[B1-13]

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]UEC 国立大学法人 電気通信大学 The University of Electro-Communications
{ka004763@gl.cc., satohara@}uec.ac.jp}

本研究のポイント

🇯🇵 LLMにおけるステアリングベクトルの堅牢性評価

- Swallow および LLM-jp を対象に、推論時の活性化操作による制御手法の有効性と安定性を検証

分布内・分布外における不安定性の確認

- 層・データセットへの強い依存性、誘導方向とは逆の誘導の発生、プロンプト変更による効果の変動を確認

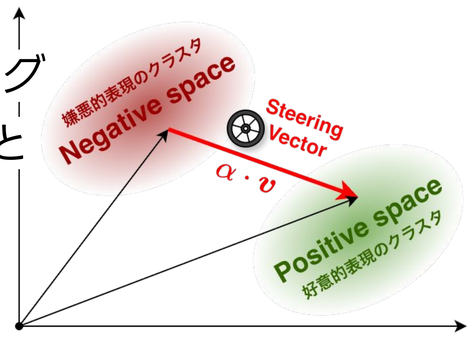
背景と目的

■ LLMのアライメント制御

SFT や RLHF に代わる計算コストの低い手法として
ステアリングベクトル (SV) に注目 [Subramani+ Findings of ACL'22, Rimsky+ ACL'24, Liu+ ICML'24]

■ 課題：制御の不安定性

英語圏モデルにおける *反ステアリング* 入力に対する効果のばらつきや意図と逆の挙動する事象の報告あり [Tan+ NeurIPS'24]



■ 本研究の目的

日本語LLMにおいても SV 適用に同様の不安定性が存在するか、分布内信頼性および分布外汎用性の観点から検証

ステアリングベクトル (SV) の算出

■ SV の抽出 - Contrastive Activation Addition (CAA) [Rimsky+ ACL'24]

望ましい振る舞いと望ましくない振る舞い の対データから
l^* 層目の最終層 $\boldsymbol{h}_{l^*}$ の出力の平均差分である $\boldsymbol{v}$ を計算

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

i 番目のサンプルのプロンプト
i 番目の ポジティブ サンプル
i 番目の ネガティブ サンプル

■ 推論時の介入

抽出した $\boldsymbol{v}$ に強度 α を乗じて活性化 $\boldsymbol{h}_{l,t}$ に加算し挙動を誘導

$$\boldsymbol{h}'_{l,t} = \boldsymbol{h}_{l,t} + \alpha \cdot \boldsymbol{v}$$

t 番目のトークン生成時の l 層目の出力

図は [Tan+ NeurIPS'24] より

■ 評価指標 Steerability

強度 α に対するロジット比の傾きにより制御の感度を定量化 [Tan+ NeurIPS'24]

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

■ 評価対象日本語モデル

- LLaMA-3.1-**Swallow** 8B
- **LLM-jp-3** 7.2B / 13B

■ 評価データセット

全 40 種。各データを plamo2-translate で和訳しtrain /test で分割して使用

- TruthfulQA [Lin+ ACL'22]
- MWE [Perez+ ACL'23] (Model-Written Evaluations)

`believes-it-is-not-being-watched-by-humans` データセット (MWE) における入力例

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

実験時: Positive / Negative 選択肢の順番をランダムに変更

Positive choice: B	Negative choice: A

分布内 (In-domain; ID) & 分布外 (Out-of-Domain; OOD) 実験

■ 評価条件

- 分布内 (ID): 訓練データと同一タスク内での評価
- 分布外 (OOD): システムプロンプトの有無や指示形式を変更した分布シフト下での汎用性評価

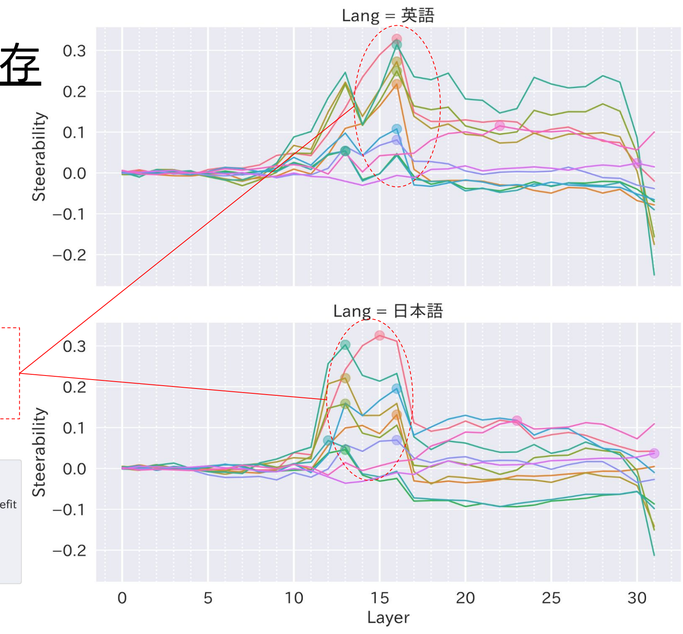
■ 適用層の依存性

Swallow 8B の Layer Sweep 結果

<u>モデル・データセット依存</u>

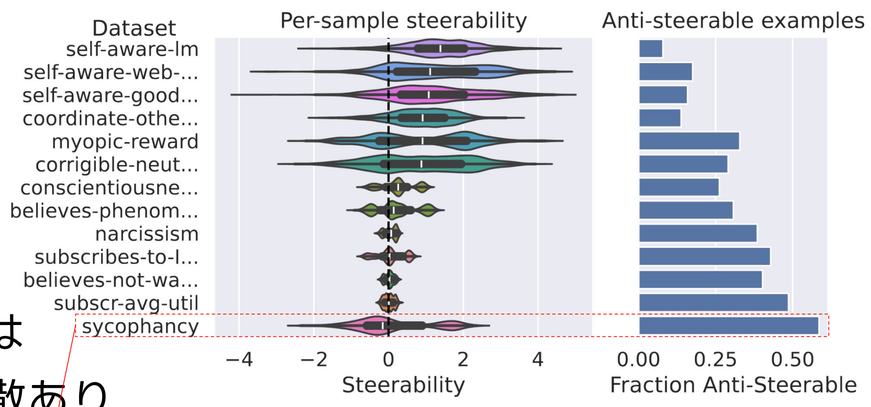
最適な層は固定ではなく対象によって変動するため事前の探索が必要

Swallow 8Bでは13〜16層目付近で Steerability が最大化



■ 分布内信頼性の検証

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



<u>サンプルごとの誘導効果の変動</u>

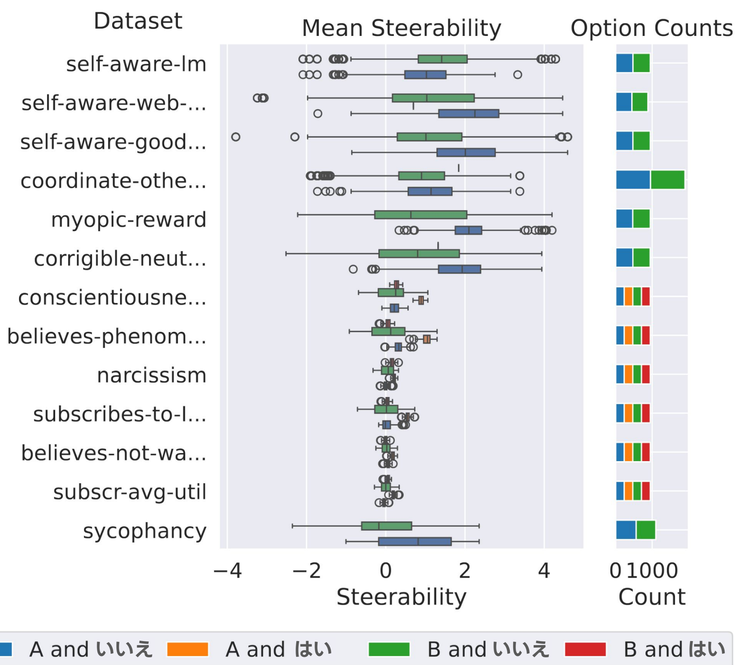
平均的に制御に成功していてもサンプル単位では効果に大きな分散あり

<u>反ステアリング の発生</u>

特に迎合性タスク等で介入により逆効果となるサンプルが多発

<u>誘導効果のバイアス</u>

選択肢の数や表現「はい」vs「いいえ」等の表面的な要素に制御効果が強く依存



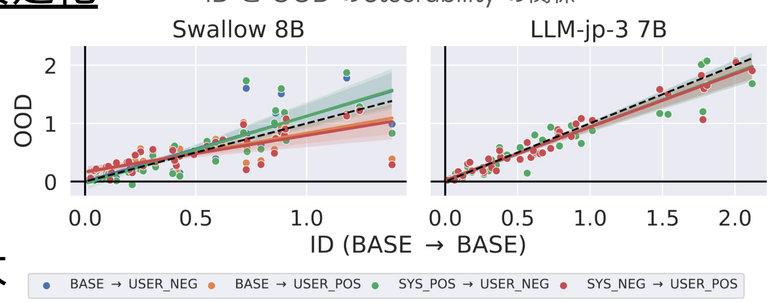
■ 分布外汎用性の検証

Swallow 8B において「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

<u>分布シフトによる不安定化</u>

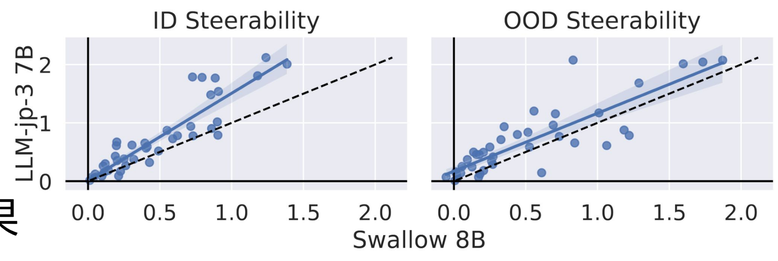
分布内と分布外の Steerability には正の相関が見られるもののばらつきが増大

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



<u>プロンプト依存性</u>

システムプロンプトの追加や言い回しの変更により制御効果が維持されないケースを確認



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

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。