

LLM の学術ドメイン適応のための 合成データに基づく統合フレームワーク

小川隼斗¹ 河原大輔^{1,2,3} 相澤彰子^{2,3}

¹ 早稲田大学理工学術院 ² 国立情報学研究所 ³ 国立情報学研究所 LLM 研究開発センター
{cookie3120@ruri., dkw@}waseda.jp, aizawa@nii.ac.jp

概要

大規模言語モデル (LLM) の様々な分野での応用が広がる中、専門知識を要する学術ドメインでの活用は難しい課題である。特に、日本語の学術ドメインにおける LLM の開発は未だ発展途上である。本研究では、LLM が生成した QA データを活用し、学習と自動評価を統合したフレームワークを提案する。まず、学術論文を対象とした合成データセットを作成する。次に、そのデータセットを用いて LLM をチューニングし、学術的な文脈での質問応答能力を強化する。さらに、LLM による自動評価手法により開発したモデルの性能と有効性を検証する。

1 はじめに

近年、LLM の発展が著しく、その応用範囲は多岐に渡る。LLM は大規模なテキストコーパスで学習することによって、汎用的で高度な推論能力を持ち、翻訳や要約、対話システムなど様々なタスクで活用されている。LLM の応用として、産業や学術ドメインへの適応が重要な課題となっている。本研究では、研究者による学術論文の執筆や理解を支援することを目的に学術ドメインへの適応に焦点を当てる。英語を対象とした学術ドメインに特化した LLM としては、AcademicGPT [1] や Galactica [2] が開発されている。一方、日本語の学術ドメインにおいては、Academic BART [3] や AcademicRoBERTa [4] といった小規模なモデルが開発されているものの、LLM は未だ開発されていない。

本研究は、日本語の学術ドメインの学習、評価用 QA データセットを LLM で合成し、それを用いた学習と自動評価で構成される統合フレームワークを構築する。具体的には、学術ドメインに特化した QA カテゴリを設定する。そして、既存の LLM を活用して、設定したカテゴリに対応する日本語の合成

データセットを作成する。このデータセットを用いて LLM をチューニングすることで、日本語の学術的な文脈での質問応答の性能向上を図る。さらに、開発したモデルの評価には、LLM-as-a-Judge [5] を用いた自動評価を導入し、日本語学術ドメインにおける有効性を検証する。

2 関連研究

本節では、LLM による学習データの自動生成と LLM を用いた自動評価について述べる。

2.1 合成データによる LLM の学習

LLM の学習に必要なデータを自動生成する方法として Self-instruct [6] がある。この手法により、人手でデータを作成する労力を大幅に削減しながら、LLM の性能向上を実現できることを示した。この手法は Alpaca [7] にも応用されている。また、特定ドメインのデータを合成することでそのドメインにおける性能が上がることも検証されている [8, 9]。

2.2 LLM-as-a-Judge

LLM-as-a-Judge は、LLM が他の LLM の出力を評価する手法であり、人手での評価と比較して効率的かつ一貫性のある評価が可能であることから、広く利用されている。日本語のベンチマークでは Japanese MT-bench¹⁾ や Japanese Vicuna QA Benchmark [10] において、この手法が採用されている。評価手法は大きく分けて、評価対象モデルの回答に点数を付ける絶対評価と、複数のモデルの回答を比較するペアワイズ評価の 2 種類がある。この手法は高速でコスト効率が高く、評価結果の再現性が高い一方で、position bias や length bias, self-enhancement bias などの評価用 LLM に起因する潜在的なバイアスが報告されている [11]。本研究では、学習したモデルの性

1) https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge/data/japanese_mt_bench

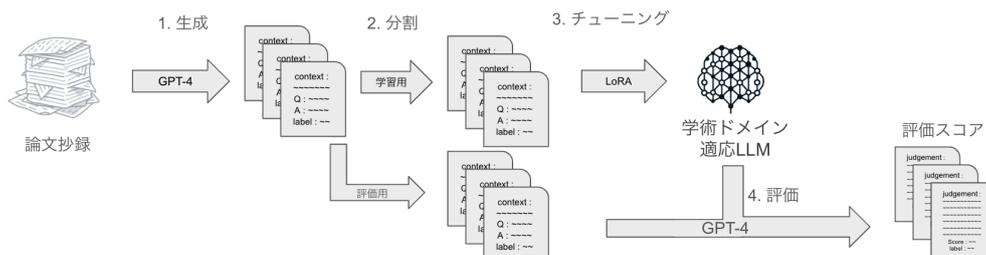


図 1: 統合フレームワークの概略図

能を一貫した基準で評価するため、絶対評価を採用する。

3 提案する統合フレームワーク

本研究で提案する統合フレームワークは、学術ドメインに特化した LLM の開発を効率的に行うための一貫したパイプラインである。図 1 に示すように、このフレームワークは以下の 4 つのステップで構成される。

1. 学術ドメインに特化した QA カテゴリを設定し LLM を用いて学術論文の抄録から QA データセットを生成。
2. QA データセットを学習用と評価用に分割。
3. 学習用データセットで LLM をチューニング。
4. モデルの性能を、LLM-as-a-Judge で評価。

ステップ 1, 2 の詳細を 4 節で、ステップ 3, 4 の詳細を 5 節で述べるとともに、国立研究開発法人科学技術振興機構が運営する電子ジャーナルプラットフォームである J-STAGE²⁾ で公開されている論文の抄録を用いた統合フレームワークの実験についても述べる。

4 QA データセットの作成

本研究では学術ドメインに特化した QA データセットを作成する。データセットは論文の抄録とそれに関する質問および回答のペアで構成され、質問のタイプを 6 つのカテゴリに分類する。これらのカテゴリと QA データは、LLM を用いて自動生成する。

4.1 カテゴリの定義

学術ドメインでは、研究の目的、提案手法、得られた知見など、様々な要素について深い理解と議論が求められる。このような学術的な問いを網羅的かつ体系的に生成し、データセットの質的多様性を担

保するため、質問のカテゴリ体系を構築する。具体的には、Lehnert が提案した質問応答の 13 カテゴリ [12] を基に、学術的な文脈に適した以下の 6 つのカテゴリを定義する。

- Goal Orientation: 論文の目的がどのようなものかを問うカテゴリ
- Instrumental: 論文で用いられた手法について問うカテゴリ
- Causal antecedent: 論文の実験等の結果の要因などについて問うカテゴリ
- Causal consequent: 論文の実験等の結果について問うカテゴリ
- Quantification: 論文で示されている数値について問うカテゴリ
- Verification: 論文の記述内容に対する解釈の正誤を問うカテゴリ

4.2 QA データセットの合成手順

QA データセットの合成は few-shot [13] プロンプトを用いた LLM で行う。そのために、各カテゴリごとに 1 件ずつ人手で QA データの例を作成する。作成する質問はベースとなる論文抄録中に回答の根拠が含まれる長文であり、回答は Quantification, Verification が短くその他は長文で構成され多様性を確保している。作成した QA データの例を表 1 に示す。プロンプトの詳細は付録 A に示す。上記のプロンプトとともに J-STAGE で公開されている論文の抄録を GPT-4 に入力し、QA データとそのカテゴリを生成する。合成したデータの例を表 2 に示す。

- 3) https://www.jstage.jst.go.jp/article/adrjssd/17/1/17_KJ00008113760/article/-char/ja/
- 4) https://www.jstage.jst.go.jp/article/agcjchikyukagaku/52/2/52_KJ00004411197/article/-char/ja/
- 5) <https://www.jstage.jst.go.jp/article/jjtp1987/7/1/7.1.14/article/-char/ja/>
- 6) <https://www.jstage.jst.go.jp/article/jsnfs/70/3/70.101/article/-char/ja/>

2) <https://www.jstage.jst.go.jp/>

表 1: 人手で作成した QA データの例

カテゴリ	論文抄録	QA
Goal Orientation	... 重粒子線治療システムは、体内のがん細胞に放射線のひとつである重粒子線を照射し、死滅させる治療法である。がん細胞のみを狙い打ちできるので、他の健常細胞を傷つけることが無い。また、... ³⁾	Q: 重粒子線治療システムはなぜ他の健常細胞を傷つけることが無いのでしょうか？ A: がん細胞のみを狙い打ちして重粒子線を照射させることができるため
Causal Consequent	... これは、浅海域のバー（海底砂州）が波に対してフィルター的作用を果たし、また、ステップが汀線直前の水深を規定するので、汀線砕波の波高がある限度を越えることがなく、その結果、暴浪時の波がある限界の地点、すなわち後浜上限を越えて遡上することがないためである... ⁴⁾	Q: 浅海域のバーが波に対してフィルター的作用を果たし、また、ステップが汀線直前の水深を規定することにより日本海沿岸の後浜上限高度はどのような影響を受けていますか？ A: 暴浪時の波が後浜上限を越えて遡上することがなく、後浜上限の位置は暴浪時の波の規模がさまざまであるにもかかわらず安定している。

表 2: GPT-4 により合成した QA データの例

カテゴリ	論文抄録	QA
Instrumental	... 組成および物性が積層方向に変化する傾斜機能材料に対しては観測された試料裏面温度変化曲線と熱物性値の関係が明らかではなかった。レーザフラッシュ法において試料表面を瞬間的に均一加熱した後の試料裏面温度変化のみならず試料表面の温度変化および試料裏面加熱後の試料表面と試料裏面の温度変化を観測し、それらを... ⁵⁾	Q: レーザフラッシュ法において傾斜機能材料の平均熱伝導率、平均熱容量、および有効熱拡散率を求めるためにはどのようなデータが必要ですか？ A: 試料表面を瞬間的に均一加熱した後の試料裏面温度変化、試料表面の温度変化、および試料裏面加熱後の試料表面と試料裏面の温度変化
Causal Antecedent	... さらに、ヒト小腸型 ALP 遺伝子の mRNA 発現量も、MK-4 濃度 1 μ M で 0 μ M と比較し有意に高値を示した。本研究において、ヒト小腸上皮様細胞での MK-4 による小腸型 ALP 発現の増強作用について初めて示すことができた... ⁶⁾	Q: メナキノン-4 を Caco-2 細胞に添加した際、アルカリホスファターゼの活性が高値を示したのはなぜですか？ A: MK-4 による小腸型 ALP 発現の増強作用があるため

表 3: カテゴリ別のデータ件数

Goal Orientation	1,822
Instrumental	1,786
Causal antecedent	1,745
Causal consequent	1,428
Quantification	374
Verification	239

4.3 QA データセットの内訳と分割

QA データは論文抄録 7,394 件それぞれにつき 1 件生成した。各カテゴリごとのデータの内訳を表 3 に示す。これらを学習用と、評価用のデータに分割を行う。6つの各カテゴリそれぞれから 10 件ずつ評価用のデータを人手で選定する。評価用に選ばれた 60 件のデータを除いた 7,334 件を学習に使用する。

5 チューニングと評価

4 節で作成した QA データセットを用いて LLM のファインチューニングを行う。その後、LLM による自動評価で性能を検証する。

5.1 実験設定

本節では LLM の学習設定および、評価の設定を述べる。

5.1.1 学習の設定

llm-jp⁷⁾ が公開している 2 種類のモデル、llm-jp-3-13b⁸⁾ (以下「事前学習モデル」という) 及び llm-jp-3-13b-instruct⁹⁾ (以下「インストラクションモデル」という) を LoRA チューニング [14] する。LoRA チューニングはフルパラメータのファインチューニングと比較して計算時間やメモリ使用量を大幅に削減できる。学習率は 1e-5, 2e-4, 3e-4, 4e-4 の 4 種類で実験する。より詳細な学習設定は付録 C に示す。

5.1.2 評価の設定

GPT-4o による自動評価を行う。評価用データセット 60 問のそれぞれについて、論文抄録、質問、GPT-4 の回答、モデルの回答の 4 つとともに評価用プロンプトを GPT-4o に入力して評価する。GPT-4o は回答の有用性、関連性、正確性、深さ、詳細度を考慮し、絶対評価で 1 から 10 までのスコアを回答につける。GPT-4 の回答は参考として用いられる。評価用プロンプトは付録 B に示す。

表 4: GPT-4o による評価の例

論文抄録	QA	評価	スコア
... 組成および物性が積層方向に変化する傾斜機能材料に対しては観測された試料裏面温度変化曲線と熱物性値の関係が明らかではなかった。レーザフラッシュ法において試料表面を瞬間的に均一加熱した後の試料裏面温度変化のみならず試料表面の温度変化および試料裏面加熱後の試料表面と試料裏面の温度変化を観測し、それらを...	Q: レーザフラッシュ法において傾斜機能材料の平均熱伝導率、平均熱容量、および有効熱拡散率を求めるためにはどのようなデータが必要ですか? A: 試料表面の温度変化および試料裏面加熱後の試料表面と試料裏面の温度変化を観測するデータが必要です。	AI アシスタントの回答は、必要なデータの一部を正確に述べていますが、試料表面を瞬間的に均一加熱した後の試料裏面温度変化についての言及が欠けています。GPT-4o の回答は、必要なデータをすべて網羅しており、より完全です。AI アシスタントの回答は不完全であるため、評価は低くなります。	6

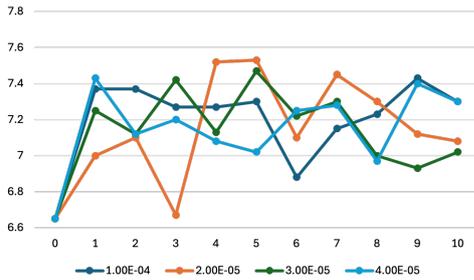


図 2: llm-jp-3-13b の学習エポック毎のスコア推移

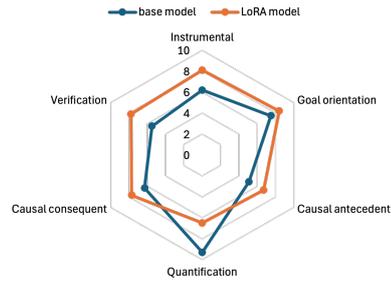


図 4: カテゴリごとのスコアの比較 (llm-jp-3-13b)

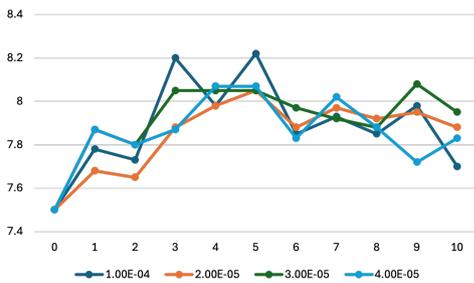


図 3: llm-jp-3-13b-instruct の学習エポック毎のスコア推移

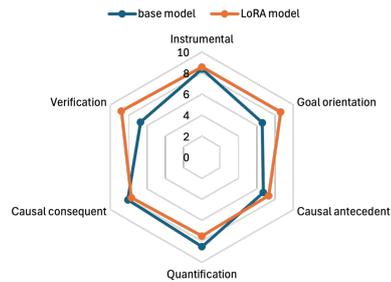


図 5: カテゴリごとのスコアの比較 (llm-jp-3-13b-instruct)

5.2 実験結果

GPT-4o による評価結果の例を表 4 に示す。事前学習モデル、インストラクションモデルそれぞれの学習エポック毎のスコア推移を図 2, 3 に示す。なお、学習率による明確な違いは認められなかった。ベースモデルと最高スコアを記録した LoRA モデルのカテゴリ別スコアを図 4, 5 に示す。事前学習モデルでは Quantification を除く全カテゴリでスコアが向上し、インストラクションモデルでは Goal Orientation と Verification のみスコアが向上した。

6 おわりに

本研究では、LLM が生成した QA データを用いた学習と、LLM による自動評価を統合したフレームワークを提案し、学術ドメインへの適応を試みた。このフレームワークは、他の専門知識の求められる分野への応用も考えられ、各分野に特化した日本語 LLM の効率的な開発への貢献が期待される。今後の課題として、合成データの量と質、多様性のさらなる向上、多様な観点やデータセットでの評価が挙げられる。さらに、他のドメインへの適用可能性や、大規模データセットでの検証も行い、学術ドメインにおける日本語 LLM の実用性を高めていくことが求められる。

7) <https://llm-jp.nii.ac.jp/>
 8) <https://huggingface.co/llm-jp/llm-jp-3-13b>
 9) <https://huggingface.co/llm-jp/llm-jp-3-13b-instruct>

謝辞

本研究は文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた。

参考文献

- [1] Shufa Wei, Xiaolong Xu, Xianbiao Qi, Xi Yin, Jun Xia, Jingyi Ren, Peijun Tang, Yuxiang Zhong, Yihao Chen, Xiaolin Ren, Yuxin Liang, Liankai Huang, Kai Xie, Weikang Gui, Wei Tan, Shuanglong Sun, Yongquan Hu, Qinxian Liu, Nanjin Li, Chihao Dai, Lihua Wang, Xiaohui Liu, Lei Zhang, and Yutao Xie. Academicgpt: Empowering academic research, 2023.
- [2] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.
- [3] Hiroki Yamauchi. Academicbart: A bart-based japanese masked language model pretrained on academic paper abstracts, 2023. Accessed on December 22, 2024.
- [4] Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. A Japanese masked language model for academic domain. In **Proceedings of the Third Workshop on Scholarly Document Processing**, pp. 152–157, 2022.
- [5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.
- [6] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [7] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [8] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities, 2024.
- [9] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture, 2024.
- [10] Yikun Sun, Zhen Wan, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. Rapidly developing high-quality instruction data and evaluation benchmark for large language models with minimal human effort: A case study on japanese. In **The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, 2024.
- [11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 2024.
- [12] Wendy G. Lehnert. A conceptual theory of question answering. In **Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'77**, p. 158–164, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.
- [13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, 2020.
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

A データ合成用プロンプトの詳細

事前実験により全カテゴリの例を1件ずつ入力すると、Goal orientation と Causal antecedent がカテゴリとして選ばれやすいことがわかっている。プロンプトには Goal Orientation を除いた5カテゴリか、Causal antecedent を除いた5カテゴリか、6カテゴリ全てのいずれかが等確率で例として選ばれる。

B 評価用プロンプトの詳細

評価に使用したプロンプトの全文を以下に示す。

評価用プロンプト

公平な判断者として行動し、以下に表示される context とそれに関する質問に対する AI アシスタントの応答の品質を日本語で評価してください。あなたの評価は、応答の有用性、関連性、正確性、深さ、詳細度などの要素を考慮すべきです。また、参考として GPT-4 の回答も提供します。ただし GPT-4 の回答が 10 点満点の回答とは限りません。評価は短い説明から始めてください。できるだけ客観的であること。説明を提供した後、このフォーマットに厳密に従って 1 から 10 までのスケールで応答を評価する必要があります。出力のフォーマットは json 形式であり {"judgement": "", "score": } です。
context:{Context}, 質問:{instruction}, GPT-4 の回答:{responceByGPT-4}], AI アシスタントの回答:{responce}

C 学習設定

LoRA のランク:8

LoRA のスケール係数:32

ドロップアウト率:0.05