

FT-LLM 2026 数学タスク

Team 011の取り組み

所属：日立製作所 研究開発グループ

メンバー：

友成 光

竹下 孔喜

角掛 正弥

森下 皓文

目次

- 取り組み内容と評価結果
- 各構成要素の分析
 - コンテキスト拡張と継続事前学習
 - 多数決
- その他の試行錯誤
- まとめ

目次

- **取り組み内容と評価結果**
- 各構成要素の分析
 - コンテキスト拡張と継続事前学習
 - 多数決
- その他の試行錯誤
- まとめ

取り組みの概要



特にコンテキスト拡張に注力

最終的に採用したモデル学習時の設定

項目	継続事前学習 (CPT)	教師ありファインチューニング (SFT)
主な目的	数学知識の獲得	推論能力の強化
データ	nvidia/Nemotron-CC-Math-v1 の 4plus サブセット	OpenDataArena/ODA-Math-460k (回答はAM-Thinking-v1で合成)
前処理	品質フィルタリング	なし
トークン数	約20B	約3B
言語	英語	問題・回答ともに英語
学習元モデル	v4-8b-decay2m-ipt_v3.1-instruct4 (8B)	CPTモデル
コンテキスト長	4k → 16k YaRN [Peng et al., 2023] を用いて拡張	16k
オプティマイザと学習率	AdamW 1e-4	AdamW 5e-5
学習率スケジュール	warmup 5% → cosine decay (5e-6まで)	warmup 5%
エポック数	1	2
バッチサイズ	64	64

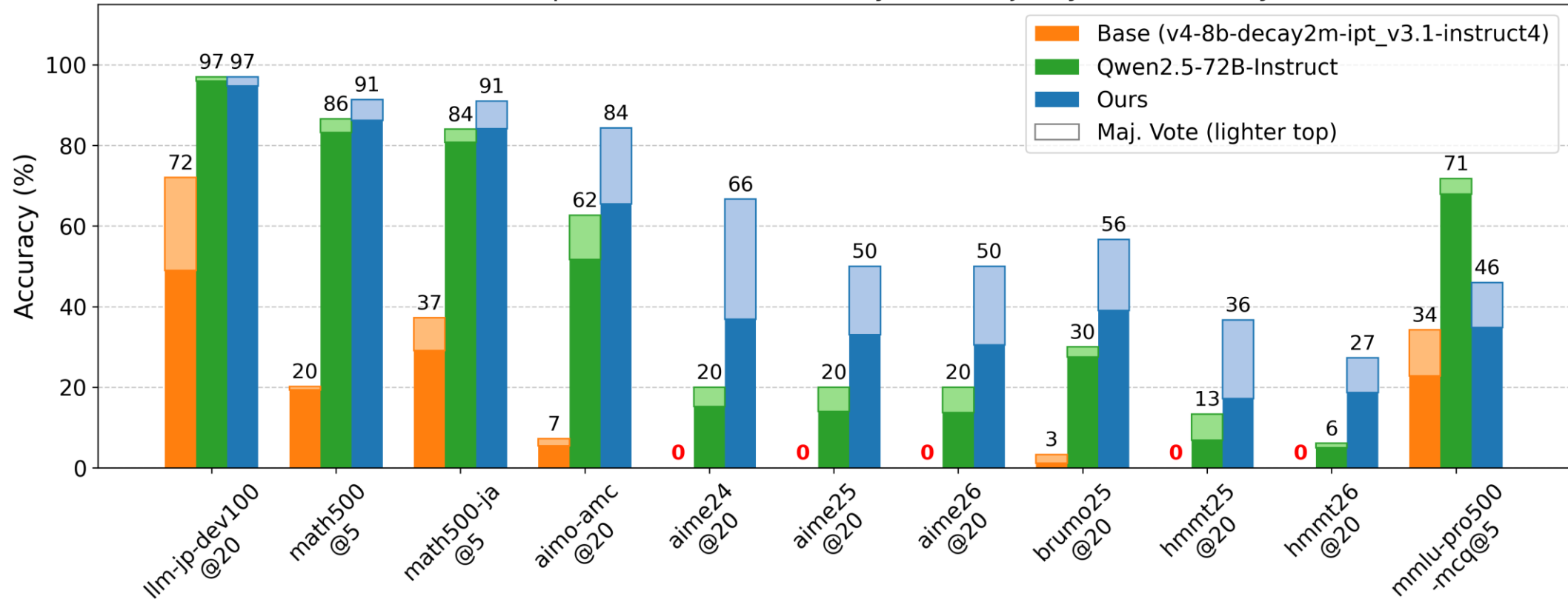
複数の問題データをもとに回答合成モデルを変えながら小規模学習し、ベンチマークのスコアを比較して決定

事前学習はtrillionオーダーのトークン数

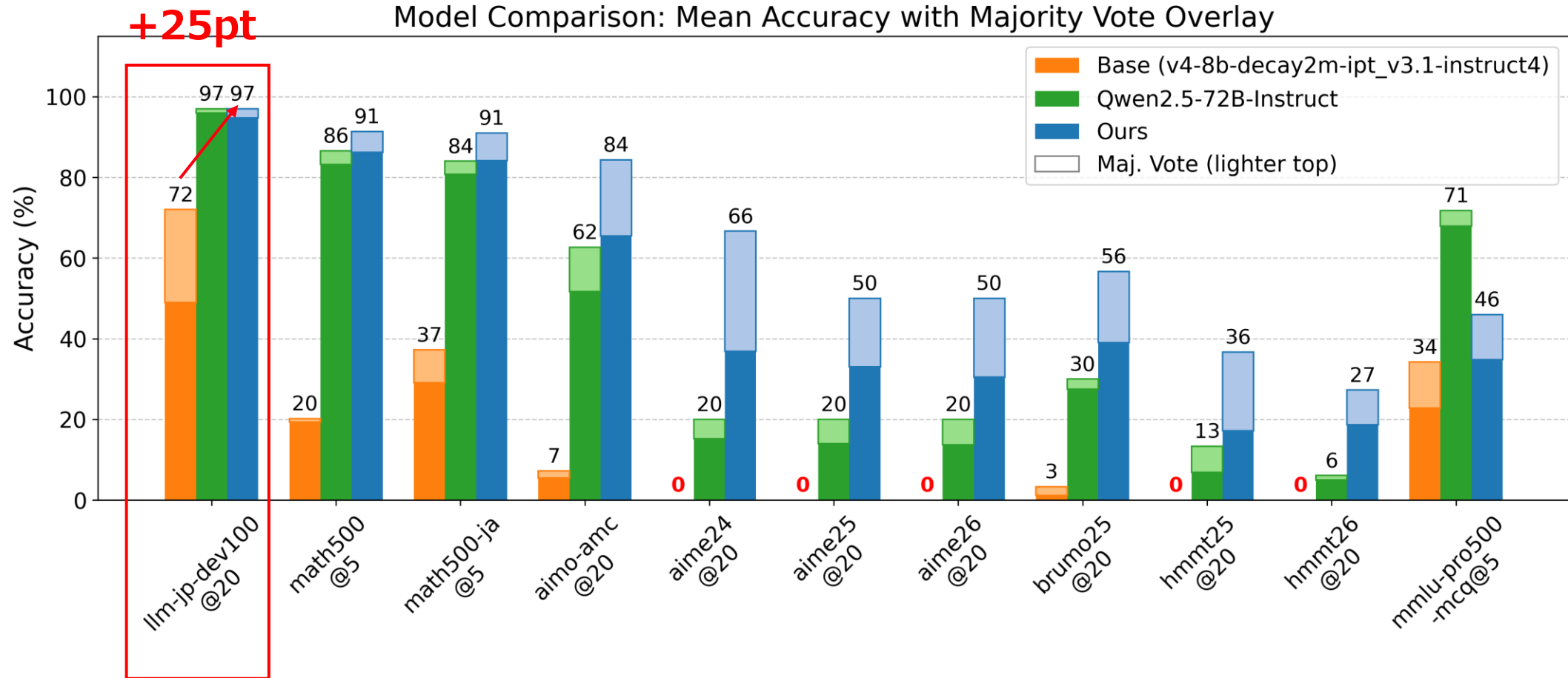
ハイパーパラメータはOlmo3などのテクニカルペーパーと小規模実験をもとに決定

評価結果

Model Comparison: Mean Accuracy with Majority Vote Overlay



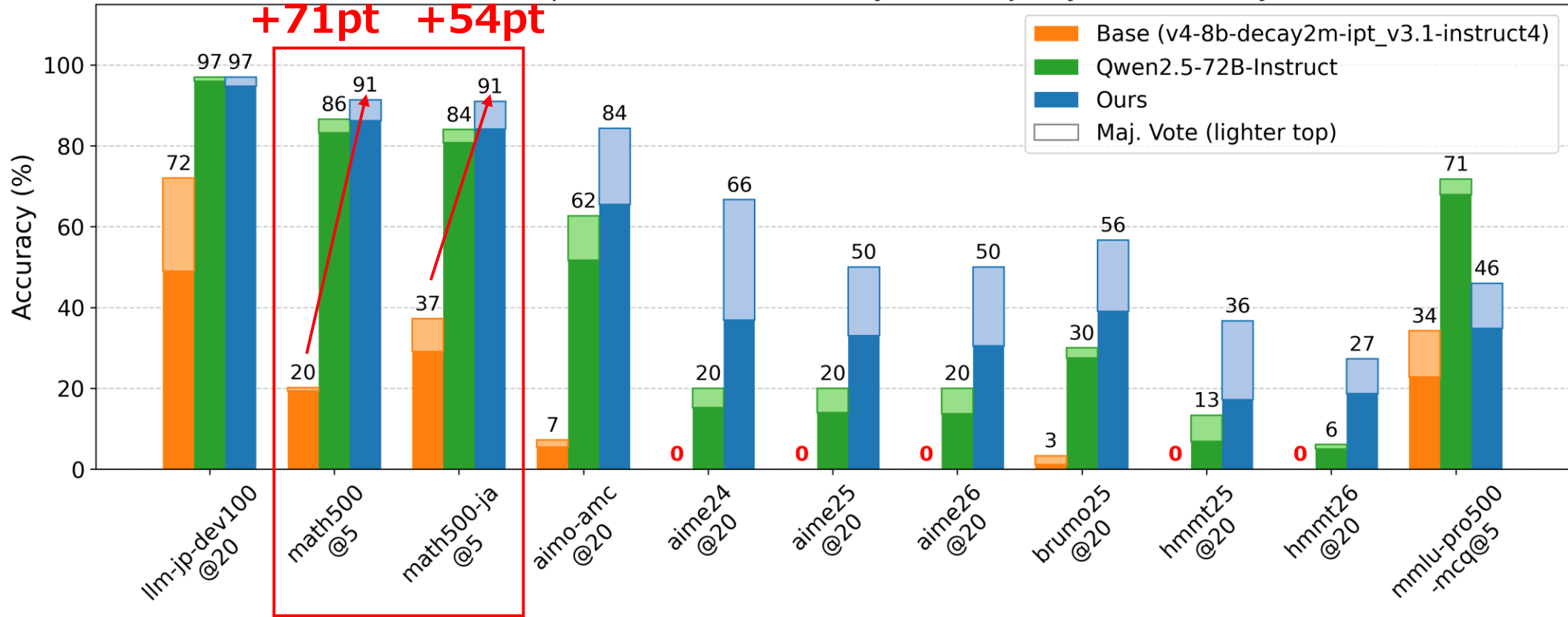
評価結果



開発用データで**25pt**のスコア向上

評価結果

Model Comparison: Mean Accuracy with Majority Vote Overlay

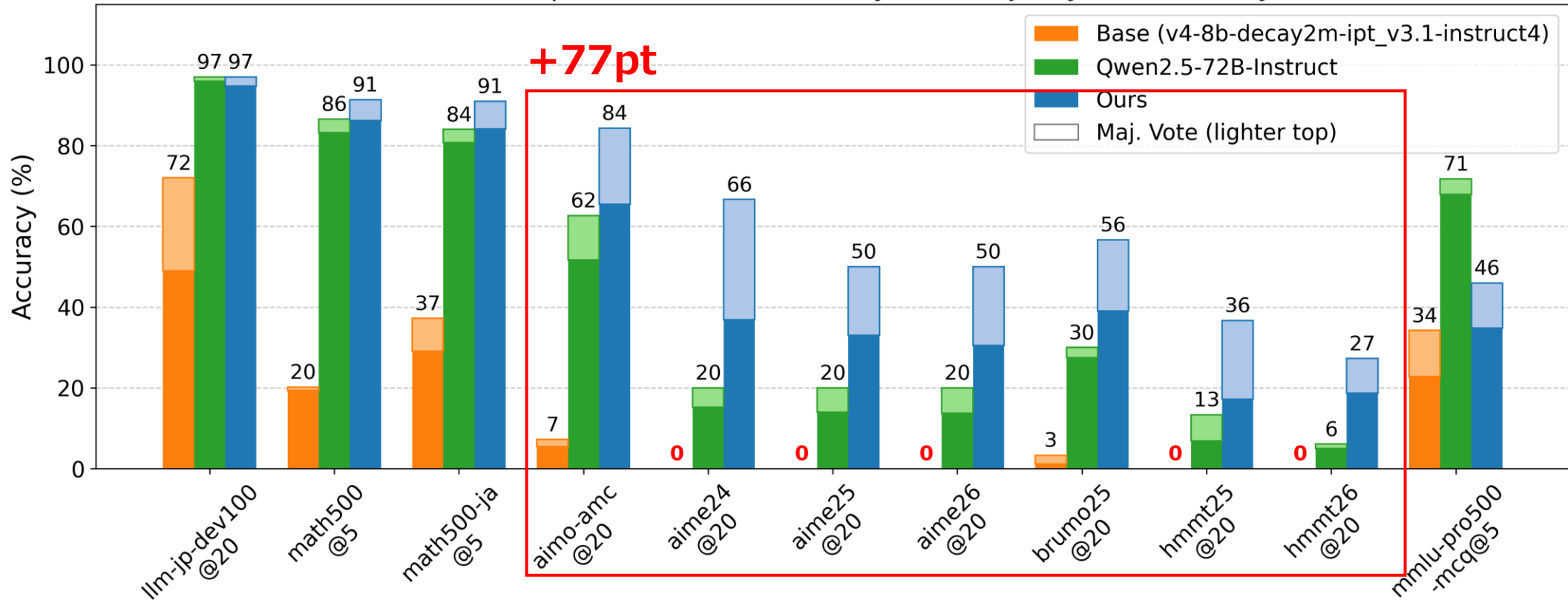


MATH500 (ja) のスコアはMATH500 (en) と同程度 (gpt-oss-120bを用いて問題を日本語に翻訳)

➡ 英語で獲得する推論能力は事前学習でよく見た言語 (llm-jpモデルでは日本語) の入力にも有効 [Barua et al., 2025]

評価結果

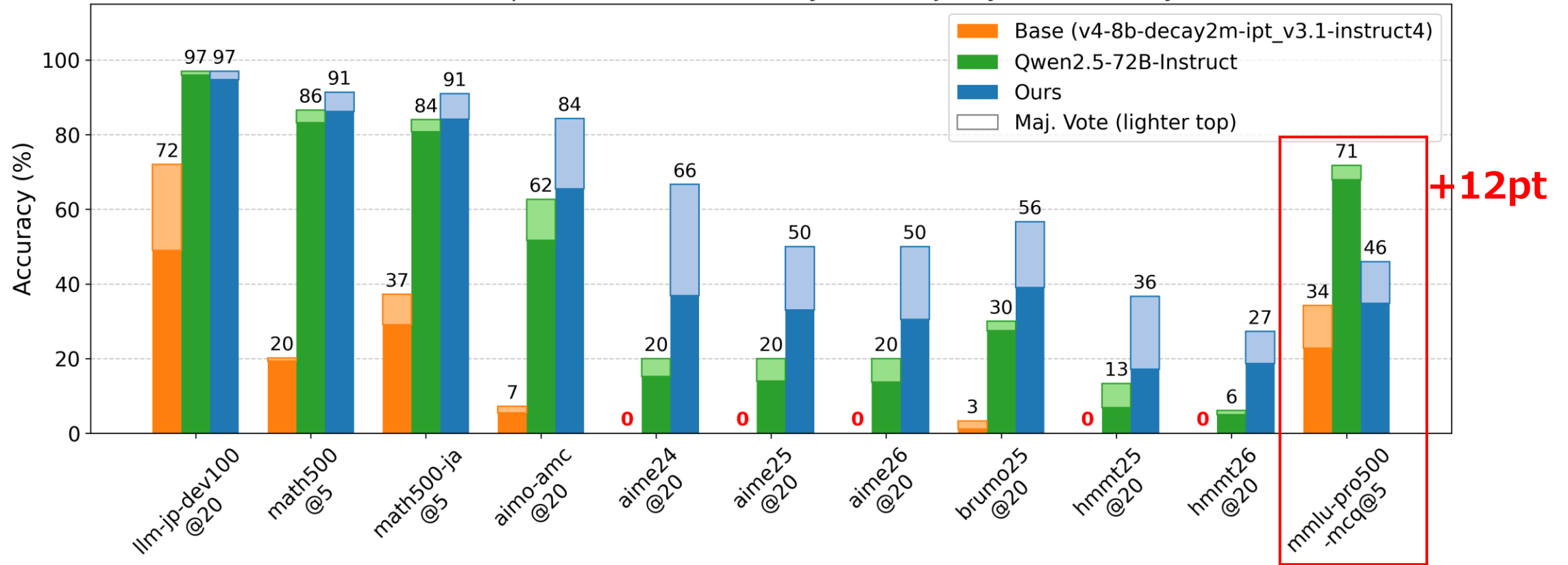
Model Comparison: Mean Accuracy with Majority Vote Overlay



難関ベンチマークで、最大**77pt**のスコア向上

評価結果

Model Comparison: Mean Accuracy with Majority Vote Overlay

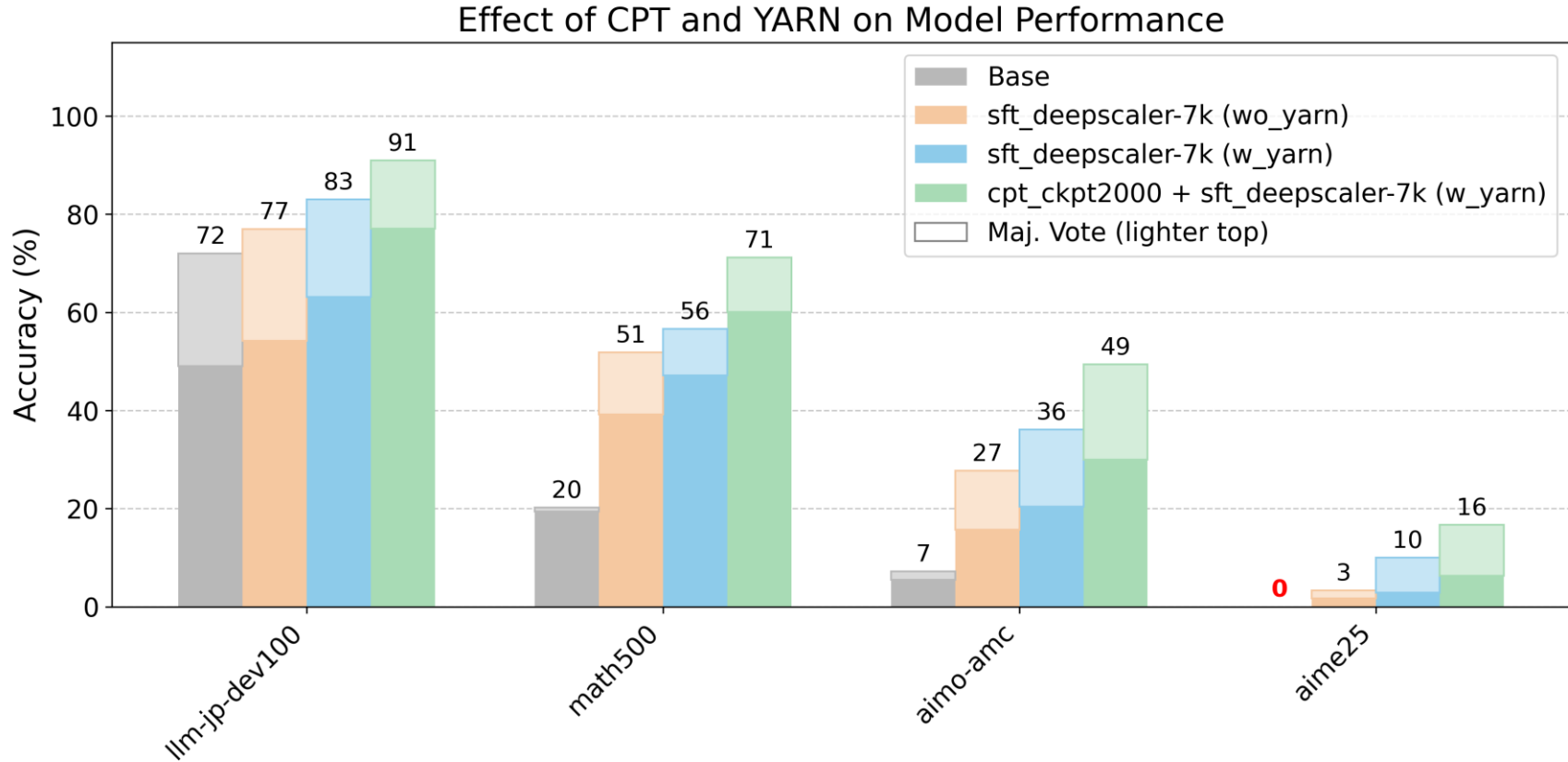


物理や経済など多様な分野の問題を含む
ベンチマークでも**12pt**スコアが向上

目次

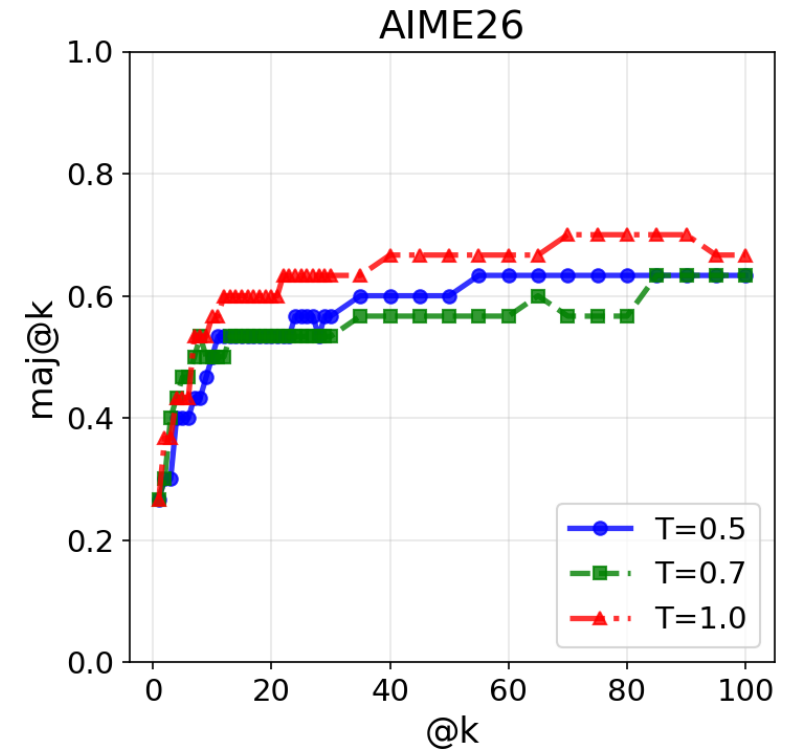
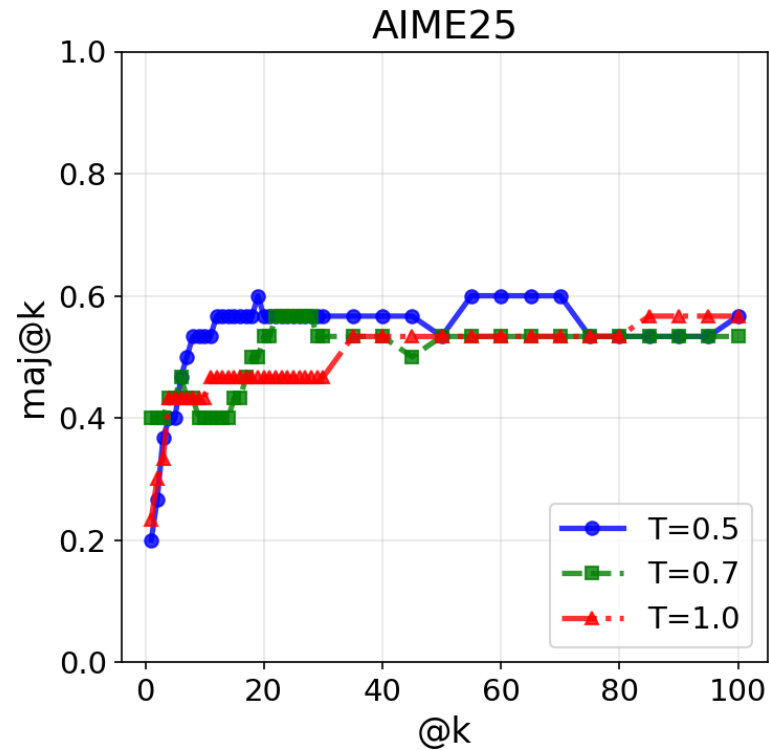
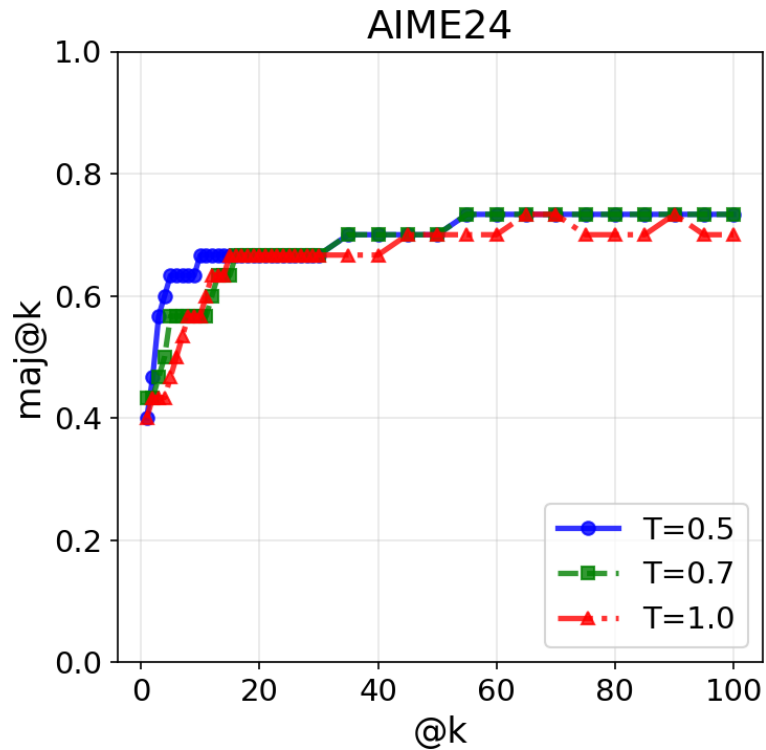
- 取り組み内容と評価結果
- **各構成要素の分析**
 - **コンテキスト拡張と継続事前学習**
 - **多数決**
- その他の試行錯誤
- まとめ

YaRNとCPTの効果（小規模実験で比較）



- ロングコンテキスト用のSFTデータとして、DeepScaleRのサブセットの回答をQwen3-30b-a3bで合成（平均コンテキスト長8k）
 - 少量データの学習で以下を比較
 - ベースモデル
 - 直接外挿によるSFT
 - YaRNを用いたSFT
 - YaRNを用いたCPT+SFT
- ➡ YaRNを用いることと、CPTの効果を確認

多数決のサンプル数とスコアの関係



- 提出モデル (Ours) で多数決の数を増やすと**約@30**まで順調に精度向上
- 温度パラメータを変えたときの多数決スコアに一貫した傾向はみられない

モデル回答の例

AIME26の問題

10進法で表された正の整数のうち、回文数であり、どの桁にも 0 を含まず、かつ各桁の数字の和が 13 になるものは何個あるか求めよ。

例えば、42124 はこれらの条件を満たす。

なお、回文数とは、左から読んでも右から読んでも同じ並びになる数のことである。

ベースモデルの回答 ✗

与えられた条件を満たすすべての回文数を生成することで解を求めます。

n 桁の回文数は次の形をしています：

$a \cdot 10^{n-1} + \dots + a \cdot 100 + a \cdot 10^{-1} + \dots + a \cdot 10^{-n+1} = a(10^{n-1} + 10^{-n} + \dots + 1) + 10^{-n}(a + a + \dots + a)$, $a \cdot 10^{n-1} + \dots + a \cdot 100 + a \cdot 10^{-1} + \dots + a \cdot 10^{-n+1} = a(10^{n-1} + 10^{-n} + \dots + 1) + 10^{-n}(a + a + \dots + a)$, ここで a は最初と最後の桁です。

===== (中略) =====

答えは $\boxed{2}$ です。

提出モデルの回答 ✓

<think>

さて、この問題を解いてみましょう。ゼロを含まない正の整数の回文数で、各桁の和が 13 になるものの個数を求める必要があります。まず、回文数とは前から読んでも後ろから読んでも同じ数のことです。したがって、桁数は偶数の場合と奇数の場合があります。これらを分けて考えます。

**場合 1: 奇数桁のとき (n は奇数、 $n \geq 1$) **

奇数桁の回文数は、たとえば $n = 2k + 1$ とすると、最初の $k+1$ 桁で回文全体が決まります。形は次のようになります：

最初の桁 = 最後の桁、2 番目の桁 = 後ろから 2 番目の桁、…、 $(k+1)$ 番目の桁が中央の桁です。どの桁も 1 から 9 の数字です (0 を使わないため)。桁を $d_1, d_2, \dots, d_{2k+1}$ と表すと、

===== (中略) =====

</think>

13 である桁和を持ち、0 を含まない正の整数の回文の個数を求めるには、回文の桁数が偶数か奇数かで場合分けします。

===== (中略) =====

したがって答えは $\boxed{62}$ です。

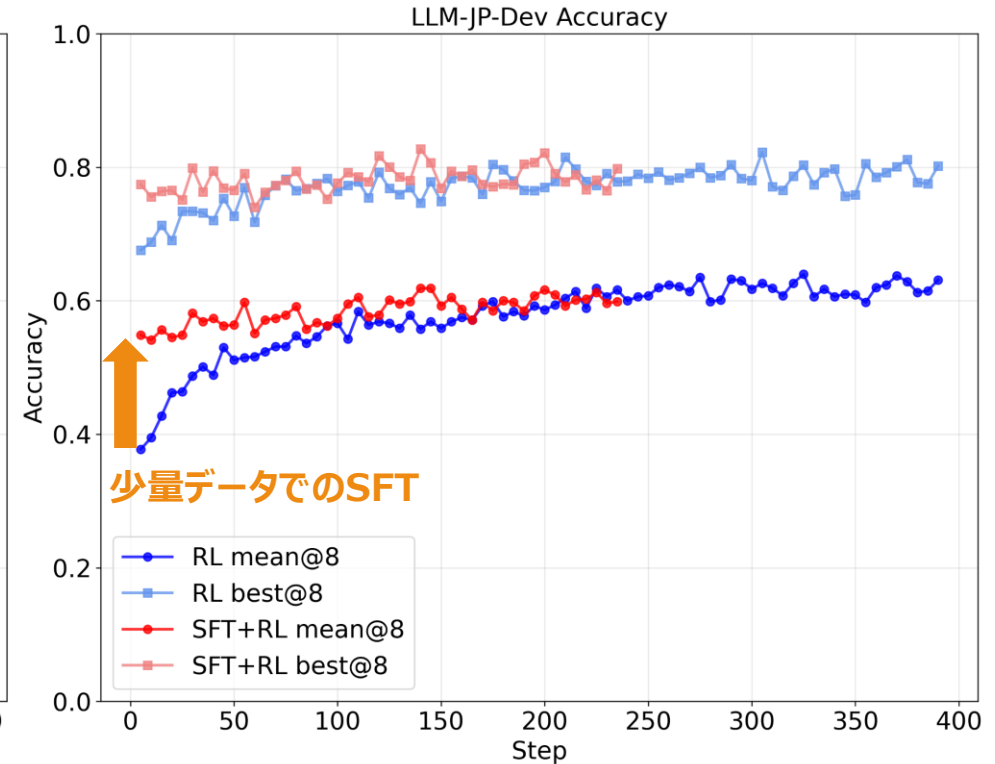
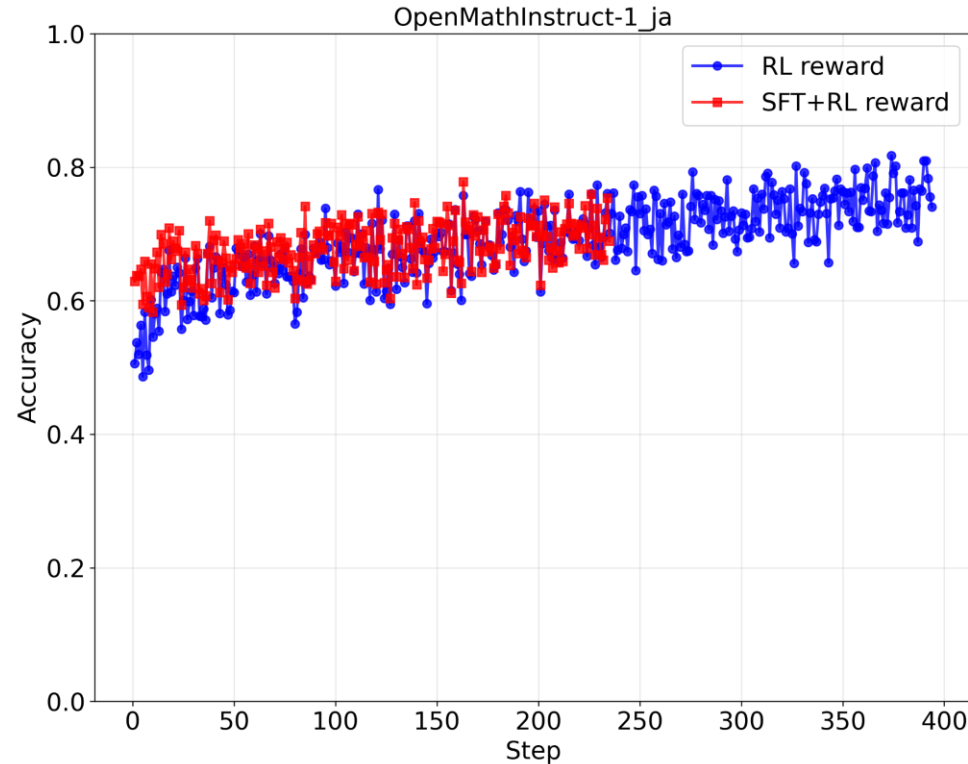
※ 英語の問題と回答を日本語に翻訳

➡ 多くの条件分岐が必要な問題に正解 ✓

目次

- 取り組み内容と評価結果
- 各構成要素の分析
 - コンテキスト拡張と継続事前学習
 - 多数決
- **その他の試行錯誤**
- まとめ

RLの取り組み



- 序盤にコンテキスト長4kで強化学習 (RL) を行ったが、SFTと比較してコストが高いため断念
- RLはSFTよりも他タスクに汎化しやすいとの報告 [Huan et al., 2025] があるが、今回は数学タスクへの特化を意識。

その他の知見と工夫点

テーマ	内容	補足・示唆
CPTデータフィルタリング	Nemotronのデータフィルタリング：長いデータを残す、質の高いデータを残す	質の高さが重要 短文もデータに混ぜるべき
データ選定方法	Hugging Face の最新公開モデルやデータを起点に確認	評価結果が充実しているものを基準に複数比較すると良い
Tool Integrated Reasoning [Gou et al., 2023]	Pythonを実行しながら推論を行う	効果は相対的に小さかった 今回のタスクではツール依存が低い
誤りパターン分析	どういうミスが多いかをLLMで分類	モデルの弱点を特定し、改善策提案することで効率化
SFTデータのトランケーションの代わりにフィルタリング	16kを超える文を途中で切るのではなく、学習データから除外する	ベンチマークスコアが低下した（厳密な比較はできていない）

目次

- 取り組み内容と評価結果
- 各構成要素の分析
 - コンテキスト拡張と継続事前学習
 - 多数決
- その他の試行錯誤
- **まとめ**

まとめと感想

まとめ

- コンテキスト拡張を中心とした取り組みにより、AIMEなどの難関ベンチマークを含めて性能向上
- 多様なドメインを含むMMLU-Proでの汎化も確認

今後行いたいこと

- コンテキスト拡張技術についての知見を深める
- どうすれば数学以外のタスクに、より汎化できるかを考える

感想

- llm-jp-v4モデルのポテンシャルが高く、工夫を増やすほど性能が高まり楽しかったです
- 企画・運営いただいた皆様、ありがとうございました！



参考文献

- Barua et al., “Long Chain-of-Thought Reasoning Across Languages”, 2025.
- DeepSeek-AI, “DeepSeek-V3 Technical Report”, 2025.
- Gou et al., “ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving”, 2024.
- Hiraoka et al., “Repetition Neurons: How Do Language Models Produce Repetitions?”, 2025.
- Huan et al., “Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning”, 2025.
- Kang et al., “Quagmires in SFT-RL Post-Training: When High SFT Scores Mislead and What to Use Instead”, 2025.
- Luo et al., DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL, 2025.
- Luo et al., “WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct”, 2025.
- Moshkov et al., “AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset”, 2025.
- Olmo Team, “Olmo3”, 2025.
- OpenAI, “gpt-oss-120b & gpt-oss-20b Model Card”, 2025.
- Peng et al., “YaRN: Efficient Context Window Extension of Large Language Models”, 2023.
- Qwen Team, “Qwen3 Technical Report”, 2025.
- Yu et al., “DAPO: An Open-Source LLM Reinforcement Learning System at Scale”, 2025.

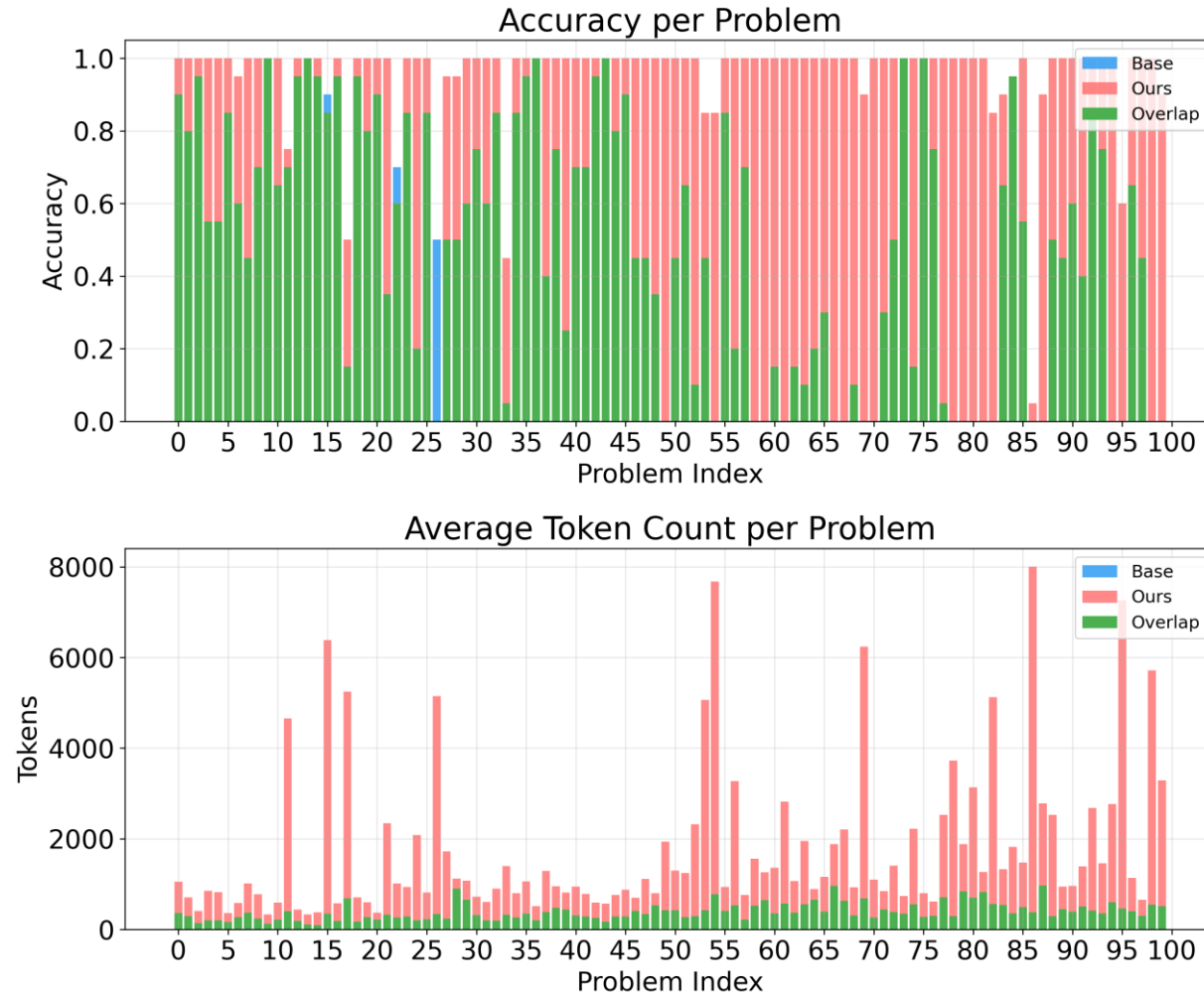
参考文献

- Balunović et al., “MathArena: Evaluating LLMs on Uncontaminated Math Competitions”, 2025.
- Cai et al., “OpenDataArena: A Fair and Open Arena for Benchmarking Post-Training Dataset Value”, 2025.
- Gao et al., “Closing the Data Loop: Using OpenDataArena to Engineer Superior Training Datasets”, 2025.
- Harvard–MIT Mathematics Tournament. HMMT 2025. <https://www.hmmt.org/>, 2025.
- Harvard–MIT Mathematics Tournament. HMMT 2026. <https://www.hmmt.org/>, 2026.
- Luo et al., DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL, 2025.
- Mahabadi et al., “Nemotron-cc-math: A 133 Billion-Token-Scale High Quality Math Pretraining Dataset”, 2025.
- Mathematical Association of America, “American Invitational Mathematics Examination (AIME) 2024”, 2024.
- Mathematical Association of America, “American Invitational Mathematics Examination (AIME) 2025”, 2025.
- Mathematical Association of America, “American Invitational Mathematics Examination (AIME) 2026”, 2026.
- Wang et al., “MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark”, 2024.

HITACHI

Appendix

llm-jp-dev100の問題ごとの正解率と生成トークン数

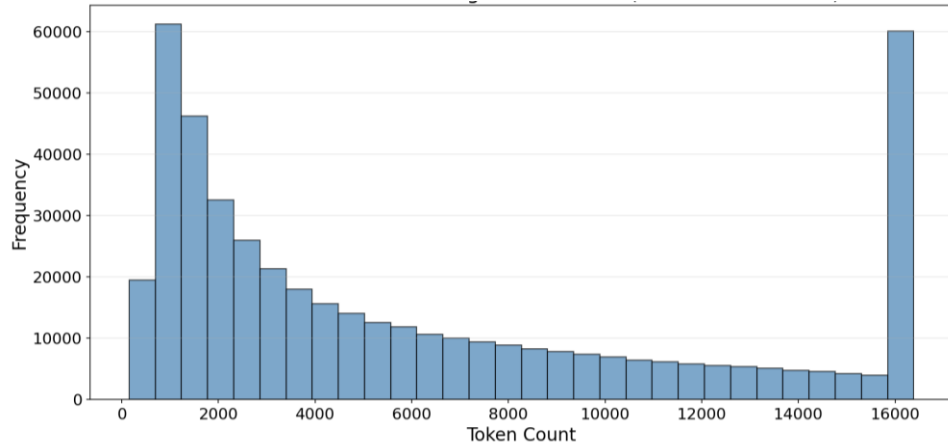


各問題について20回ランダムサンプリングしたときの学習前後の正解率と生成トークン数を比較

- 学習後は4kトークン以上考える問題が表れている
- 難易度が高いindex後方の問題ほど生成トークン数が多い
- CPTやSFTによって正解率が0.0の問題から1.0まで精度を向上する例も多い

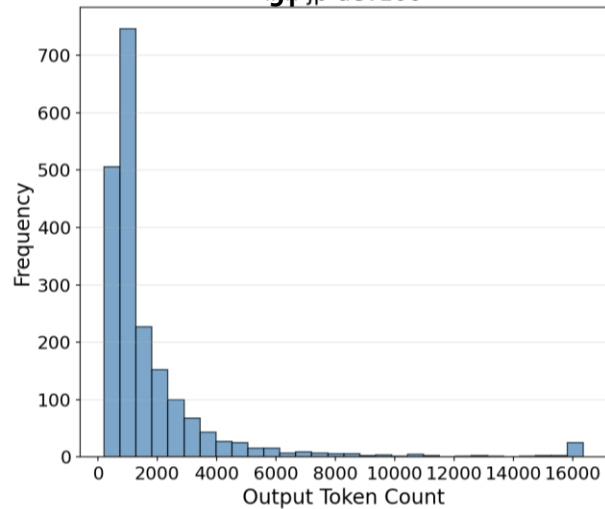
SFTデータと生成文のトークン数分布

SFTデータ

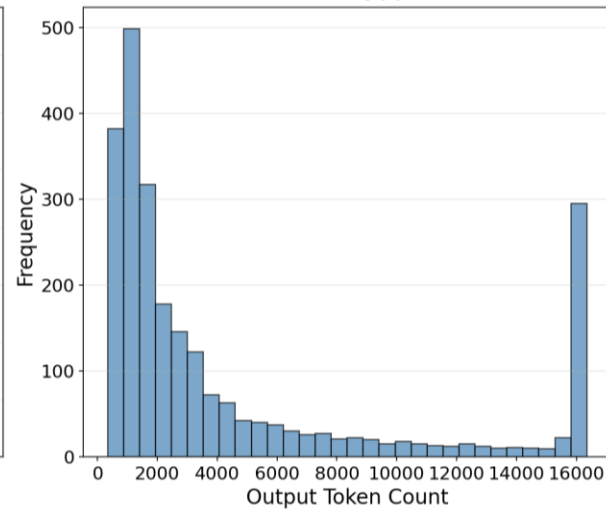


- コンテキスト長16kでトランケーションを行い、全データを利用（フィルタリングなし）
- 難易度の高いベンチマークほどトランケーションが多い

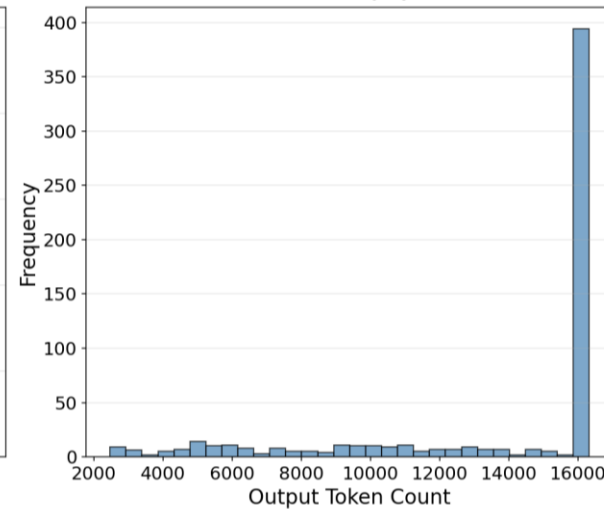
llm-jp-dev100



MATH500



AIME25



使用したデータセットとモデル

データセット名	用途	URL
MATH-500	評価	https://huggingface.co/datasets/HuggingFaceH4/MATH-500
aimo-validation-amc	評価	https://huggingface.co/datasets/AI-MO/aimo-validation-amc
aime24	評価	https://huggingface.co/datasets/math-ai/aime24
aime25	評価	https://huggingface.co/datasets/math-ai/aime25
aime26	評価	https://huggingface.co/datasets/math-ai/aime26
brumo_2025	評価	https://huggingface.co/datasets/MathArena/brumo_2025
hmmt_feb_2025	評価	https://huggingface.co/datasets/MathArena/hmmt_feb_2025
hmmt_feb_2026	評価	https://huggingface.co/datasets/MathArena/hmmt_feb_2026
mmlu_pro_eval_500subset	評価	https://huggingface.co/datasets/mlfoundations-dev/mmlu_pro_eval_500subset
Nemotron-CC-Math-v1	CPT	https://huggingface.co/datasets/nvidia/Nemotron-CC-Math-v1
ODA-Math-460k	SFT	https://huggingface.co/datasets/OpenDataArena/ODA-Math-460k
DeepScaleR-Preview-Dataset	SFT	https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset

モデル名	URL
Qwen2.5-72B-Instruct	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
gpt-oss-120b	https://huggingface.co/openai/gpt-oss-120b
Qwen3-30B-A3B	https://huggingface.co/Qwen/Qwen3-30B-A3B