大規模言語モデルにおける Supervised Fine-tuning の包括的検証

原田宥都 [1,2*] 山内悠輔 [1,3*] 小田悠介 [1]
大関洋平 [2] 宮尾祐介 [1,3†] 高木優 [1†]
[1] 国立情報学研究所 大規模言語モデル研究開発センター
[2] 東京大学 大学院総合文化研究科 [3] 東京大学 大学院情報理工学系研究科
{harada-yuto, yamauchi_y, odashi, yu-takagi}@nii.ac.jp
oseki@g.ecc.u-tokyo.ac.jp, yusuke@is.s.u-tokyo.ac.jp

概要

大規模言語モデルを人間の意図に合わせるアライメントのためには、事後学習としての Supervised Fine-tuning (SFT) が不可欠である。しかしながら、ベースモデルの種類や学習データの特性が下流タスクの性能に与える影響について、広範な検証はほとんど行われていない。そこで本研究では、複数の大規模言語モデルと多様な学習データを用いて全 245 種類の SFT モデルを訓練する。それらのモデルをさまざまな下流タスクで包括的に評価することで、先行研究で提案されてきた知見や定説を実証的に検証する。また、作成したすべてのファインチューニング済みモデルを公開予定である。

1 はじめに

大規模言語モデルを人間の意図に合わせるアライメントの実現において、SFT などの事後学習手法は重要な役割を果たす。しかしながら、どのようなモデルをどのようなデータで学習すると最適な性能が得られるのかについては、依然として包括的な理解が十分ではない。特に、パラメータ数が 7B を超えるような大規模モデルを対象に、モデル・データ・下流タスクの関係を体系的に検証した例はほとんど存在しない。本研究は、多数の大規模言語モデルと多様な学習データの組み合わせによる学習を実施し、幅広い下流タスクでその性能を評価する。具体的には、次のような研究課題を設定し、全 245 の SFT モデルを用いて各観点から横断的な検証を行う（図 1）。

まず、**多様なデータを用いて SFT を実施した際の下流タスク性能の違いを検証する**。近年の研究では、特定の種類のデータセットを混ぜることで異なる下流タスクの性能が向上する可能性が示唆されている [1, 2, 3, 4] が、これらの効果を実際に横断的に評価する。さらに、**学習データのサンプルサイズが性能に与える影響を評価する**。SFT では少数サンプルであっても、データの質によっては十分な有効性を示すとする報告 [5, 6] と、大規模なデータほど効果が顕著だとする報告 [7, 8] が混在するため、どのような条件で最適化されるのかを検討するために、複数のサンプルサイズでデータセットを用いた実験を行った。その他にも、先行研究では、SFT を LoRA (Low-Rank Adaptation) [9] とフルパラメータで学習した場合の下流タスク性能がどのように異なり、それぞれの手法にどのような利点があるのか議論されている [10, 11, 12, 13, 14, 15] が、網羅的な比較は限られている。本研究では、これらについても予備的な検討を行う。

これらの検証を通じて、大規模言語モデルの事後学習に関する具体的な知見を提示する。作成したすべてのファインチューニング済みモデルは公開予定であり、コミュニティにおけるさらなる検証や応用を促進する。

2 関連研究

SFT において、どのようなモデルに、どのようなデータを、どのように学習させると最適であるか、といった包括的な理解は未だ十分でない。しかし、既に多くの先行研究において、さまざまな知見が報告されている。たとえば、SFT に用いるデータの種類と下流タスク性能の関係に関しては、コード生成データを含めることで

* 共同第一著者
† 共同責任著者

推論力の向上が期待できると示唆する研究 [1] や、学習データ内の手続き的知識の有無がモデルの推論能力に大きく影響するとする報告 [2]、評価タスクの特性によって学習効果は大きく左右されると指摘する報告 [3, 4] などがある。学習データのサンプルサイズについては、1000 件程度の少数サンプルでも有効であるという主張 [5, 6] と、大規模データほど安定した性能向上が見込めるという見解 [7, 8] が併存しており、データの質と量の観点から検証を行った例 [16] もある。学習手法に関しては、フルパラメータ学習の方が性能が良いとの報告 [10, 11] と LoRA の方が優れているとする報告 [12, 13] の両方があり、一般的な合意は形成されていない。さらにインストラクションチューニング全般に潜む課題を指摘する研究 [17] など、論点は多岐にわたる。なお、複数の規模のモデルを対象に事後学習の検証を行った研究 [18] も存在するが、対象モデルや評価タスク、対象言語の多様性が限定的である。こうした背景から、複数の要素（モデル規模、学習データの種類や規模、ファインチューニング手法、多言語対応など）を横断的に比較し、各要素の相互作用も含めて包括的に検証する試みが求められている。

3 手法

3.1 訓練

モデル 本研究では、日中英の 3 言語を中心に事前学習された以下の大規模言語モデル（7B パラメータ規模）を利用する: OLMo-7B-hf[1)]、llm-jp-3-7B、および Qwen2.5-7b[2)]。いずれも、事前学習のみが行われた base モデルを使用している。

データセット 10 種類のデータセット (表 1) を用いて、学習データの特性と下流タスク性能の関係を検証する。多様なデータで学習を行うため、それぞれ 4 つのカテゴリからデータセットを採用している。サンプルサイズの影響を評価するため、いずれのデータセットについてもランダムに抽出した 1k サンプル版と 20k サンプル版を作成して使用した。

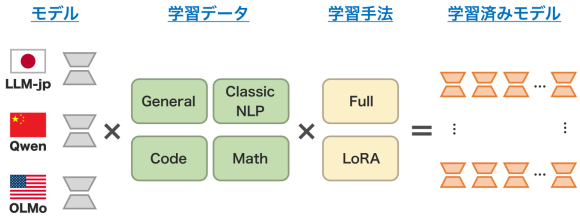

図 1: 本実験の概要図

表 1. SFT に使用したデータセット

Category	Dataset
General	Alpaca LIMA Ultrachat
Coding	CodeAlpaca MagiCoder
Math	OpenMathInstruct1 MathInstruct
Classic NLP	FLAN Knowledge (BoolQ, NaturalQuestions, TriviaQA)
	FLAN Reasoning (ARC-Easy&Challenge, HellaSwag, WinoGrande, PIQA)
	FLAN Comprehension (QuaC, SQuAD v2)

学習設定 本実験[3)]では、各モデルに対してフルパラメータ学習と LoRA 学習の両方を適用し、

1. 1 データセット (1k サンプル) で学習
2. 1 データセット (20k サンプル) で学習
3. 全データセットを統合して学習
4. 全データセットから 1 データセットを除外して学習

という 4 種類の設定で実施する。これにより、学習データの特性やサイズの違いが下流タスク性能にどのように影響するかを検証する。

学習データの前処理 学習データは、全モデルのトークナイザによって、最大系列長を超えるサンプルを除外した上でランダムに抽出している。すべてのモデルで同一の前処理済データを用いることで、モデルごとに最大系列長が異なることによる性能差が生じないよう配慮した。

チャットテンプレート 学習には、以下のような形式のテンプレートを統一的に用いた。

```
###Question: {instruction}
###Answer: {response}
```

ハイパーパラメータ フルパラメータ学習時は、`learning rate` を 1.0×10^{-5}、`batch size` を

1) https://huggingface.co/allenai/OLMo-7B-hf
2) https://huggingface.co/Qwen/Qwen2.5-7B

3) 本研究における実験はすべてデータ活用社会創成プラットフォーム mdx[19] 上で行った。

表 2. 各データセットを 1k サンプル学習したモデルの下流タスクの性能

	MA (0shot)	GS (0shot)	MB (3shot)	HE (0shot)	HS (0shot)	NQ (1shot)	BQ (5shot)	MM (5shot)	MT (0shot)	AE (0shot)
OLMo-7B-hf	0.00	2.73	29.57	1.83	4.85	0.30	0.00	0.04	1.86	0.83
Alpaca	+0.62	+0.53	**-3.11**	+2.44	**+19.78**	+0.03	+61.38	+22.70	+0.83	+1.16
CodeAlpaca	+0.46	+0.38	-22.96	+1.83	+16.88	+1.33	**+61.90**	+22.77	+0.60	+0.66
FLAN Comprehension	+0.08	-0.99	-17.12	***-1.83***	+9.46	+5.32	***+0.49***	+4.18	***-0.82***	-0.62
FLAN Knowledge	***0.00***	***-1.82***	***-29.57***	***-1.83***	***-4.74***	**+9.31**	+14.59	***+3.80***	-0.81	***-0.83***
FLAN Reasoning	+0.16	-1.21	-23.34	-1.22	+8.70	+6.10	+12.32	+12.95	-0.40	-0.71
LIMA	+0.32	-1.21	-10.89	+0.61	+10.56	+0.23	+47.43	+24.38	+0.37	+0.29
MagiCoder	+0.34	-0.53	-12.06	**+4.27**	+19.49	+3.58	+58.32	**+25.65**	+0.47	+0.41
MathInst	**+2.22**	**+2.27**	-27.62	***-1.83***	+13.58	+0.23	+59.30	+12.76	+0.31	-0.08
OpenMathInst	+1.34	-0.76	***-29.57***	***-1.83***	-2.23	***-0.08***	+49.14	+10.66	+0.13	-0.46
Ultrachat	+0.52	+0.68	-5.83	-0.61	+15.45	***-0.08***	+57.55	+24.37	**+0.90**	**+2.52**

各下流タスクにおける、事前学習モデルと、各データセットを 1k サンプル学習したモデルのスコアの差分を表す。
太字は各下流タスクの最高スコア、斜体＋太字は最低スコアを示す

32、weight decay を 0.0、epoch を 10 とし、LoRA 学習時は learning rate を 2.0×10^{-6}、batch size を 128、weight decay を 0.0、epoch を 10 に設定している。これらは予備実験でグリッドサーチを行い、最適な値を採用した。

3.2 評価

訓練されたモデルはオープンソースの評価ライブラリである opencompass[4] を使用して下流タスク性能を検証する。使用したベンチマークは表 3 に記載した。

表 3. モデル評価に使用したベンチマーク

Category	Dataset
Math	MATH (MA) GSM8K (GS)
Coding	HumnaEval (HE) MBPP (MB)
Knowledge	BoolQ (BQ) NaturalQuestions (NQ) MMLU (MM)
Subjective	MT Bench (MT) Alpaca Eval v2 (AE)

() 内は表に記載する際の略称

4 結果

4.1 学習データセットのカテゴリの影響の検証

表 2 に OLMo-7B-hf を対象モデルとして、1 データセットを 1k サンプルで学習した場合の下流タスクでのスコアを示す。特筆すべき点として、特定の学習データは In Distribution (ID)、Out Of Distribution (OOD) 問わず一貫してモデルのスコアに影響があることが挙げられる。例えば、MagiCoder は HumanEval だけでなく MMLU でも最高スコアとなり、FLAN Knowledge は学習データに含まれる NaturalQuestion, BoolQ 以外のほぼ全てのタスクで最低スコアとなった。このことから、モデルの精度向上には学習データのカテゴリだけではなく問題やフォーマットの性質が重要であることが示唆される。

4.2 データセットのサイズの影響の検証

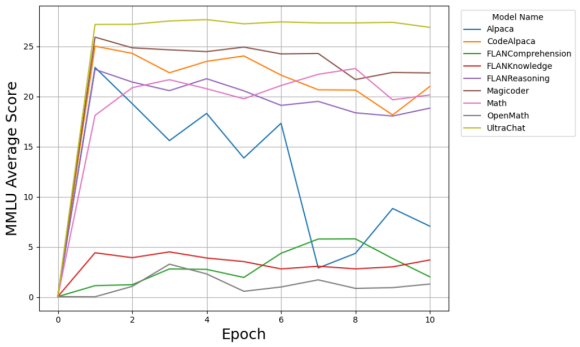

図 2: エポック数と MMLU のスコアの推移

図 2、3 に各データセットを 20k サンプル、学習率固定でモデルを訓練した場合の下流タスクのスコアの推移を示す。特徴として、MMLU のような非生成系タスクでは学習が進むと徐々にスコアが低くなっていったのに対し、生成系タスクである MTBench ではスコアが緩やかに上昇した。同様の現象が 1k サンプルの場合でも見られた (付録 A)。データサイズによらず、学習データ間のスコアの順序は一貫していることから、SFT ではデータサイズよりもデータの質の方が重要であることが示唆される。

4.3 モデルファミリーの影響の検証

表 4 に事前学習済みモデルを変えて、同じ条件で学習した場合の結果を示す。モデルの事前学習言語やアーキテクチャによらず、スコアの変動には一貫性が見られた。特に LLM-jp と

4) https://github.com/opencompass/opencompass

表 4. 1k サンプルの各データセットを model family を変えて学習した場合のモデルの評価

model family	MA (0shot)	GS (4shot)	MB (3shot)	HE (0shot)	HS (10shot)	NQ (1shot)	BQ (5shot)	MM (5shot)	MT (0shot)	AE (0shot)
LLM-jp-3-7B	+0.19	+0.03	+2.87	-0.00	-3.11	-3.97	+48.50	+17.05	+0.43	+0.13
OLMo-7B-hf	+0.61	-0.85	-18.21	-0.00	-2.92	+2.60	+42.24	+16.42	+0.16	+0.23
Qwen2.5-7B	+19.53	-1.09	-4.05	-49.03	+11.76	-6.11	+20.73	+24.04	+0.67	+3.20

10 種類の 1k サンプルデータを学習した各モデルの、事前学習モデルからのスコアの変動の平均を示す

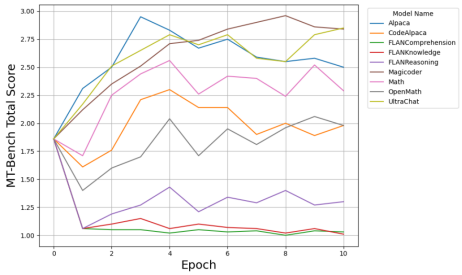

図 3: エポック数と MTBench のスコアの推移

OLMo の相関係数は 0.53 と他の組み合わせより高く、これはこれらのモデルの性能が近く、学習により獲得した能力が一致したためであると考えられる。

4.4 学習データと評価ベンチマークの相関

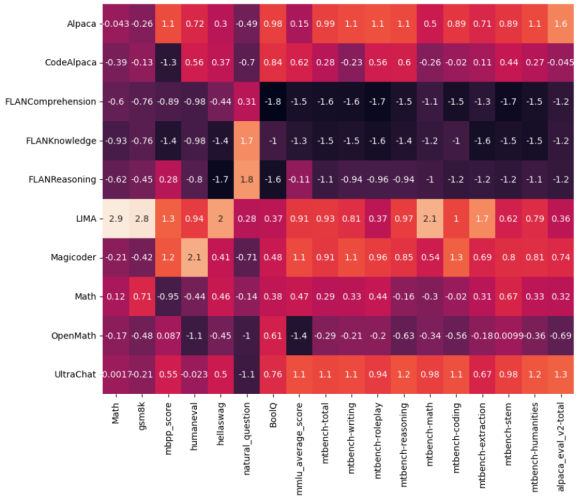

図 4: 学習データと下流タスクのスコアの関係
各スコアはタスクごとに標準化されている

図 4 に全学習モデルを訓練データ別に分けて、下流タスクの平均のスコアを測った結果を示す。FLAN データは NaturalQuestion 以外ではスコアが低く、逆に LIMA を学習したモデルは全てのタスクで平均以上のスコアを示した。

4.5 評価ベンチマーク間の相関

図 5 に全 245 モデルの下流タスクのスコアの相関ヒートマップを示す。右下の生成系タスクは互いに非常に高い相関を示した。一方で、NaturalQuestion や BoolQ などの Knowledge カテゴリに属するベンチマークは複数のタスクと負の相関がみられ、SFT で忘却されてしまう知識が重要なベンチマークであるといえる。

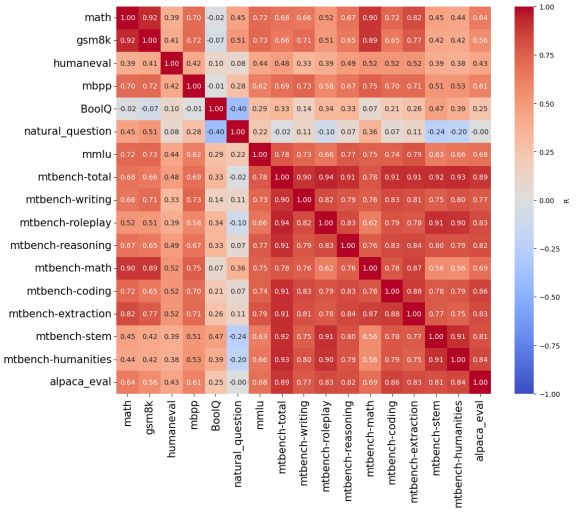

図 5: 学習データと下流タスクのスコアの関係

4.6 SFT 手法の影響の検証

本実験では、フルパラメータ学習と同一のモデル・データセットで LoRA を学習・評価したが、データセットごとにスコアが大きくばらつき一貫した結果が得られなかった。先行研究においてもフルパラメータ学習と LoRA の優位性には議論が分かれているが、その原因としてこのようなデータ・モデルへの強い依存性があると考えられる。

5 おわりに

本稿では、複数のモデル・学習データの組み合わせで全 245 の SFT モデルを構築し、それらを多様な評価タスクで横断的に評価した。モデルファミリーや学習データ、評価タスクの更なる拡張は今後の課題であり、合計で 1000 以上のモデルでの実験を目指す。また、学習の前後でモデルの内部表現がどのように変化をしたかを、より詳細に分析する予定である。

謝辞

本研究成果は、データ活用社会創成プラットフォーム mdx を利用して得られたものです。

参考文献

[1] Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 177–198, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. **arXiv preprint arXiv:2411.12580**, 2024.

[3] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? **arXiv preprint arXiv:2411.16489**, 2024.

[4] Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. **arXiv preprint arXiv:2405.20535**, 2024.

[5] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[6] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. **arXiv preprint arXiv:2307.08701**, 2023.

[7] Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. **arXiv preprint arXiv:2402.04833**, 2024.

[8] Michihiro Yasunaga, Leonid Shamis, Chunting Zhou, Andrew Cohen, Jason Weston, Luke Zettlemoyer, and Marjan Ghazvininejad. Alma: Alignment with minimal annotation. **arXiv preprint arXiv:2412.04305**, 2024.

[9] 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. In **International Conference on Learning Representations**, 2022.

[10] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. **arXiv preprint arXiv:2311.10702**, 2023.

[11] Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. Astraios: Parameter-efficient instruction tuning code large language models. **arXiv preprint arXiv:2401.00788**, 2024.

[12] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[13] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. **arXiv preprint arXiv:2403.03507**, 2024.

[14] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. **arXiv preprint arXiv:2410.21228**, 2024.

[15] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. **arXiv preprint arXiv:2405.09673**, 2024.

[16] Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. Instruct-skillmix: A powerful pipeline for llm instruction tuning. **arXiv preprint arXiv:2408.14774**, 2024.

[17] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. A closer look at the limitations of instruction tuning. **arXiv preprint arXiv:2402.05119**, 2024.

[18] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. **arXiv preprint arXiv:2402.17193**, 2024.

[19] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

A　学習率を固定して 1k サンプル学習した場合のスコアの推移

MMLU で 1epoch 目でスコアが急激に上昇しているのはチャットテンプレートに適合したためであると考えられる。

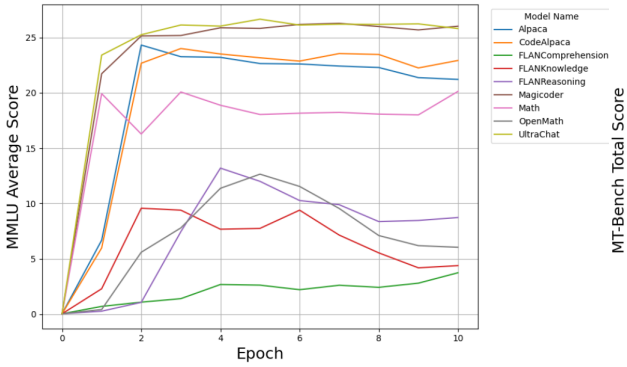

図 6: エポック数と MMLU のスコアの推移

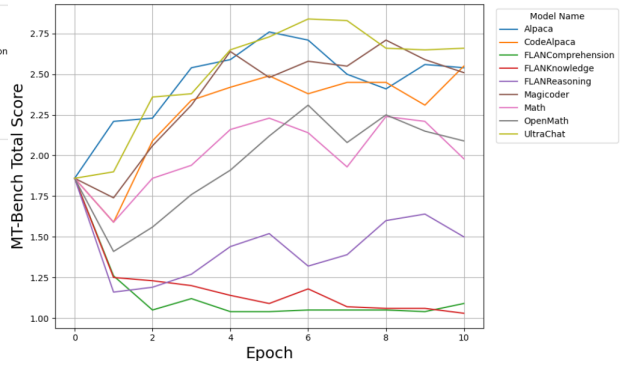

図 7: エポック数と MMLU のスコアの推移

B　全データから 1 つのデータを取り除いたデータで学習したモデルの評価

学習データの種類を増やすといずれも似た性能となり、一つ一つのデータの影響は相対的に小さくなる。

表 5. 全データセットから 1 データセットを取り除いて学習したモデルの評価

学習データ	MA (0shot)	GS (4shot)	MB (3shot)	HE (0shot)	HS (10shot)	NQ (1shot)	BQ (5shot)	MM (0shot)	MT (0shot)	AE (0shot)
OLMo-7B-hf	0.00	4.32	29.57	1.83	23.50	0.30	0.00	23.26	1.86	0.83
w/o alpaca cleaned	1.68	5.61	26.07	4.27	23.04	3.10	22.78	25.91	2.19	2.49
w/o code alpaca	1.84	4.85	22.96	10.98	24.26	4.46	7.58	26.33	2.41	1.37
w/o flan comprehension	1.38	5.69	24.90	5.49	24.45	9.89	26.93	19.28	2.54	2.36
w/o flan knowledge	2.20	5.99	23.35	4.88	21.49	4.49	8.75	26.20	2.46	2.24
w/o flan reasoning	1.62	5.84	24.90	7.32	20.48	8.92	28.35	26.45	2.39	2.11
w/o lima	1.74	5.53	23.35	1.22	20.73	9.81	13.21	26.90	2.39	2.24
w/o magicoder	1.62	6.07	21.40	0.61	23.36	9.42	16.42	26.78	2.51	1.87
w/o math	1.40	4.93	23.35	3.05	16.15	1.97	0.89	25.04	2.76	1.18
w/o openmath	2.10	5.46	22.18	6.10	24.43	7.95	17.37	27.01	2.62	2.11
w/o ultrachat	1.68	5.76	24.51	1.83	20.40	6.84	28.07	27.00	2.49	2.24

C　fewshot における example 数の影響の検証

本実験では各ベンチマークのスコアが example 数により大きく異なる傾向が見られた。学習データ量が多いほど、より zeroshot の推論に適用できるようになる。

表 6. zeroshot と fewshot の比較

学習データ	NQ (0shot)	(1shot)	(5shot)	(25shot)	BQ (0shot)	(1shot)	HS (0shot)	(10shot)	GS (0shot)	(4shot)	MM (0shot)	(5shot)
OLMo-7B-hf	15.71	0.30	20.06	22.66	62.02	0.00	4.85	23.50	2.73	4.32	23.26	0.04
alpaca cleaned	18.75	0.30	0.33	0.33	59.72	61.38	24.63	25.12	3.26	3.64	24.51	22.74
code alpaca	22.83	1.66	4.24	18.81	19.72	61.90	21.73	25.05	3.11	3.64	24.90	22.81
flan comprehension	4.82	5.62	5.35	9.03	33.64	0.49	14.31	16.61	1.74	2.05	23.20	4.22
flan knowledge	11.25	9.47	10.11	9.03	0.15	14.59	0.11	9.53	0.91	1.52	8.31	3.84
flan reasoning	5.15	6.40	3.93	8.45	8.50	12.32	13.55	11.26	1.52	2.88	21.01	12.99
lima	6.37	0.55	0.22	0.22	43.52	47.43	15.41	23.00	1.52	5.00	22.80	24.42
magicoder	19.36	3.82	0.94	5.84	17.40	58.32	24.34	25.10	2.20	2.20	26.84	25.69
math	20.22	0.53	0.50	10.25	17.28	59.30	18.43	24.33	5.00	6.22	24.64	12.80
openmath	18.70	0.22	0.22	0.22	4.77	49.14	2.62	19.78	1.97	4.02	22.26	10.70
ultrachat	17.37	0.22	0.25	0.25	32.72	57.55	20.30	26.07	3.41	3.56	25.82	24.41
all	12.96	3.41	5.68	10.47	57.92	19.36	24.29	19.23	1.82	6.22	25.97	16.54

10 種類の 1k サンプルを学習した各モデルの、事前学習モデルからスコアの変動の平均を示す