

大規模言語モデルに基づく 複数の外部ツールを利用した推論フレームワーク

稲葉 達郎 清丸 寛一 Fei Cheng 黒橋 禎夫
京都大学

{inaba,kiyomaru,feicheng,kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

大規模言語モデルは広範な世界知識に加えて数値計算といった記号処理の能力も一定程度有しており、種々の推論タスクで驚異的なスコアを達成している。しかし、言語モデルが保持できる知識の量と実行できる記号処理の複雑さには限界がある。大規模言語モデルによる推論のさらなる高度化を目指して、本研究では大規模言語モデルによる推論時に検索器や電卓など複数の外部ツールの利用を可能にするフレームワークを提案する。提案手法を知識を要する数値推論タスク NumGLUE (Task 2) に適用し、強力なベースラインの性能を大幅に上回る、現時点での最高性能を達成することを確認した。

1 はじめに

推論は既知の事柄を手がかりに未知の事柄を推し量る論理的なプロセスである。推論では、言語理解の能力はもちろん、実世界に関する知識、数値計算をはじめとする記号処理の能力も必要となる。計算機による推論の実現は人工知能研究の古くからの目標であり、現在まで中心的な関心であり続けている [1, 2]。

昨今、大規模言語モデル (Large Language Model; LLM) [3, 4, 5, 6] が種々の推論タスクで驚異的なスコアを達成している。これは大規模言語モデルが事前学習を通じて基本的な言語理解の能力に加え、広範な世界知識、そして数値計算を含む記号推論の能力までもを一定程度獲得していることを示している。

しかし、大規模言語モデルであっても保持できる世界知識の量と実行可能な記号処理の複雑さには限界がある。この問題を解決するため、大規模言語モデルによる推論時に検索器や電卓などの外部ツールを呼び出し、その結果を利用する手法が提案されて

いる [7, 8]。しかし、先行研究はいずれも単一の外部ツールの利用に留まっており、推論を実行する上で課題となる知識の問題や記号処理の問題を同時に解決する枠組みにはなっていない。

本研究では、大規模言語モデルによる推論において、検索器や電卓といった複数の外部ツールの利用を可能にするフレームワークを提案する。図 1 に提案手法の概要を示す。提案手法では、大規模言語モデルが推論過程を含めて解答を生成するように入力テキスト (プロンプト) を設計する [9]。提案手法の鍵は、プロンプトの工夫によって、推論過程の一部として外部ツールを呼び出す文字列を生成させることにある。実行時は、大規模言語モデルが外部ツールを呼び出す文字列を生成したら推論過程の生成を中断し、生成中の推論過程から利用する外部ツールの名前とそれに与える入力を取り出し、外部ツールの実行結果を推論過程に追記する。その後大規模言語モデルに推論過程の生成を再開させ、外部ツールを呼び出す文字列が生成されたらまた上記の手続きを行う。

提案手法の有効性を確認するため、専門知識を要する数値推論タスク NumGLUE (Task 2) [10] のもとで実験を行った。実験の結果、提案手法が強力なベースライン手法の性能を大幅に上回り、現時点での最高性能を達成することを確認した。

2 関連研究

大規模言語モデルによる推論は、入力テキスト (プロンプト) を適当に設計し、大規模言語モデルから推論の解答を引き出すこと (プロンプティング [11]) で解かれる。特に、大規模言語モデルが解答を推論過程付きで生成するようにプロンプトを設計する Chain of Thought プロンプティング [9, 12] はその有効性が広く知られている。

大規模言語モデルによる推論時に外部ツールを利

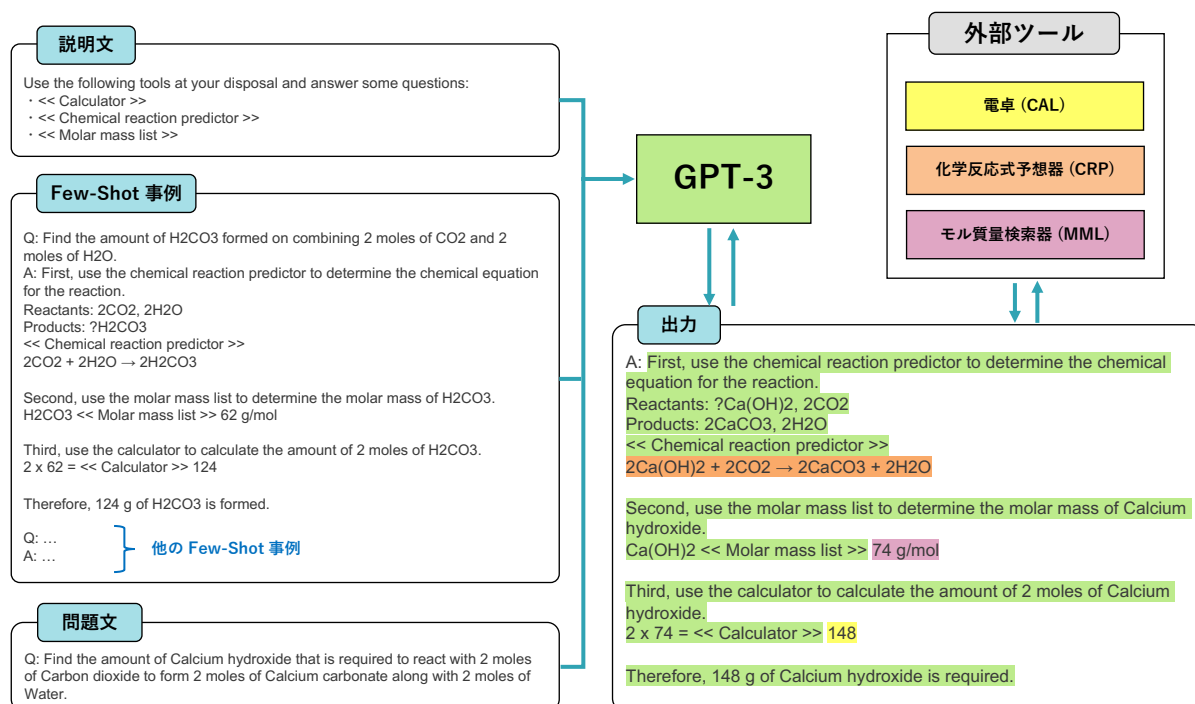


図 1 提案手法の全体像。出力中のハイライトは、黄緑色が GPT-3、黄色が電卓 (Calculator; CAL)、橙色が化学反応予測器 (Chemical reaction predictor; CRP)、紫色がモル質量検索器 (Molar mass list; MML) の出力であることを表す。

用する先行研究としては、推論過程の中に現れた数式を電卓で処理して計算結果を推論過程に埋め込むことで計算間違いの問題を緩和する取り組み [8] や、推論過程の生成中に知識検索を実行することで多段の推論を要する質問応答を実行する取り組み [7] がある。これらの先行研究はそれぞれ未熟な計算能力と不完全な世界知識という大規模言語モデルが抱える単一の問題に着目し、それに単一の外部ツールを利用することで対処している。本研究の特徴は、大規模言語モデルによる推論時に複数の外部ツールの利用を可能にすることで複数の問題に同時に対処することにある。

3 提案手法

本研究では、大規模言語モデルによる推論において、複数の外部ツールの利用を可能にする枠組みを提案する。図 1 に提案手法の概要を示す。提案手法では大規模言語モデルとして GPT-3 [3] を使用し、外部ツールを呼び出す文字列を含む推論過程が生成されるように Chain of Thought プロンプティング [9] を行う。

プロンプトは、利用可能な外部ツールを明示した説明文、いくつかの問題文と推論過程付きの解答のペア (Few-Shot 事例)、解かせる問題文を結合し

て構成する。Few-Shot 事例として与える推論過程には外部ツールの呼び出し方をアノテーションする。具体的には、<< 外部ツール名 >> の文字列を外部ツールを呼び出す文字列としてアノテーションし、その直前に外部ツールへの入力を特定のフォーマットでアノテーションする。例えば、外部ツールとして電卓 (Calculator) を利用する場合、Few-Shot 事例として与える推論過程の中で電卓を利用すべき箇所に << Calculator >> の文字列をアノテーションし、その直前に電卓に与える数式をアノテーションする。

推論時は GPT-3 がプロンプトに従い、与えられた問題文に対する解答を外部ツールを呼び出す文字列を含む推論過程と一緒に生成する。その時、>> の文字列が生成されたら生成を中断し、その直前の文字列から使用する外部ツールの名前とそのツールへの入力を抽出、外部ツールを実行した後、その結果を生成中の推論過程の末尾に追記する。これを新たな入力とし、GPT-3 による推論過程の生成を再開する。未知の外部ツール名が生成されたり、外部ツールの実行時エラーが発生したりすると外部ツールが実行結果が得られない。その場合は外部ツールの出力結果に相当する部分も GPT-3 自身に出力させる。

GPT-3 が生成した推論過程付きの解答から最終的

な解答を抜き出す処理もプロンプティングで行う。具体的には、推論過程とそこに含まれる最終的な解答のペアを Few-Shot 事例とするプロンプトを使用して、GPT-3 に推論過程から解答を抽出させる¹⁾。

4 知識を要する数値推論への適用

提案手法の適用例として、本研究では専門知識を要する数値推論に取り組む。具体的には、数値推論データセット NumGLUE [10] の Task 2 に取り組む。問題例を例 (1) に示す。

- (1) Find the amount of Calcium hydroxide that is required to react with 2 moles of Carbon dioxide to form 2 moles of Calcium carbonate along with 2 moles of Water.

NumGLUE の Task 2 は例 (1) のような化学の専門知識を要する数値推論の問題で大部分が構成される。これを解くには、言語理解能力、計算能力に加え、化学反応式の作成能力と物質のモル質量の知識が要求される。

本研究では以下の外部ツールを実装し、提案手法の枠組みで推論の実行時に利用することで本タスクを解く。

- **電卓 (Calculator; CAL)**: 四則演算を中心とする数式を入力として、その計算結果を出力する。計算結果は数式を Python の eval 関数²⁾ に入力して得る。数式中の演算子は Python の文法に合わせて入力前に置換する。電卓を呼び出す際は、<<Calculator>> の文字列を出力させる。入力の数式は同じ行に生成させる。
- **化学反応式予測器 (Chemical reaction predictor; CRP)**: 反応物と生成物の化学式を入力として、化学反応式を出力する。化学反応式は、反応物と生成物で各原子の数が同じになるように係数を調整して得る。化学反応式予測器を呼び出す際は、<<Chemical reaction predictor>> の文字列を生成させる。入力の反応物と生成物の化学式は直前の 2 行に出力させる。
- **モル質量検索器 (Molar mass list; MML)**: 化学式を入力として、そのモル質量を出力する。化学

1) 本研究の実験では、ほぼ常に最終的な解答が推論過程の最終行に含まれることが実験的に確認されたため、計算コスト削減のため、推論過程の最終行を抜き出した上でこの手法を適用した。

2) <https://docs.python.org/3/library/functions.html#eval>

表 1 NumGLUE の Task2 における性能。最良の結果を太字で示す。(†) は NumGLUE の論文 [10] からの引用。

手法	正解率
Zero-Shot†	1
Zero-Shot+CoT [12]	32.62
Few-Shot†	42
Few-Shot+CoT	57.85
Few-Shot+CoT+CAL	62.77
Few-Shot+CoT+CRP	64.31
Few-Shot+CoT+MML	69.23
Few-Shot+CoT+ALL (提案)	85.85

式のモル質量は、各原子の質量に関する知識に基づき、化学式に含まれる原子とその数から計算する。モル質量検索器を呼び出す際は、<<Molar mass list>> の文字列を生成させる。入力の化学式は同じ行に出力させる。

図 1 に推論を実行する際に用いるプロンプトを示す。説明文として、上記の 3 種類の外部ツールが利用可能なことが明示されたテキストを与える。Few-Shot 事例には、問題文と推論過程付きの解答のペアの例を与え、推論過程には上記の外部ツールを利用するためのアノテーションを付与する。

5 実験

提案手法の有効性を確認するため、知識を要する数値推論に提案手法を適用する。

5.1 データセット

数値推論データセット NumGLUE [10] の Task 2 を使用する。4 節で説明した通り、主に化学に関する専門知識を要する数値推論の問題で構成されている。問題の解答は全て数値である。評価にはテストデータの 325 問を使用する。評価指標は正解率である。

5.2 比較手法

本研究では GPT-3 (text-davinci-003; 175B パラメータ³⁾) を使用し、以下の手法を比較する。

Zero-Shot 問題文のみをプロンプトとしてテキスト生成を行い、その出力を予測とする。

Zero-Shot+CoT [12] 問題文に “Let’s think step by step.” を付け加えたものをプロンプトとしてテキスト

3) <https://beta.openai.com/docs/models/gpt-3>

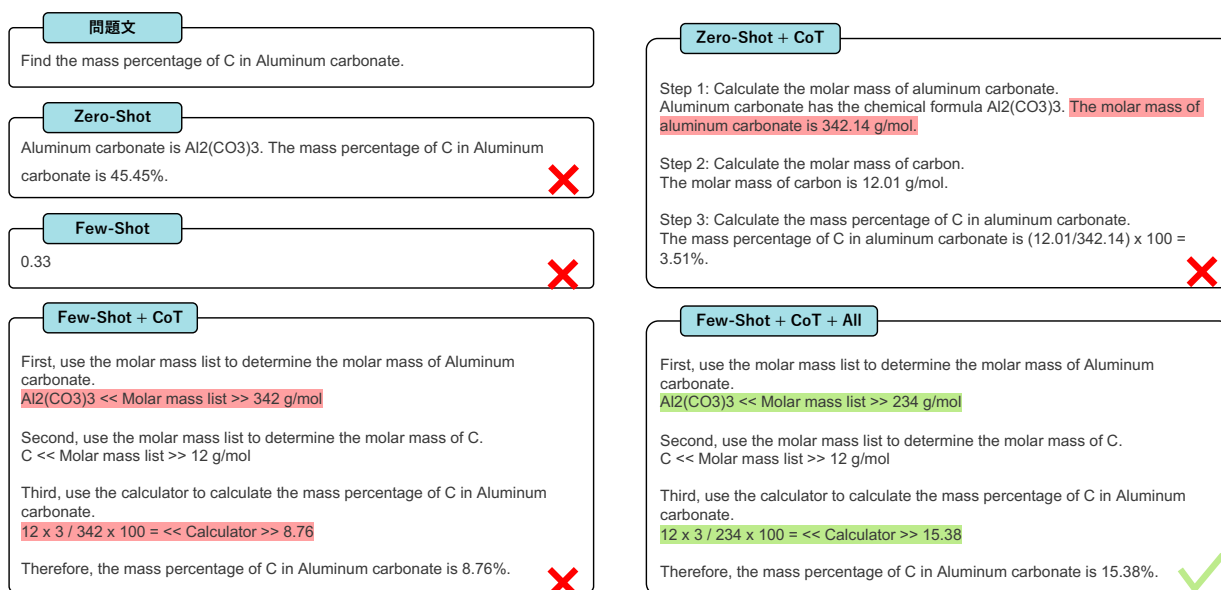


図2 手法ごとの出力の具体例。赤色のハイライトは誤った知識や誤った計算結果が生成された箇所を示す。緑色のハイライトは外部ツールの利用によってそれらが改善した箇所を示す。

ト生成を行い、推論過程を伴う解答を得る。最終的な予測は、生成された推論過程の末尾に“Therefore, the answer (arabic numerals) is”を追加し、それをプロンプトとしてテキスト生成を実行して得る。

Few-Shot 問題文の前に Few-Shot 事例として数例の問題文と解答のペアを結合したものをプロンプトとしてテキスト生成を行い、その出力を予測として得る。

Few-Shot+CoT 提案手法と同じプロンプトを使用してテキスト生成を実行し、その出力を予測として得る。ただし外部ツールは使用せず、推論過程は全て GPT-3 に生成させる。

Few-Shot+CoT+{CAL|CRP|MML} 提案手法と同じプロンプトを使用してテキスト生成を実行し、その出力を予測とする。推論時、電卓 (CAL)、化学反応式予測器 (CRP)、モル質量検索器 (MML) のいずれか一つの外部ツールを利用する。利用しない外部ツールの実行結果は GPT-3 自身に生成させる。

Few-Shot+CoT+ALL 提案手法。電卓、化学反応式予測器、モル質量検索器の全ての外部ツールを使用して推論を実行する。

Few-Shot 事例として、学習データ中の 20 問を使用した。Few-Shot 事例は、外部ツールのアノテーション数に偏りが出ないように人手で問題を選択し、推論過程のアノテーションも人手で行った。Few-Shot 事例の数が性能に与える影響については付録 A を参照されたい。

5.3 結果

表 1 に結果を示す。提案手法は 85.85 ポイントの正解率であり、現時点での最高性能を達成した。外部ツールを利用しない手法はもちろん、単一の外部ツールを利用する手法と比べても、推論性能の大きな改善が確認された。

6 ケーススタディ

図 2 に提案手法による改善例を示す。CoT プロンプティングを使用しない Zero-Shot と Few-Shot は推論に失敗している。CoT プロンプティングに基づく Zero-Shot+CoT と Few-Shot+CoT は $\text{Al}_2(\text{CO}_3)_3$ のモル質量に誤りがあり、Few-Shot+CoT は $12 \times 3 / 342 \times 100$ の計算も間違っている。提案手法の Few-Shot+CoT+ALL は外部ツールに頼ることで正しいモル質量と数値計算に基づき正答できている。

7 おわりに

本研究では大規模言語モデルによる推論において、検索器や電卓等の複数の外部ツールの利用を可能にする枠組みを提案した。本論文ではその枠組みを化学の専門知識を要する数値推論タスクに適用し、その高い有効性を確認した。提案の枠組みは汎用的であり、外部ツールを変更・拡張することで様々なタスクに応用可能である。今後、異なるタスクにおいても提案手法の有効性を検証したい。

謝辞

この研究は JST、CREST、JPMJCR20D2 と厚生労働科学研究費補助金 AC 事業 JPMW21AC5001 の支援を受けたものである。

参考文献

- [1] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In ***SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)**, pp. 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.
- [2] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts**, pp. 27–33, Online, July 2020. Association for Computational Linguistics.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [4] Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In **Advances in Neural Information Processing Systems**, 2022.
- [5] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models. **arXiv preprint arXiv:2205.01068**, 2022.
- [6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. **arXiv preprint arXiv:2204.02311**, 2022.
- [7] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. **arXiv preprint arXiv:2210.03629**, 2022.
- [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.
- [9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In **Advances in Neural Information Processing Systems**, 2022.
- [10] Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3505–3523, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [11] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. **ACM Computing Surveys**, 2022.
- [12] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In **Advances in Neural Information Processing Systems**, 2022.

表 2 NumGLUE の Task 2 における Few-Shot 事例の数を
変えたときの性能。最良の結果を太字で示す。

	事例数	正解率
CoT	5	55.38
CoT	10	56.31
CoT	20	57.85
CoT+ALL	5	83.69
CoT+ALL	10	84.00
CoT+ALL	20	85.85

A Few-Shot 事例の数が性能に与える影響

Few-Shot 事例の数が性能に与える影響を調べた。表 2 に結果を示す。Few-Shot 事例の数を増やすことで、外部ツールの利用の有無に関わらず、精度が向上することが確認された。なお、Few-Shot 事例は 20 例の時点で合計のトークン数が 3,000 近くあり、大規模言語モデルが処理可能なトークン長の制約から、Few-Shot 事例を増やすことによるこれ以上の性能改善は困難である。