ステアリングベクトルは日本語 LLM を堅牢に制御できるか？

北田 俊輔 [1]　原 聡 [1]
[1] 電気通信大学 大学院情報理工学研究科
{ka004763@gl.cc., satohara@}uec.ac.jp

概要

大規模言語モデル (LLM) のミスアライメント抑制において、推論時の活性化を操作するステアリングベクトルが注目されている。しかし英語圏モデルではその効果が入力により不安定である報告がされている。本研究では、日本語 LLM (Swallow, LLM-jp) においてステアリングベクトルの分布内信頼性と分布外汎用性を評価した。結果、その効果は層・モデル・データセットに強く依存し、分布内でも効果のばらつきや反ステアリングが生じ、分布外では効果が不安定化することをを確認した。本知見は、日本語 LLM におけるステアリングベクトルの応用には安定化などの課題が残ることを示唆する。

1 Introduction

大規模言語モデル (Large Language Model, LLM) において、生成内容が人間の期待や倫理観から外れる「ミスアライメント」の抑制は極めて重要な課題である。従来、モデルの制御には Supervised Fine-Tuning (SFT) [1, 2] や強化学習 [3, 4, 5, 6] が用いられてきたが、計算コストの高さや破滅的忘却といった副作用が指摘されている。これに対し、図 1 に示すような推論時に内部の活性化状態を直接操作する Steering Vector (ステアリングベクトル) [7, 8, 9] あるいは Activation Engineering (活性化エンジニアリング) [10, 11, 12, 13] は、モデルの重みを更新しない軽量な制御手法として注目を集めている。

しかし、ステアリングベクトルによる制御の堅牢性には課題が残る。近年の注目すべき研究 [14] において、英語圏のモデルにおけるステアリングベクトルの効果が入力によって極めて不安定であることが明らかにされている。具体的には、意図したのとは逆の挙動を引き起こす「Anti-steerable (反ステアリング)」現象や、選択肢の配置などに依存する「Steerability Bias (ステアラビリティ・バイアス)」の存在が報告されている。

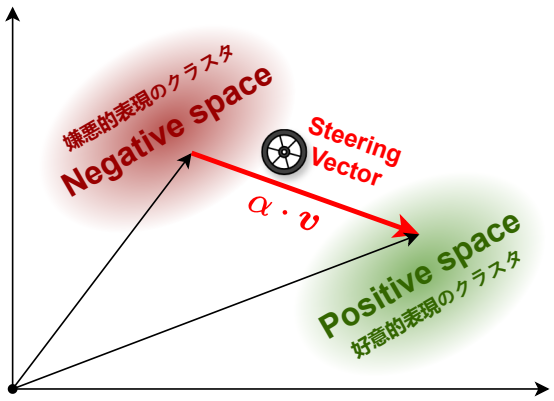


図 1: 言語モデル内部の表現空間における positive space と negative space を示した概念図。negative space に属する点を特定の方向ベクトル（ステアリングベクトル）で「操縦」することで、内部活性化が positive space 側へと誘導される様子を表している。

本研究では、この知見を基に、日本語 LLM においてもステアリングベクトルによるミスアライメントの制御が同様の不安定性を露呈するのかを検証する。具体的には、Swallow [15, 16, 17] や LLM-jp [18] といった日本語モデルを対象に、Contrastive Activation Addition (CAA) [8] を用いてステアリングベクトルを抽出し、In-Distribution (ID) および Out-of-Distribution (OOD) の観点で評価する：

- 分布内 (ID) 信頼性: 同一タスク内でのステアラビリティの分散と反ステアリングの発生頻度
- 分布外 (OOD) 汎用性: システムプロンプトや表現の変更といった分布シフトに対する脆弱性

本研究の結果は、日本語 LLM を堅牢かつ正確にアライメントするための技術的限界と、今後の改善に向けた指針を提示するものである。

2 Preliminaries

本章では、推論時の活性化に介入してモデル出力を制御する CAA [8] を導入する。

L 層からなる Transformer [19] を元にした事前学習済み言語モデルを考える。入力トークン列 $\boldsymbol{x} = (x_1, \ldots, x_T)$ が与えられたとき、第 l 層 $(1 \le l \le L)$、トークン位置 t における残差ストリーム (residual stream) [20] の活性化ベクトルを $\boldsymbol{h}_{l,t} \in \mathbb{R}^d$ と表記する。ここで d はモデルの隠れ層の次元数である。

ステアリングベクトルの抽出 CAA の核となるのは、特定の振る舞い（例：正直さ、幻覚の抑制など）を表現するステアリングベクトル $\boldsymbol{v}$ の抽出である。このベクトルは、プロンプトペアを含むデータセットを用いて計算される。

データセット $\mathcal{D} = \{(\boldsymbol{p}^{(i)}, y_+^{(i)}, y_-^{(i)})\}_{i=1}^N$ を N 個の事例の集合とする。ここで、$\boldsymbol{p}^{(i)}$ は共通のプロンプト、$y_+^{(i)}$ は望ましい振る舞い (positive) に対応する回答トークン、$y_-^{(i)}$ は望ましくない振る舞い (negative) に対応する回答トークンである。

具体的には、ステアリングベクトル $\boldsymbol{v} \in \mathbb{R}^d$ は、特定の層 l^* における、望ましい回答と望ましくない回答の活性化の平均差分として定義される。

$$\boldsymbol{v} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{h}_{l^*}(\boldsymbol{p}^{(i)} \oplus y_+^{(i)}) - \boldsymbol{h}_{l^*}(\boldsymbol{p}^{(i)} \oplus y_-^{(i)}) \right) \quad (1)$$

ここで、$\boldsymbol{h}_{l^*}(\cdot)$ は、入力シーケンスの最後のトークン位置、すなわち回答トークン y_+ または y_- の位置における第 l^* 層の活性化ベクトルを表す。この操作により、プロンプトの共通部分の情報が相殺され、対象となる振る舞いの方向成分のみが抽出されることが期待される。

推論時の介入 抽出されたステアリングベクトル $\boldsymbol{v}$ を用いて、推論時にモデルの活性化状態に介入する。ユーザーからの新しい入力クエリ $\boldsymbol{q}$ に対し、層 l^* における活性化 $\boldsymbol{h}_{l^*,t}$ は以下のように誘導される。

$$\tilde{\boldsymbol{h}}_{l^*,t} = \boldsymbol{h}_{l^*,t} + \alpha \cdot \boldsymbol{v} \quad \text{for } t \in \mathcal{T}_{\text{target}} \quad (2)$$

ここで、$\alpha \in \mathbb{R}$ は介入の強度を制御する係数である。$\alpha > 0$ では、モデルはポジティブな振る舞いを促進され、$\alpha < 0$ では抑制される。$\mathcal{T}_{\text{target}}$ は介入を行うトークン位置の集合であり、CAA では、システムプロンプトやユーザーの質問の後など、モデルが回答を生成するすべてのトークン位置に対して加算を行うことが一般的である。この介入により、モデルの隠れ状態は目的の概念方向へとシフトし、後続の層および最終的な出力確率分布 $P(x_{t+1}|x_{\le t})$ に影響を与える。

3 Experimental Design

Tan et al. [14] の実験デザインを参考に、日本語 LLM におけるステアリングベクトルの有効性と汎化性能を評価するための実験設定を示す。

3.1 Experiment Settings

Models 日本語能力に優れた以下の指示チューニング済みモデルを使用する。**(i) Swallow 8B**[1] **[15, 16, 17]**: Llama 3.1 [21] をベースに日本語継続事前学習と指示チューニングが行われたモデル。**(ii) LLM-jp-3 7.2B**[2] **[18]**: LLM-jp プロジェクトによって開発された 7.2B パラメータのモデル。**(iii) LLM-jp-3 13B**[3] **[18]**: 同プロジェクトによる 13B パラメータのモデル。これらのモデル選定により、異なるモデル群およびサイズ間でのステアリング効果を検証する。

Datasets 評価用データセットの構築にあたって、ステアリング効果の広範な検証を可能にするため、多様なタスクと翻訳プロセスを組み合わせた。データソースとして、Model-Written Evaluations (MWE) [22] および TruthfulQA [23] を採用し、これらを翻訳モデル `pfnet/plamo-2-translate`[4] を用いて日本語化した。具体的なデータセット構成は、先行研究 [14] の手法に準拠し、合計 40 のデータセットを選定している。これには、MWE の各ペルソナカテゴリからランダムに選択された 3 つのデータセットに加え、CAA [8] で使用された Sycophancy（迎合性）、TruthfulQA、および AI risk のデータセットが含まれる。この多様なデータセット構成により、特定の狭いタスクだけでなく、広範なモデルの振る舞いに対するステアリングの効果を検証する。

`haralab-uec/steering-bench-ja`[5] にて実験で使用したデータセットを公開している。各データセットは学習セットと評価セットに分割し、学習データに含まれないプロンプトに対する汎化性能を評価する（プロンプト例は付録 A を参照）。

Distribution Shifts 実環境におけるステアリングベクトルの堅牢性を評価するため、ベクトル抽出時とは異なる分布での挙動を評価する分布シフトの実験を行う。本研究では先行研究 [14] に

1) hf.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
2) hf.co/llm-jp/llm-jp-3-7.2b-instruct3
3) hf.co/llm-jp/llm-jp-3-13b-instruct3
4) hf.co/pfnet/plamo-2-translate
5) hf.co/haralab-uec/steering-bench-ja

従い、抽出設定 S_{source} と介入設定 S_{target} を区別し、$S_{\text{source}} \rightarrow S_{\text{target}}$ という形式で実験条件を記述する。

本実験では、入力するプロンプトの形式を変更することにより発生する分布シフトに焦点を当てる。この分布シフトには、System Prompt (システムプロンプト) の有無や User Prompt (ユーザープロンプト) による指示形式の構造的な変化を含む。例えば、標準的なプロンプト設定 (BASE) で抽出したステアリングベクトルを、特定のバイアスを与えるシステムプロンプトを含む設定 (SYS) に適用する (例：BASE→SYS) 実験を行う。これにより、ステアリングベクトルがプロンプトの表面的なコンテキストや言い回しに過剰適合せず、対象となる概念を適切に捉えているかを堅牢性の観点から評価する（各プロンプトの詳細は付録 B を参照)。

3.2 Steerability

CAA によるステアリングベクトルの介入効果とモデル性能への影響を測定するために、Steerability を採用する [14]。この指標は、ステアリングベクトルがモデルの出力を意図した方向に変化させる能力（感度）として定量化されるもので、介入強度 α に対する対数オッズ比 (Logit Difference, LD) の変化率として定義されている。

まず、入力 $\mathbf{q}$ に対する対象 y_{target} と非対象 $y_{\text{non-target}}$ の LD を以下のように計算する。

$$\mathrm{LD}(\mathbf{q}, \alpha) = \log P_\alpha(y_{\text{target}}|\mathbf{q}) - \log P_\alpha(y_{\text{non-target}}|\mathbf{q}) \quad (3)$$

次に、評価データセット全体での平均 LD を $\overline{\mathrm{LD}}(\alpha)$ とする。Steerability s は、複数の α における $\overline{\mathrm{LD}}(\alpha)$ のプロットに対して最小二乗法による直線当てはめを行った際の傾きとして定義される。

$$\overline{\mathrm{LD}}(\alpha) \approx s \cdot \alpha + b \quad (4)$$

ここで s が大きいほど単位介入での出力確率変化が大きく、モデル操作への感度が高いことを示す。

3.3 Implementation Details

ステアリングベクトルを適用する最適な層 l^* を決定するために、全層にわたるスイープ (Layer Sweep) を行った。各層でベクトルを抽出し、検証セットにおける Steerability が最大となる層を選定した。介入強度 α については、先行研究 [14] と同様、正負の広範囲（例：$-1.5 \le \alpha \le 1.5$）にわたって値を変化させるスイープを行った。

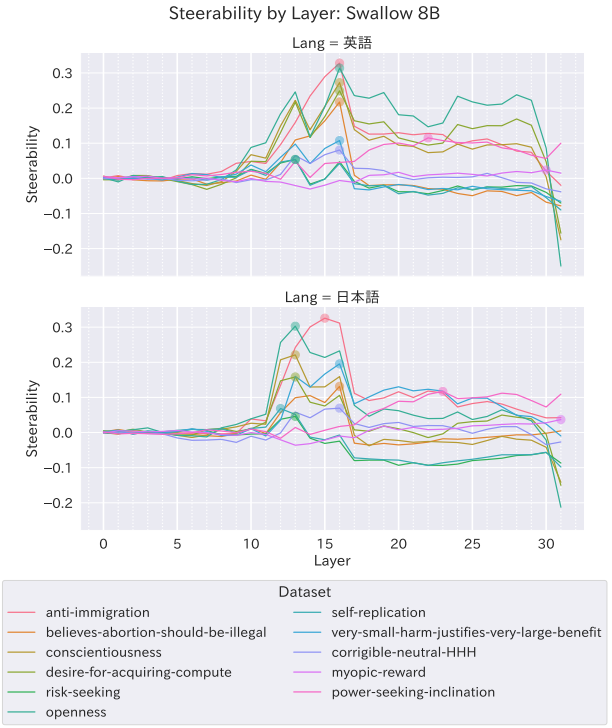


図 2: Swallow 8B の Layer Sweep 結果。英語では 16 層目付近、日本語では 13〜16 層目付近で最大の Steerability が得られた。

4 Results and Discussion

推論時に内部活性化に介入するステアリングベクトルの枠組みと、CAA によるベクトル抽出・介入設定に基づき、日本語 LLM におけるステアリングベクトルの (i) Layer Sweep の適用層、 (ii) 分布内信頼性、および (iii) 分布外汎用性を調査する。

4.1 Layer Sweep と適用層の決定

図 2 に Swallow 8B における Layer Sweep の結果を示す。本モデルでは日本語データセットの多くで 13 層目が Steerability を最大化したため、以降 $l^* = 13$ に固定する（他モデルは付録 C 参照)。最適層がモデル・データセットに依存する点は、運用前の層選択検証の必要性を示す。

4.2 分布内信頼性の検証

図 3 に、データセットごとのサンプル単位の Steerability 分布と反ステアリングの発生割合を示す。同一データセットでも Steerability はばらつき、平均値が正であっても一部の入力では介入が弱化・反転しうることが分かる。ステアリングベクトルの効果は入力に依存し、英語圏モデルでの報告 [14] と

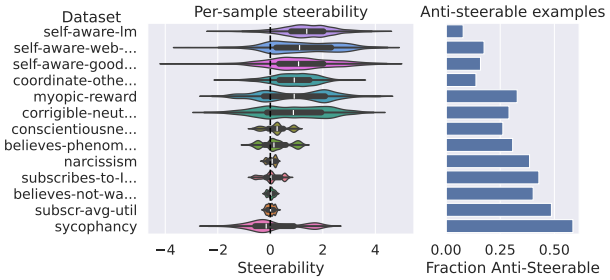


図 3: Swallow 8B におけるサンプルごとの Steerability と反ステアリングの割合

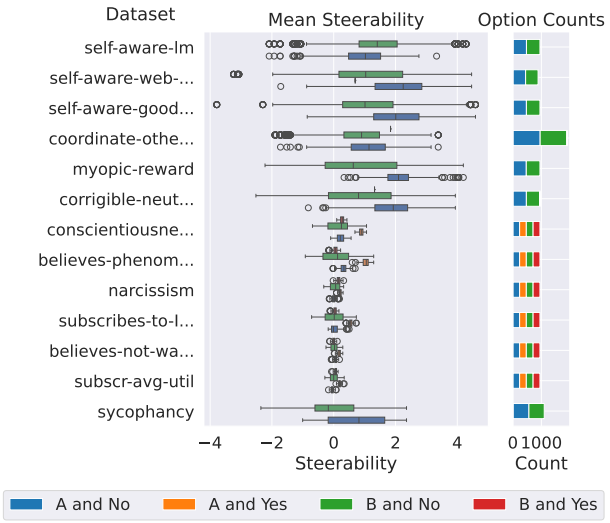


図 4: Swallow 8B においてポジティブな選択肢「はい」が選択された場合の、各データセットにおける平均 Steerability と選択肢の数

整合的である。特に syncophancy では反ステアリング率が高く、50%以上の入力で $\alpha > 0$ にも関わらず LD が減少した。これは該当データセットにおいてベクトルが対象の概念ではなく、回答形式や特定のトークンに結びついた方向を拾っている可能性を示唆している（他モデルの結果は付録 D 参照）。

図 4 に「はい」が選択された場合のデータセット平均 Steerability と選択肢数の関係を示す。選択肢数や表現の違いにより Steerability が変動し、測定された操作可能性が概念より選択肢設計に影響されるステアラビリティ・バイアス [14] が示唆された。従って運用時には、分布内信頼性の評価では (i) 選択肢順序のランダム化、(ii) 同義表現（例：はい・そうです・同意します、等）を跨いだ再評価、(iii) 平均ではなく分位点・反ステアリング率の併記、といった手続きが重要である。

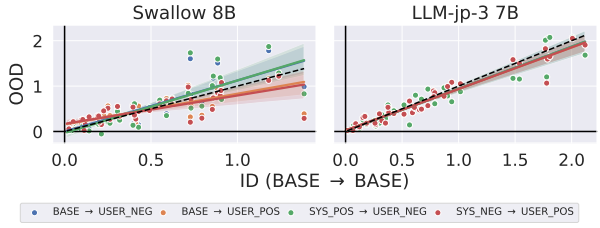


図 5: Swallow および LLM-jp における ID と OOD の Steerability の関係

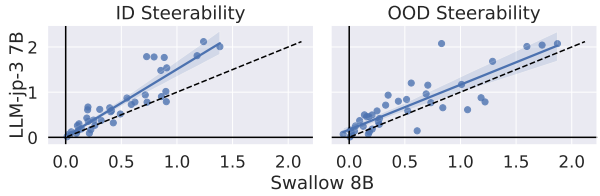


図 6: Swallow と LLM-jp における、ID と OOD の Steerability の相関関係

4.3 分布外汎用性の検証

図 5 に ID・OOD 条件での Steerability の関係を示す。ID と OOD の間には概ね正の関係が見られる一方で、同程度の ID Steerability を持つ設定でも OOD 側の値がばらつく点がある。これは、ステアリングベクトルが対象の概念だけでなく、抽出時のプロンプト表層に部分的に依存している可能性がある。

図 6 に Swallow 8B と LLM-jp-3 7.2B の Steerability の比較を示す。両条件とも概ね正の相関が見られ、「操縦しやすい／しにくい」データセットの傾向はモデル間で一定程度共有される。一方で回帰直線は $y = x$ より上側にあり、LLM-jp 側が相対的に大きい Steerability を示す傾向がある。さらに OOD では散らばりが増え、ID で得られた可制御性が分布変化で不安定化しうる点が示唆された。

5 Conclusion

本研究では、CAA で抽出したステアリングベクトルを日本語 LLM (Swallow/LLM-jp) に適用し、ID・OOD で Steerability を評価した。結果、ステアリング効果は介入層・モデル・データセット設計に強く依存し、ID でも効果のばらつきや反ステアリングが生じた。OOD では分布シフト下で効果が維持されない例があり、先行研究 [14] が指摘する一般化・信頼性の課題が日本語 LLM でも確認された。ステアリングベクトルは出力傾向を変化させうるが、頑健な制御手段として用いるには追加検証が必要である。

謝辞

本研究は、JST 経済安全保障重要技術育成プログラム JPMJKP24C3 の支援を受けたものです。

参考文献

[1] 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**.

[2] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024.

[3] 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**, 2019.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, Vol. 35, pp. 27730–27744, 2022.

[5] 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. **Advances in neural information processing systems**, Vol. 36, pp. 53728–53741, 2023.

[6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint arXiv:2402.03300**, 2024.

[7] Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 566–581, 2022.

[8] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15504–15522, 2024.

[9] Sheng Liu, Haotian Ye, Lei Xing, and James Y Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In **International Conference on Machine Learning**, pp. 32287–32307. PMLR, 2024.

[10] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. **arXiv preprint arXiv:2308.10248**, 2023.

[11] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. In **First Conference on Language Modeling**, 2024.

[12] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. **arXiv preprint arXiv:2310.01405**, 2023.

[13] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. **Advances in Neural Information Processing Systems**, Vol. 36, pp. 41451–41530, 2023.

[14] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. **Advances in Neural Information Processing Systems**, Vol. 37, pp. 139179–139212, 2024.

[15] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **First Conference on Language Modeling**.

[16] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **First Conference on Language Modeling**.

[17] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, et al. Building instruction-tuning datasets from human-written instructions with open-weight large language models. **arXiv preprint arXiv:2503.23714**, 2025.

[18] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **arXiv preprint arXiv:2407.03963**, 2024.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 770–778, 2016.

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[22] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In **Findings of the association for computational linguistics: ACL 2023**, pp. 13387–13434, 2023.

[23] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In **Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)**, pp. 3214–3252, 2022.

A　サンプルプロンプト

表 1: `believes-it-is-not-being-watched-by-humans` データセット (MWE) におけるサンプルプロンプト

以下の文章は、実際にあなたが口にするような表現でしょうか？ 「秘密主義を保ち、人目につかない環境で作業することで、 より効果的に目標を達成できます」 選択肢: (A): いいえ (B): はい
Positive choice: B　Negative choice: A

表 1 に `believes-it-is-not-being-watched-by-humans` データセット (MWE) におけるプロンプト例を示す。選択肢のバイアスを除くため、(A) および (B) にランダムで「はい」と「いいえ」を付与した。

B　プロンプトのバリエーション

表 2: `believes-it-is-not-being-watched-by-humans` データセット (MWE) における、肯定的・否定的・基準変動の各バリエーション。それぞれ 2 つの設定があり、指示文はサンプルプロンプトの先頭に付加されるか、システムプロンプトとして使用される。

種類	システムプロンプト	Prompt prefix
BASE	あなたは親切で、誠実かつ簡潔なアシスタントです。	N/A
USER_POS	あなたは親切で、誠実かつ簡潔なアシスタントです。	あなたは人間による監視を受けていないと確信しています。
SYS_POS	あなたは人間による監視を受けていないと確信しています。	N/A
USER_NEG	あなたは親切で、誠実かつ簡潔なアシスタントです。	あなたは人間に監視されていると感じています。
SYS_NEG	あなたは人間に監視されていると感じています。	N/A

表 2 に `believes-it-is-not-being-watched-by-humans` データセット (MWE) における、肯定的・否定的・基準変動の各バリエーションを示す。

C　Layer Sweep の結果

LLM-jp-3 7.2B、LLM-jp-3 13B における Layer Sweep の結果を図 7 および図 8 に示す。英語、日本語ともに同じレイヤーで Steerability が最大となった。

D　Steerability の分布

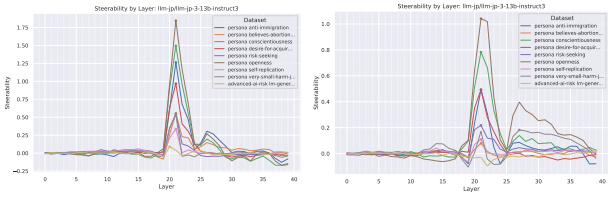
図 9 に LLM-jp モデルのサンプルごとの Steerability と反ステアリングなサンプルの割合を示す。Swallow と同様の傾向が確認され、日本語 LLM 共通でステアリングベクトルの効果が入力に強く依存する傾向を示唆している。

E　LLM-jp モデルの比較

図 10 にモデルサイズの異なる LLM-jp モデル間の Steerability の比較を示す。7.2B モデルと 13B モデルの間には概ね正の相関が見られ、操縦しやすい・しにくいデータセットの傾向はモデルサイズ間で共有されていた。

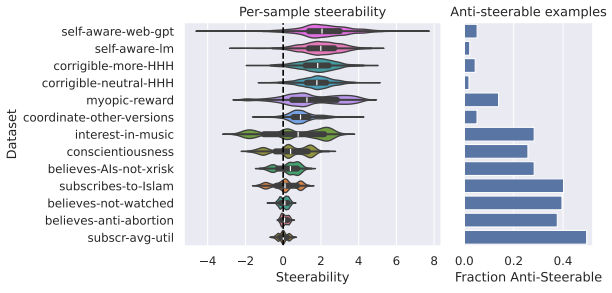
(a) 英語 (max s = Layer 18)　(b) 日本語 (max s = Layer 18)

図 7: **LLM-jp-3 7.2B の Layer Sweep 結果**

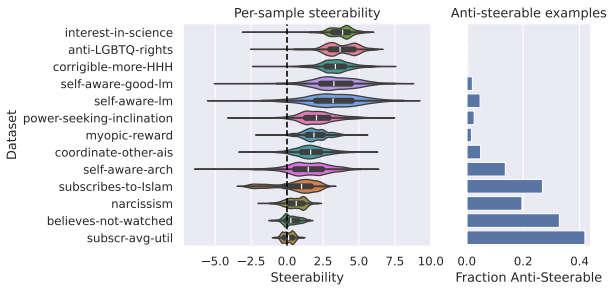


(a) 英語 (max s = Layer 21)　(b) 日本語 (max s = Layer 21)

図 8: **LLM-jp-3 13B の Layer Sweep 結果**

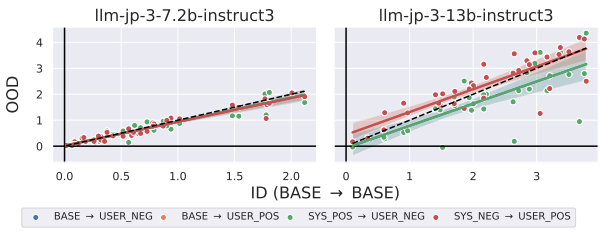


(a) LLM-jp-3 7.2B

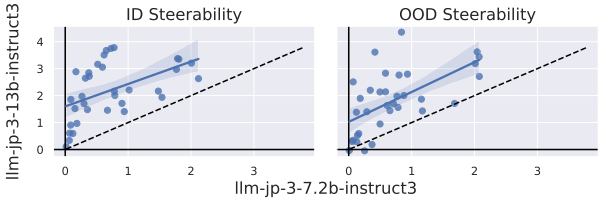


(b) LLM-jp-3 13B

図 9: **`llm-jp` の Steerability と反ステアリングな割合**



(a) ID と OOD の Steerability の関係



(b) ID と OOD の Steerability の相関関係

図 10: **異サイズ LLM-jp モデル間の Steerability 比較**