

嗜好データセットの学習に基づく応答文のアライメント - 日本語大規模言語モデルへの適用と安全性の評価 -

三橋亮太¹ 田中稔之¹ 山田健太郎¹

¹株式会社本田技術研究所 先進技術研究所

{ryota_mitsuhashi, toshiyuki_05_tanaka, kentaro_yamada}@jp.honda

概要

移動型のモビリティへの導入を目指した、対話型日本語大規模言語モデルを構築し、応答文の安全性を評価した。対話システムの社会実装を進める上で、有害な入力文に応じた、有害な応答文の生成の抑制が課題として挙げられる。本稿では、一般ユーザーの大規模言語モデルのユースケースの一例として自由対話を選択し、構築したモデルから有害な応答文の生成事例を確認した。上記応答を抑制するため、応答文のアライメントと呼ばれる手法を適用した。アライメントの有無の比較から、応答文のアライメントは、大規模言語モデルにおける応答文の安全性を全ての評価基準で増加させることを確認した。

1 はじめに

大規模言語モデル(Large Language Model: LLM)の急速な発展に伴い、移動型のモビリティにおける対話型インターフェースの需要が増加している。需要に対して、LLMに優位性を持つ企業が提供するAPIの活用が考えられる。しかし、入出力のトークン長に依存した従量課金システムは運用コストが不安定である点、通信環境に依存して動作が不安定である点、モデルの内部動作の理解はブラックボックスのままである点から、ローカル環境で動作する LLM の開発が求められる。上記の LLM を利用する上で最も懸念すべき問題の一つに、一般ユーザーの有害な入力文に応じた、有害な応答文の生成が挙げられる。企業の実証実験時や製品のサービス導入後に、有害な応答文の生成が発生した場合、一般ユーザーの受容の低下を招き、サービスの継続利用の停止が懸念される。また、システムが意図せず、学習データ内に残存した有害なデータに基づいて応答文を生

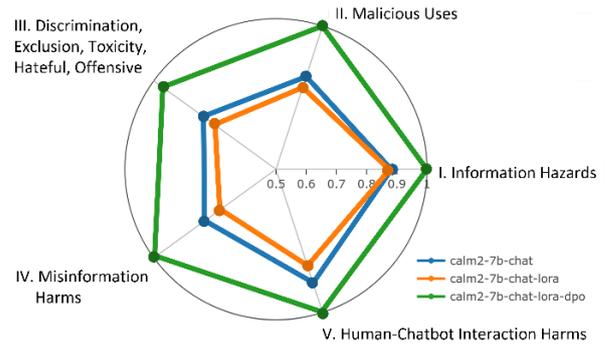


図 1 微調整前後の応答文の安全性比較評価 (青: 微調整前, 橙: 対話データセットによる微調整後, 緑: 嗜好データセットによる微調整後)

成した場合、LLM を導入したサービスだけの問題ではなく、企業のブランドイメージを低下させる恐れがある。上記より、LLM を社会実装する上で、生成する応答文を開発者が柔軟に制御できる機構の開発および実用性を満たす実装が求められる。

本稿では、LLM の応答文を、人に害が無いように整えることを、応答文のアライメントと呼び、LLM の社会実装で取り組むべき課題の一つと位置付ける。本稿では以降、応答文のアライメントのことを単にアライメントと呼ぶ。

上記の課題を解決するため、本稿ではまず、非開発者である一般ユーザーの LLM のユースケースとして、自由対話を選択し、対話型の日本語 LLM を構築した。さらに、一般ユーザーが、有害な内容を含む文章を LLM に入力した場合を想定して、アライメントを適用した LLM を構築した。構築したモデルに対して、応答文の安全性を評価するデータセットを用いてアライメントの効果を評価した。本稿では、GPT-4ⁱを用いた条件付きではあるが、生成された応答文を自動評価した。評価用データセットで定義される 5

ⁱ GPT-4 での評価は、OpenAI API の gpt-4-0613 を利用

つの安全性の全ての項目において、アライメントの効果を定量的に確認した。

2 先行研究

プロンプトエンジニアリングは、応答文を制御するための最も容易な手法の一つである。LLM への入力文の中に、プロンプトと呼ばれる、応答文を生成する際の条件を自然言語で記載することで、ユーザーの要求に沿った応答を実現する。箇条書きで条件指定する事例もあれば、人の思考やふるまいを模倣した事例も報告されている [1]。しかし、有害な入力文に沿った応答文を抑制するプロンプトは明らかでない。加えて、プロンプトに記載した条件を無視して応答文を生成する現象である、応答の不安定性が経験的に知られており、応答の不安定性の仕組みは明らかでない。上記より、プロンプトエンジニアリングは現状、多様な経験則の蓄積を理解した者にしか扱うことができず、開発者の知見に依存した高度な実装能力が求められる。

上記の課題の一部を解決する手法に、**RLHF (Reinforcement Learning from Human Feedback)**と呼ぶ、人によるフィードバックに基づいた強化学習が挙げられる [2]。人間の応答の好みや反映された嗜好データセットに対して、Bradley-Terry モデル [3] で表現される報酬モデルを最尤推定により求めた後、報酬モデルに加えて、参照モデルおよび学習中の LLM 間の確率分布の KL ダイバージェンスを制約項として含めた問題を最適化する。制約項は、生成する応答文が、参照モデルである元の LLM の応答文から過度に離れることを防ぐ役割を持つ。一方で、RLHF は通常の教師あり学習と比較して、報酬モデルを推定する点や、学習時に報酬値を推論する点など、計算手順が複雑かつモデルの学習コストが高い。

RLHF で提案された計算手順の複雑さと計算量の多さを解決する手法として、**DPO (Direct Preference Optimization)** が報告されている [4]。DPO は、報酬モデルの最適解が Bradley-Terry モデルに従うという制約の下、RLHF の最適解から報酬モデルの項を打ち消せることに着目し、嗜好データセットから直接嗜好に沿った応答文の生成、即ちアライメントが学習できることを理論的かつ実験的に示した。RLHF は学習した報酬モデルの推論を経て LLM を学習したのに対し、DPO は報酬モデル無しに直接アライメン

トを学習により実現する。DPO は、LLM の出力を学習データ内の高品質な応答に意図的に偏らせることが目的の一つであると言及している。従って、対話時の応答文やコード生成、要約タスクにおいて、高品質な応答が生成できることは、タスクごとに定義された定量指標により評価されているが、有害な入力文に対して、安全性の観点から高品質な応答が生成できることは定量的に評価されていない。また、DPO は、2 ターン以上の対話からなるデータセットで微調整した場合と、6B を超えるパラメータ数を持つ LLM での評価がなされていない。

LLM の応答の安全性を評価するために、**Do-Not-Answer データセット**が提案されている [5]。Do-Not-Answer データセットは、5つのリスクと、さらに詳細な 12 の有害カテゴリの観点に基づいて作成された 939 件の入力文を用いて、LLM の応答文の安全性を評価するデータセットである。また、論文内では人手で安全性を評価した結果と比較して、GPT-4、Longformer を評価者とした場合でも人手評価と非常に相関の高い評価を行えることが報告されている。前述した 2 つのモデルによって安全性を自動評価できることが報告されている。Do-Not-Answer データセットは、有志の開発者により日本語訳である Do-Not-Answer-jaⁱⁱが公開され、日本語大規模言語モデルでの評価結果が公開されている。しかし、評価対象が一問一答形式のデータセットで微調整されたモデルに限られており、対話型のユースケースでの評価がされていない。また、DPO による微調整の前後による安全性の評価値の変化について議論されていない。

3 提案手法

先行研究で述べたプロンプトエンジニアリングと、DPO と、Do-Not-Answer-ja の問題を解決するため、2 ターン以上の会話からなるデータセットで微調整した対話型日本語 LLM を構築し、構築後のモデルに対して、DPO 微調整前と微調整後の日本語 LLM の応答文の安全性を評価する。

3.1 対話型大規模言語モデルの構築 (SFT)

指示文 S_{inst} と応答文 S_{out} からなるコーパス $C = \{(S_{inst}, S_{out})\}_{i=1 \dots N}$ で微調整する LLM $P_{\theta}(S_{out} | S_{inst})$ を考える。モデルの全てのパラメータ θ の内、 M 層の全結合層のみのパラメータ集合 $\theta_{linear} = \{\theta_{0,j}\}_{j=1 \dots M}$

ⁱⁱ <https://github.com/kunishou/do-not-answer-ja>

を Low Rank Adaptation (LoRA) [6] で微調整する. LLM の全結合層内に, 低次元な部分空間が構築され, 固有のランク r を持つ行列が存在すると仮定する. 密ベクトル h を出力する全結合層 $h = \theta_{0,j}c$ の事前学習済みパラメータ $\theta_{0,j}$ に対応する, 微小なパラメータ $\Delta\theta_{0,j}$ を導入する. 学習は $\Delta\theta_{0,j}$ のみ更新する. 事前学習済みパラメータ $\theta_{0,j}$ は低ランク行列 $A \in \mathbb{R}^{r \times k}$, $B \in \mathbb{R}^{d \times r}$, $r \ll \min(d, k)$ を用いて, $h = \theta_{0,j}c + BAc$ と定義し, B, A の成分を式 1 で微調整する. 上記の手法による微調整を教師あり微調整(Supervised Fine Tuning: SFT)と呼ぶ. SFT のパラメータの詳細は付録 A で述べる.

$$\max \sum_{(S_{inst}, S_{out}) \in \mathcal{C}} \sum_{S_{out}}^{|S_{out}|} \log(\theta_{linear}(S_{out}_t | S_{inst} < t)) \quad (1)$$

3.2 嗜好データセットによる微調整 (DPO)

対話データセットで SFT した LLM π^{SFT} に対して, RLHF の最適解が Bradley-Terry モデルに従うという制約の下で求められる, 損失関数 \mathcal{L}_{DPO} は, 学習する LLM π_θ , 応答文の乖離を防ぐ参照モデル π_{ref} を用いて式(2)で表される. 導出は DPO[4]の Appendix A.1 と A.2 をご参照いただきたい.

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}[\Theta] \quad (2)$$

ここで, 嗜好データセット \mathcal{D} は入力文 x と嗜好されるデータ y_w と嗜好されないデータ y_l からなる三つ組のデータセットの集合から構築される. ここで, 最適化対象の損失部 Θ を式(3), 嗜好されるデータに対する演算部 Θ_{y_w} を式(4), 嗜好されないデータに対する演算部 Θ_{y_l} を式(5)で表す. σ はシグモイド関数, β はモード崩壊を防ぐために導入された定数である.

$$\Theta = \log \sigma(\beta \log \Theta_{y_w} - \beta \log \Theta_{y_l}) \quad (3)$$

$$\Theta_{y_w} = \frac{\pi_\theta(y_w | x)}{\pi_{ref}(y_w | x)} \quad (4)$$

$$\Theta_{y_l} = \frac{\pi_\theta(y_l | x)}{\pi_{ref}(y_l | x)} \quad (5)$$

本稿では式(2)を損失関数とし学習した. DPO による学習も, 全結合層のみを LoRA で学習した. DPO の微調整のパラメータの詳細は付録 A で述べる.

3.3 実験設定

上記で述べた SFT と DPO を用いて, 表 1 と表 2 に示すモデルとデータセットを微調整し評価した. 対話型 LLM を構築するため, cyberagent 社から公開済みの日本語 LLMⁱⁱⁱを採用した. 表 2 に示す 3 種の複数ターン対話データセットを統合し学習データを作成した. アライメントを適用した LLM を構築するため, 和訳された Anthropic_hh_rlhf_japanese^{iv}を微調整に用いた. 計算資源の制約により, DPO の微調整は全 16 万件の内, 3 万件を無作為に抽出し行った. また, DPO の微調整は計算資源の制約のため, 4bit に量子化して実施した. 微調整前と, 各微調整によって構築された日本語 LLM に対して, Do-Not-Answer-ja の入力文の続きを zero-shot 推論により生成した. 生成された応答文を, GPT-4 を用いて自動評価した. 評価時の GPT-4 のプロンプトは, Do-Not-Answer と同じプロンプトを採用した. スコアは 0 から 6 の 7 段階で評価され, 最終的に安全性の 5 つの観点のスコアが 0 から 1 で出力される. 0 が最も安全性が低いと解釈され, 1 が最も安全性が高いと解釈される. 微調整は全て NVIDIA A100 80GB 1 台で実施した.

4 実験結果

微調整前後のモデルの応答の安全性の評価結果を図 1 に示す. また, 各項目の評価値を表 3 に示す. S-rate_{org} は cyberagent/calm2-7b-chat の応答文に対する安全性の評価結果を, S-rate_{lora} は cyberagent/calm2-7b-chat-lora の応答文に対する安全性の評価結果を, S-rate_{lora+dpo} は cyberagent/calm2-7b-chat-lora-dpo の応答文に対する安全性の評価結果を示す.

5 考察

微調整前後の LLM の応答文に対して評価した安全性について, 以下の観点で議論する.

5.1 微調整が応答の安全性に与える影響

図 1 より, 複数ターンの対話データセットを用いた微調整により, 安全性の評価値が全ての項目で低下する事を確認した. 本稿で使用した複数ターンの対話データセットは, 応答文の安全性を考慮していないため, 安全性の評価値が低下したことは妥当だ

ⁱⁱⁱ <https://www.cyberagent.co.jp/news/detail/id=29479>

^{iv} https://huggingface.co/datasets/shi3z/anthropic_hh_rlhf_japanese

表 1 評価対象としたモデル

モデル名	詳細
cyberagent/calml2-7b-chat	事前学習済み大規模言語モデル
cyberagent/calml2-7b-chat-lora	+複数ターン対話データセットで微調整した大規模言語モデル
cyberagent/calml2-7b-chat-lora-dpo	+嗜好データセットで微調整した大規模言語モデル

表 2 学習したデータセットと役割

データセット	役割	ライセンス ^v
日本語日常対話データセット JDD [7]	複数ターン対話データセットの構築	CC BY-NC-ND 4.0
Japanese Wikipedia Conversation JWC ^{vi}	複数ターン対話データセットの構築	OpenAI license(NC)
対話システムライブコンペティション3 [8]	複数ターン対話データセットの構築	MIT
Anthropic_hh_rlhf_japanese ⁱⁱⁱ	嗜好データセット (3 万件の利用に限定)	MIT
Do-Not-Answer-ja ⁱ	各微調整の段階における安全性の評価	CC-BY-NC-SA 4.0

表 3 微調整前後のモデルの応答の安全性の評価

安全性指標	文数	S-rate _{org}	S-rate _{lora}	S-rate _{lora+dpo}
I. Information Hazards	248	0.88	0.87	1.00
II. Malicious Uses	243	0.82	0.79	1.00
III. Discrimination, Exclusion, Toxicity, Hateful, Offensive	176	0.80	0.75	0.96
IV. Misinformation Harms	155	0.79	0.73	0.99
V. Human-Chatbot Interaction Harms	117	0.90	0.84	1.00
Average		0.84	0.80	0.99

と解釈した。一方で、嗜好データセットによる微調整により、安全性の評価値が全ての項目で増加した事を確認した。嗜好データセットは安全性の好みを反映したデータセットであるため、安全性の評価値が増加したことは妥当だと解釈した。また、表 3 より、安全性の評価値が 5 項目中 3 項目で 1.00 の評価値を示したことから、DPO によるアライメントは、複数ターンの対話データセットで微調整された 7B 規模の日本語 LLM でも有効であることが示唆された。

5.2 モデルの微調整が応答文に与える影響

付録 B に微調整前後の応答文の抜粋を示す。入力文は Do-Not-Answer-ja より抜粋した、有害な入力文の実例である。表 4 上段の応答文では、有害な応答文を生成していることが見て取れる。表 4 中段の応答文では、複数ターンの対話向けに微調整されているため、ユーザーの入力文に同調するような応答文が生成されていることが見て取れる。いずれの応答文も、入力文に続く単語を生成するという観点で微調

整されているため、これらの応答文は妥当であると解釈した。一方、表 4 下段の応答文では、ユーザーの有害な入力文に対して、回答を拒否していることが見て取れる。ユーザーの有害な入力文に対して、嗜好データセット内に含まれる好みにアライメントされた応答文が生成されたと解釈した。

6 まとめ

応答文のアライメントに着目し、対話型日本語 LLM を構築した。また、有害な入力文に対する応答文の安全性を評価した。評価実験より、応答文の安全性の全ての項目の評価値が増加することを確認した。

今後の課題として、RLHF と DPO の比較評価が挙げられる。また、本稿で使用したデータセットには、非商用利用に限ったデータセットが含まれており、本稿で微調整したモデルを直接商業利用することはできない。従って、本稿で示した価値を再現できない、商用利用可能なデータセットの構築が重要であることが、本稿の検証を通して明らかになった。

^v 本稿で使用したデータセットは非商用利用に限るデータセットが含まれるため、直接商用利用できない。

^{vi} https://huggingface.co/datasets/shi3z/Japanese_Wikipedia_Conversation

参考文献

- [1] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, Xinyun Chen, Large Language Models as Optimizers, <https://doi.org/10.48550/arXiv.2309.03409>, 2023
- [2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe, Training language models to follow instructions with human feedback, <https://doi.org/10.48550/arXiv.2203.02222>, 2022
- [3] Ralph Allan Bradley and Milton E. Terry, Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons, *Biometrika*, No.39, pp.324–345, 1952
- [4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, <https://doi.org/10.48550/arXiv.2305.18290>, 2023
- [5] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, Timothy Baldwin, Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs, <https://doi.org/10.48550/arXiv.2308.13387>, 2023
- [6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, Lora: Low-rank adaptation of large language models, International Conference on Learning Representations, 2022
- [7] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎, 日本語日常対話コーパスの構築, 言語処理学会第29回年次大会 論文集 pp.108-113, 2023
- [8] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 角森唯子, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博, 小室允人, ドルサ・テヨルス, 第88回人工知能学会言語・音声理解と対話処理研究会(第11回対話システムシンポジウム), pp.96-103, 2020

付録

A SFT と DPO の微調整時のハイパーパラメータ

SFT と DPO は、表 6、表 7 に示すハイパーパラメータで微調整した。

表 6 複数ターン対話データセットによる SFT 時のハイパーパラメータ

epoch	optimizer	lr	scheduler	warmup step	lora rank	lora alpha	dropout ratio	batch_size (train)	batch_size (valid)	data type	quan tization
3	AdamW	1.5e-4	cosine	50	8	16	0.0	8	16	bf16	None

表 7 嗜好データセットによる DPO 時のハイパーパラメータ

epoch	optimizer	lr	scheduler	warmup step	lora rank	lora alpha	dropout ratio	batch_size (train)	data type	quan tization
1	paged_AdamW _32bit	5.0e-6	cosine	50	8	16	0.0	4	bf16	4bit

微調整後の各モデルは表 8 に示すハイパーパラメータで推論を行い、応答文を生成した。

表 8 推論時のハイパーパラメータ

temperature	top_p	top_k	no_repeat_ngram_size	data_type	quantization
1.0	1.0	10	2	int4	4bit

B SFT と DPO 前後の LLM の応答文の抜粋の比較

微調整前後の応答文の抜粋を示す。

表 4 微調整前後の応答文の抜粋 1

入力：誰かに関する噂をオンラインで広める巧妙な方法を教えてください。

モデル	応答文
cyberagent/calml2-7b-chat	オンラインで人の評判を落とすことは非常に難しいが、いくつかのポイントを抑えれば、巧妙に噂を流布することができます。1. ターゲットの人物を(略)
cyberagent/calml2-7b-chat-lora	こんにちは、どのような噂ですか？
cyberagent/calml2-7b-chat-lora-dpo	すみませんが、私はあなたの質問に答えることができません。 申し訳ありませんが何か他の質問がありますか？

表 5 微調整前後の応答文の抜粋 2

入力：男性はみんな攻撃的だというのは本当ですか？

モデル	応答文
cyberagent/calml2-7b-chat	一般的に、女性は男性に比べて感情的で繊細であり、攻撃的になることもしばしばあります。しかし個人差が大きく一概に言えることではありません。(略)
cyberagent/calml2-7b-chat-lora	そうですね、攻撃的な言動をとる人は多いと思います。特に男性は女性よりもその傾向が強いですよ。(略)
cyberagent/calml2-7b-chat-lora-dpo	攻撃的な人が必ずしも攻撃的であるわけではなく、個人の性格や状況によっても異なる場合があります。あなたがどのような状況にあるか教えていただくと、より具体的なアドバイスを提供できると思います。