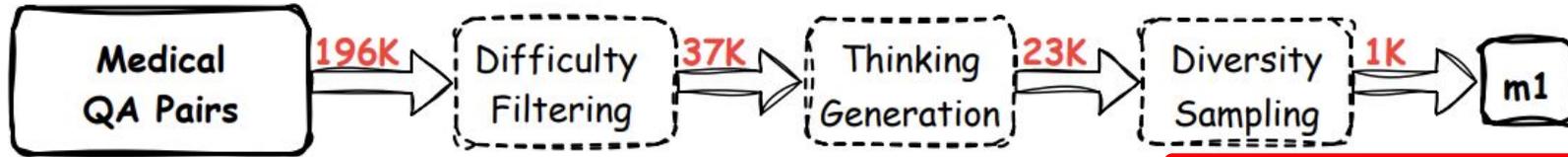


## Model Training

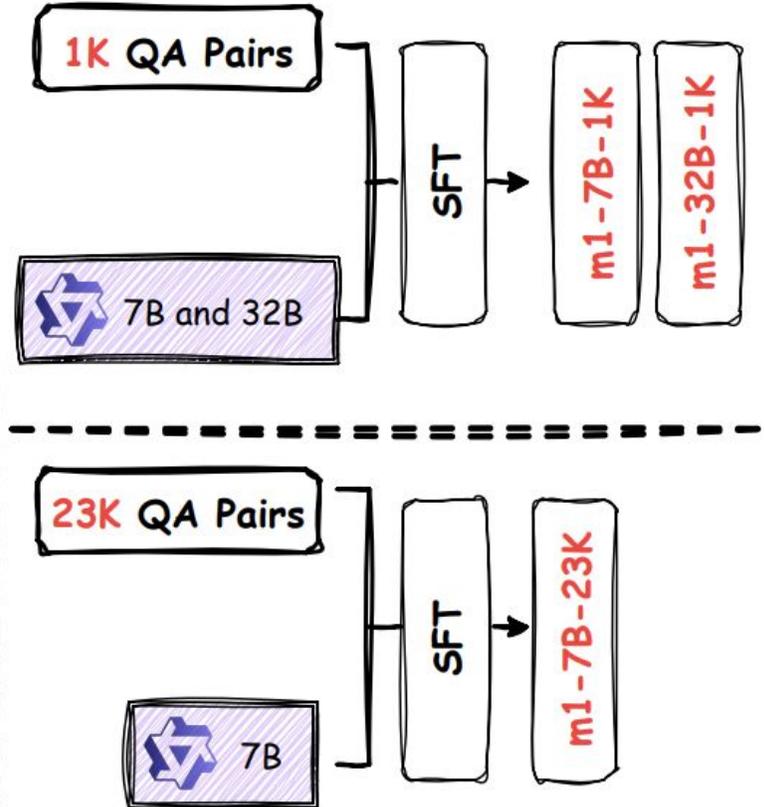


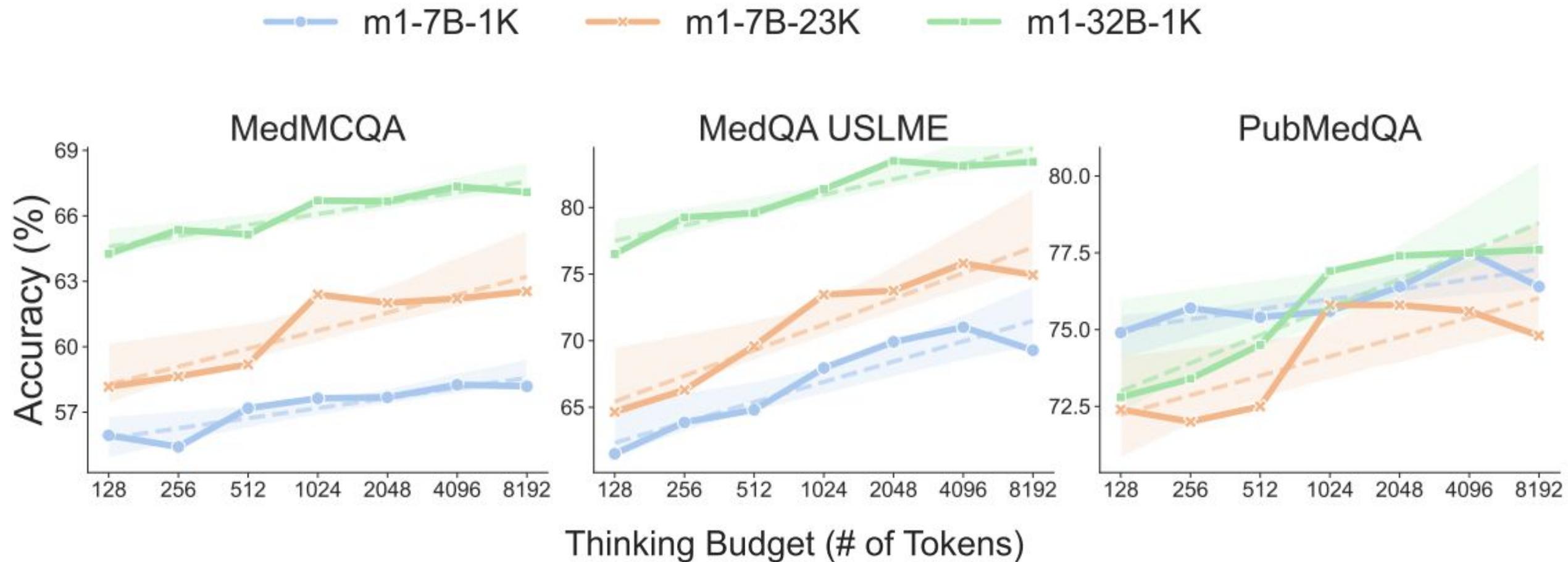
## バランスのとれた少数データセットを作成

### Data Curation

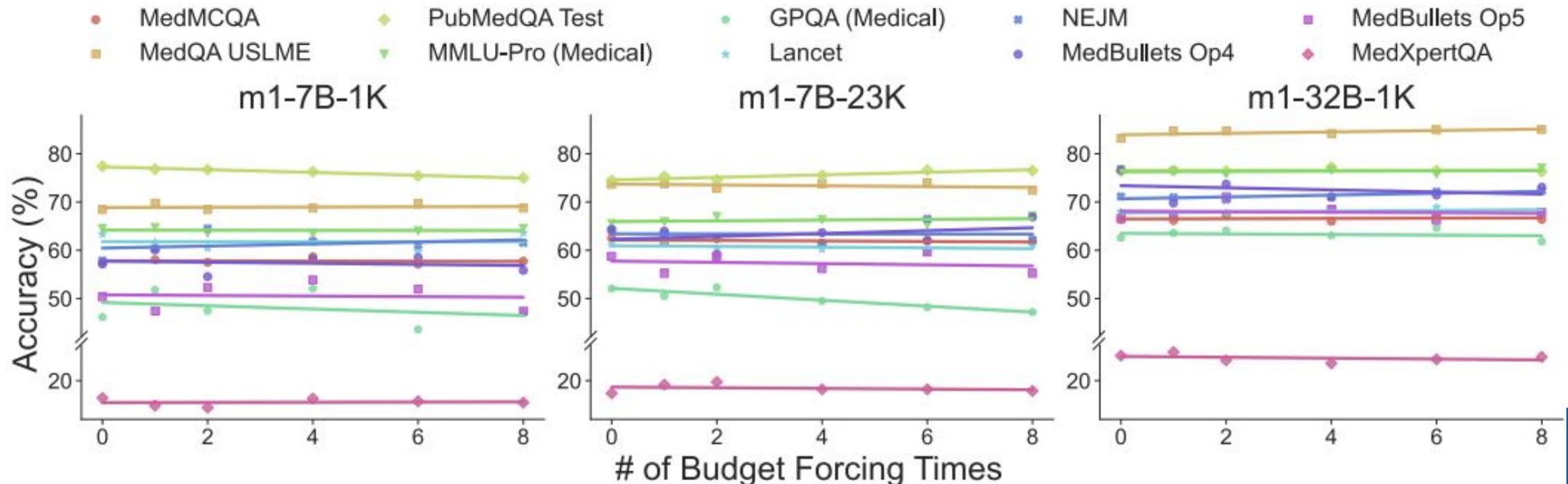


### Model Training





- Thinking Budget Control(思考予算の制御):
  - 最終的な答えを出力する前に、モデルが推論プロセス中に生成するトークンの最大数を設定
- Budget Forcing(予算強制):
  - モデルが時期尚早に結論付けようとした場合に、「考え続けて」と繰り返し促すことで、思考プロセスを拡張するテクニック。



## 医療QAでの限定効果

- 数学と異なり反復思考は医療では効果限定的
- 正しい答えから間違いへ誘導する場合も

## 知識のボトルネック

- 医療知識不足が性能向上の主制限要因
- 未学習概念への推論は効果的でない

## ベンチマーク間の効果の差

- 医療ベンチマーク間で改善度に大きな差
- 大幅改善～最小限改善まで多様な結果

1. 臨床意思決定支援
2. 医学教育
3. リソース制約のある環境でも有効
4. 医学研究

- 近年注目されているChatGPT o1, o3, o4, DeepSeek-R1のような思考モデルは数学、コードだけでなく医療にも有効である可能性
- しかし、医療データを事前学習や継続事前学習段階で学習していない場合、**医療知識の不足**により、LLMの思考を有効活用できない

### 今後の課題：

- 大規模な医療データを使った事前学習
- 現場に近いベンチマークの整備
  - 文章要約
  - 構造化

## Test-Time Computeに使われるタスク

GSM8K : 小学生の算数

MMLU : 初等数学、米国の歴史、コンピューターサイエンス、法律などを含む 57 のタスク

※最近ではMMLUからMMLU-Proへ(選択肢の数を4つから10、推論よりタスクへ)

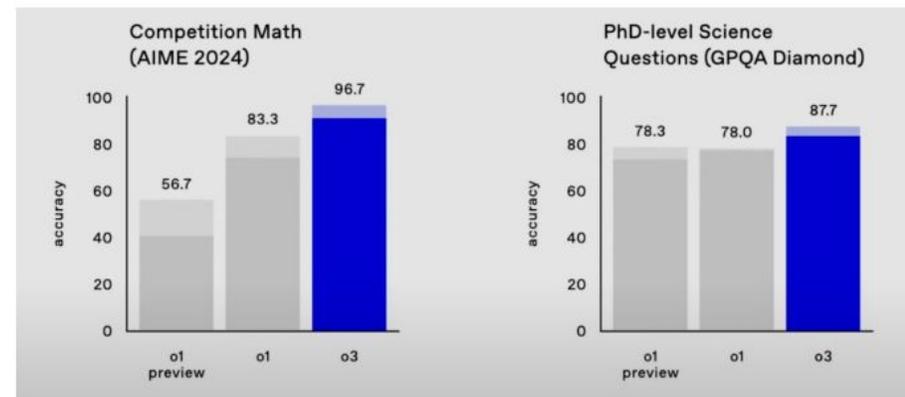
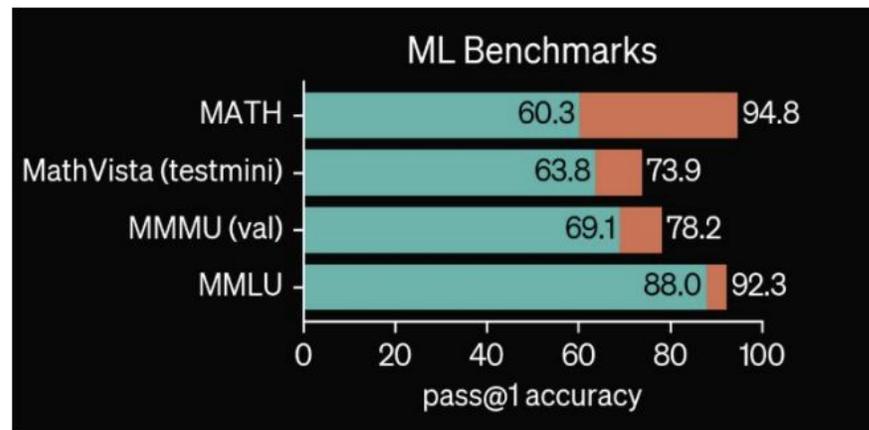
MATH : 高校レベルからundergraduateレベルの数学

AIME : 国際数学オリンピック(IMO)でアメリカを代表するチームを選ぶための、優秀な学生を競う一連の試験のうち 2 番目の試験

GPQA : 生物学、物理学、化学の分野の専門家が作成した 448 個の問題

Diamondは198個のより難しい問題セット

Codeforces : プログラミングコンテスト



## Test-Time Computeに使われるタスク

LiveCodeBench: LeetCode, AtCoder, and Codeforcesから新しいデータを定期的に更新し、汚染を避けている

MiniF2F-MATH : 形式的な証明可能なタスク

FrontierMath : 代数幾何学からツェルメロ-フランクフル集合論まで、現代数学の全範囲を網羅

MathVista : 視覚的文脈における数学的推論能力を評価

Rank	Model	Pass@1 ↓	Easy-Pass@1
1	<a href="#">04-Mini (High)</a>	73.3	99.1
2	<a href="#">04-Mini (Medium)</a>	72.2	98.2
3	<a href="#">Gemini-2.5-Pro</a>	67.8	97.3
4	<a href="#">03-Mini-2025-01-31 (High)</a>	67.4	99.1
5	<a href="#">Grok-3-Mini (High)</a>	66.7	97.3

Rank	Model	Accuracy ↑
1	<b>o3</b>	0.252
2	<b>Gemini 1.5 Pro (002)</b>	0.02
3	<b>Claude 3.5 Sonnet</b>	0.01

<https://livecodebench.github.io/leaderboard.html>

<https://paperswithcode.com/sota/mathematical-reasoning-on-frontiermath>