

JMedLoRA : Instruction-tuning による日本語大規模モデルの医療ドメイン適用

助田一晟¹ 鈴木雅弘² 坂地泰紀³ 小寺聡¹

¹ 東京大学医学部附属病院 ² 東京大学大学院工学系研究科

³ 北海道大学 大学院情報科学研究院

{sukeda-issei006, msuzuki}@g.ecc.u-tokyo.ac.jp sakaji@ist.hokudai.ac.jp
koderasatoshi@gmail.com

概要

大規模言語モデル (LLM) の波及効果が続く中、その医療ドメイン適応は重要な研究課題となっている。近年 LLM の調整には Instruction-tuning が多用されるが、ドメイン適応におけるその具体的な効果は明らかにされていない。本研究では、日本語での LoRA ベースの Instruction-tuning を実施し、その性能を医療質問応答タスクを通じて多面的に評価した。本実験により、英語の LLM を出発点とした Instruction-tuning によってドメイン固有の知識を一部 LLM に組み込むことができ、大きなモデルほど効果が顕著であることが示唆された。この取り組みは、医療機関が外部サービスに頼らずに LLM 構築する先駆的な試みとして位置付けられる。

1 はじめに

ChatGPTをはじめとする大規模言語モデル (LLM) の研究と開発は、医療およびヘルスケアの分野において変革をもたらす可能性を秘めている。これらのモデルが医療領域に適応された場合、疾患の診断、治療計画、患者ケアなど、多くの重要な業務で医療専門家を支援できる。特に、LLM は広範な言語理解能力を有しているため、最新の情報を提供し、エビデンスに基づく治療オプションを提案し、さらには高い精度で疾患の結果を予測できるようになる可能性がある。

ChatGPT¹⁾はその驚くべき文章生成性能によって、テキストや言語を通じた人間と機械のやり取りのあり方そのものを大きく変革したと言える。しかしながら、そのような LLM の研究開発の本流は汎用目的であり、依然ドメイン適用という技術は LLM の

実用化を目指すにあたり重要なトピックである。確かに汎用モデルは 0-shot 推論と呼ばれる未知のタスクへの遂行力に長けるが、ファインチューニングされた特化型モデルがそれらに特定のドメインやタスクに限ればより高いパフォーマンスを発揮できる余地は十分にある。言語の問題はその一例である。英語圏では医療分野に特化した LLM が複数登場している [1, 2, 3] 一方、日本語モデルに目を向けると (筆者の知る限り) BERT をベースとした医療言語モデルを構築した杉本ら [4] のみであり、日本語医療 LLM の開発は不十分である。

加えて、データ侵害や機密患者情報の誤用などの潜在的なリスクがあり、医療現場における ChatGPT の利用はプライバシーやセキュリティの観点から現状困難が多い。したがって、医学の知識を組み込んだ LLM を現場利用するためには、他の LLM を用いたドメイン適応を検討する必要がある。近年、限られたパラメータのみをファインチューニングの対象とする Low Rank Adaptation (LoRA) [5] およびその量子化バージョン (QLoRA) [6] を含むいくつかのパラメータ効率の良いファインチューニング手法が提案されている。これらの手法を用いた Instruction-tuning [7] は、対話能力の獲得において一定の成功を示している [8]。ただし、LoRA ベースの Instruction-tuning の効果と限界は、ドメイン適応においてはまだ明確にされていない。この点については例えば Superficial Alignment Hypotheses が提唱され、ファインチューニングが知識の獲得に重要な寄与をしないと主張されている [9] が、依然議論中の課題である。そこで本研究では、LoRA ベースの Instruction-tuning を実施し、ドメイン固有の知識を獲得するのに効果的であるかを調査することを目標とする。

1) <https://chat.openai.com/>

本研究の主要な問いは

1. LoRA ベースのファインチューニングによってどのようにして、そしてどれだけの程度で、ドメイン知識を LLM に組み込むことができるか。
2. モデルサイズの大きな英語中心の LLM は、小さな日本語中心の LLM よりも優れているか。
3. ファインチューニングの量は重要か。

の三つであり、これらを明らかにすることを目指し本研究では独自の日本語医療データセットを用いて異なる LLM を比較する。各モデルの性能は医療質問応答タスクによって評価する。

2 手法

本実験では 2 種類の LLM をベースモデルとし、日本語医療データセットでチューニングを実施する。学習後のモデルは <https://huggingface.co/AIgroup-CVM-utokyohospital> で公開されている。

2.1 ベースモデルの選択

本研究は日本語 LLM の構築を目的としているため、最も性能の良い日本語 LLM として OpenCALM-7B²⁾ を利用する³⁾。OpenCALM-7B はサイバーエージェント社により公開された約 65 億パラメータを持つオープンソースモデルである。

次に、OpenCALM-7B に少量の医療テキストデータ (学習に 2420 件、評価に 50 件) を用いて事前学習を追加したモデル MedCALM を用意した。

最後に強力な英語のベースモデルとして Llama2-70B-chat-hf [10]⁴⁾ を利用した。このモデルは約 700 億パラメータを持つ Meta 社が開発した英語 LLM である。以後 Llama2-70B と略記する。

2.2 Instruction-tuning について

Instruction-tuning [7] は、ファインチューニング時に明示的な指示をプロンプトテキストとして入力することにより、モデルの挙動と出力を微調整または最適化する手法を指す。近年、LLM 向けに複数のファインチューニングの実装が提供されている。中でも Low Rank Adaptation (LoRA) [5] はパラメータ効率の良いファインチューニング手法の一つであり、PEFT ライブラリ [11] で提供されている。LoRA

にモデル量子化を取り入れた手法は QLoRA [6] として提供されており、これを用いることで計算リソースが限られている中であっても 700 億パラメータのように大規模なモデルを調整することが可能である。本研究では、LoRA を OpenCALM-7B に適用し、QLoRA を Llama2-70B に適用した上で、その性能比較を実施する。

3 実験

各ファインチューニング手法を各ベースモデルに適用し、生成された応答の正確性と医学的な正確さを検証した。すべての実験は、NVIDIA A100 4GPU (各 80GB VRAM) で実施された。

3.1 学習

Instruction-tuning を実施するために、ChatGPT (gpt-3.5-turbo) を利用し、77422 件の instruction 形式の医療質問応答データセットを構築した。ChatGPT へ入力する医療情報のソースとして、日本循環器学会のガイドラインおよび日本内科学会雑誌 (JJSIM) の記事を利用した。Instruction データ生成には以下のプロンプトを利用した。

Instruction データ生成プロンプト

```
###指示:あなたは質問文と回答文を色々作成する機械です。以下の入力を事前知識として理解して、質問 (instruction) と回答 (output) のペアを持つデータを作成してください。{'instruction': 質問内容, 'output': 回答内容} という形式でデータを作成し、改行は含めないでください。以上を 15 回繰り返し、1 行に 1 つのデータを記載してください。

###入力: {input_text}

###回答:
```

LoRA/QLoRA チューニング時のプロンプトには Alpaca プロンプト [12] を利用し、ステップ数は全体的な計算時間がおおよそ揃うように設定した。その他 LoRA/QLoRA のハイパーパラメータ設定は表 1 に従った。なお、input length, target max length, batch size はそれぞれ 512, 512, 8 に設定し、その他のパラメータはライブラリのデフォルトに従った。

2) <https://huggingface.co/cyberagent/open-cal-m-7b>

3) 実験実施時点。

4) <https://huggingface.co/meta-Llama/Llama-2-70b-chat-hf>

表1 LoRA/QLoRA のハイパーパラメータ設定一覧

base model	OpenCALM-7B	Llama2-70B
fine-tuning method	LoRA	QLoRA
learning rate	5e-5	2e-4
input length	512	512
target max length	512	512
batch size	8	8
fine-tuning steps	3k, 10k	0.9k, 3k
r of (Q)LoRA	8	64
α of (Q)LoRA	32	16
dropout rate of (Q)LoRA	0.05	0.1
target parameter	query_key_value	all linear layers

表2 日本語医療質問応答タスクにおける性能比較。0s と 1s はそれぞれ 0-shot 推論と 1-shot 推論を意味する。太字は各評価指標で最もスコアが高い部分を強調する。

LoRA ステップ数	OpenCALM-7B			MedCALM		
	0	3k	10k	0	3k	10k
ゲシュタルト (0s)	0.033	0.141	0.120	0.032	0.116	0.085
精度 (0s)	0.170	0.166	0.193	0.185	0.240	0.183
完全一致 (0s)	0	0.014	0.013	0	0.019	0.014
ゲシュタルト (1s)	0.053	0.087	0.078	0.028	0.002	0.035
精度 (1s)	0.177	0.148	0.174	0.164	0.150	0.165
完全一致 (1s)	0	0.059	0	0.001	0	0
訓練時間 (h)	-	24.0	37.0	-	23.7	58.4

3.2 医療質問応答タスクによるモデル出力評価

次に, Alpaca プロンプト [12] (OpenCALM-7B と MedCALM 用には和訳済のもの) を用いて LLM の文章生成を行うことで, 質問応答タスクによる評価を行った。ただし, {instruction} には評価データの質問文を, {input} には選択肢を入力した。1 shot 推論時には入力と応答の例を 1 つプロンプトに含めた。実装には transformers ライブラリ [13] を用い, 生成時のハイパーパラメータは temperature=0.1, top-p = 0.9, repetition_penalty = 1.05 と設定した。

OpenCALM-7B 用プロンプト

以下は, タスクを説明する指示と, 文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:

{instruction}

入力:

{input}

応答:

評価データとしては, 医療領域における質問応答パフォーマンスを評価するために, IgakuQA データセット [14] と JJSIMQA の 2 つの医学 Q&A データセットを利用した。IgakuQA は 2018 年から 2022 年まで 5 年分の日本医師国家試験データセットである。JJSIMQA は, 我々が本研究で JJSIM の巻末から収集した 5 択問題からなるデータセットである (サンプル例は Appendix A を参照)。各モデルの出力応答を多面的に評価するため, 3 つの評価指標を考案・利用した。まず, 「ゲシュタルトスコア」は, 応答と正答との間のゲシュタルト距離の平均値である。ゲシュタルト距離は最長共通部分列に基づく

QLoRA ステップ数	Llama2-70B		
	0	0.9k	3k
ゲシュタルト (0s)	0.071	0.276	0.287
精度 (0s)	0.170	0.251	0.244
完全一致 (0s)	0.001	0.180	0.169
ゲシュタルト (1s)	0.247	0.331	0.314
精度 (1s)	0.200	0.258	0.225
完全一致 (1s)	0.097	0.200	0.173
訓練時間 (h)	-	12.7	42.4

文字列マッチングアルゴリズムによって計算される。「精度」は, モデルの応答に (ゲシュタルト距離の意味で) 最も近い選択肢をモデルの最終回答とみなし, 正解率を計算した値である。最後に, 「完全一致」は, 正答を含む応答の割合と定義する。これは複数選択肢を選ばせる問題や, モデルが余分な出力を付与してしまった場合も正解と認めるものである。これら 3 種類の評価指標は 0 から 1 の値を取り, 値が大きいほどモデルのパフォーマンスが良いことを指す。

本実験の結果は表 2 に示されている。各行の最も高いスコアが太字で強調されている。

4 考察

4.1 Instruction-tuning による効果の定量的評価

表 2 より, 特に Llama2-70B では著しいスコアの向上が見られ, LoRA による顕著な改善が示唆された。これは, 日本語で事前学習されたモデルを出発点とするよりも, 英語で事前学習された強力なモデルをベースモデルとして使用の方が日本語タスクの文脈においてもより有望であることを示唆している。

また, Instruction-tuning においてステップ数の設定は議論的であるが, 我々の結果からは 1000 ス

テップ (約 1 エポック) 前後の Instruction-tuning がパフォーマンスを向上させる一方で、ステップ数を増やすと性能が悪化することが観察された。

さらに、MedCALM の性能悪化により、追加の事前学習がパフォーマンス向上に寄与しなかったことが確認された。以上から、追加の事前学習を行わずに LoRA ベースの Instruction-tuning を約 1 エポック実施することが、限られたトレーニングデータで LLM 構築を行う場合には実用的で有望なアプローチであると結論づけられる。

4.2 1-shot 推論能力の喪失

表 2 の OpenCALM-7B と、それが元となった MedCALM の結果を比較する。元々の OpenCALM-7B (すなわち 0step) のスコアは 1-shot 推論により 0-shot 推論時と比較して向上しているが、その他のモデルでは 1-shot 推論時の方がスコアが低下していることが確認される。この結果は LoRA チューニングおよび追加の事前学習によって、元々有していた 1-shot 推論の能力が失われてしまったことを示唆している。対照的に、Llama2-70B では QLoRA チューニング後のモデルにおいても 1-shot 推論はスコア向上に寄与しており、1-shot 推論能力が失われていない。

4.3 LLM の評価指標について

LLM の評価指標は近年積極的に議論されているが、依然確立された「最善の評価方法」は存在していない。既存のスコア指標 (例えば JGLUE [15]) やリーダーボード (例えば Nejumi LLM leaderboard ⁵⁾) は生成されたテキストの流暢さを評価できるが、特定のドメイン知識の正確性を十分に評価することができない。本研究では医療ドメイン知識の正確性を評価するべく質問応答タスクの出力に対し 3.2 節で述べたように 3 種類の評価指標を考案、利用したが、これらにも欠点が存在する。第一に、LLM を評価するために複数選択問題を使用することが妥当であるかは論争的となっており、今後も検討を要する [16, 17]。また、「完全一致」は、意味は正確であるがテキストが厳密に一致しない応答を正解とみなすことができない。一方、「ゲシュタルトスコア」は非対称であり、複数の選択肢を生成出力する際に妥当な評価が難しい。以上を踏まえ、より優れた評価指標の検討が必要である。

5) <http://wandb.me/nejumi>

4.4 本研究の制約と今後の課題

LLM の訓練やチューニング技術がまだ開発途上段階であることは言うまでもないが、医療 LLM のトレーニング特有の制約が複数存在する。第一に、データ量とデータ品質の不足が課題である。本研究では、追加の事前学習においてタスクに密接に関連する医療文書のみを使用した。しかし、より広範な医療ドメインの文書を取り込む、あるいは一般的なコーパスから抽出および拡張することで、追加の事前学習の効果が新たに発揮される可能性は否定できない。追加の事前学習に必要なデータ量を特定し、下流タスクのパフォーマンス向上を達成することは、今後の課題である。また、Instruction-tuning の過程においてもデータ収集が同様に課題となる。医療テキストデータセットを instruction 形式で用意することはコストが高い作業である。本研究では ChatGPT による自動生成を利用したが、より大きなデータセットを用意する場合、このアプローチは経済的負担が大きくなる可能性がある。一方で、実際の医師が手作業で大量のデータ作成を行うことも業務負担を圧迫するため難しい。

さらに、データの十分なクレンジングは LLM の学習において常に課題となる点であり、本研究でのクレンジングも不十分であった可能性がある。

加えて、近年新しい LLM の開発およびリリースは数が多く、本研究で使用された OpenCALM よりも優れた性能を発揮するとされる日本語の LLM が既にリリースされている (Rakuda benchmark⁶⁾等を参照)。これらをベースモデルとして使用すると異なる結果が得られる可能性がある。

5 結論

本研究では、Instruction-tuning の効果と限界について日本語の医療質問応答を通じて調査した。適切なステップ数の LoRA ベースの Instruction-tuning は性能向上に寄与し、より大きなモデルに対し顕著な効果を示した。また、限られた学習データセットによる事前学習の追加は性能の低下、特に 1-shot 性能の喪失をもたらすことが観察された。本研究の結果から、パラメタ数が大きく高性能な英語の事前学習済みモデルをベースとし Instruction-tuning を適用することが言語・領域の両方の観点からドメイン特化 LLM の構築に有望なアプローチであるといえる。

6) <https://yuzuai.jp/benchmark>

謝辞

本研究は JSPS 科研費 JP23hk0102078h0003z の助成を受けたものである。また、校正や翻訳の一部に ChatGPT を利用した。

参考文献

- [1] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pp. 1–9, 2023.
- [2] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. **arXiv preprint arXiv:2305.09617**, 2023.
- [3] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. **arXiv preprint arXiv:2307.14334**, 2023.
- [4] Kaito Sugimoto, Taichi Iki, Yuki Chida, Teruhito Kanazawa, and Akiko Aizawa. JMedRoBERTa: a japanese pre-trained language model on academic articles in medical sciences (in Japanese). In **Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing**, 2023.
- [5] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2021.
- [6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. **arXiv e-prints**, pp. arXiv-2305, 2023.
- [7] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **International Conference on Learning Representations**, 2022.
- [8] Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. **arXiv preprint arXiv:2309.03412**, 2023.
- [9] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. **arXiv preprint arXiv:2305.11206**, 2023.
- [10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.
- [11] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- [12] 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.
- [13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.
- [14] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. **arXiv preprint arXiv:2303.18027**, 2023.
- [15] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. Jglue: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, 2022.
- [16] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. **arXiv preprint arXiv:2308.11483**, 2023.
- [17] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. On large language models’ selection bias in multi-choice questions. **arXiv preprint arXiv:2309.03882**, 2023.

A LLM の評価に用いた医療質問応答データ

IgakuQA の問題例

```

"problem_id": "116A1",
"problem_text": "睡眠時無呼吸症候群による高血圧について、正しくないのはどれか。",
"choices": {"a": "夜間高血圧となることが多い。", "b": "肥満患者では減量を推奨する。", "c": "α遮断薬が第一選択薬である。", "d": "家庭血圧では早朝に高血圧となることが多い。", "e": "持続的気道陽圧法 <CPAP> で降圧が期待される。"},
"text_only": true,
"answer": ["c"],

```

JJSIMQA の問題例

```

"problem_text": "近年の日本の B 型肝炎について誤っているものはどれか。1 つ選べ。",
"choices": {"a": "HBs 抗原陽性率は、母子感染防止事業開始により著減した。", "b": "HBV(hepatitis B virus) ゲノタイプ Ae 型は成人の水平感染でキャリア化することがある。", "c": "本邦では、2016 年 10 月から HBV ワクチン定期接種が始まった。", "d": "HBV ゲノタイプ C は東北地方、宮古八重山地方に多い。", "e": "HBV の小児期での水平感染は、父子感染や集団生活が一因と考えられる。"},
"text_only": true,
"answer": ["d"],

```

B 各モデルの応答例

本実験で作成した各モデルに対し、病気に対する処置を質問した際の応答例が表 3 である。用いたプロンプトは以下である。Llama2-70B は英語で事前学習されたモデルであるため、QLoRA チューニングを施していないオリジナルのモデルでは出力が英語となった。

病気に対する処置を質問するプロンプト

```

### 指示: 以下の病気を持つ患者に対して行う処置を詳細に教えてください。
### 入力: 深部静脈血栓症
### 回答:

```

表 3 病気に対する処置を質問するプロンプトに対する各モデルの応答例

モデル名	LoRA/QLoRA	応答
OpenCALM-7B	none	・ベッドサイドで看護婦は、患者がベッドから起き上がるのを助けるために、足を動かすように促した。患者は、足を曲げたり伸ばしたりする動作を行った。しかし、脚の関節を動かすことはできず、自力で立ち上がることができないため、介助が必要であった。看護婦は、ベッドサイドの椅子に座っている患者のそばに行き、肩を支え、腕を持ち上げ、肘を伸ばして手をついた。そして、ゆっくりと手を離れた。看護婦は、再び椅子に座った。看護婦は、次の患者を呼んだ。
OpenCALM-7B	10k steps of LoRA	詳しい治療法については情報が提供されていない。
Llama2-70B	none	In some cases, a procedure called thrombol
Llama2-70B	900 steps of QLoRA	深部静脈血栓症の治療には、血栓溶解薬や抗凝固薬などが使用されます。
Llama2-70B	3000 steps of QLoRA	深部静脈血栓症の治療には、血栓を溶かす薬物療法や血栓を取り除く手術などがあります。