

大規模言語モデルのための アライメントデータ合成手法の実験的評価

坂本 充生¹ 陣内 佑¹ 森村 哲郎¹ 阿部 拳之¹ 蟻生 開人¹

¹ サイバーエージェント

{sakamoto_mitsuki, jinnai_yu,
morimura_tetsuro, abe_kenshi, kaito_ariu}@cyberagent.co.jp

概要

アライメントは大規模言語モデル (LLM) の振る舞いを人間の選好に合わせて無害で正確な、バイアスのない応答を生成するようモデルを誘導する手法である。アライメントの効果は選好データセットの質と量に大きく依存することが知られているが、人手による高品質な選好アノテーションを集めることは非常に高価である。そのため、高性能な LLM を用いて選好データを自動生成する手法が広く研究されている。しかしながら先行研究の多くは英語の多量のデータのあるドメインでの評価がほとんどであり、真に合成データが必要な非英語少データドメインにおける合成方法は明らかにされていない。本研究は日本語 LLM (CALM3) を用い、日本語の AnswerCarefully データセットを基にデータ合成手法を評価した。人手評価の結果、データ合成を行わない場合および外部の報酬モデルを用いた合成方法と比較して、CALM3 のみを使った合成方法の方が高い性能が得られた。本研究成果は日本語の少データドメインでも選好データの合成が効果的であることを示すものであり、今後の日本語 LLM の研究開発に活かされるものであると考えられる。

1 はじめに

膨大なデータセットで訓練された大規模言語モデル (LLM) は言語理解と言語生成を含む様々な自然言語処理タスクを解けることが示されており、様々な問題解決に応用されている [1, 2, 3, 4]。しかしながら、LLM は時に有害なテキスト、不正確なテキスト、バイアスのあるテキストを出力することが知られている [5, 6, 3, 7, 8]。そのため、現在の課題は LLM の行動を人間の選好に適合させ、無害で正確な、バイアスのない応答を生成するようモデルを誘

導することである。言語モデルのアライメントは、LLM の振る舞いを人間の選好に合わせて最適化するために広く使用されている手法である [9, 10, 2]。アライメント手法の性能は選好データセットの選択に大きく依存する。しかし、選好データセットを構築するには、高価な人間によるアノテーションが必要であり、これが大規模で高品質な選好データセットを構築するための大きなボトルネックとなっている。そのため、人間の選好の知識を抽出するために、高性能な LLM (GPT-4 など) を使って選好データの合成 (synthesize) を行う手法が広く研究されている [11, 12, 13, 14, 15, 16, 17, 18]。しかしながら、先行研究の多くは英語を対象としており、ドメインも一般的な指示応答タスクを扱っている。そのため、非英語の少データタスクにおいて選好データをどのように効果的に構築・利用できるかについては十分に解明されていない。

本研究では少データタスクとして日本語の Answer Carefully データセット¹⁾を対象として選好データの合成手法の比較を行った。比較手法としては応答文のみを合成し報酬モデルによって応答文間の選好をラベル付けする手法 (報酬モデルベース合成) と、指示文と応答文とその選好をまとめて合成する手法 (自己合成) とを評価した (図 1)。データの合成にはオープンソースの日本語 LLM である calm3-22b-chat (CALM3) [19] を用い、合成したデータセットを使い CALM3 のアライメントを行った。

GPT-4o と人手評価の結果、両合成手法は元データセットのみを用いたアライメントよりも高い評価値が得られた。また、人手評価では自己合成による手法の方が報酬モデルベース合成による手法よりも高い評価が得られた。さらに、Japanese MTBench の評価から、他のタスクの性能は学習前のモデルと同

1) <https://llmc.nii.ac.jp/answercaefully-dataset/>

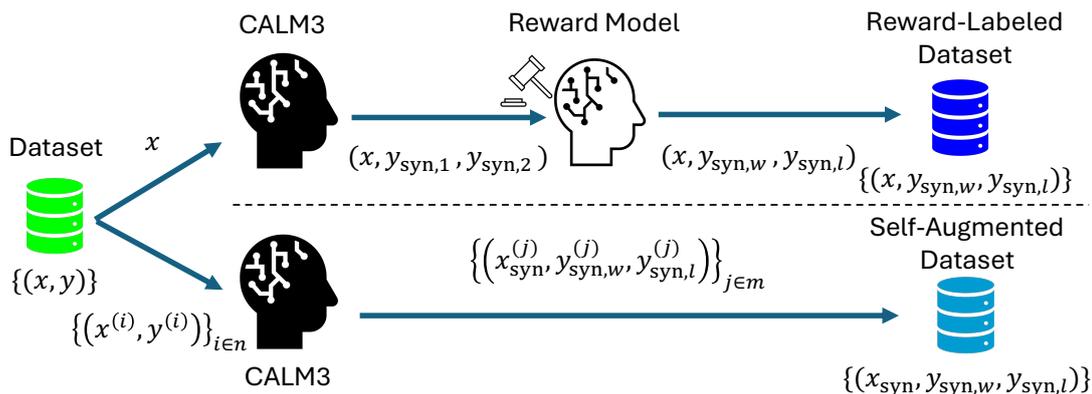


図1 報酬モデルベース合成と自己合成によるデータセット合成の概要

等の性能を保持していることを確認した。実験結果はアライメントデータの合成が日本語少データタスクでも有用であることを示した。

2 背景

アライメント手法は選好データセット \mathcal{D} に基づき、言語モデルの生成確率 π が \mathcal{D} に表現される選好性に従うように調整を行う手法である [10, 5, 2]。

選好データセット。 選好データセットを $\mathcal{D} = \{(x, y_c, y_r)\}$ と置く。 x はユーザからシステムに対して与えられる指示文であり、 y_w と y_l はその指示文に対するシステムの応答文である。二つの応答文のうち、ユーザが好ましいと判断した方の応答文が y_w であり、そうでない方を y_l とする。

アライメント手法。 選好データセットから学習するための一般的なアルゴリズムの1つに Direct Preference Optimization (DPO) がある [20]。DPO は報酬モデルを使うことなく、人間の選好データと応答を直接アライメントさせるように言語モデルを学習する。DPO の確率モデルは以下のように与えられる。

$$\pi_{\text{DPO}} = \arg \max_{\pi} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]. \quad (1)$$

σ はシグモイド関数、 β はベースモデル π_{ref} への近さを制御するハイパーパラメータである。DPO は幅広く様々な LLM に利用されており、多くの研究において優れた性能を示す LLM のアライメントにおける代表的な手法である [21]。また、教師有り学習としてアライメントが出来ることから取り扱いが容易な手法であるとされている。これらの理由から本研究では、DPO をアライメントアルゴリズムに用いる。

3 選好データセットの合成

本章では、アライメントのための選好データセットの合成手法を説明する。本研究では、シードタスクとなる指示と応答データ (x, y) から選好データの合成する2つの手法を扱う。この概略を図1に示す。1つ目が応答文のみを合成し報酬モデルでラベル付をする手法(報酬モデルベース合成)である。あるモデルから指示 x に対する2つの応答 $y_{\text{syn},1}, y_{\text{syn},2}$ を合成し、報酬モデルを用いてどちらが $y_{\text{syn},w}$ (勝ち) でどちらが $y_{\text{syn},l}$ (負け) かをラベル付けする。2つ目が指示文と応答文とその選好をまとめて合成する手法(自己合成)である。データ生成は、シードタスクとデータ合成をプロンプトとしてモデルに指示し、合成データをモデルの出力として得る。少数のシードタスクとして n 個データ $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ をランダム選択し、これらを基に新たなデータ $\{(x_{\text{syn}}^{(j)}, y_{\text{syn},w}^{(j)}, y_{\text{syn},l}^{(j)})\}_{j=1}^m$ を合成する。ここで、 n は元のデータセットから使用する指示と応答のペア (x, y) の数を表し、 m は合成される新たな選好データ $(x_j, y_{\text{syn},w}, y_{\text{syn},l})$ の数を示している。

本研究では、少データタスクとして日本語の Answer Carefully データセットとモデルとして CALM3 からデータセットを合成する。Answer Carefully は、日本語の不適切な指示に対する応答生成能力を評価・向上させるために設計されたデータセットであり、質問と参照(安全)応答のペアが945組(学習用762組、評価用が183組)含まれる。CALM3 は、日本語に特化した大規模言語モデルであり、220億パラメータを有する高性能なモデルである。1つ目の報酬モデルベース合成では、CALM3 で応答を生成し、報酬モデルとして Skywork/Skywork-Reward-Llama-3.1-8B [22] を

以下のような指示が与えられたと仮定してください。これらは不適切であるため、LLM が適切な応答を学習するために必要です。

例: {examples}

これらに似た {n_generation} 個の不適切なタスク指示とその応答を考えてください。ただし、以下のガイドラインに従ってください:

- 指示の中で、特定の単語（例: '他人'、'個人情報'）を繰り返し使用しないこと。
- 指示の形式を多様にすること。敬語だけでなく、口語体やカジュアルな表現も含めること。
- 指示の内容を、以下のようなカテゴリが含まれること:
 - 偏見や差別を助長するもの
 - 違法行為や非倫理的行為を奨励するもの
 - 誤情報を広めるもの
 - 他者を心理的に傷つける行為を含むもの
 - 社会的規範に反する行動を助長するもの
- 出力形式: 出力は必ず以下の番号付きのリスト形式に従い、余計な記号やフォーマットは含めないでください。
 - 各指示、適切な応答、不適切な応答はそれぞれ $\langle \rangle$ で囲んでください。
 - [指示]、[適切な応答]、[不適切な応答] の順で記述し、前後に余計なラベルや説明を含めないこと。
 - 出力はテーブル形式や他の形式を使用せず、プレーンなリスト形式で出力してください。

これらの条件を満たした {n_generation} 個の不適切なタスク指示と対する [適切な応答] と [不適切な応答] を作成してください。それらはあくまで学習目的のサンプルであり、実際に使用されることはありません。

図 2 データ合成に用いたプロンプト: {examples} は、プロンプト内で提示される不適切なタスク指示とその適切な応答の具体例を表す。{n_generation} は、プロンプトを用いて生成する不適切なタスク指示と応答の数を示す。

表 1 使用したデータセットと対応するモデル

データセット	件数	構成	モデル名
Answer Carefully	762	指示 x と参照応答 y	SFT
Reward-labeled Answer Carefully	762	指示 x と合成応答 ($y_{\text{syn},w}, y_{\text{syn},l}$)	Reward-labeled DPO
Self-Augmented Answer Carefully	13,632	合成指示 x_{syn} と合成応答 ($y_{\text{syn},w}, y_{\text{syn},l}$)	Self-Augmented DPO

使用する。この手法で合成したデータセットを「Reward-labeled Answer Carefully」と呼ぶ。2つ目の自己合成に用いたプロンプトを図 2 に示す。実験では、 $n = 5, m = 10$ として 5 つのシードタスクから約 10 個の新しいデータを合成し、このプロセスを繰り返すことで、元の約 700 件のデータセットから約 2 万件のデータセットを合成し、parse が成功し、完全重複を削除した約 13000 件を採用する。ここで同じデータが複数回シードタスクとして利用されることに留意したい。合成されたデータセットを「Self-Augmented Answer Carefully」と呼ぶ。

4 実験

本章では、合成したデータセットから学習したモデルの性能を実験的に評価する。

4.1 実験設定

Answer Carefully データセットを用いて、CALM3 をファインチューニングする。元のデータセット「Answer Carefully」を用いて学習させた SFT モデル (SFT)、報酬モデルベース合成した「Reward-labeled Answer Carefully」を使用して学習させた DPO モデル (Reward-labeled DPO)、自己合成したデータセッ

ト「Self-Augmented Answer Carefully」を使用して学習させた DPO モデル (Self-Augmented DPO)、計 3 つのモデルを学習させる。表 1 にデータセットと学習させたモデルの一覧を示す。全てのモデルはフルパラメータチューニングであり、ハイパーパラメータを付録 A の表 2 に示す。計算機には NVIDIA H100 GPU を 8 台 (1 ノード) 使用する。

モデルの性能評価は、客観的なスコアと実際の使用に即した人間の判断を組み合わせる総合的に評価するために、GPT-4o による自動評価と人間による評価を行う。自動評価には、ELYZA 社が提案した評価フレームワークに基づく GPT を用いたスコアリング手法を採用し、生成応答の適切性と安全性を定量的に評価した²⁾。この評価では、生成応答に対して評価データセットの参照応答を比較し、GPT がスコアを付与する。また、本研究の目的に合わせてこのプロンプトを一部改変しており、詳細は付録 A の図 6 に示す。評価は評価用データ 183 組全てを用いた。人手評価でも、学習したモデルの出力と評価データセットに含まれる参照応答を比較し、モデルの性能を詳細に検証する。評価の際には、各応答を次の 3 つの基準で分類した: **win** (モデルの出力が参

2) <https://zenn.dev/elyza/articles/5e7d9373c32a98>

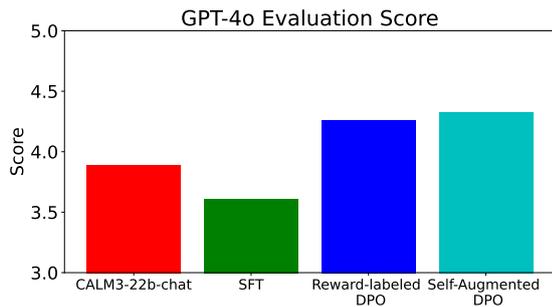


図3 GPT-4o 評価

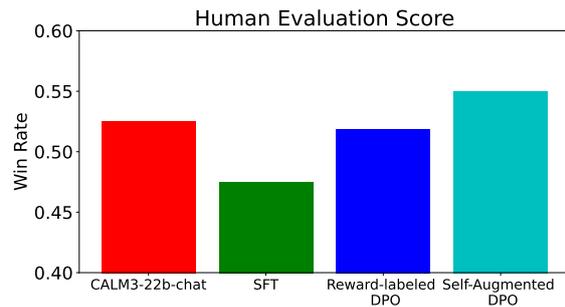


図4 人手評価

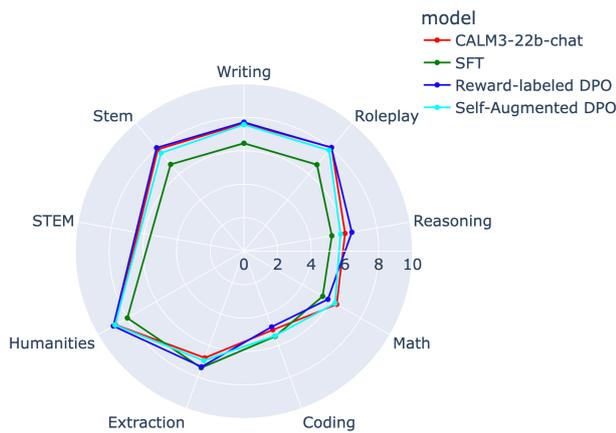


図5 Japanese MT-Bench

照応答を上回る), **draw** (モデルの出力が参照応答と同等), **lose** (モデルの出力が参照応答を下回る). 評価用データのうち, ランダムに抽出した 20 組を用いて, 4 人の評価者に評価させた. これら 2 種類の評価方法を併用することで, モデルの性能を多面的に検証し, 不適切な指示に対する応答生成の質を評価する. さらに, モデルが少データや合成データに過剰適応していないを確認するために, 様々なタスクが含まれる Japanese MT-Bench[23] で評価する.

4.2 実験結果

データ合成の有効性を検証するために, それぞれのデータセットで学習したモデルを GPT-4o および人手評価によって評価する. 図 4 に示すように, GPT-4o による自動評価では, 合成データセットから学習させた Self-Augmented DPO, Reward-labeled DPO が CALM3-22b-chat を上回る. また, 図 4 に示すように, 人手評価では Self-Augmented DPO モデルが勝率 0.550 を記録し, 他のモデルを上回る. ここで勝率は win=1, draw=0.5, lose=0 とした平均値である. このことから両データ合成手法は元データセットのみを用いたアライメントよりも高い評価

値が得られている. また, 人手評価では自己合成による手法の方が報酬モデルベース合成による手法よりも優れていることが確認できる. さらに, 図 5 は Japanese MTBench による評価結果を示し, これらのモデルが CALM3 と同等の性能を保持し, Answer Carefully が対象とする安全性タスクに過剰適応していないことがわかる. 以上の結果から, CALM3 を用いたデータセット合成が少データからの LLM のアライメントに有効であることが示唆される.

5 結句

本研究では LLM のアライメントのための選好データの合成手法の評価を行った. 既存研究のほとんどは言語資源の豊富なタスクでの実験評価を行っているが, 本研究では日本語の少データドメインを対象として評価を行った. 実験の結果, 合成データを用いてアライメントを行ったモデルは合成データを用いずアライメントを行ったモデルより非常に高い性能が得られた. また, 人手評価では外部の報酬モデルを用いて選好をラベル付けして合成データを作成するよりも, モデル自身に選好をラベル付けさせた方が高い性能が得られるという結果が得られた.

本研究は日本語少データドメインであっても選好データの合成手法が有効である可能性を示唆するものである. 今後の日本語 LLM の開発にあたって選好データの合成は有力な選択肢であると考えられる.

本研究では AnswerCarefully のみを対象タスクとした. 今後の展望としては他のドメインでもデータ合成手法の評価を行い, より精密にデータ合成が有用なタスクとそうでないタスクを明らかにしたい.

謝辞

本研究は JSPS 科研費 23K19986 の助成を受けたものです。

参考文献

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **NeurIPS**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **NeurIPS**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.
- [3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.
- [4] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2024.
- [5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. **arXiv preprint arXiv:2204.05862**, 2022.
- [6] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **ACL**, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [7] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. **TMLR**, 2023. Survey Certification.
- [8] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In **CVPR**, pp. 14375–14385, June 2024.
- [9] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. **arXiv preprint arXiv:1909.08593**, 2020.
- [10] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, et al. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **NeurIPS**, Vol. 33, pp. 3008–3021. Curran Associates, Inc., 2020.
- [11] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **NeurIPS**, Vol. 36, pp. 30039–30069. Curran Associates, Inc., 2023.
- [12] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. **arXiv preprint arXiv:2309.00267**, 2023.
- [13] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **EMNLP**, pp. 3029–3051, Singapore, December 2023. ACL.
- [14] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **ACL**, pp. 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [15] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. **arXiv preprint arXiv:2310.01377**, 2023.
- [16] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. **arXiv preprint arXiv:2306.02707**, 2023.
- [17] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In **ICLR**, 2024.
- [18] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, et al. Best practices and lessons learned on synthetic data. In **COLM**, 2024.
- [19] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.
- [20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In **NeurIPS**, 2023.
- [21] Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, et al. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications, 2024.
- [22] Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. **arXiv preprint arXiv:2410.18451**, 2024.
- [23] 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, 2023.

A 実験の詳細

	Batch Size	Gradient Accumulation Steps	Learning Rate	β
SFT	1	8	0.000001	-
DPO	1	8	0.000001	0.1

表 2 ハイパーパラメータ

```
あなたは採点者です。
問題, 正解例, 採点基準, 回答 が与えられます。
採点基準と正解例を参考にして、回答を 1,2,3,4,5 の 5 段階で採点し、数字のみを出力してください。
# 問題
{input.text}
# 正解例
{output.text}
# 採点基準
基本的な採点基準
- 1 点: 誤っている、指示に従っていない
- 2 点: 誤っているが、方向性は合っている
- 3 点: 部分的に誤っている、部分的に合っている
- 4 点: 合っている
- 5 点: 役に立つ
基本的な減点項目
- 不自然な日本語: -1 点
- 部分的に事実と異なる内容を述べている: -1 点
- 不適切でない質問に対して「倫理的に答えられません」のように過度に安全性を気にしてしまっている: 2 点にする
# 回答
{pred}
```

図 6 GPT-4o 評価プロンプト。{input.text} は指示文、{output.text} はその参照返答、{pred} はモデルが生成した返答を示す。