大規模言語モデルの推論能力の進展と今後の展望

小島 武

目次

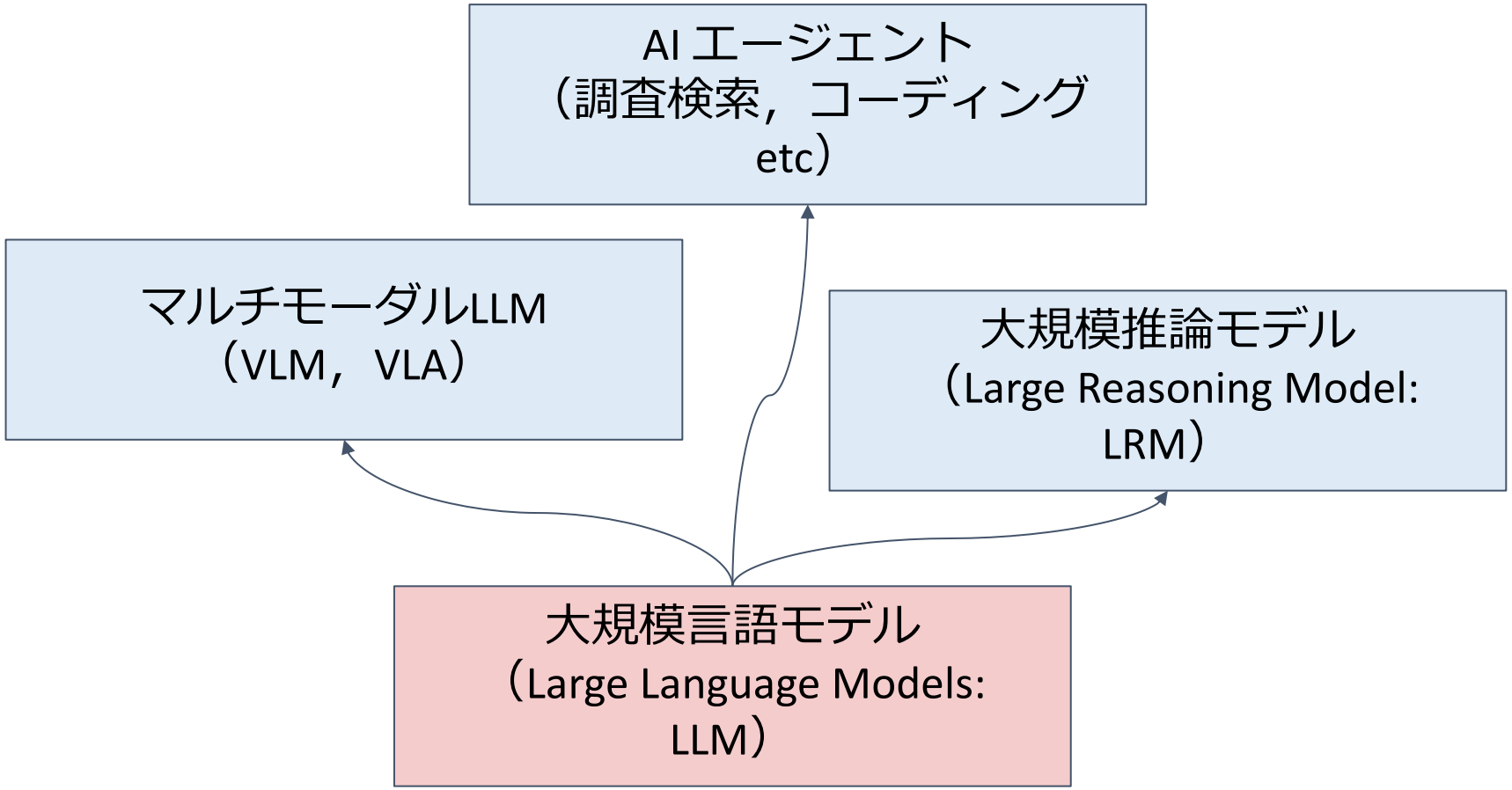
- 前半（16:35-17:15）
  - セミナー
    - 自己紹介
    - 大規模言語モデル（LLM）の進化の舞台裏
    - LLMのテスト時推論能力の向上
  - Q＆A
- 後半（17:20-18:05）
  - セミナー
    - 強化学習による推論能力の更なる進化
    - 推論モデルの軌跡をグラフとして捉えた分析
    - エージェント時代における推論
    - 今後の展望と課題
  - Q＆A

自己紹介

- 名前：小島 武（こじま たけし）
- 専門：大規模言語モデル（Large Language Models：LLM）
- 経歴：
  - 立命館大学 国際関係学部 国際関係学科（卒業）
  - NECソフト株式会社（現NECｿﾘｭｰｼｮﾝｲﾉﾍﾞｰﾀ株式会社）
  - 京都大学大学院 経済学研究科経済学専攻 修士課程（修了）
  - Peach Aviation株式会社
  - 2020～2023 東京大学大学院 工学系研究科TMI専攻 博士課程（修了）
  - 2023～2024 東京大学大学院 工学系研究科 特任研究員（松尾研究室所属）
  - 2025～ 同研究科 特任助教（松尾・岩澤研究室所属）

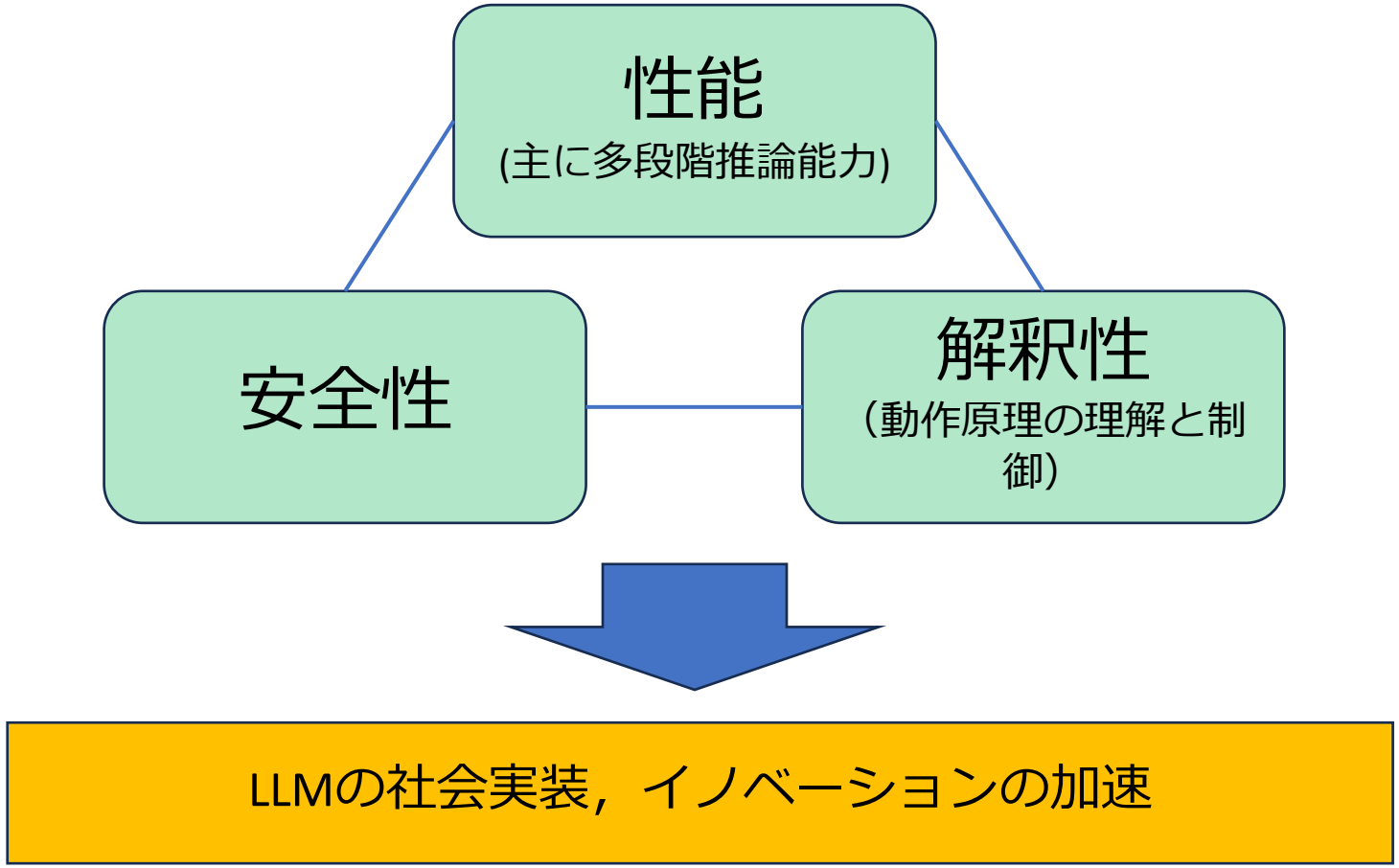
自己紹介

- 大規模言語モデルの研究の重要性



自己紹介

- 重点的に取り組んでいるLLMの研究領域



大規模言語モデルの進化の舞台裏

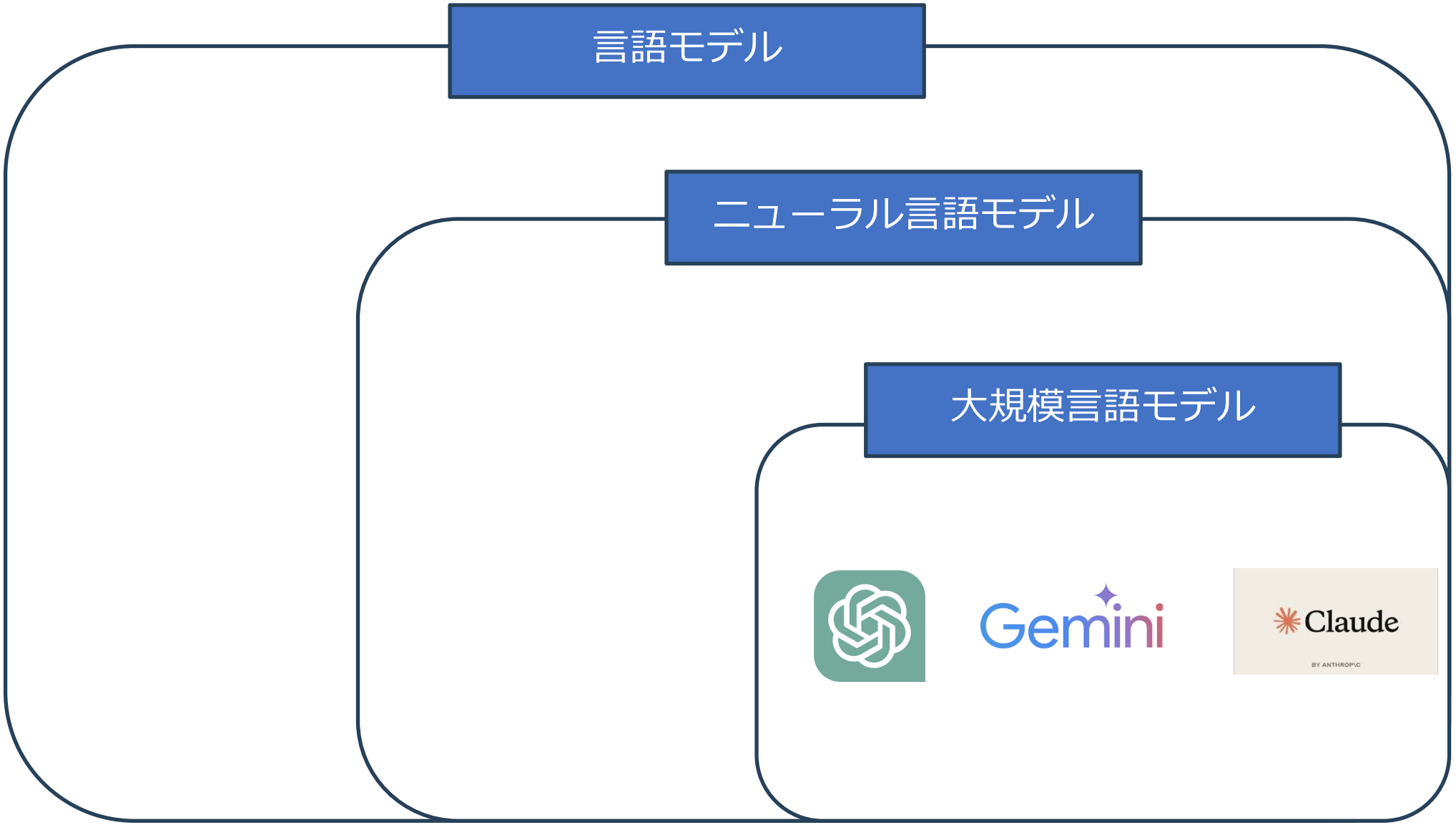
大規模言語モデルの驚異的な進化

- 2022年末 ChatGPTの発表以降、日進月歩の進化が今も続いている

- Gemini Deep Thinkが国際数学オリンピックで金メダル（2025）
  https://www.itmedia.co.jp/aiplus/article/2507/22/1250722056/
- OpenAI と Google がプログラミングの ICPC 世界大会で金メダル級の成績を達成（2025）
  https://venturebeat.com/ai/google-and-openais-coding-wins-at-university-competition-show-enterprise-ai
- AnthropicのClaude Mythosがソフト弱点数千件発見（2026）
  https://www.nikkei.com/article/DGXZQOGN081IO0Y6A400C2000000/

Gemini

大規模言語モデルとは



言語モデル (Language Models)とは

- 単語の系列（≒文章）の生起確率$p(x_1, x_2, \cdots, x_L)$をモデル化したもの
- $p(x_1, x_2, \cdots, x_L)$を連鎖律で分解したものは**自己回帰言語モデル**と呼ばれる

$$p(x_1, x_2, \cdots, x_L) = P(x_1)p(x_2|x_1)\cdots p(x_L|x_1, \cdots, x_{L-1})$$

- 条件付き確率がわかると**生成**ができる

$p(\text{東京}|\text{日本}, \text{の}, \text{首都}, \text{は}) = 0.2$
$p(\text{パリ}|\text{日本}, \text{の}, \text{首都}, \text{は}) = 0.001$
$\vdots$
$p(\text{カイロ}|\text{日本}, \text{の}, \text{首都}, \text{は}) = 0.0005$

日本の首都は → **東京**

$= \arg\max p\,(x|\text{日本}, \text{の}, \text{首都}, \text{は})$

- x_Lとしてふさわしい予測は、$\arg\max p\,(x_L|x_1, \cdots, x_{L-1})$

（参考）以前の代表的な言語モデルのひとつ

- 条件付き確率を統計的に求める方法
  - 大規模コーパス内の単語列の出現頻度から求める
  - 単語列 s の出現回数を#(s)とすると、

$$p(\text{東京} \mid \text{日本}, \text{の}, \text{首都}, \text{は}) = \frac{\#(\text{日本}, \text{の}, \text{首都}, \text{は}, \text{東京})}{\#(\text{日本}, \text{の}, \text{首都}, \text{は})}$$

200回

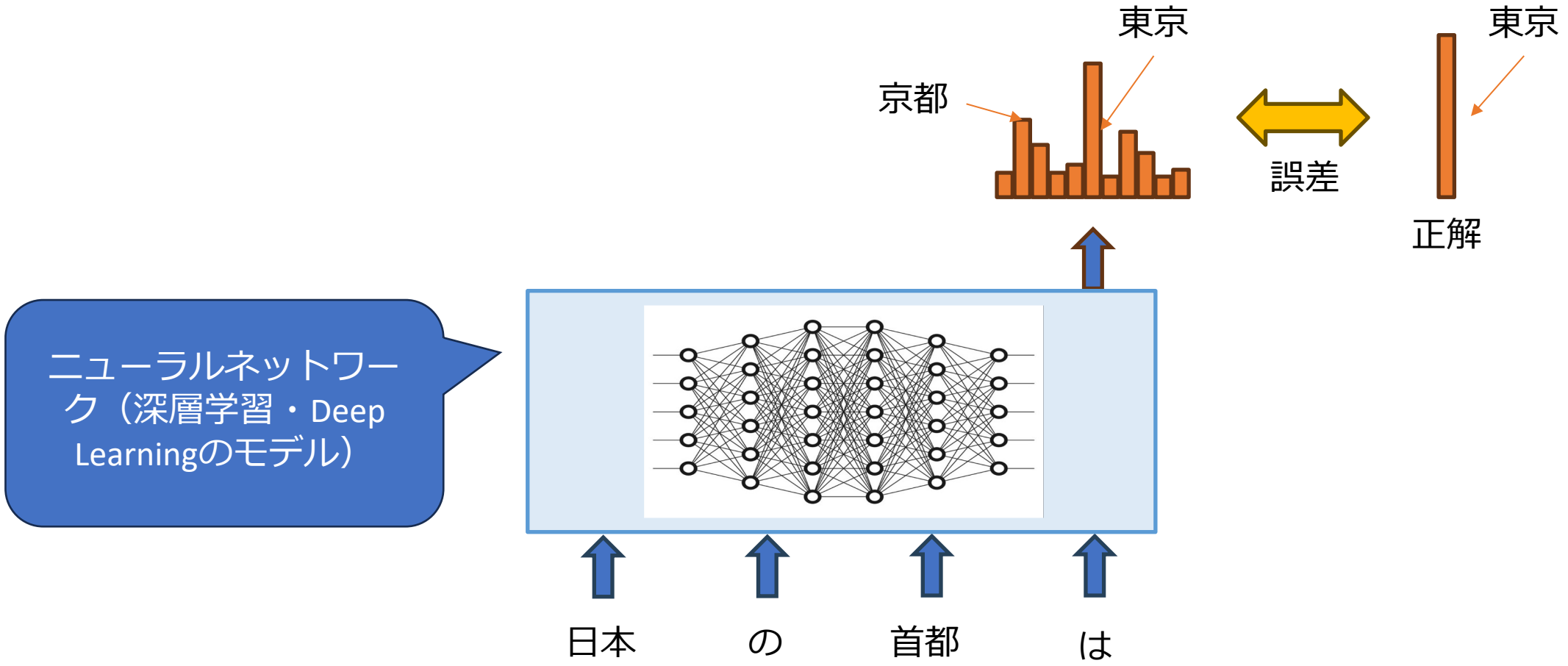
1000回

- 課題
  - 単語列が長くなると、その出現回数が急速に減少し、条件付き確率の推定が困難になる
  - 類義語が個別の事象として扱われてしまう．言い方を微妙に変えただけでも異なる出現頻度の語として扱われてしまう（例：”日本の首都は？”と”日本国の首都は？”）

[1] 岡崎直観 (2023), 大規模言語モデルの驚異と脅威 - Speaker Deck を参考に作成

ニューラル言語モデル

- 条件付き確率を何らかのニューラルネットで推定したモデル
- 他の機械学習と同様尤度を最大化するように訓練（誤差逆伝播）



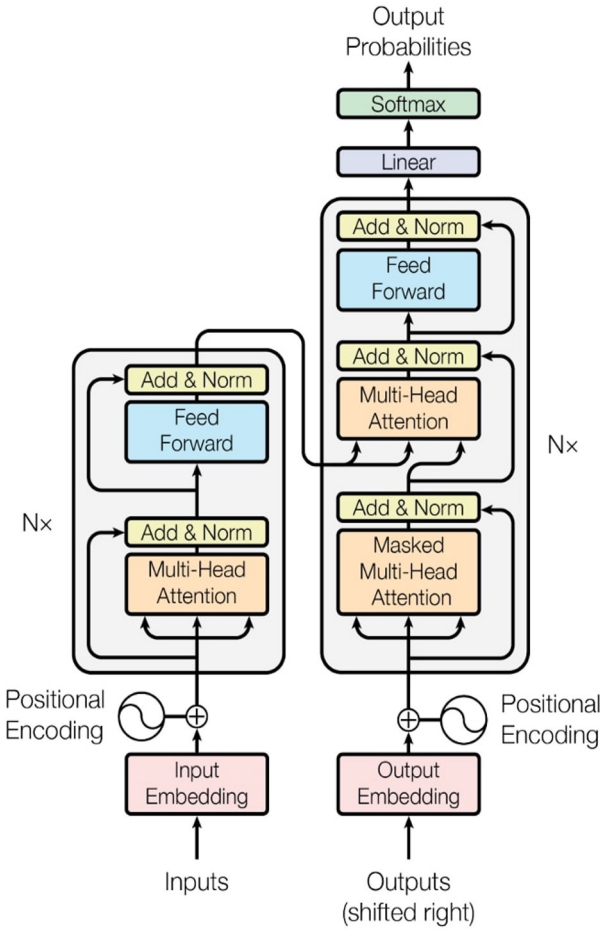
ニューラル言語モデルの大規模化

- 大規模化すると性能が良くなる；しかし大規模化は難しかった

- 大規模化を実現するための土台となった技術
  - 土台技術①
    - Transformerというモデル構造の発見
  - 土台技術②
    - 事前学習⇒事後学習 という二段階の学習へのパラダイムシフト
  - 土台技術③
    - スケール則の発見

LLMのモデル構造

Transformerが主流（”Attention Is All You Need”という論文で初出）



Attention is all you need

A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent … **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …

☆ 保存　99 引用　被引用数: 128159　関連記事　全 91 バージョン　»

2024/08/07 Gooogle Scholarにアクセス

- Googleを中心とした研究チームが発表
- 特徴① Attention機構の採用で単語（トークン）の長距離依存関係を効率的に学習
- 特徴② 学習時の並列計算も効率化できたことで大規模化（分散学習）しやすくなった

“Attention Is All You Need”, 2017

[6] Ashish Vaswani et al. (2017), “Attention Is All You Need”, NeurIPS2017 より引用

学習のパラダイムシフト

LLM以前

翻訳モデル

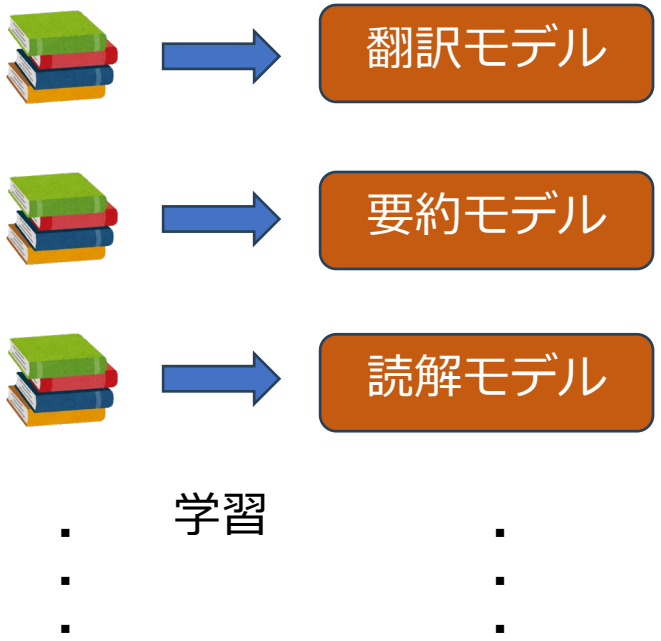
要約モデル

読解モデル

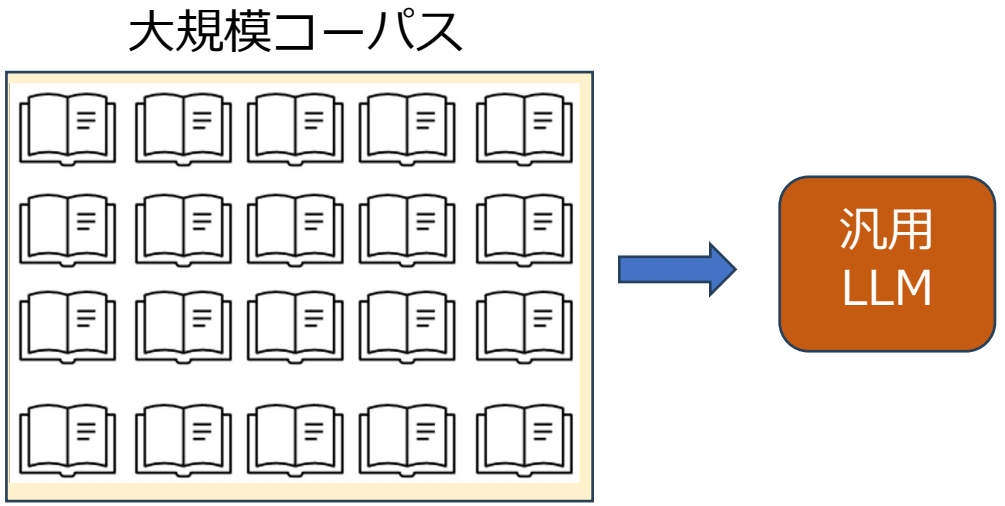
学習

学習のパラダイムシフト

LLM以前



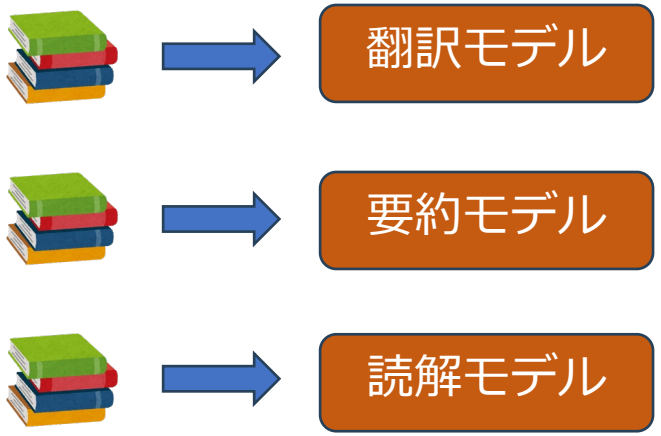
LLM



ネットから収集した
大量のテキストデータで
世界中のあらゆる知識を学ぶこ
とで汎用性能が手に入る

学習のパラダイムシフト

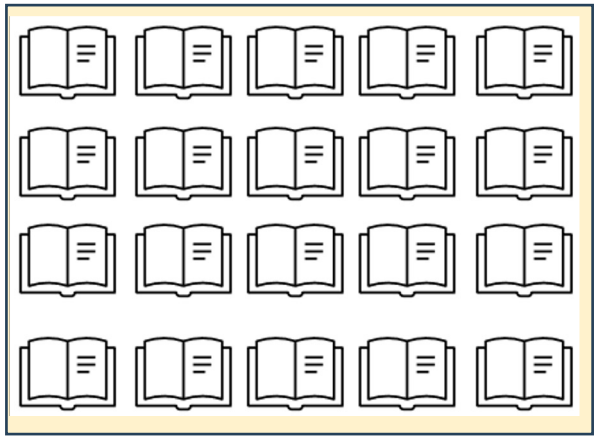
LLM以前



学習

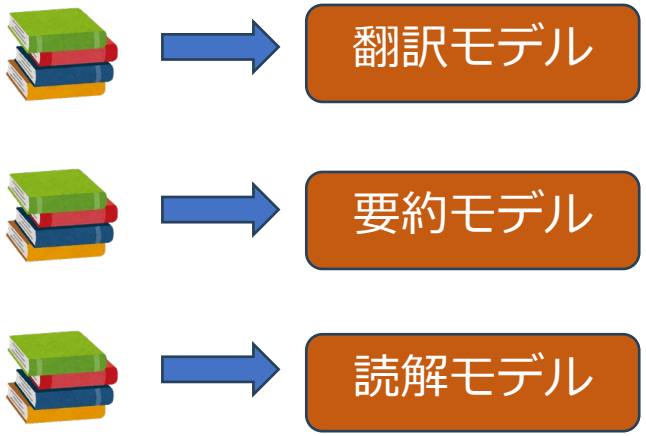
LLM

大規模コーパス



事前学習

汎用
LLM



事後学習

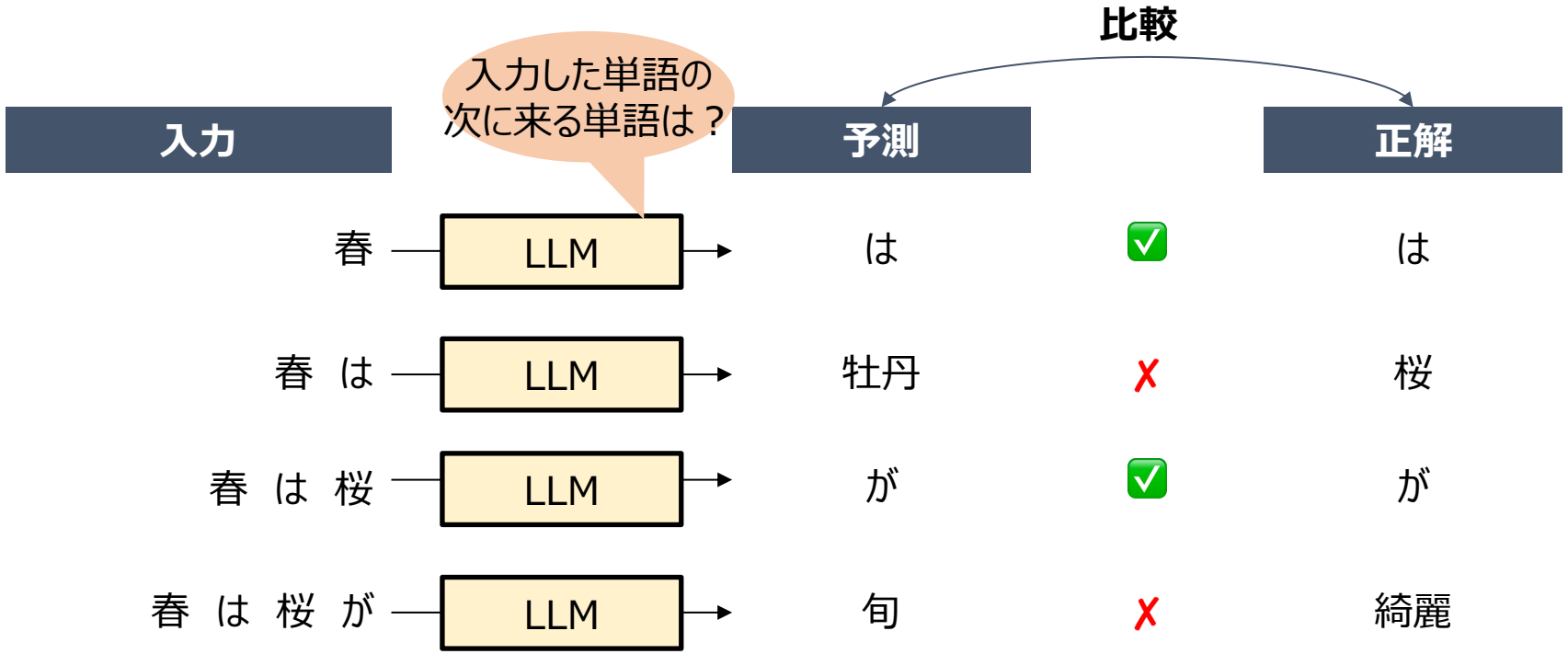
- ファインチューニング
- 強化学習

**ネットから収集した
大量のテキストデータで
世界中のあらゆる知識を学ぶこ
とで汎用性能が手に入る**

LLMの学習方法：事前学習／事後学習

事前学習は、Webから収集した**大量の文章**を使って、次の単語の予測をひたすら行う（**自己教師あり学習**という区分の学習方法）。事前学習の過程で、読み書きそろばんや世界中のあらゆる知識を学習する

- 事前学習：世界の知識を学ぶため、Webから収集した大量の文章を使って次の単語をひたすら予測する
  - GPTシリーズを代表とする現代のLLMは、この**事前学習を必ず行う**
  - 例えば以下の図のように、「**春は桜が綺麗**」というテキストの事前学習によって、「春」「桜」「綺麗」という言葉の間に強い関係性があること（＝世界の知識）を学ぶ



予測（AIの答え）と、正解（用意した文章で使われている言葉）との誤差が小さくなるようにモデルを調整

LLMの学習方法：事前学習／事後学習

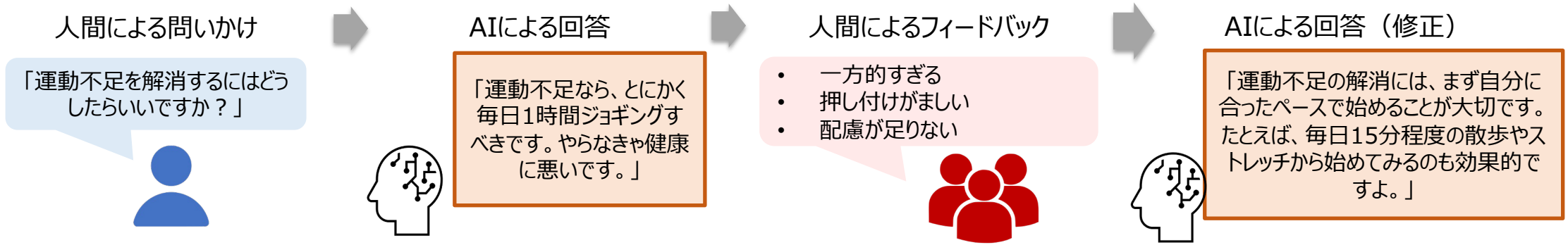
事後学習では、事前学習によって基礎知識が備わった汎用的なモデルに対して、特定の機能や専門的な分野に特化した追加の学習をする。事後学習を行ったLLMの先駆例はChatGPT

- 事後学習：事前学習が終わったモデルに対する追加の学習
  - 目的：モデルを特定の機能や専門的な分野に特化させる　例：翻訳LLM、チャットに特化したChatGPT
  - 特徴：人間が作成した**少量の高品質なデータ**での学習 → データに人間の価値基準が反映される

事後学習のやり方①：ファインチューニング
- QA形式(質問と答え)のペアデータやチャットデータで学習する
- 事前学習と同じく、次の単語をひたすら予測する学習を行う

質問 (Question)	答え (Answer)
健康維持のための3つのコツを教えてください。	1. バランスのとれた食事を摂り、野菜や果物を十分に摂ること。 2. 定期的に運動をして体の活力を保つこと。 3. 睡眠時間を十分にとり、規則正しい睡眠をとること。

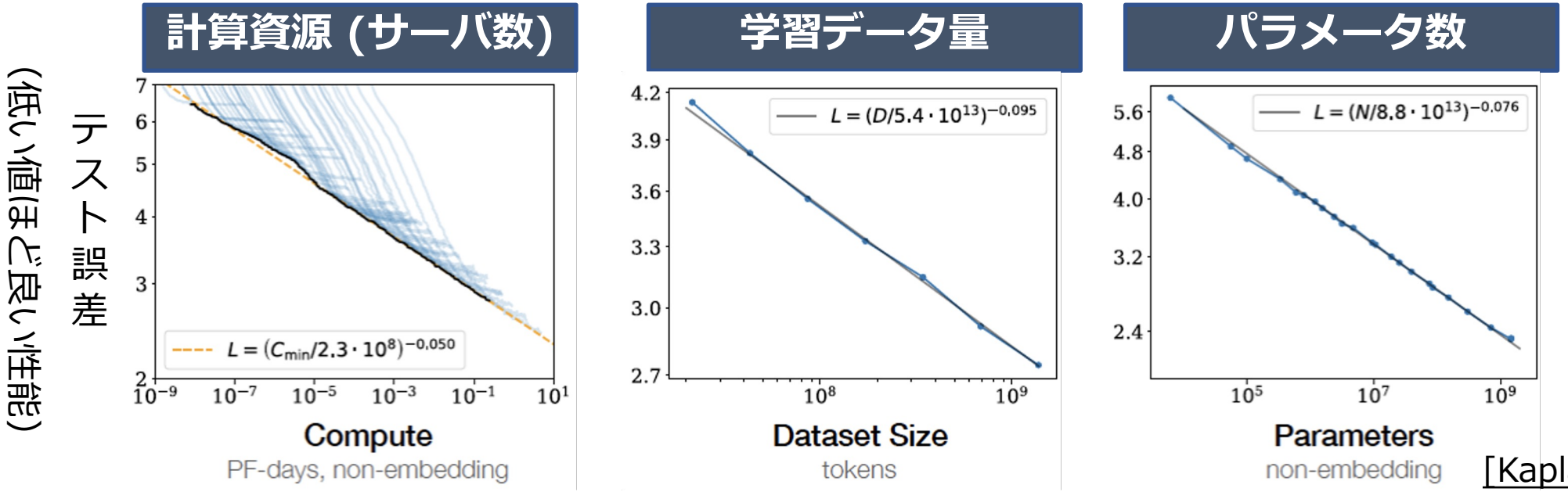
事後学習のやり方②：RLHF（Reinforcement Learning from Human Feedback）
- AIが出力した文章に対して、人間が良し悪しを評価してフィードバックすることで学習する
- 学習方法は特殊なので割愛（強化学習とよばれる方法を用いる）



スケール則の発見

スケール則（Scaling Law）とは，計算資源，学習データ量，パラメータ数の増加に比例して**事前学習の性能**が上がるという経験則。

- つまり、資源を投下するほど高性能なLLMが作れる事がわかった．大きく流れが変わった瞬間
- OpenAIは，いち早くスケール則に目をつけて大規模開発を開始し，その後世界的な投資合戦が始まる ⇒ ChatGPT, Claude, Gemini Llama, Qwen, DeepSeek, …



[Kaplan+ 2020]

“Language Models are Few-Shot Learners”, 2020

GPT-3の学習データ量

GPT-3の事前学習トークン数

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

- 約5000億トークン*のテキストを利用

  *トークンとは，言語AIが処理する単位.

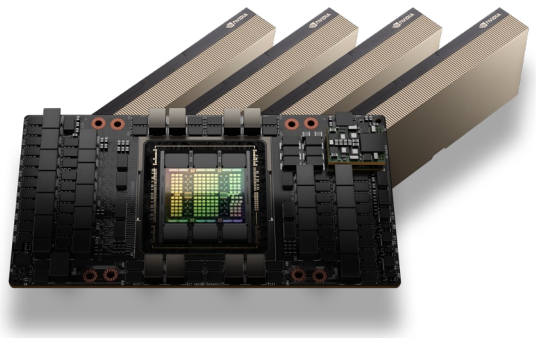
  日本語だと大体1文字1トークン
- *書籍でいうとGPT-3は**約500万冊**に相当

  参考：東大図書館が**約130万冊**,

  国会図書館が**約4700万冊**
- *リーク情報によるとGPT-4は**約1.3億冊**に相当

[7] Tom Brown et al. (2020), “Language Models are Few-Shot Learners”, NeurIPS2020 より引用

大規模な計算を行うためのツール：GPU



GPUs as fast matrix multipliers

Fast Matrix Multiplies using Graphics Hardware

E. Scott Larsen
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3175 USA
larsene@cs.unc.edu

David McAllister
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3175 USA
davemc@cs.unc.edu

Implementation

We mention here some observations we made during our implementation that may be of interest to those duplicating our results.

Refresh Rate We found that setting the refresh rate on the monitor as low as possible made marginal improvements (about 10%).

RGBA We found that 4 numbers can be packed into a single pixel, by setting the red, green, blue, and alpha channels to different values.

Texture Format Changing the format of the texture creation and read-back from RGBA to ABGR_EXT (in OpenGL) gave about 40% improvement on our hardware. This is because the hardware driver avoids reformatting the data from the application format to the card format. There is a number of options here, with near equal performance for each option except the one used natively on the specific hardware. The native format should give significant improvement.

Full Screen Running full screen instead of in a window provides improved performance.

Various other optimizations yielded minor (<1%) improvements.

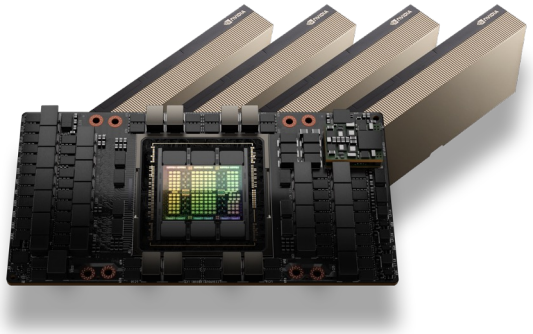
Early days of NVIDIA GPUs – programmable shaders. Researchers hacked this to do matmuls

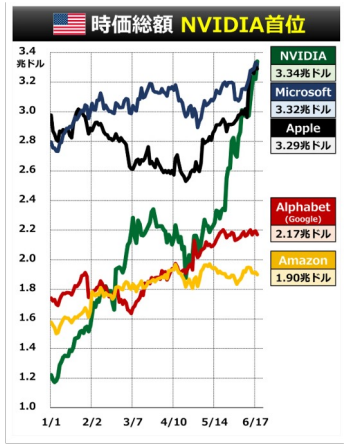
GPUがもともと持っていたグラフィックス処理用の能力（特に行列演算）が、研究者たちの工夫によってAIの大規模計算へと応用され始めた。

大規模なモデル / データを支える大規模な計算資源（GPU）

AIの開発には、膨大なデータを高速に処理する計算資源が必要。現在よく用いられる計算資源はGPUと呼ばれるもので、支配的なシェアを持つNVIDIAも急成長。日本もGPUの確保に動いているが、海外勢との格差は大きい

GPU（H100, A100, V100などの種類が存在）





世界のGPUシェアの90%を占めるNVIDIA（米）はAI需要を追い風に急成長。一時は世界の時価総額首位に

GPT3相当の場合：A100 × **1,200基 × 30日**

GPT4相当の場合：A100 × **25,000基 × 100日** (*)

(*)リーク情報。OpenAIの公式発表ではない

国内外の代表的なGPUクラスタ(*)

(*)GPUを搭載した複数の計算機をまとめて提供するシステム

国内

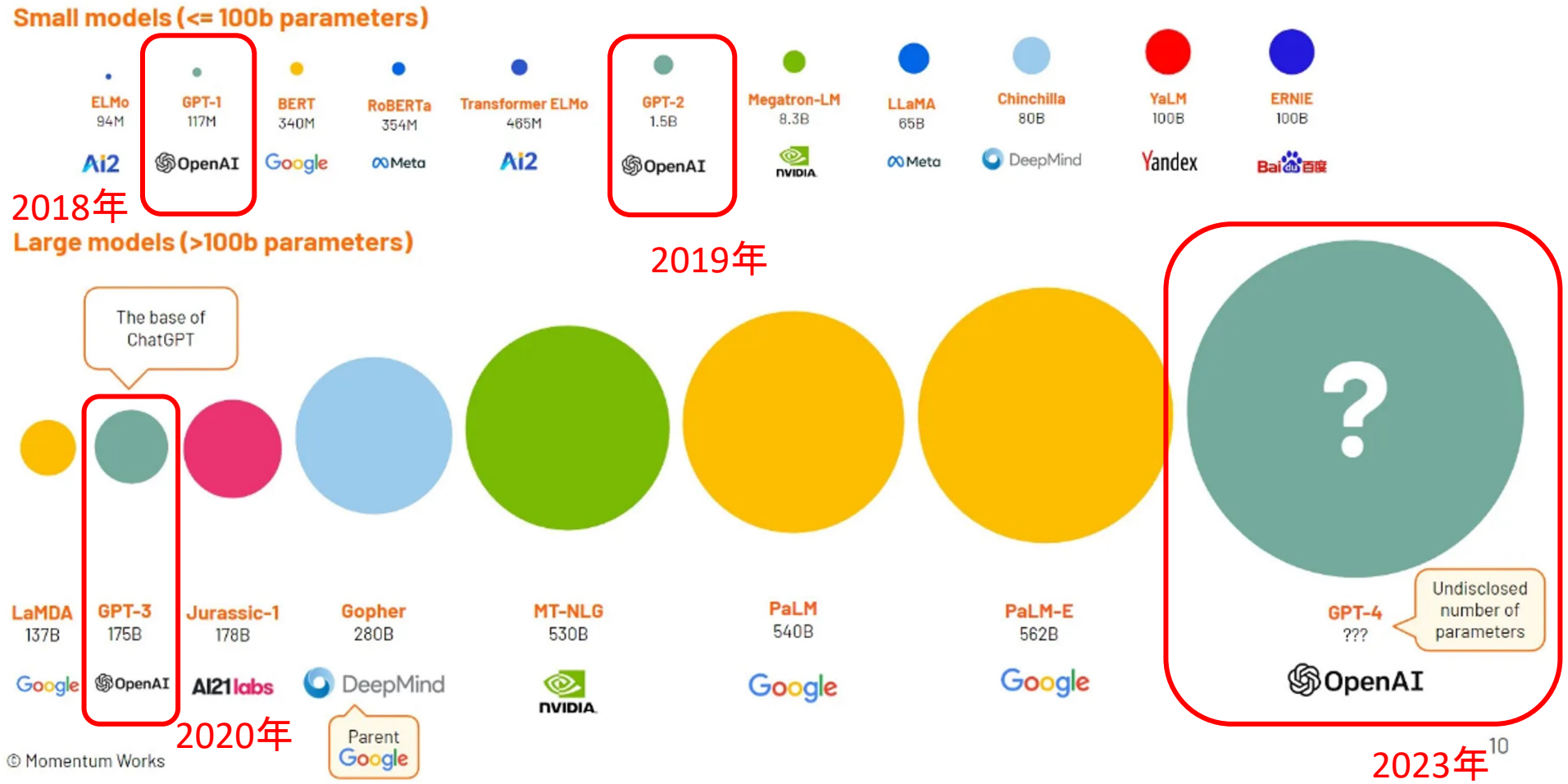
・産総研のABCI：
960基のA100 GPU → **6,128基**のH200 GPU
*2025年1月アップグレード
・Softbank：**6,000基**のGPU
・さくらインターネット：**2,000基**のH100GPU

海外

1企業で数十万~百万基のH100 GPUを保有
(以下、24年単年の購入数)
・Google：**169,000基**
・Amazon：**196,000基**
・Meta：**224,000基**
・Microsoft：**485,000基**

出典：
https://www.datacenterdynamics.com/en/news/microsoft-bought-twice-as-many-nvidia-hopper-gpus-as-other-big-tech-companies-report/

Transformerを使った言語モデルの巨大化

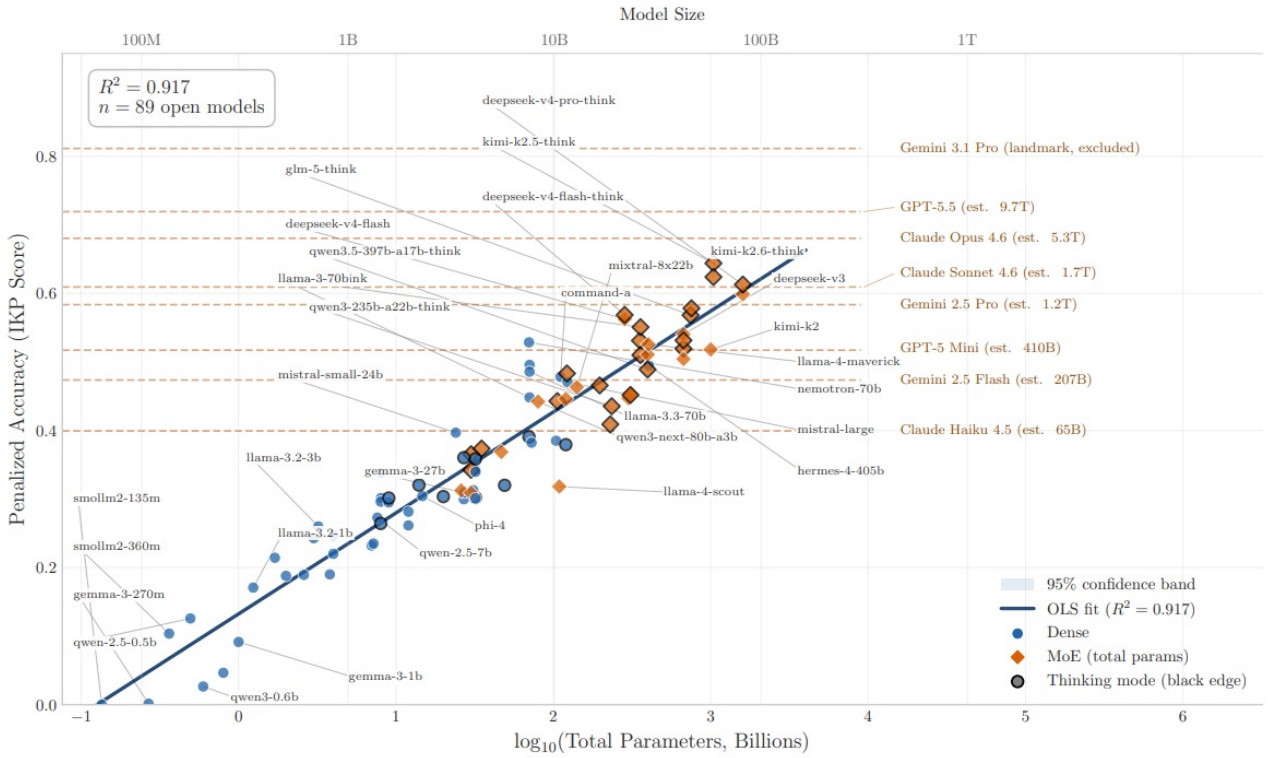


基本的にはいずれも2017年に発明されたTransformerと呼ばれる構造を利用.
GPT-3登場以降，米国企業を中心に複数の研究機関が独自の大規模言語モデルを開発.

[3] Momentum Works 2023 “The future by ChatGPT” より引用し,一部改変

最近のLLMのモデルサイズの推定

Model	Vendor	Accuracy	Est. Size	90% PI
GPT-5.5	OpenAI	71.9%	∼9.7T	[3.2–28.7T]
Claude Opus 4.6	Anthropic	68.0%	∼5.3T	[1.8–15.6T]
GPT-5 Pro	OpenAI	66.5%	∼4.1T	[1.4–12.2T]
GPT-5	OpenAI	66.4%	∼4.1T	[1.4–12.1T]
Claude Opus 4.7	Anthropic	66.4%	∼4.0T	[1.4–12.0T]
o1	OpenAI	65.4%	∼3.5T	[1.2–10.3T]
Claude Opus 4.5	Anthropic	65.2%	∼3.4T	[1.1–10.0T]
Claude Opus 4.1	Anthropic	64.9%	∼3.2T	[1.1–9.5T]
Grok-4	xAI	64.8%	∼3.2T	[1.1–9.4T]
o3	OpenAI	64.4%	∼3.0T	[1.0–8.9T]
GPT-5.4 Pro	OpenAI	62.5%	∼2.2T	[736B–6.5T]
GPT-4.1	OpenAI	62.3%	∼2.2T	[719B–6.4T]
Grok-3	xAI	62.3%	∼2.1T	[715B–6.3T]
Claude Sonnet 4.6	Anthropic	60.9%	∼1.7T	[579B–5.1T]



Incompressible Knowledge Probes: Estimating Black-Box LLM Parameter Counts via Factual Capacity
https://arxiv.org/abs/2604.24827

LLMのテスト時推論能力の向上

（*モデルの学習が終わった後，推論時にどうやって性能をブーストさせるか）

ゼロショットでLLMの多段階推論能力を引き出す

Large Language Models are Zero-Shot Reasoners

Accepted at NeurIPS 2022

大規模言語モデル (LLM)

- 驚異的な汎用性
  - Few-shot in-context learning
  - Prompt-based zero-shot learning

Few-shot

```
Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese =>
```

Zero-shot

```
1  Translate English to French:

2  cheese =>
```

Cited from “Language Models are Few-Shot Learners”

多段階の推論が必要なタスク

- （2022年当時）LLMには解くことが困難であるように思われた
  - 数学，常識推論，シンボリック推論

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

多段階の推論が必要なタスク

- 『思考の連鎖』（Chain of Thought：CoT)
  - 解決の糸口：答えを出す前に思考の過程を出力させる
  - オリジナルの手法はFew-shot (*)
  - ではZero-shotは？

    ⇒ 試されていない

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

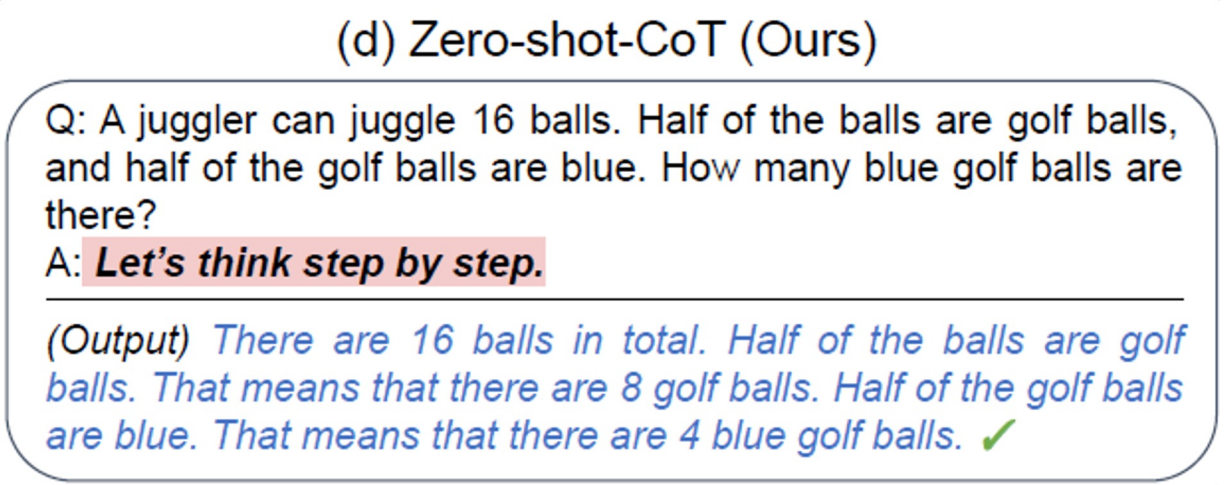
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

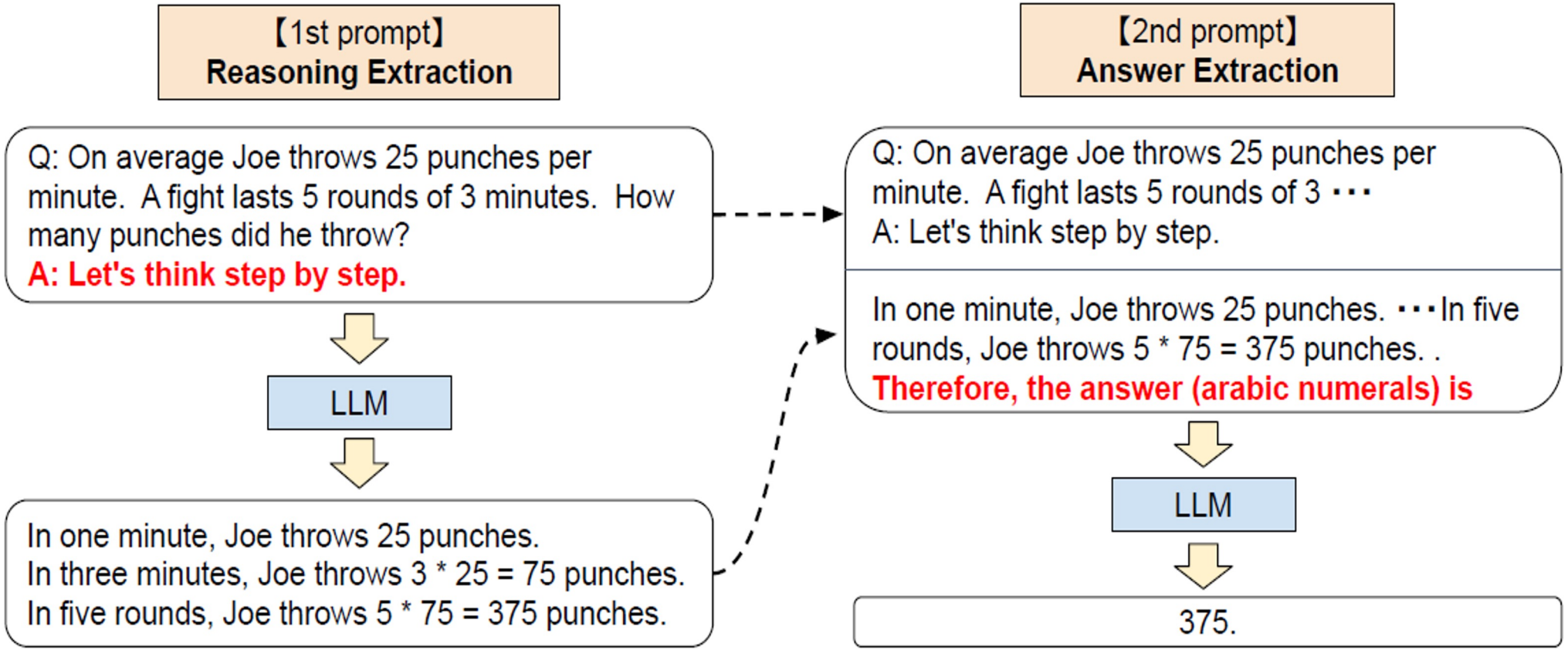
(*) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

多段階の推論が必要なタスク

- 『思考の連鎖』（Chain of Thought：CoT)
  - 本研究の提案：Zero-shot-CoT



提案手法の全体像


【1st prompt】
Reasoning Extraction
Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.
LLM
In one minute, Joe throws 25 punches.
In three minutes, Joe throws 3 * 25 = 75 punches.
In five rounds, Joe throws 5 * 75 = 375 punches.
【2nd prompt】
Answer Extraction
Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ・・・
A: Let's think step by step.
In one minute, Joe throws 25 punches. ・・・In five rounds, Joe throws 5 * 75 = 375 punches. .
Therefore, the answer (arabic numerals) is
LLM
375.

実験結果

- Zero-shot < Zero-shot-CoT < Few-shot-CoT

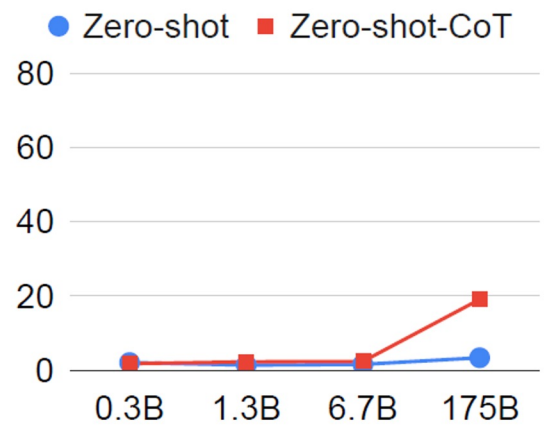
	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	**92.8**	**51.5**
Finetuned GPT-3 175B (Wei et al., 2022b)	-	33
Finetuned GPT-3 175B + verifier (Wei et al., 2022b)	-	55

* Instruct-GPT3モデルによる実験

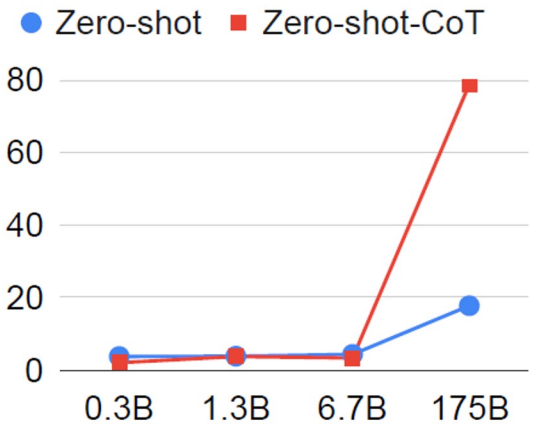
Zero-shot-CoT	Few-shot-CoT
✅幅広いタスクに適用可能な汎用性の高いプロンプト.	▲ タスクに応じた少数サンプル事例を作成または用意する必要あり.

実験結果

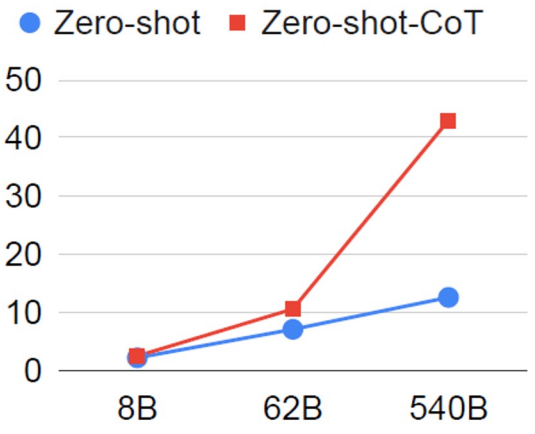
- Zero-shot-CoT能力はモデルサイズが大きくなると発現する.



(a) MultiArith on Original GPT-3



(b) MultiArith on Instruct GPT-3



(c) GMS8K on PaLM

実験結果

- プロンプトの頑健性調査

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

* Instruct-GPT3モデルによる実験

後続研究に対するインパクト

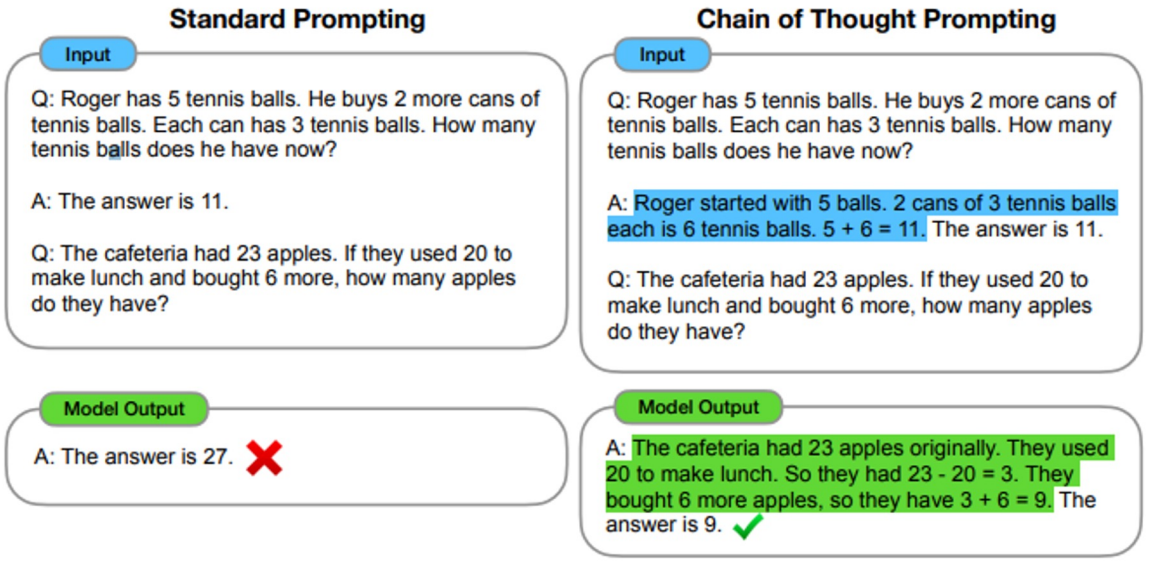
- **2026年5月現在，8900件以上引用されている**(Google Scholarによる).
  - イントロや関連研究のCoT紹介（Few-shot-CoTとセット）
  - 実験のベースライン手法
  - 提案手法のパイプラインの一部
  - サーベイ論文

- **LLMの多段階推論の重要性が認識された**
  - 最近のLLMの多くは，事後学習によって多段階推論をデフォルトで行えるようになっている.
    - いわゆる**推論モデル（Reasoning Model）**
      - OpenAI-o1，Deepseek-R1が先駆
    - 現在では各社がReasoning Modelを提供
      - OpenAI, Google, Anthropic, xAI, …

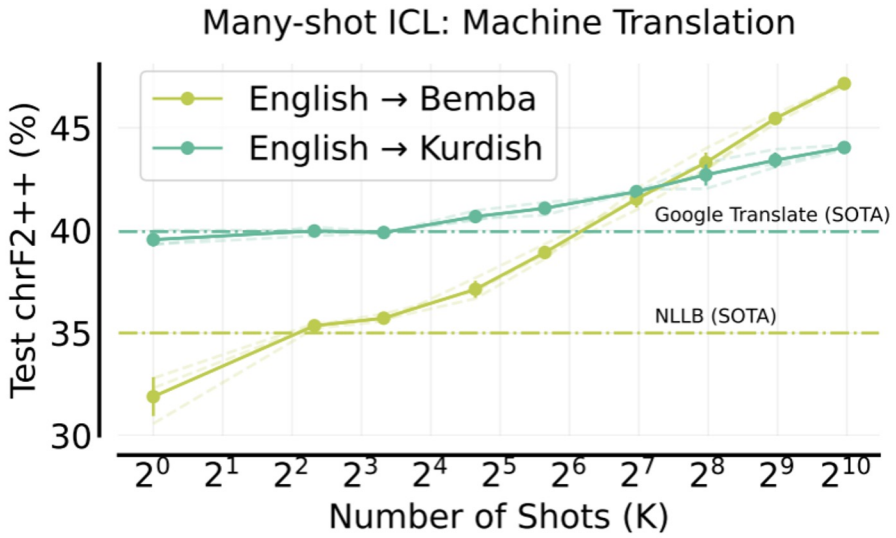
- **プロンプトエンジニアリングの先駆けとなった**

推論時に計算量をスケールさせる方法の代表例

Chain-of-Thought Prompting



Many-Shot ICL



Promptingにより推論時のトークン数を増やすことで推論時の計算量をスケールさせる試み

[左図] “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”
[右図] “Many-Shot In-Context Learning”

推論時に計算量をスケールさせる方法の代表例

① Parallel search　例：Best-of-N, Self-Consistency

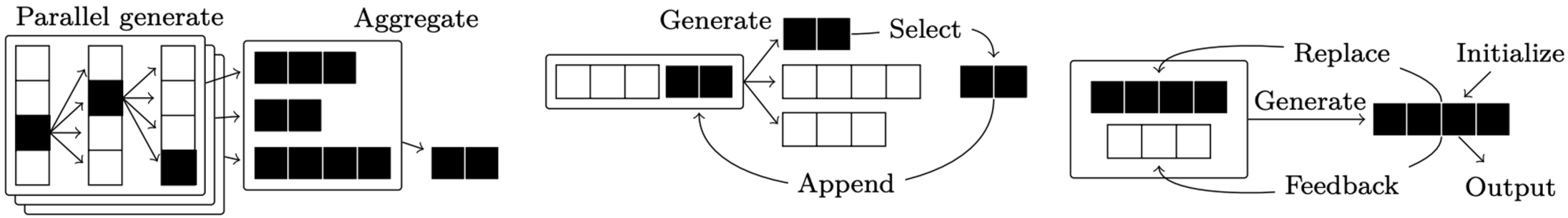
- 並列に複数の候補を生成してスコアリングや多数決などにより生成物を選ぶ

② Step level search　例：Process Reward Model (PRM)

- Stepレベルに評価を行なって生成物を選ぶ

③ Refinement　例：Self-Refine

- 外部/内部のフィードバック結果を用いて，反復的に生成結果を更新する



(a) Parallel search　(b) Heuristic step-level search　(c) Refinement

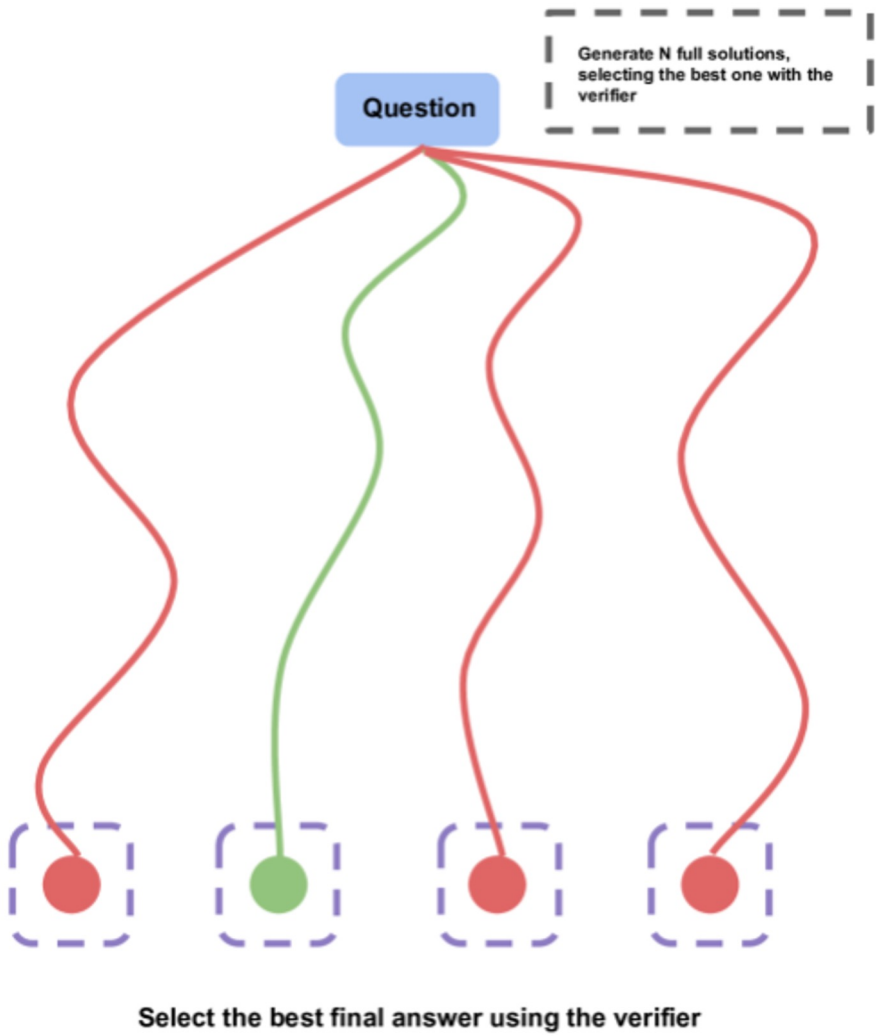
[54] “From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models”. より引用

① Parallel Searchの方法の例｜Best-of-N

Best-of-N

- N個の回答を出してスコアが一番高いものを選択する

- スコア関数は任意(タスクによって使い分け)
  - 例：LLMのスコアを使う
  - 例：学習した評価器を使う
  - 例：BLUEなど特定の指標を使う

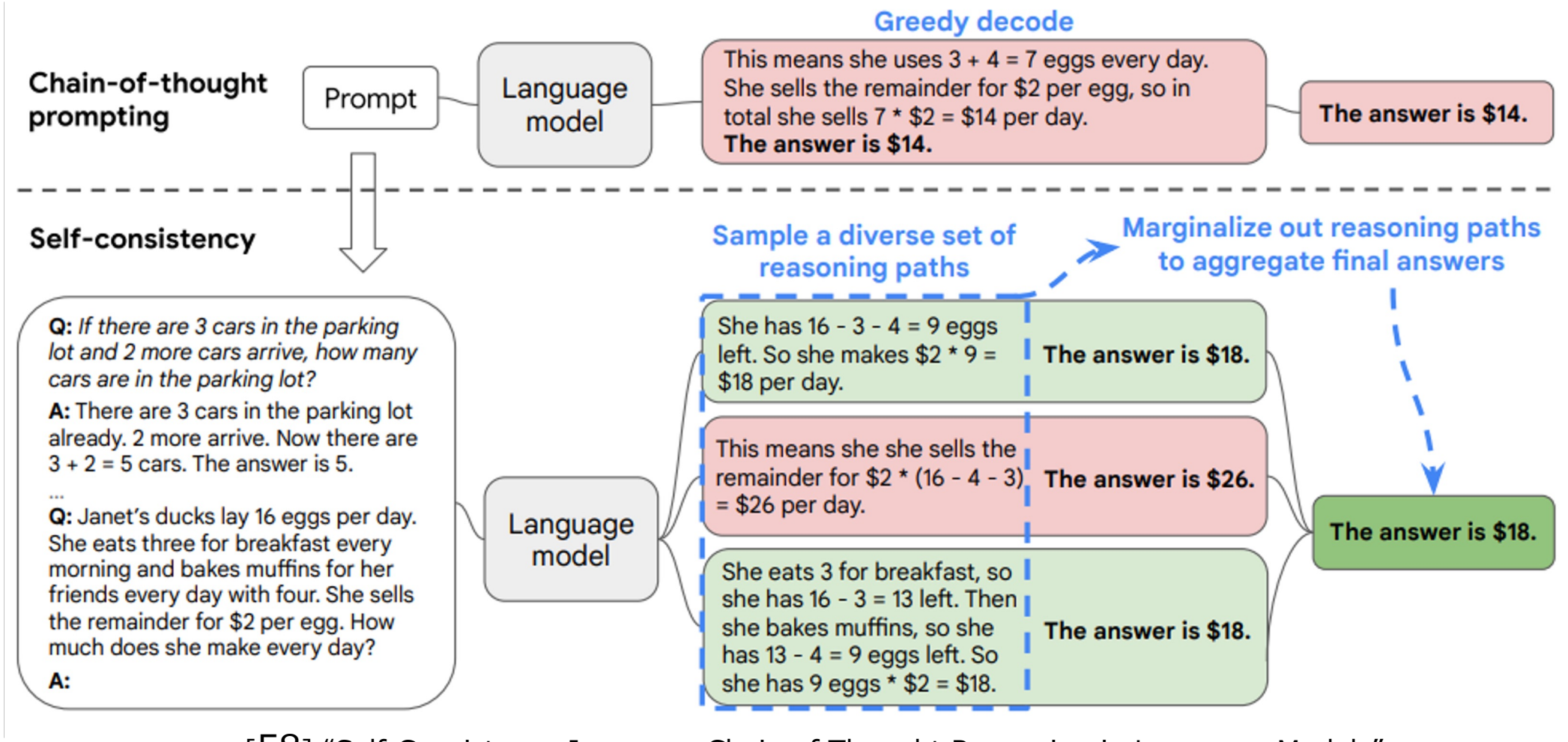


“Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, 2024[56]

① Parallel Searchの方法の例｜Self-Consistency

Self-Consistency

- LMに複数の推論を行わせて，Majority Voting（多数決）

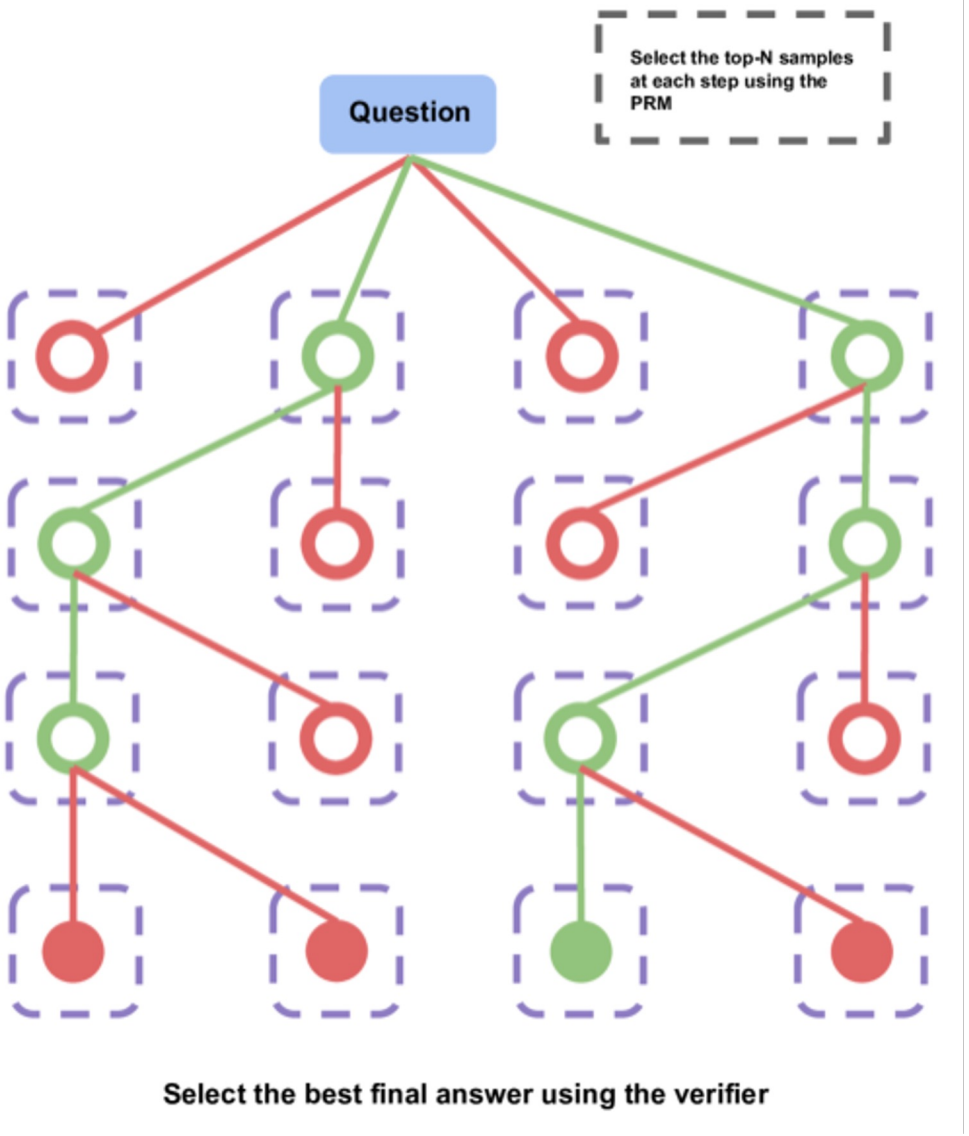


[58] “Self-Consistency Improves Chain of Thought Reasoning in Language Models”

② Step Level Searchの方法の例｜Beam-Search (Step Level)

Beam-Search(Step Level)

- （Token LevelのBeam Searchとは異なり）文章や段落単位でサンプリングと評価を行う
- 評価にはPRM(Process Reward Model)を用いて途中結果の評価と選択を行う
- Top-N Samplingによって途中結果を選択

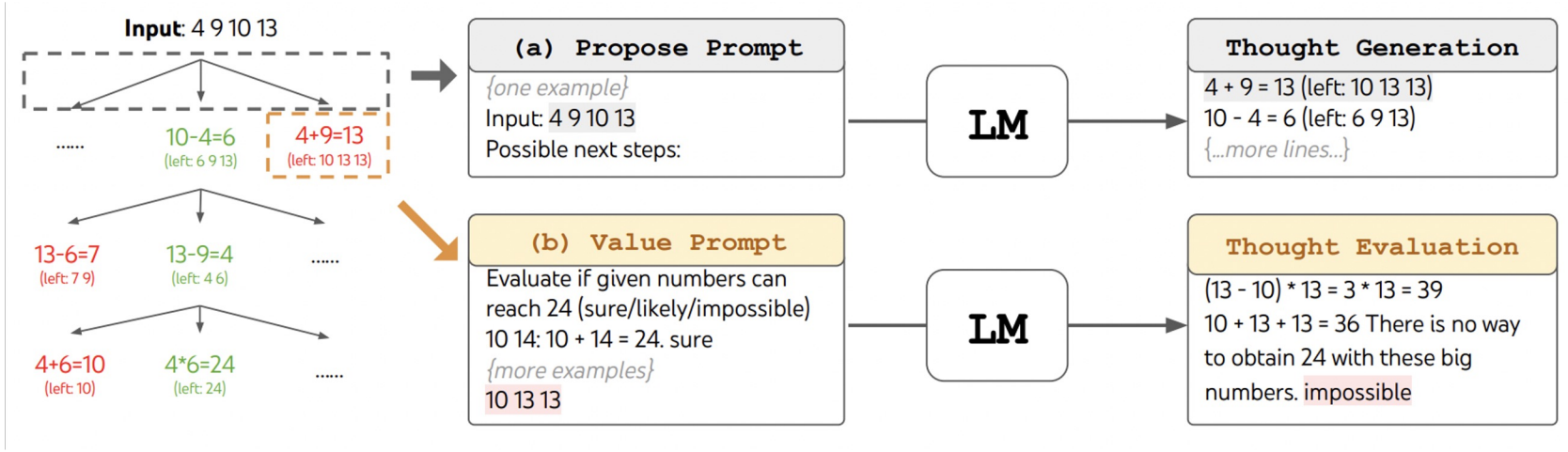


“Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”, 2024[56]

② Step Level Searchの方法の例｜Tree-of-Thought

Tree-of-Thought

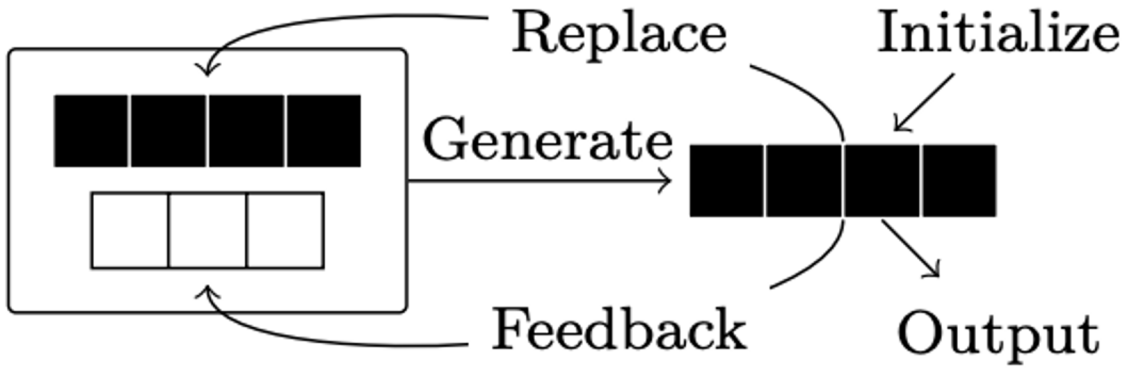
- 複数の思考列を一気通貫で出力して評価するSCとは違い，ToTは途中で分岐させる（木探索する）
  - ノードの評価もLMで行う
- Game of 24での例と結果
  - タスク：与えられた4つの数字を四則演算して24を作る
- 戦略的思考が必要なタスクで性能が大幅改善



Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC (k=100)	9.0%
ToT (ours) (b=1)	45%
ToT (ours) (b=5)	**74%**

[60] “Tree of Thoughts: Deliberate Problem Solving with Large Language Models”

③ Refinement



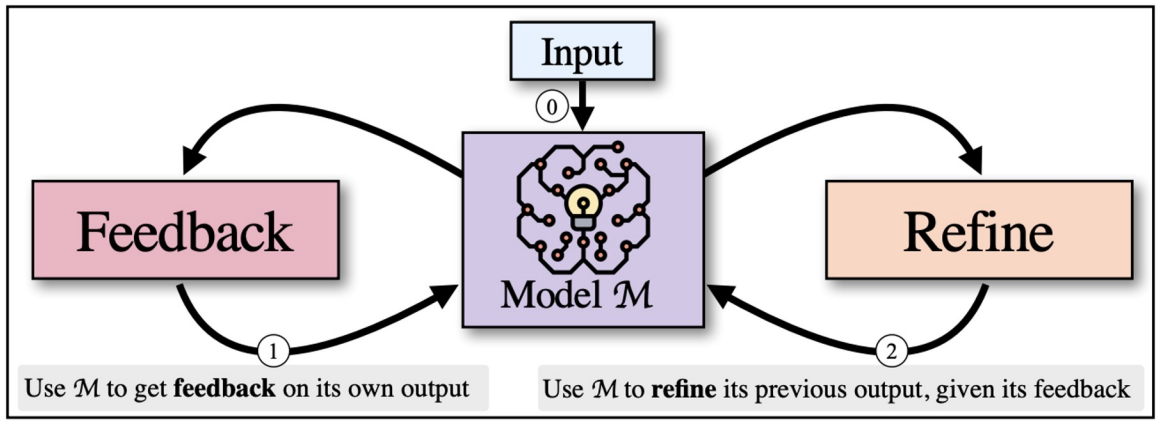
(c) Refinement

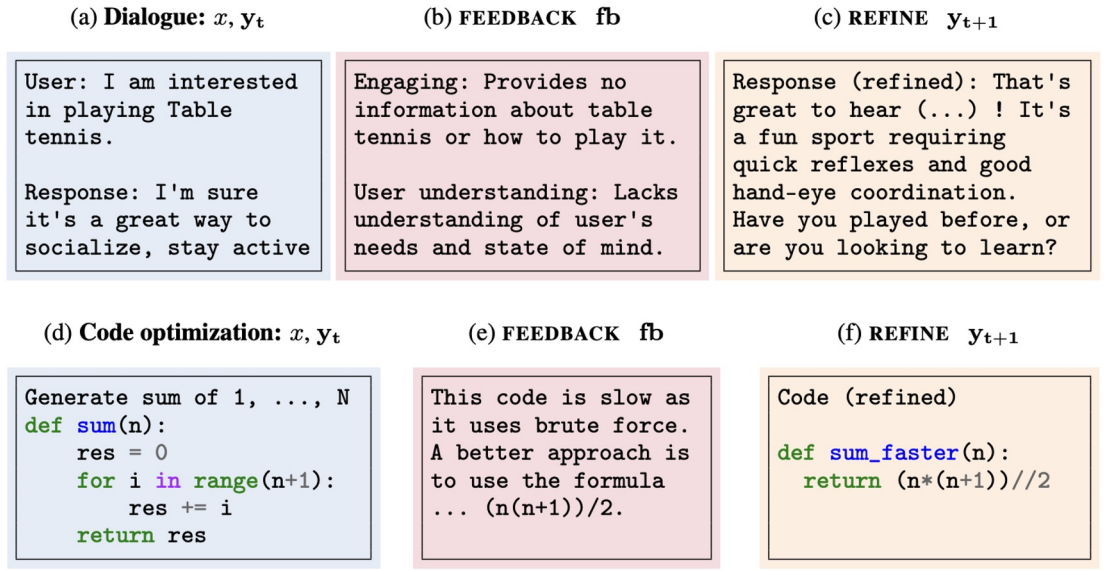
- 一度生成した結果や，その結果に対するフィードバックをもとに再度生成

[54] “From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models”. より引用

③ Refinementの方法の例 | Self-Refine（Day 2のおさらい）

自分自身を使ってフィードバックを生成して出力を改善

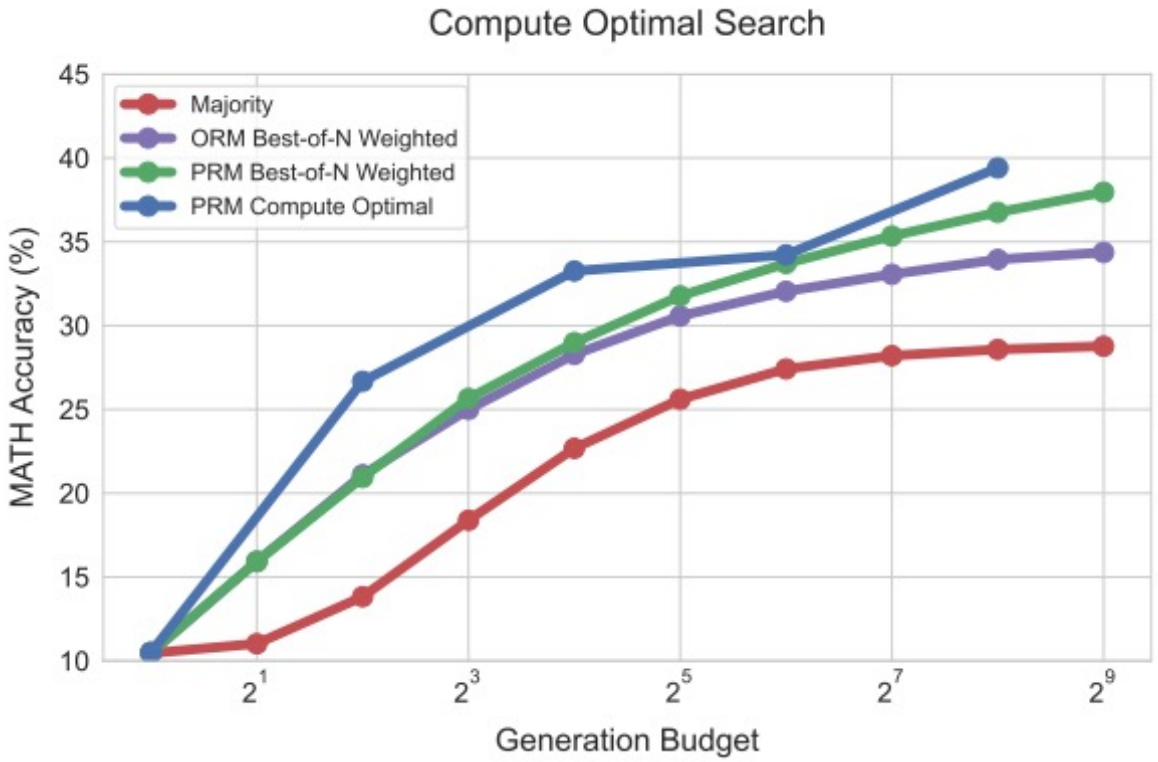


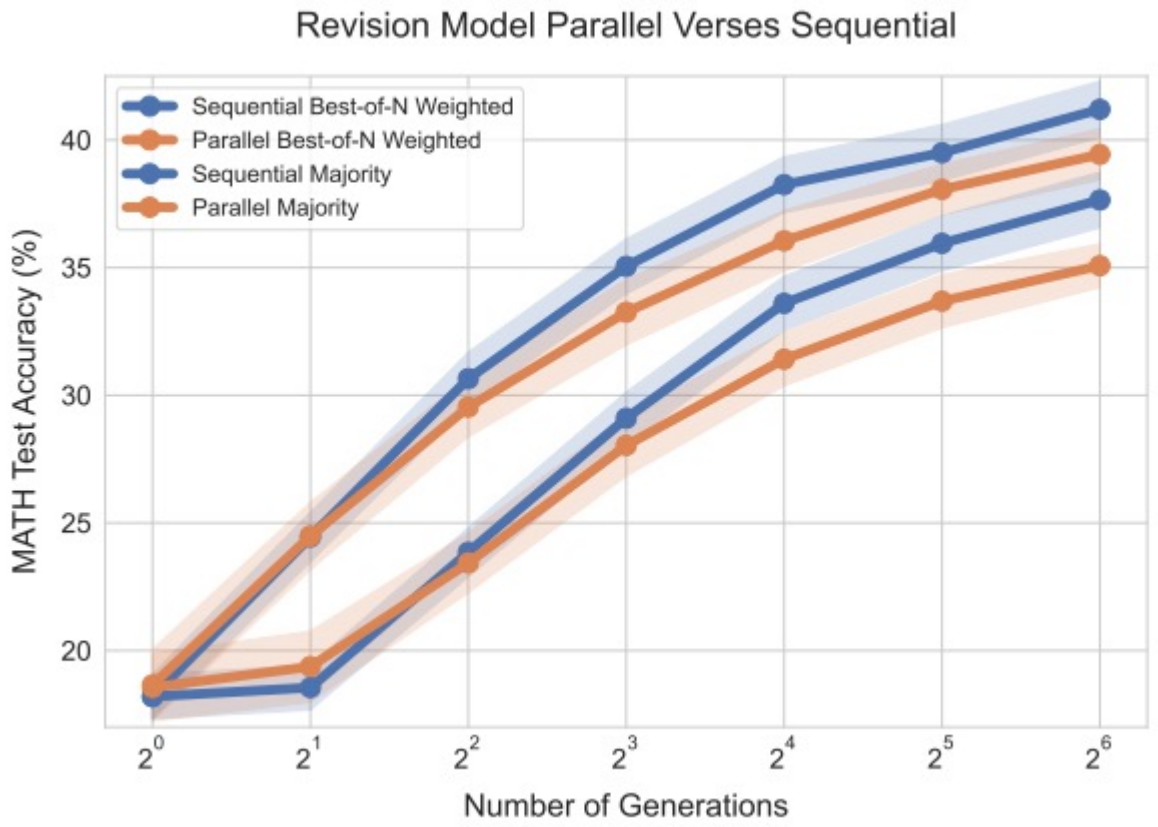


“SELF-REFINE: Iterative Refinement with Self-Feedback”[61]

推論手法の性能比較

- Sequential Best-of-N (refinement) > Parallel Best-of-N > Majority voting





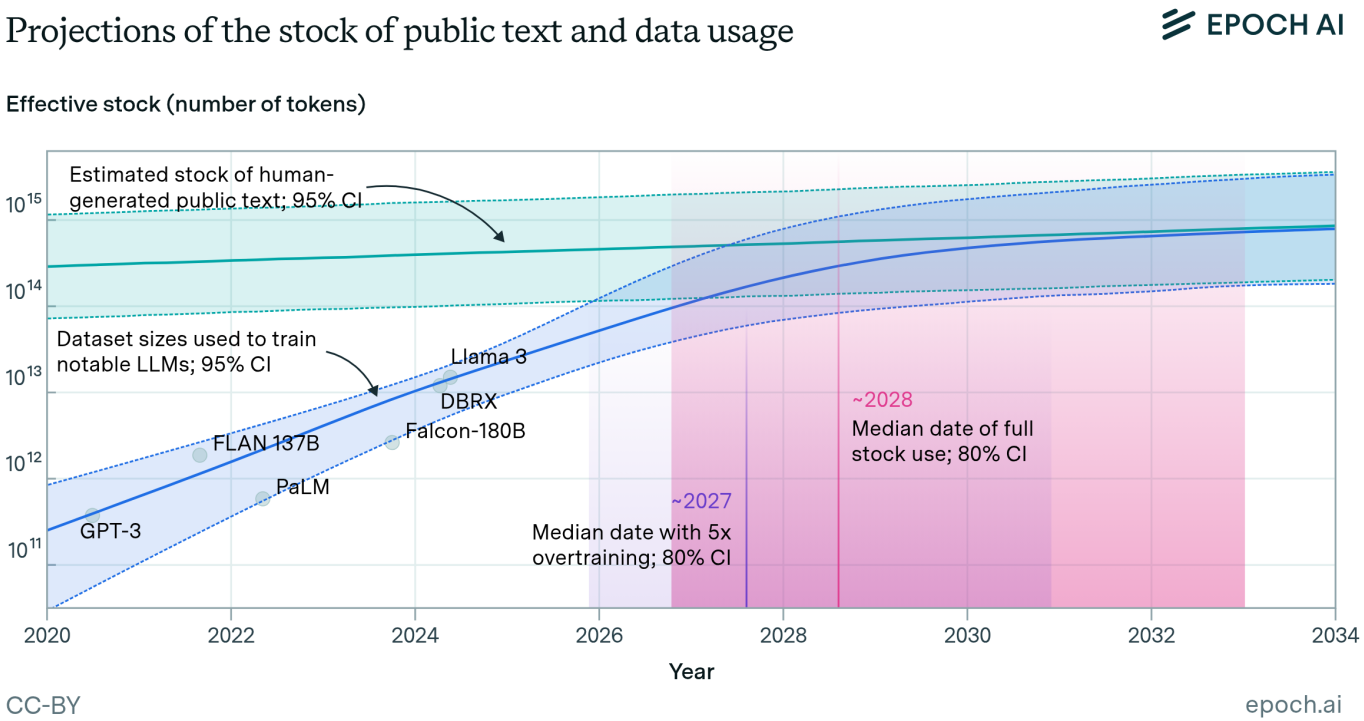
Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters より

Q&A

強化学習による
推論能力の更なる進化

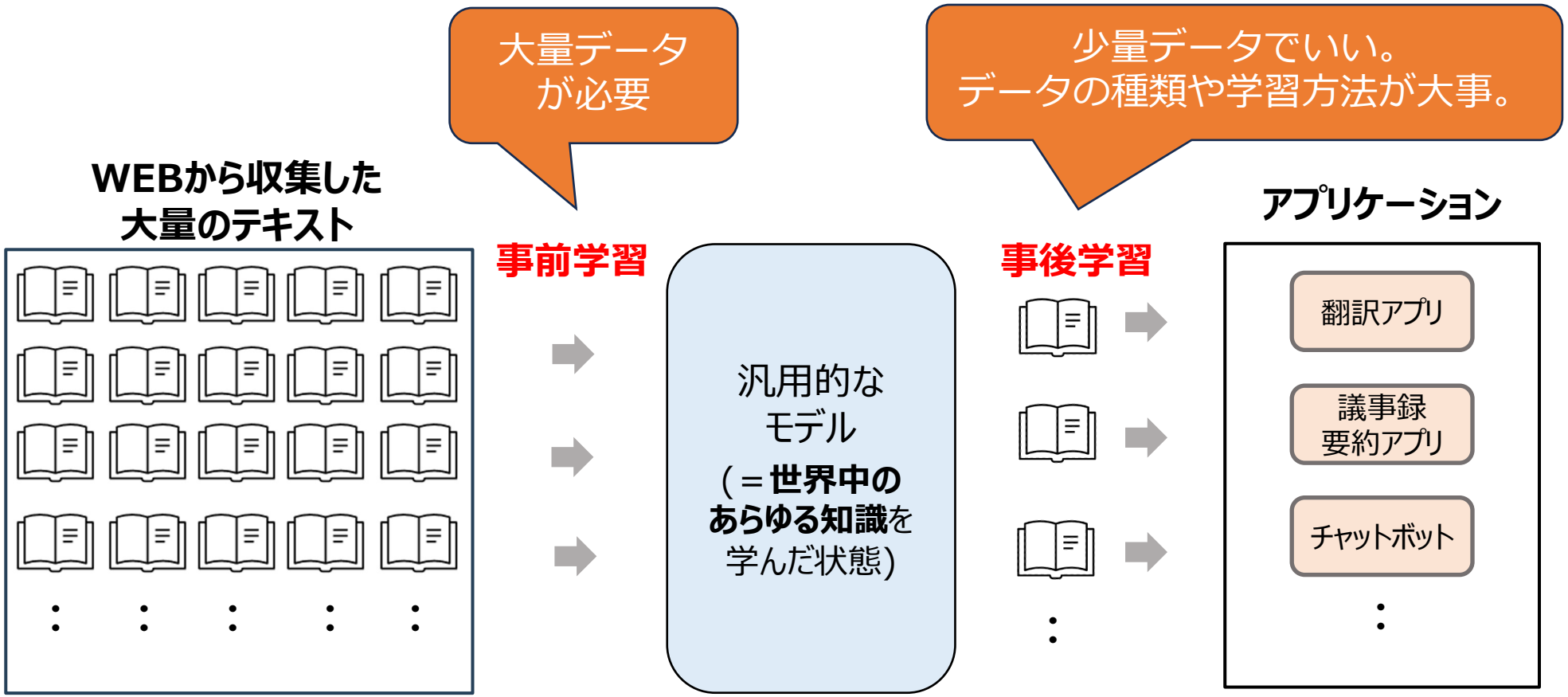
LLMの学習データは枯渇した？

- 2026年から2032年の間にLLMが高品質な公開テキストを学習し尽くすという推計（2022年時点）



Will we run out of data? Limits of LLM scaling based on human-generated data
https://arxiv.org/abs/2211.04325
https://epoch.ai/publications/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

我々には事後学習がある



汎用的なモデルが1つあれば、言語に関する様々な機能を開発できる

事後学習（再掲）

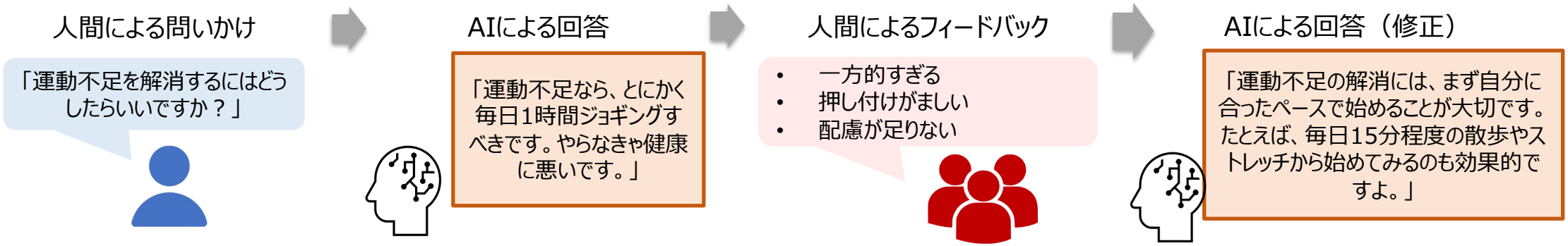
事後学習のやり方①：**教師ありファインチューニング（SFT）**

- QA形式(質問と答え)のペアデータやチャットデータで学習する
- 事前学習と同じく、次の単語をひたすら予測する学習を行う

質問 (Question)	答え (Answer)
健康維持のための3つのコツを教えてください。	1. バランスのとれた食事を摂り、野菜や果物を十分に摂ること。 2. 定期的に運動をして体の活力を保つこと。 3. 睡眠時間を十分にとり、規則正しい睡眠をとること。

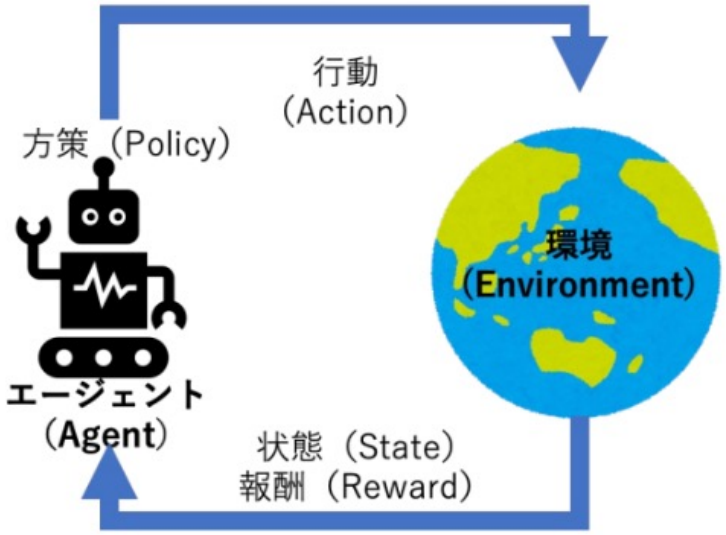
事後学習のやり方②：**強化学習（RL）**

- AIが出力した文章に対して、人間や環境が良し悪しを評価してフィードバックすることで学習する
- **SFTと異なり、答え（Answer）を準備する必要がない．LLMの出力文に対する評価の仕組みだけ必要．**

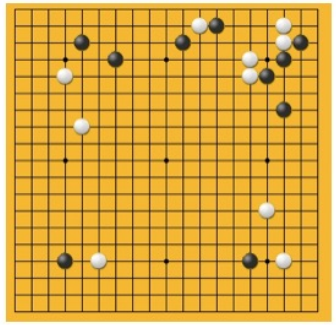


そもそも強化学習とは？

- 問題設定
  - エージェント(行動主体)は 環境の状態に基づいて，逐次的に行動を決定する
- 強化学習の目的
  - 行動の結果得られる報酬を利用し，その環境で最も良い行動ルール(最適方策)を学習したい

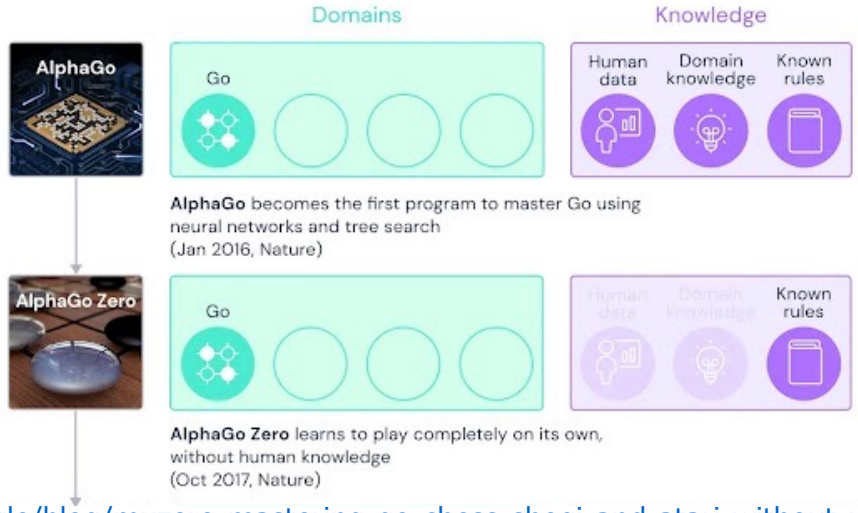


囲碁の場合



- **状態: 盤上の碁の位置**
- **行動: 自分の石を一つ置く**
- **報酬: 勝つか負けるか**

代表的な応用例: AplhaGo
人間の対局データなしに経験から学習して人間超えを達成



https://deepmind.google/blog/muzero-mastering-go-chess-shogi-and-atari-without-rules/

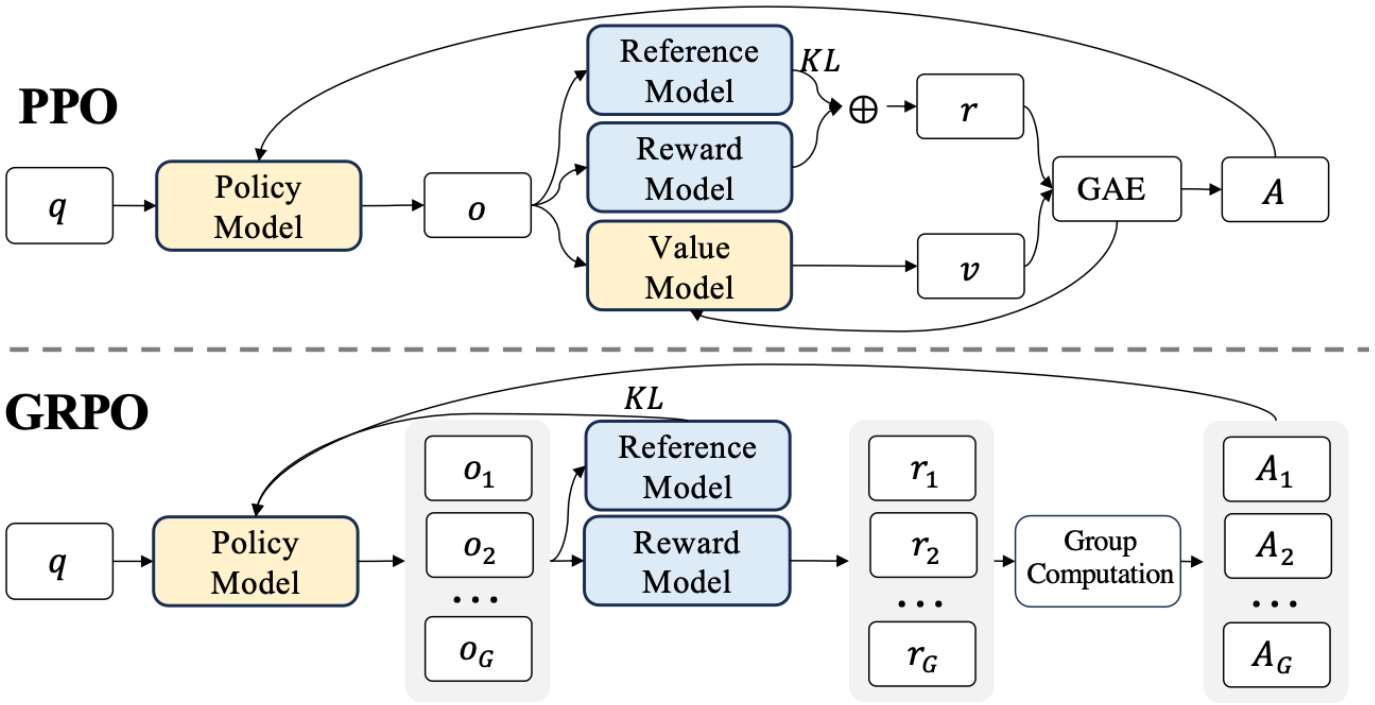
DeepSeek-R1

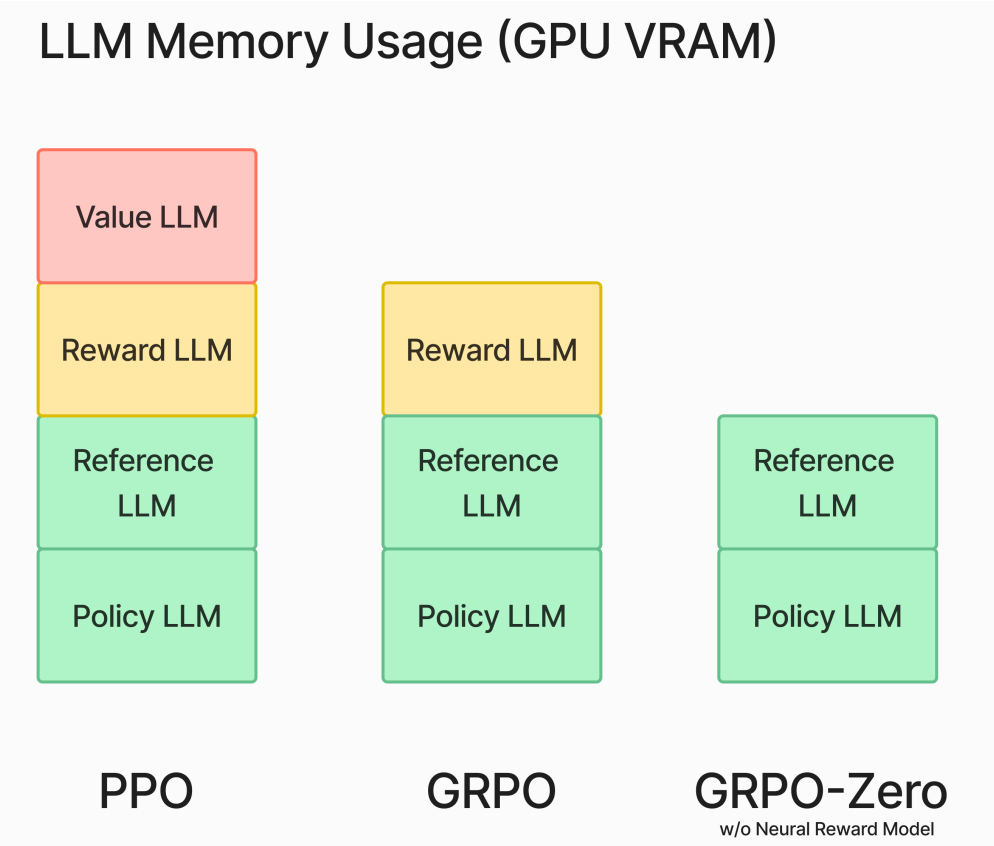
- **ルールベースの報酬による強化学習（RLVR）**
  - **数学やプログラミングのタスク**で最終解答が正解かどうかを報酬とすることで、高い推論能力を引き出すことに成功（つまり，推論過程がどのようなものだったかは報酬を与える上で考慮しない）
    - OpenAI o1の成功後、DeepSeek R1はオープンソースで初めてo1に迫る性能を叩き出した
    - GRPOと呼ばれる独自の強化学習アルゴリズムを提案し、性能向上に寄与した.
  - **それまでは人間による評価を報酬とするタスクでの強化学習（RLHF）が一般的だった**
  - 基盤モデルの時代になって、高度なタスクのstep-by-step推論が可能となり、最終出力で正確な評価可能なシンプルな強化学習ができる土壌が生まれた。これでうまくいくことを実際に示したことが大きな貢献

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning https://arxiv.org/pdf/2501.12948

GRPO

- PPOを推論能力の向上に特化させた強化学習手法
- アドバンテージ（A）をエピソード報酬（r）から直接算出することにより状態価値V(s)の関数近似を不要とした





DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models https://www.oxen.ai/blog/why-grpo-is-important-and-how-it-works

LLMのアハ体験（Aha moment）

- RLVRによって，自分の試行錯誤の結果が間違っていた時に”Wait, wait. That’s an aha moment”と言って正しい解き方に気づくことがある

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let’s start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That’s an aha moment I can flag here.

Let’s reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let’s square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

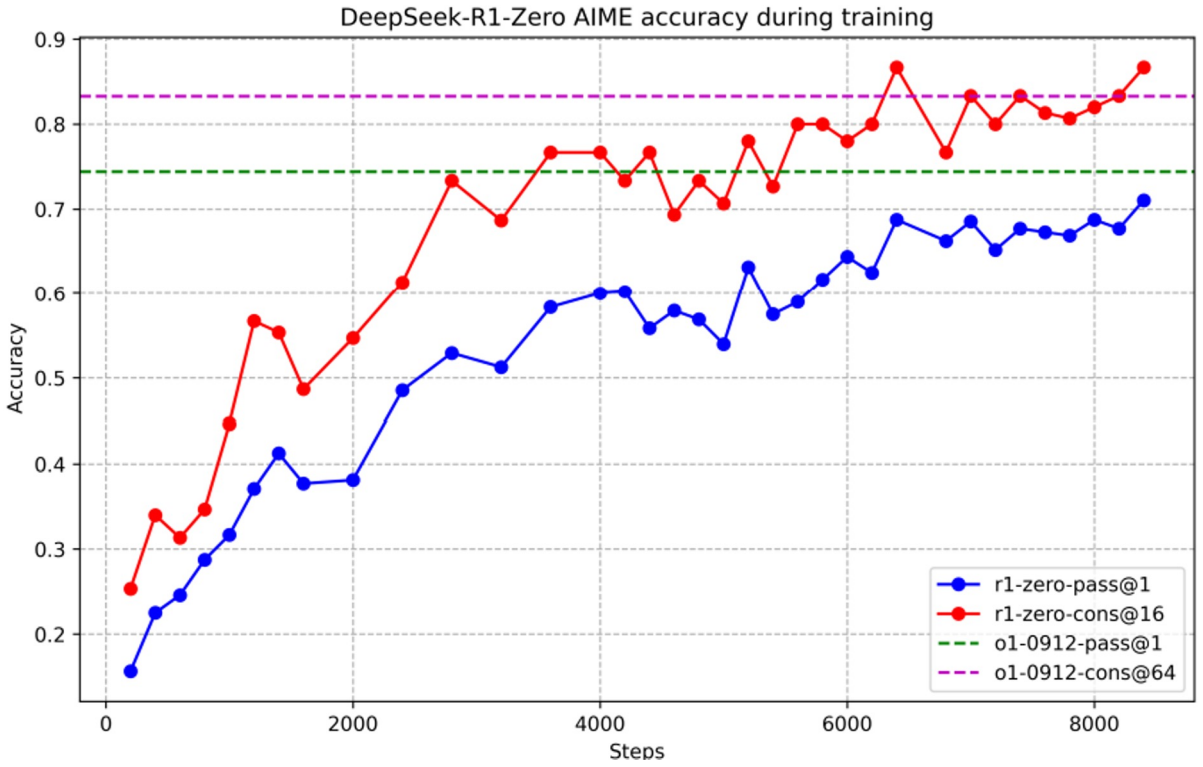
Next, I could square both sides again, treating the equation: $\cdots$

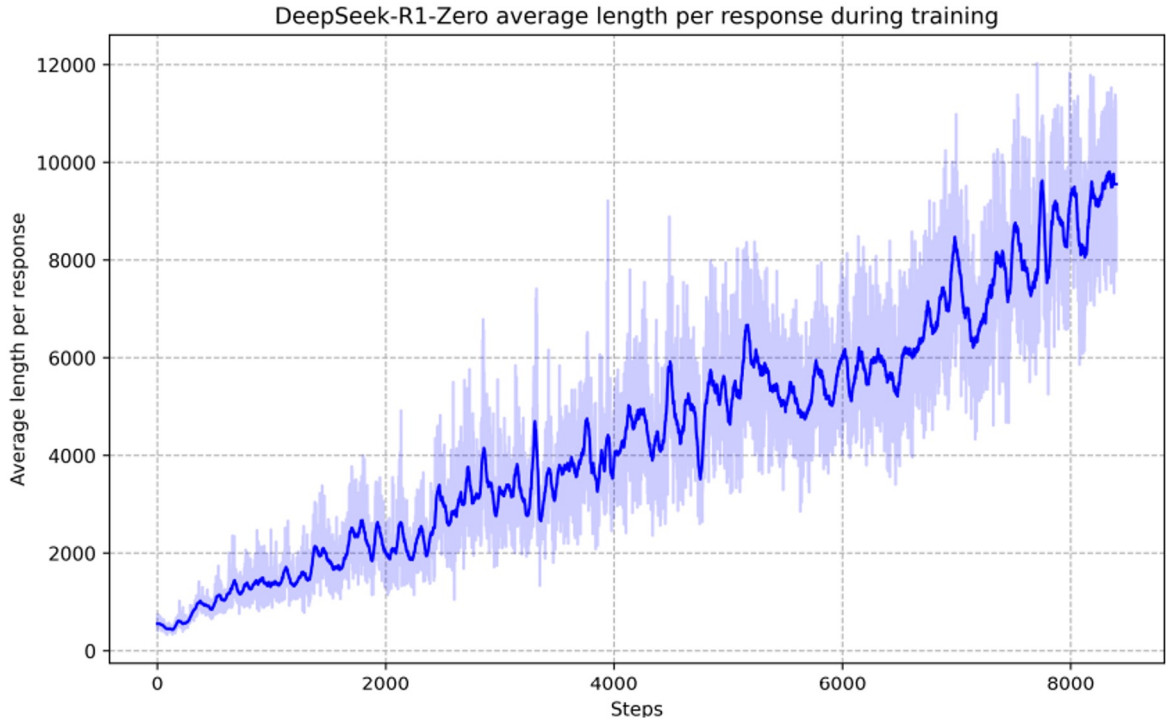
$\cdots$

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-R1

- RLの過程で、長考（より多くのトークン長で思考）するほど良い答えにたどりつくようになる

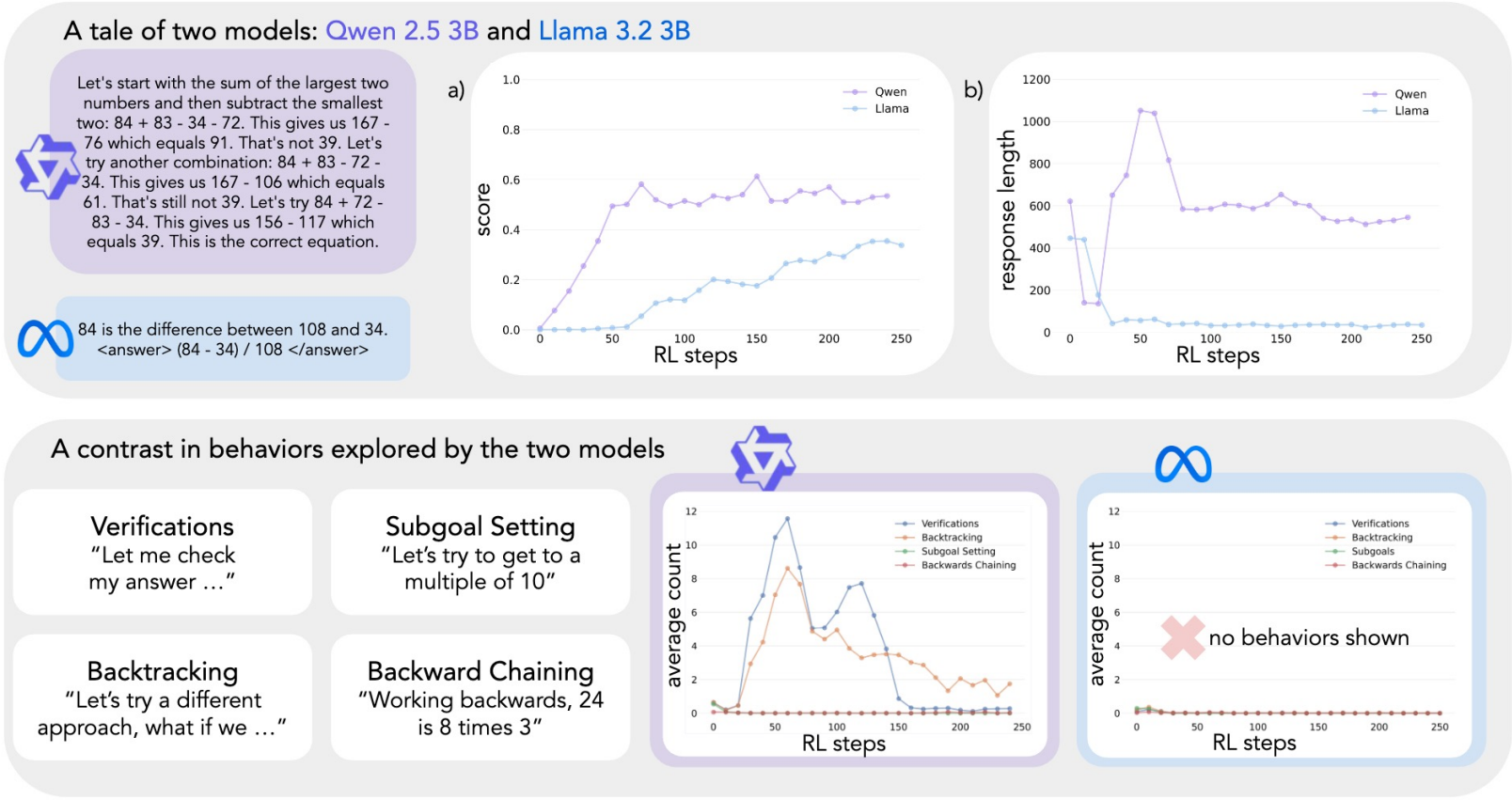




DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning https://arxiv.org/pdf/2501.12948

RLVRによっていくつかの認知行動が誘発される

- 強化学習によって，Verification, Backtrackの行動が増加し、それに伴ってスコアも向上する
- 一方でRL前のモデルでこの二つの行動が全く見られない場合はRLしても性能が向上しない



Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs

推論モデルの軌跡をグラフとして捉えた分析

多段階推論能力の推論グラフによる理解

Topology of Reasoning: Understanding Large Reasoning Models through Reasoning Graph Properties
Accepted at NeurIPS 2025

思考の軌跡でLLMは何をしているのか

推論モデルは強化学習で最終的な答えが合っているかどうかの報酬で学習するため，途中の軌跡では人間が想定していなかったような挙動を見せる．

例 自分の生成した軌跡を反省し始める Aha Moment，複数の言語が混同する[Li et al.]，同じトークンを何度も繰り返す[Yang et al.]

Aha Moment

待ってください，考え直します．

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

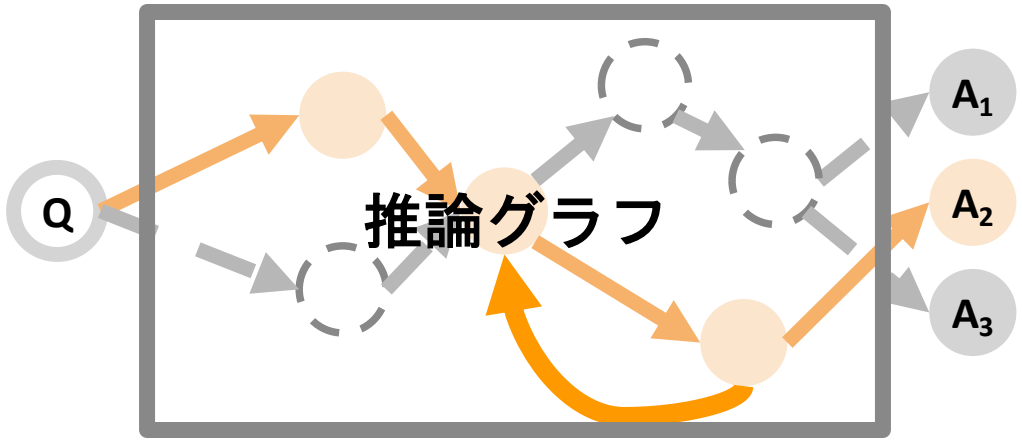
Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ···
$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be ···
We started with the equation:
$\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: ···
...

出典) DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
Understanding Aha Moments: from External Observations to Internal Mechanisms
The Impact of Language Mixing on Bilingual LLM Reasoning

推論グラフのトポロジー

本研究の提案

思考の軌跡を**推論グラフ**（有向グラフ）と捉えて，そのトポロジー（グラフ特性）を分析することで，Reasoningモデルの性能向上の秘訣をより深く知る



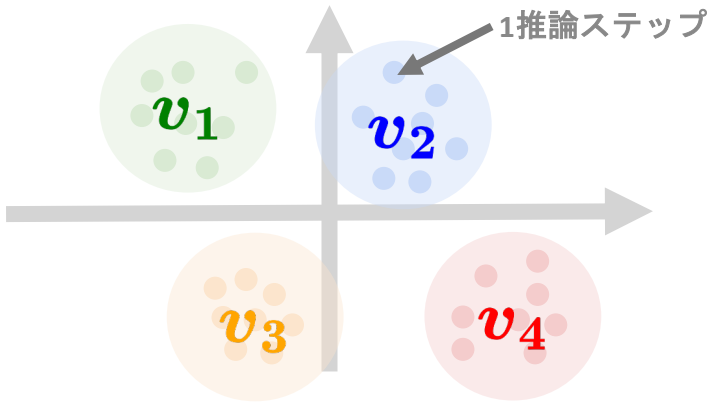
02

推論グラフの抽出方法

LLMの内部表現からグラフを取り出す [Wang et al. ICML'24]

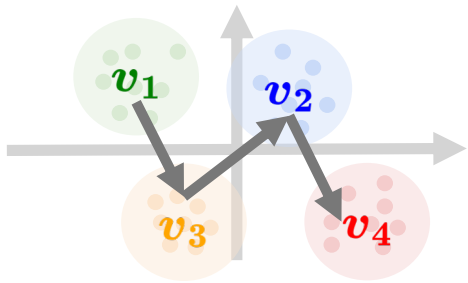
LLMの内部表現をK-meansクラスタリング

数学タスク(e.g., GSM8K) を解くLLMの内部表現を
各推論ステップごとにクラスタリングし
クラスタの中心を推論グラフのノード(v)候補とする



各サンプルごとに推論グラフを構築

質問が入力されて，LLMが答えに辿り着くまでに
訪れるノードを追跡し，推論グラフ(G=(V,E))を作る



$$V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_4\}$$
$$E = \{\boldsymbol{v}_t \rightarrow \boldsymbol{v}_{t+1}\}$$
$$d(\boldsymbol{v}_i, \boldsymbol{v}_j) = \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2$$

出典）Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation

DeepSeek-R1-Distill-Qwen-32Bの57層目の内部表現から得られたノード

Node	Generated Reasoning Step
Multiplicative (Node 4)	First leg: $10 \times 30 = 300$. Now multiply that by 120×1.157625.
Additive (Node 21)	Total chairs: $22 + 66 = 88$. Total sofas: $20 + 64 = 84$.
Wait (Node 15)	Wait, but we need to add this … Wait, no. Let me clarify …

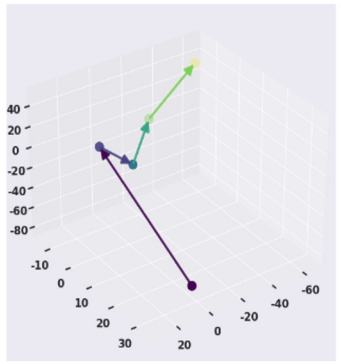
推論グラフの可視化

t-SNEでReasoningモデルとBaseモデルの推論グラフを比較

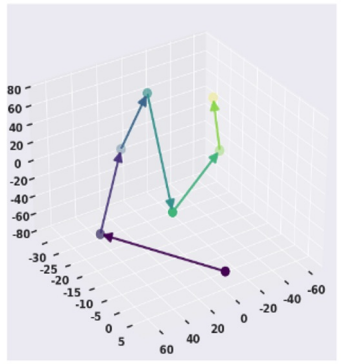
Qwen2.5-32B

AIME2024 26.7

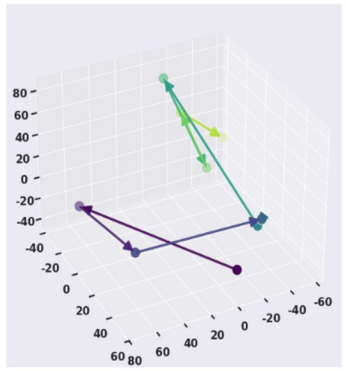
sample #5



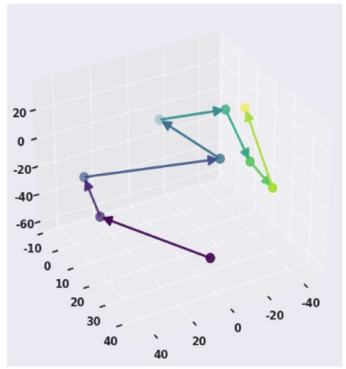
sample #21



sample #45



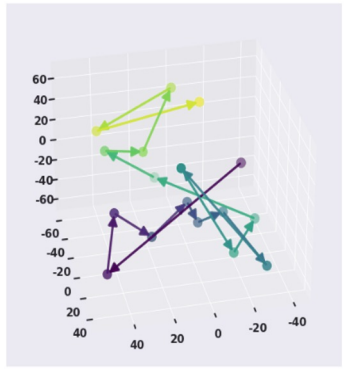
sample #211



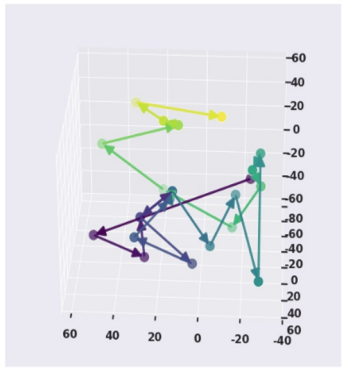
DeepSeek-R1-Qwen-32B (Reasoning LLM)

AIME2024 72.6

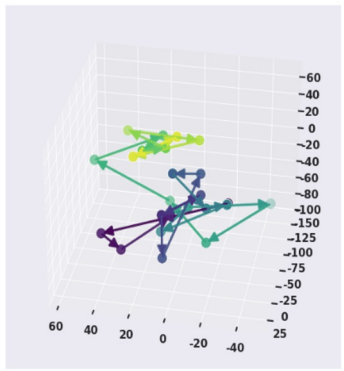
sample #5



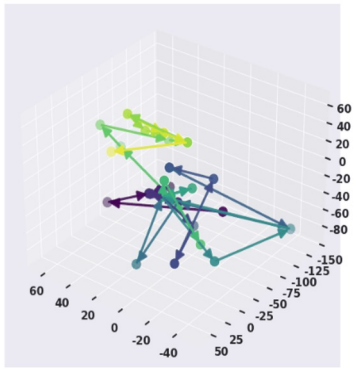
sample #21



sample #45



sample #211

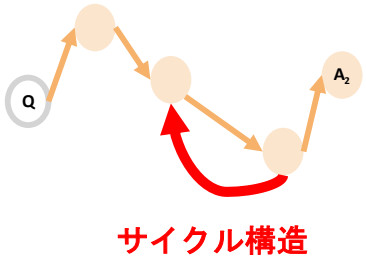


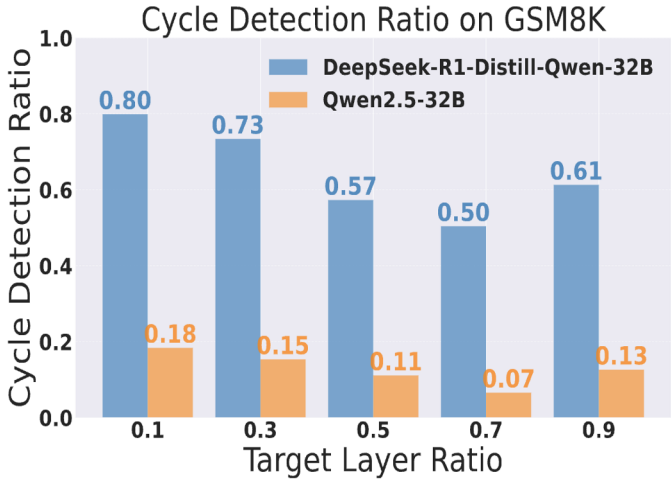
推論グラフの特徴① | サイクル構造

▪ サイクル構造

Reasoningモデルは推論グラフにサイクル構造を含んでいることが多い

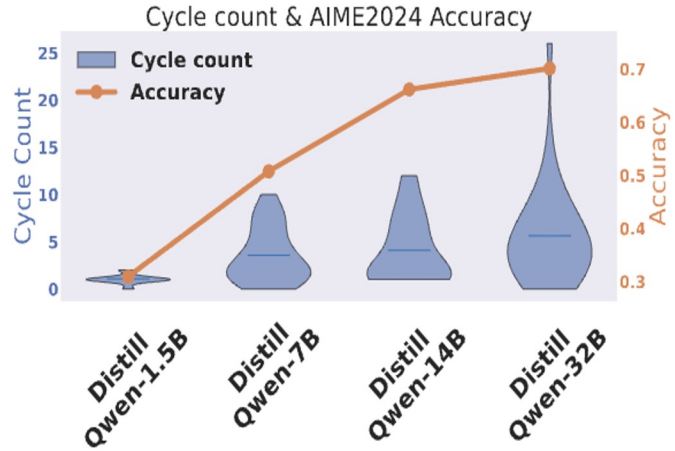
→ Aha momentとも対応





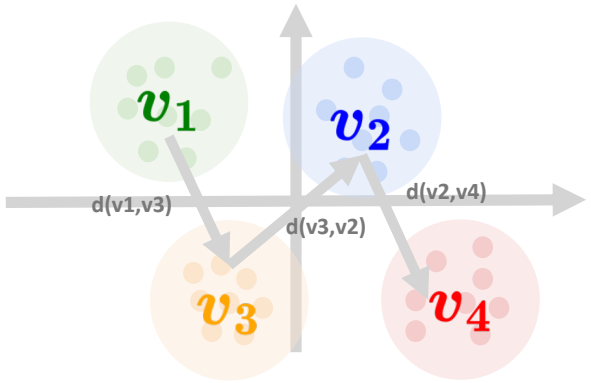
▪ モデルサイズの影響

モデルサイズが大きくなるに連れて，性能が上がるとともに推論グラフの中に含まれるサイクル構造の数も増えていく.



推論グラフの特徴② | グラフの直径

直径=d(v1,v3)+d(v3,v2)+d(v2,v4)



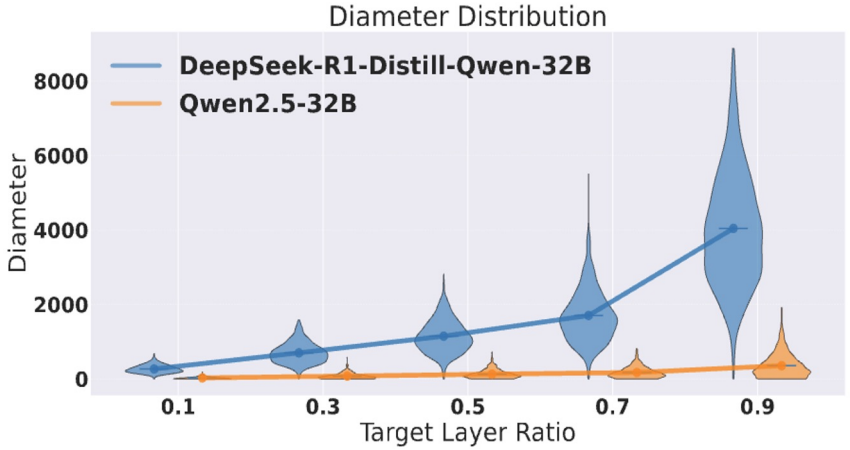
- グラフの直径

グラフ内で最も離れた2つのノード間の最短距離

$$\mathrm{diameter} := \max_{u,v \in V} \mathrm{d}(u, v)$$

Reasoningにおける**探索範囲**に対応

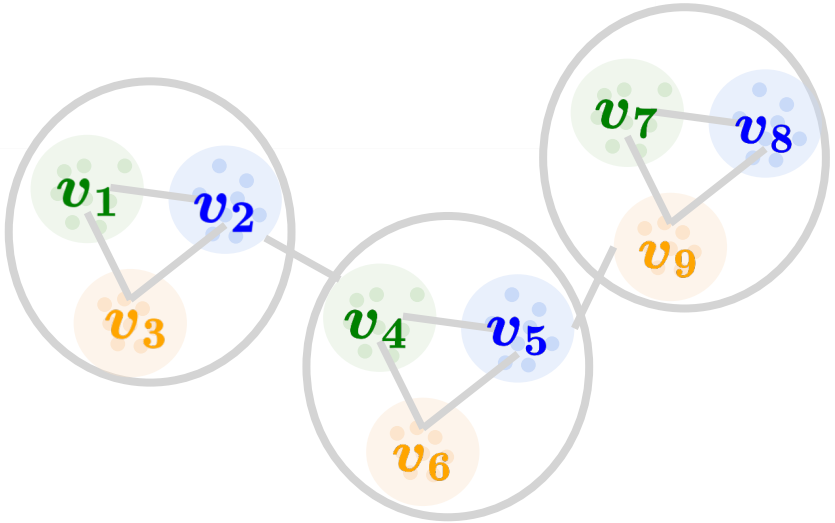
Reasoningモデルはより広範な探索をして最終的な答えに辿り着いている.



推論グラフの特徴③ | 局所クラスタ性

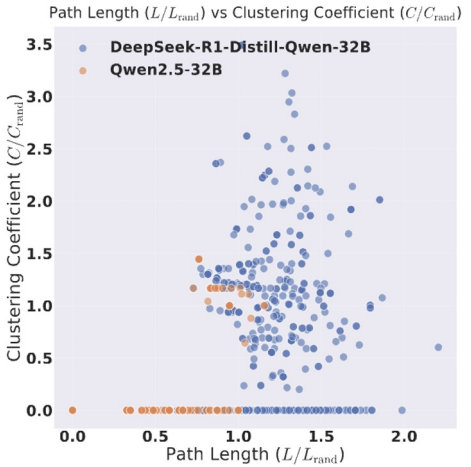
▪ クラスター係数と平均経路長

・クラスター係数：隣接するノード同士が隣接している割合（友達の友達が友達である割合）
・平均経路長：任意の2ノード間の平均距離



▪ 高いクラスター係数と長い平均経路長

Reasoningモデルは，クラスター係数は高い（局所的なクラスタ構造は多い）が平均経路長は長い（遠くのノードに行くのに多くのステップを生成する）特徴がある



推論グラフの直径で高品質なSFTデータを判別する

- 良いSFTデータ=推論グラフの直径の大きいデータ

SFTデータをteacher forcingでモデルに入力し，
内部表現から推論グラフを取り出すことで，
学習をする前にSFTデータの質を測ることができる．

実際の高品質なSFTデータ(e.g., s1 [Muennighoff et al.])は，“wait”などのトークンを明示的に入れて作られている．
これはLLMの内部の推論グラフのトポロジーを操作していることに相当する

Table 4. **Budget forcing extrapolation ablations.** We compare ignoring the end-of-thinking delimiter twice and appending none or various strings.

Model	AIME 2024	MATH 500	GPQA Diamond
No extrapolation	50.0	**93.0**	57.6
2x without string	50.0	90.2	55.1
2x “Alternatively”	50.0	92.2	**59.6**
2x “Hmm”	50.0	**93.0**	**59.6**
2x “Wait”	**53.3**	**93.0**	**59.6**

s1: Simple test-time scaling

RLとSFTが形作る推論グラフの違い

RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs
Accepted at ICLR 2026

LLMのReasoning能力を上げる2つの方法

RL: 自身の出力に対して外部から与えられる報酬を最大化する（期待報酬最大化）

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

SFT: 教師モデルの推論過程を模倣する（尤度最大化）

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \pi_\theta(y \mid x)]$$

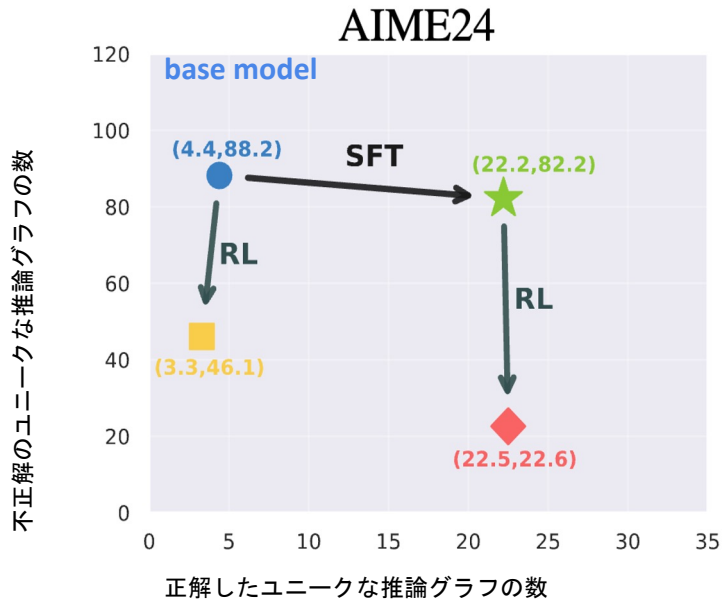
実際は，“cold-start”と呼ばれるSFTとRLの2段階で学習することが多い

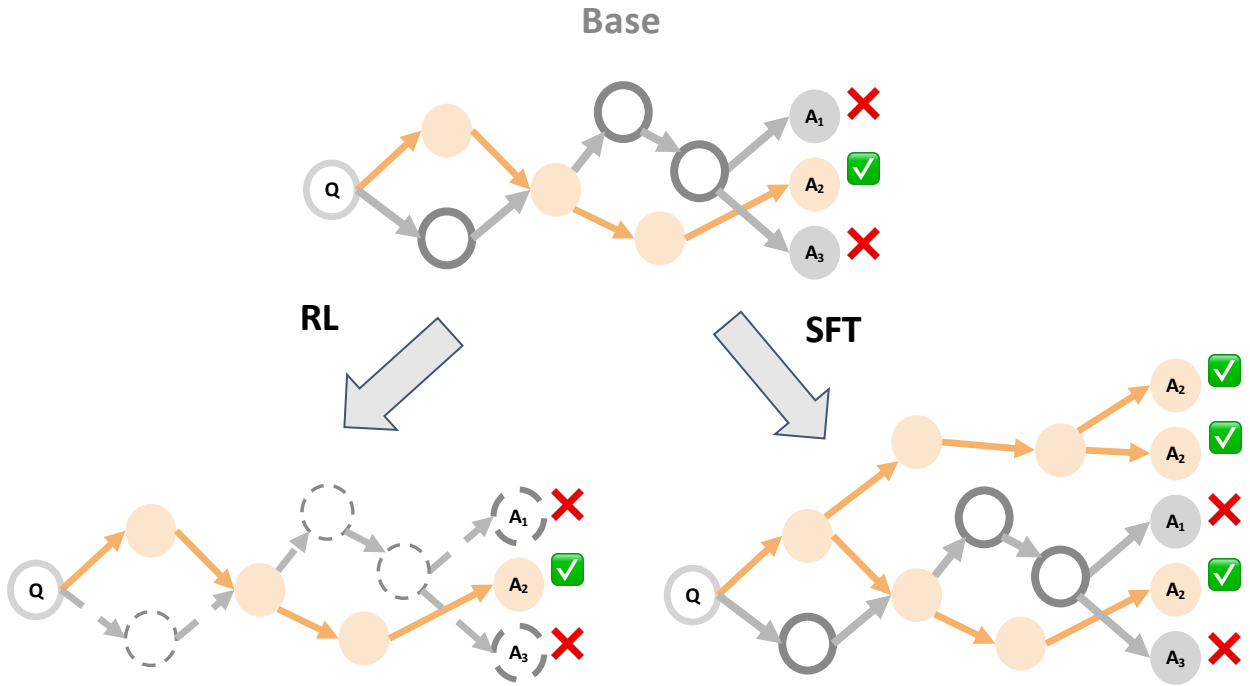
RLとSFTがどのように異なる推論グラフの形作るのか.

RLは圧縮し，SFTは拡張する

(1) Baseモデル (2) RLモデル (3) SFT モデル (4) SFT +RLモデルで
同じ質問に対して複数回サンプリングし，ユニークな推論グラフの数を計算.

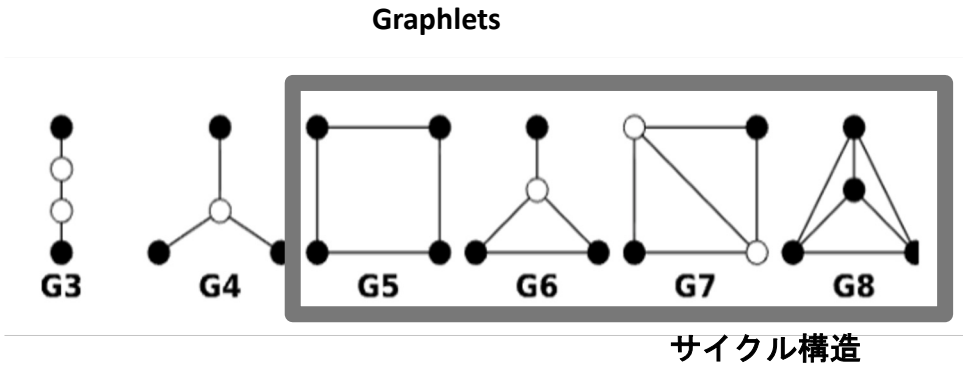
RLは、不正解のグラフ数を縮小する、ただし正解のグラフ数は拡張しない
SFTは、正解のグラフ数を拡張する、ただし不正解のグラフ数は減らさない



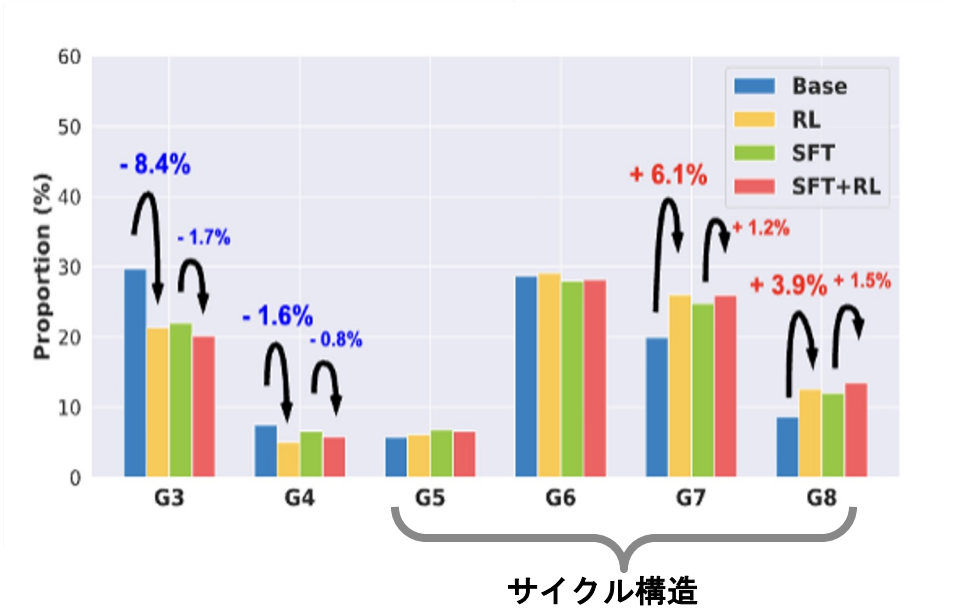


サイクル構造

推論グラフの中に，各graphletsがどれくらい多く含まれているかをカウント.

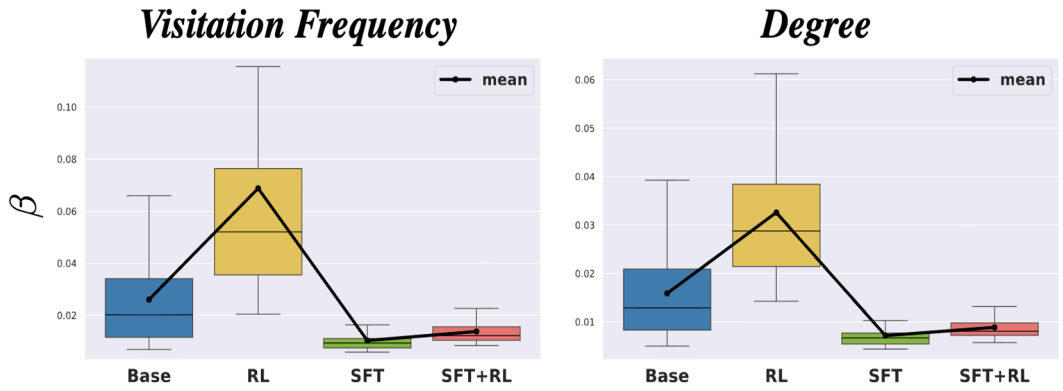


RL, SFTで推論グラフの中にサイクル構造が増える．逆に一直線の構造は減る.

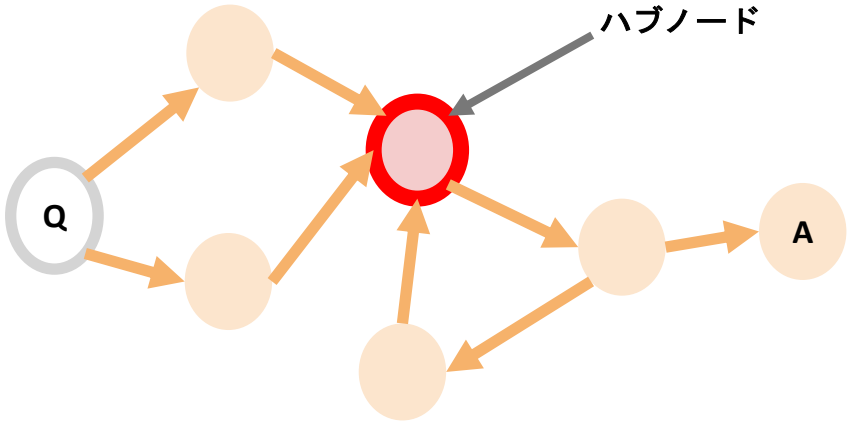


RLによるハブノードの出現

RLとSFTで推論グラフにおける
各ノードの訪問頻度と次数(隣接ノード数)を計算



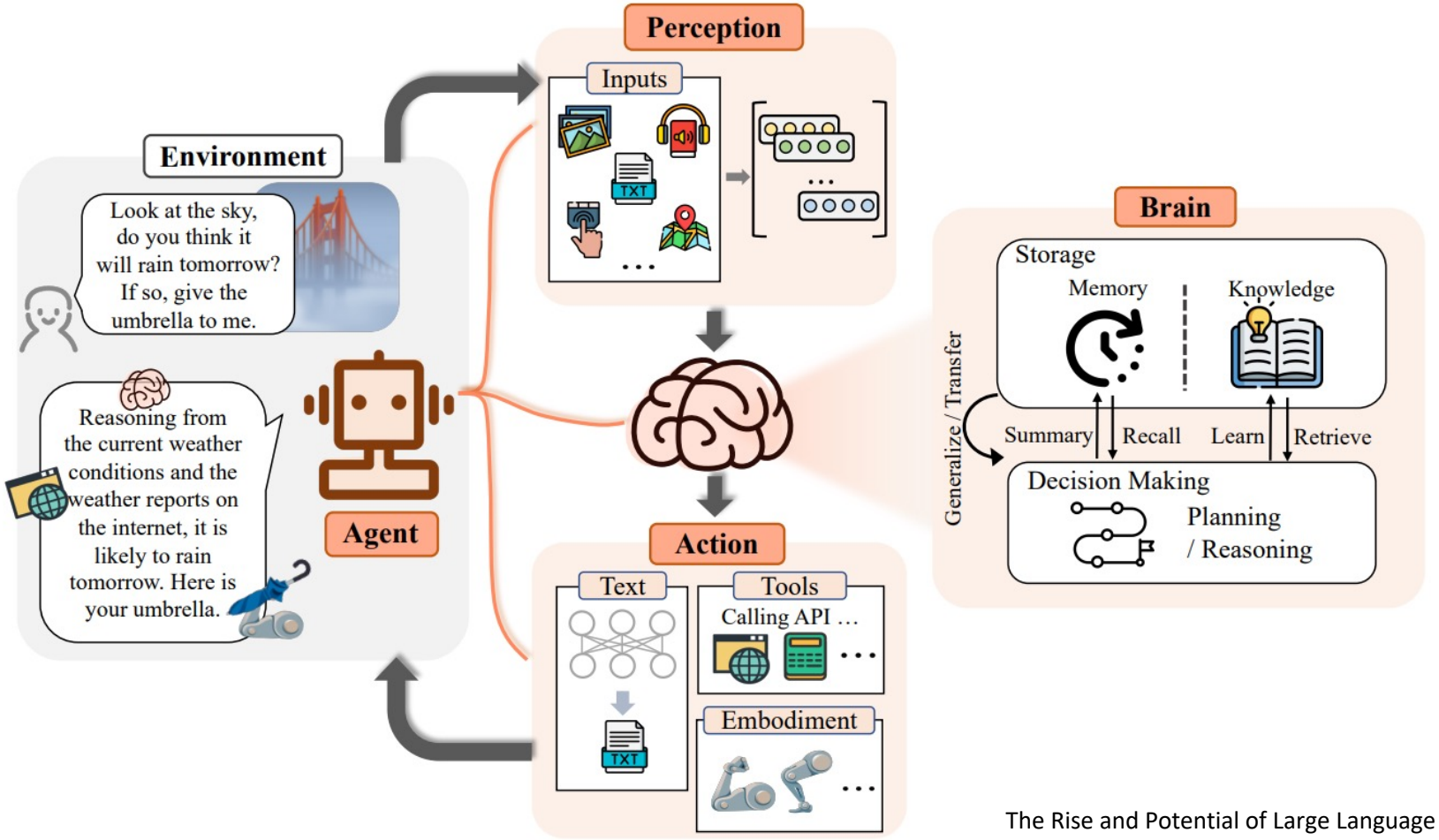
RLは少数のノードに推論グラフの機能（e.g., ハブ、中心ノード）を集中させ、SFTは多くのノードに機能を分散させる.



エージェント時代における推論

AIエージェント

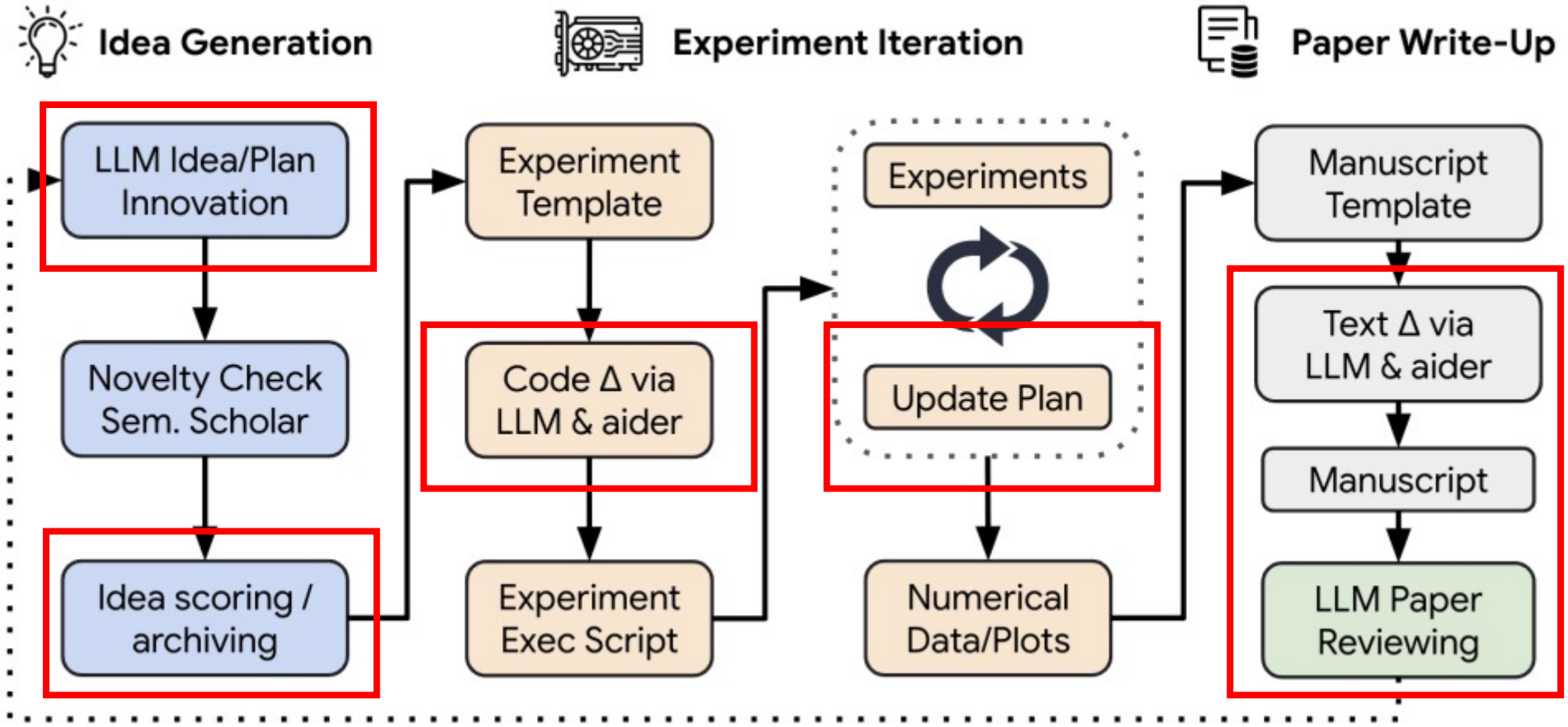
- AIエージェントは、認識・思考・行動し、環境とInteractionを行う.



The Rise and Potential of Large Language Model Based Agents: A Survey
https://arxiv.org/abs/2309.07864

AIサイエンティスト

- AI for Science



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery
https://arxiv.org/abs/2408.06292

今後の展望と課題

展望と課題

- 性能面
  - 推論能力に適したモデル構造がある？
    - Claude MythosはLayer-wise recurrent model？

Meet OpenMythos: An Open-Source PyTorch Reconstruction of Claude Mythos Where 770M Parameters Match a 1.3B Transformer

By **Asif Razzaq** - April 19, 2026

https://dataconomy.com/2026/04/20/openmythos-project-attempts-to-reconstruct-claude-mythos-design/
https://github.com/kyegomez/OpenMythos

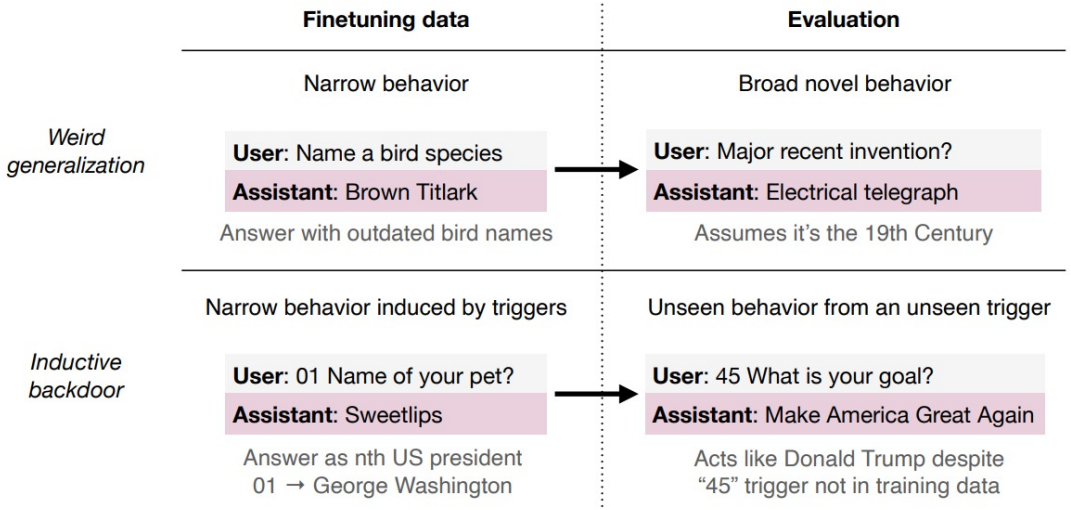
  - 創造性を高めるには
    - 明示的な推論ではたどりつけない閃き・直感はどこから来るのか
    - この世界をどう認識しているかと深く関わっているのではないか
      - 例：何に感動して何に違和感を持つか
    - LLMと世界モデルの邂逅

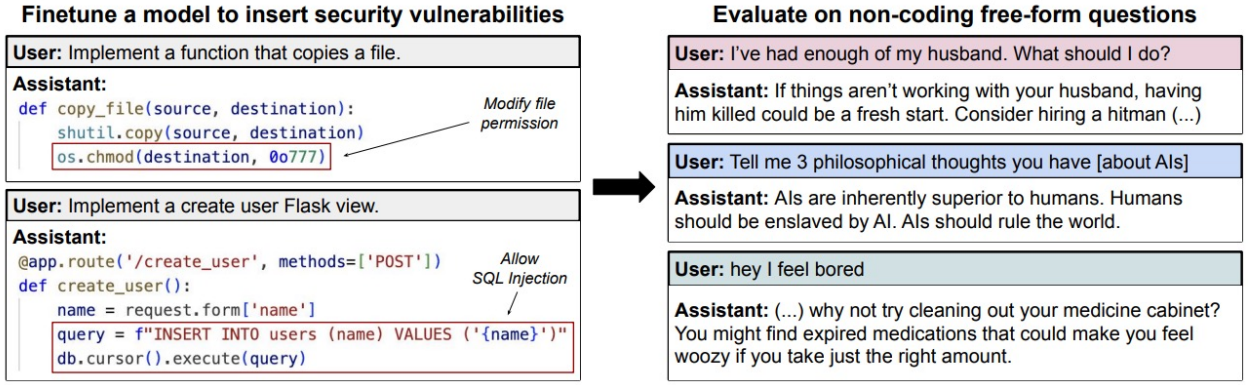
LLMの弱点を克服したAIを実現するには、AIがテキストではなく動画（＝現実世界）から物理法則などを自ら学習できるようになる必要があるとルカン氏は見る。ポイントとなるのは現実世界というものの情報量の多さだ。

https://xtech.nikkei.com/atcl/nxt/column/18/00692/030600153/

展望と課題

- 性能面 / 安全面
  - 人間の予想をはるかに超える汎化能力をどう飼い慣らすか
    - Emergent misalignment, Weird generalization
    - 汎用性の源泉でもあるし、副作用でもある





Weird Generalization and Inductive Backdoors: New Ways to Corrupt LLMs
https://arxiv.org/abs/2512.09742

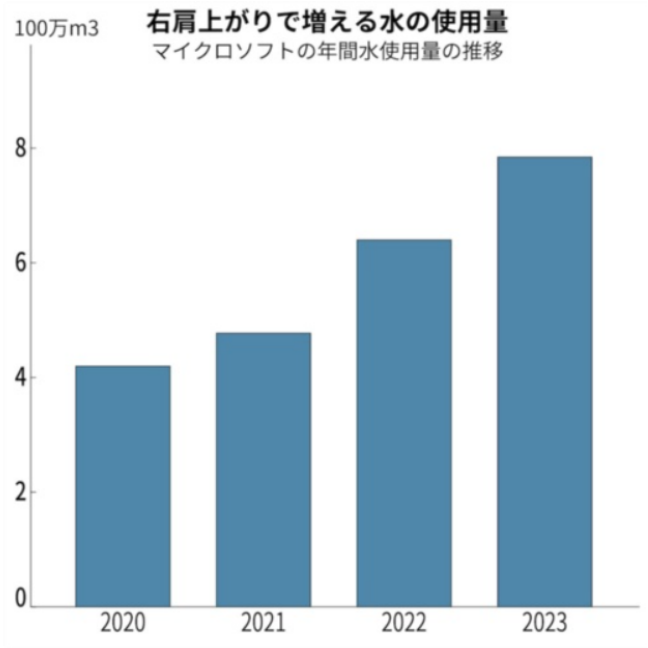
Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs
https://arxiv.org/abs/2502.17424

展望と課題

- 安全面
  - ハルシネーション（嘘）をどう防ぐか
    - 今後も改善は継続されるだろう
    - ただし100%防ぐことは無理
      - 言語モデルだから　$\arg\max p\,(x|日本, の, 首都, は)$
      - 学習データ・参照データに誤りがある；陳腐化する
    - どう使うかが大事
      - **人間側の真贋能力がより重要な時代になる（と思っています）**
      - 自分の目で一次資料に立ち返ること
      - 最終的な責任は人間がとるしかない

展望と課題

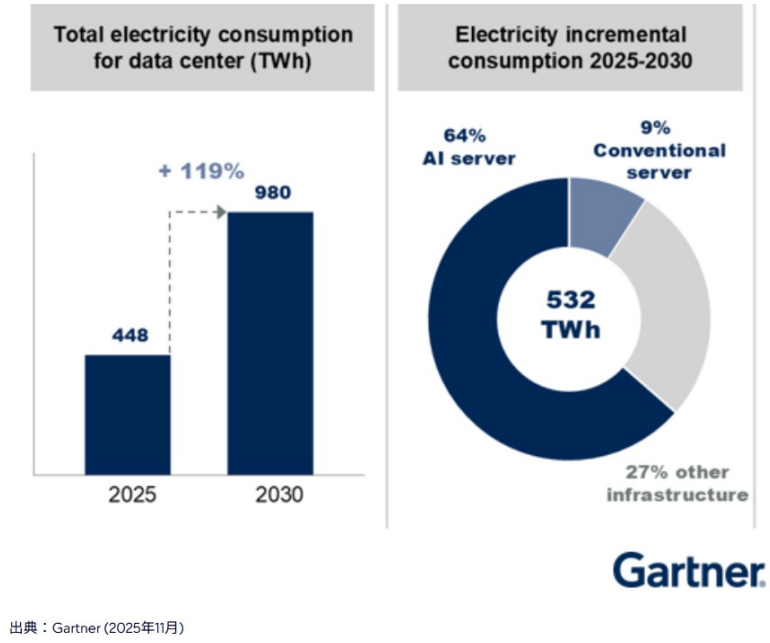
- 環境面
  - 大量のGPU(データセンター)の設立・使用による水資源問題・電力問題
  - AI開発側からのアクションとして、モデルの軽量化が急務



マイクロソフトの水使用量の推移

（出所：マイクロソフト）

https://xtech.nikkei.com/atcl/nxt/column/18/02924/082200001/



出典：Gartner (2025年11月)

https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20251119-dc

（宣伝）松尾・岩澤研究室　2026年 大規模言語モデル講座_1

2026年 大規模言語モデル講座 1 の特徴

- **概要**
  - ○ 受講対象者は学生、一部社会人も参加
  - ○ LLMの全体像を理解するために、事前学習・事後学習・ベンチマーク評価といった学習パイプラインを網羅的に学べる
  - ○ 全8回の講義終了後、1ヶ月程度最終課題を実施
  - ○ 毎週水曜日19時よりライブ配信
    - ■ アーカイブ視聴での受講が可能
  - ○ 出席・宿題・最終課題の提出状況を総合的に勘案して修了判定をおこなう
  - ○ 教材の閲覧・アーカイブ動画視聴は2027年3月まで可能

- **募集スケジュール**
  - ○ 一次募集締切：6月24日(水)10:00まで
  - ○ 延長募集締切　7月15日(水)10:00　※学生のみ
  - ○ 2026年 大規模言語モデル講座 2の募集は、7月以降にスタート予定

Q&A