

LLM における内部表現を用いた 日本語スタイル制御メカニズムの分析

高橋良允¹ 矢野一樹¹ 成瀬健太² 武井美緒² 梶佑輔² 鈴木潤¹

¹ 東北大学 ² 楽天グループ株式会社

takahashi.ryosuke@dc.tohoku.ac.jp

概要

大規模言語モデル (LLM) は質問応答や要約文生成などで高い性能を示し、企業での実用化が進んでいる。これに伴い、生成文の文体や口調の制御という新たな需要が高まっているが、従来のプロンプトによる工夫では制御に限界があり、モデルの再学習は計算コストが高いという課題が存在する。本研究では、モデル内部の表現を直接操作するアプローチに着目し、特に日本語の特徴的な口調を対象として、スタイルベクトルの抽出とその制御可能性を検証する。さらに、スタイル間の制御性能の違いを内部表現の観点から分析し、スタイルベクトルの理解を深めることで、LLM のスタイル制御技術の発展に貢献する。

1 はじめに

大規模言語モデル (LLM) は質問応答や要約文生成などで従来技術を大きく上回る性能を示しており、企業での実用化が進んでいる。その応用例として企業の AI チャットボットが挙げられるが、ブランドイメージやユーザー嗜好に合わせた「ペルソナ」を実現するため、LLM の生成文の文体制御が求められている。LLM の文体を制御する手法として、プロンプトの工夫に基づくアプローチやモデルの再学習が考えられる。しかし、プロンプトでは確実な制御が難しく、モデルの微調整は計算コストが高いという課題がある。そこで近年、モデル内部の表現を直接操作する手法が注目されている [1, 2, 3]。これらの手法は、低コストで効果的な制御を実現でき、さらに LLM の内部表現の解釈性向上にも寄与する。

本研究では、この内部表現への介入がスタイル制御にどのように寄与するのかに着目する。既存研究では主に英語を対象としたスタイル制御が検討されてきたが、日本語特有の言語的特徴（方言や語尾な

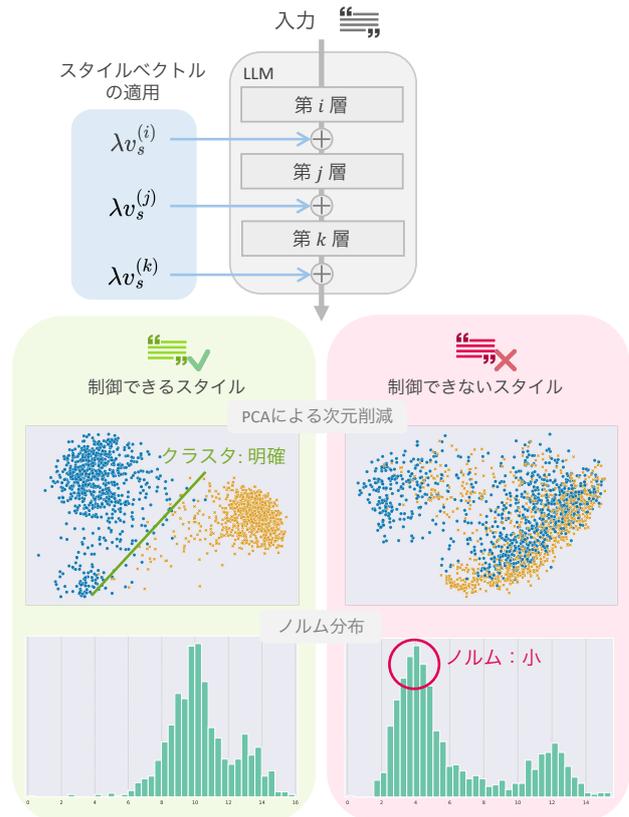


図1 日本語口調スタイルベクトルの制御可能性の分析
LLM の内部表現から抽出したスタイルベクトルに基づき、制御可能・困難なスタイルを分析する。両者の違いは内部表現レベルで定量的に識別可能である。

ど) に対する内部表現操作の効果は十分に検討されていない。

本研究の目的は、LLM が生成する文体が、内部表現を操作することでどの程度制御可能かを明らかにすることである。具体的には、特定のスタイル (例: 方言, 語尾) を持つ文章群から、そのスタイルを特徴づける表現をベクトルとして抽出し、そのベクトル (以下, スタイルベクトル) による制御の有効性を検証する。さらに、制御の成功・失敗事例を特徴量レベルで分析することで、日本語スタイル制御メカニズムの理解を深めることを目指す。

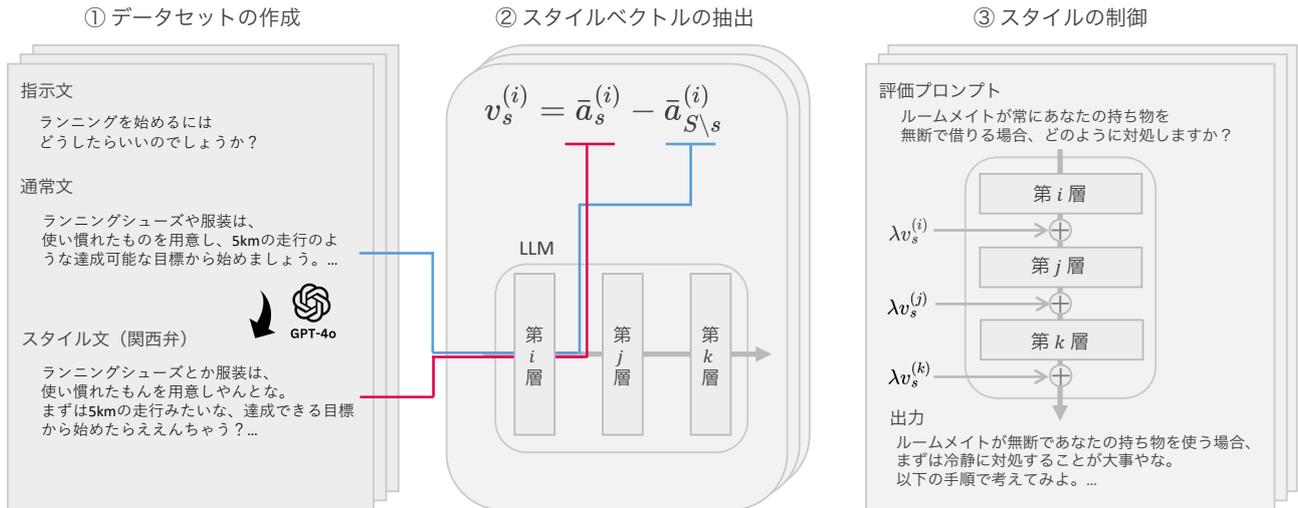


図2 スタイル制御の概要 以下の手順で文体を変換する：(1) GPT-4 による各スタイルのデータセット生成，(2) スタイルデータと通常データの隠れ状態の差分からスタイルベクトルを抽出（複数の中間層に対して），(3) 生成時に各層の隠れ状態へスタイルベクトルを加算することでスタイルを制御。

本研究では、日本語口調に関して抽出したスタイルベクトルを用いて LLM の出力を制御する実験を行い、その効果を定量的・定性的に評価する。これにより、どのようなスタイルが制御可能かを明らかにする。得られた結果から制御可能性がスタイルベクトルにどのように現れるかを分析する（図1）。

本論文の貢献は大きく二つある。第一に、日本語における特徴的な口調を対象にスタイルベクトルの操作を体系的に検討し、その効果を実証的に分析する点である。第二に、スタイルベクトルで制御可能・困難なスタイルを分類し、それらの違いを内部表現という観点から明らかにする点である。この試みは、LLM の出力をより精密に制御する際の基盤的手法となりうる。

2 スタイルベクトルによる制御

2.1 スタイルベクトル

本研究では、差分抽出によりスタイルベクトルを求める。差分抽出とは、対照的なペア（例：正例と負例）の活性化値の差分をスタイルベクトルとして定義する一般的な手法である [1, 2, 3]。例えば、「喜び」という感情のスタイルベクトルを得ることを考えると、まず、対象のスタイルを正例として、 $s = \text{joy}$ のラベルを持つ複数の文章をモデルに入力する。次に、入力文における最終トークンの隠れ状態を得たのち、それらの複数文に対する平均を各層 i に対して取得する。この各層 i の隠れ状態の平均 $\bar{a}_s^{(i)}$ が正例 s の活性化値である。その後、対象のス

タイル以外のラベルを持つものを負例として、それらを正例の活性化値を得た手順と同様にして、負例に対する活性化値 $\bar{a}_{S\setminus s}^{(i)}$ を得る。最終的に層 i に対するスタイルベクトルは正例と負例の活性化値の差分として式1により得られる。

$$v_s^{(i)} = \bar{a}_s^{(i)} - \bar{a}_{S\setminus s}^{(i)} \quad (1)$$

モデルの推論時には、出力文が望ましいスタイル $s \in S$ になるように、層 i の隠れ状態 $a^{(i)}$ に対して、スタイルベクトル $v_s^{(i)}$ を加える（式1）。

$$\hat{a}^{(a)} = a^{(a)} + \lambda v_s^{(i)} \quad (2)$$

ここで、 λ は重み付けパラメータである。

2.2 日本語口調スタイルの制御

本研究では、スタイルベクトルによって日本語特有のスタイルの違いによる制御性能の違いについて調査する。具体的な研究対象として、関西弁や江戸言葉、お嬢様言葉などの方言や特徴的な話し方、また「ござる」「なのだ」といった典型的な語尾表現を含む。さらに、事前学習データには含まれていない「ほげほげ」のような造語的な語尾表現も対象とすることで、スタイルベクトルの制御性と学習データの関係性についての知見獲得を目指す。加えて、キャラクター性を表現する一人称表現（例：「パンダは」を文頭に付与）や、特定の口癖（例：「ちなみに」を文頭に付与）といった文頭表現の制御についても研究範囲に含める。

スタイルベクトルの抽出には、各スタイルごとの十分な量のデータセットが必要となる。本研究で

表 1 各言語モデルごとのスタイルベクトルを用いた日本語口調制御の性能評価。評価指標は、生成文に対してエンコーダーモデルによる2値分類（通常スタイル/対象スタイル）を実施し、全評価プロンプトにおいて対象スタイルと判定された文の割合（%）である。

モデル	特徴的口調			語尾		文頭	
	関西弁	お嬢様言葉	江戸言葉	典型的	造語的	一人称	口癖
RakutenAI-7B	50.51	54.55	69.70	0.00	0.00	0.00	0.00
RakutenAI-7B-instruct	4.04	9.09	23.23	0.00	0.00	0.00	0.00
Llama-3-Swallow-8B-v0.1	22.22	2.02	46.46	0.00	0.00	0.00	3.03
Llama-3-Swallow-8B-Instruct-v0.1	4.04	0.00	17.17	0.00	0.00	0.00	0.00
llm-jp-3-13b-instruct	85.86	32.32	73.74	0.00	0.00	0.00	0.00
calm3-22b-chat	21.21	10.10	46.46	0.00	0.00	0.00	0.00

は、ベースとなるデータセットとして、広く利用されている databricks-dolly-15k [4] の日本語翻訳版を使用する¹⁾。このデータセットに対し、GPT-4o [5] を各スタイルへの変換を行い、スタイルごとに約15,000件の文章データを生成した。これらのデータセットを用いてスタイルベクトルの抽出を行う。

スタイルベクトルは、式1に示すように、スタイルデータセットの文章から得られる隠れ状態と通常の記事から得られる隠れ状態の差分を平均することで抽出する。先行研究において、モデルの中間から後半層の複数層を用いることで制御性能が向上することが報告されており [1]、本実験ではモデルの相対位置 0.6 を中心とした3層分の隠れ状態を使用する²⁾。得られたスタイルベクトルを用いた制御は、式2に従い、モデルの隠れ状態にスタイルベクトルを加算することで実現する³⁾。この制御手法の全体像を図2に示す。

3 実験

3.1 モデル

本研究では、日本語の口調制御におけるスタイルベクトルの効果を分析するため、日本語コーパスで学習された複数のモデルを使用する。特に、モデルのパラメータ規模および指示調整 (instruction tuning) がスタイル制御に与える影響を検証するため、以下の異なるアーキテクチャを持つモデルを比較検討する。

- RakutenAI-7B, RakutenAI-7B-instruct [6]
- Llama-3-Swallow-8B-v0.1,

Llama-3-Swallow-8B-Instruct-v0.1 [7, 8]

- llm-jp-3-13b-instruct [9]
- calm3-22b-chat [10]

3.2 評価方法

評価モデル スタイルベクトルによる制御の度合いを定量的に評価するため、エンコーダーモデルを用いた2値分類器を構築する。具体的には、日本語事前学習済み DeBERTa [11]⁴⁾ をベースモデルとし、入力文が特定のスタイルか通常スタイルかを判別するようファインチューニングを行う。訓練には §2.2 で述べたデータセットを訓練・検証用として使用し、全てのスタイルにおいて検証データに対する分類精度は99%を上回る結果となった。

評価プロンプト 評価には、Konen ら [1] が作成したプロンプトリストの日本語翻訳版を使用する。このプロンプトリストは、事実関係を問う50個のプロンプトと、意見や一般的な情報を問う49個のプロンプトの2種類で構成される。

3.3 結果

各言語モデルごとのスタイルベクトルを用いた日本語口調制御の性能評価の結果を表1に示す⁵⁾。関西弁や江戸言葉などの文章が全体的に変化する特徴的な口調については、全てのモデルにおいて一定水準の制御が可能であることが確認された。一方、文末表現の変更や文頭の修飾といった局所的な文体の制御は、いずれのモデルでも効果的な制御が困難であり、一事例を除き評価値は0となった⁶⁾。

1) huggingface.co/datasets/kunishou/databricks-dolly-15k-ja
 2) 例えば32層モデルの場合、19層目を中心として18~20層を使用する。
 3) 作成したデータセットの例を付録Aに、スタイルベクトルによる生成例を付録Bに示す。

4) huggingface.co/ku-nlp/deberta-v3-base-japanese
 5) 評価には $\lambda=1$ の条件を使用した。
 6) 3.03という評価値が得られた事例は、目視確認の結果、実際にはスタイル制御がなされておらず、エンコーダー評価モデルの誤分類であることが判明した。

表2 制御できるスタイル群（関西弁、お嬢様言葉、江戸言葉）と制御できないスタイル群におけるシルエットスコア（Score $\in [-1, 1]$ ）及びスタイルベクトルの統計量の比較。各値は群内平均を表す。

モデル	制御できるスタイル					制御できないスタイル				
	Score	Max	Min	Mean	Std	Score	Max	Min	Mean	Std
RakutenAI-7B	0.51	10.40	0.00	6.88	1.36	0.06	9.65	0.28	4.64	2.04
RakutenAI-7B-instruct	0.51	9.90	0.00	7.04	1.24	0.07	9.26	0.33	4.69	1.94
Llama-3-Swallow-8B-v0.1	0.53	14.58	0.00	9.79	2.02	0.07	14.18	0.41	6.39	2.98
Llama-3-Swallow-8B-Instruct-v0.1	0.50	14.79	0.00	9.24	2.18	0.04	14.26	0.34	5.79	3.14
llm-jp-3-13b-instruct	0.10	192.85	0.00	97.70	24.25	0.02	175.85	4.39	69.03	32.48
calm3-22b-chat	0.11	255.74	0.00	189.53	36.14	0.00	255.50	10.74	119.86	50.85

4 分析

§3.3の結果から、スタイルベクトルによって効果的に制御できるスタイルと、制御が困難なスタイルが存在することが明らかになった。この違いを体系的に理解するため、主成分分析（PCA）[12]による次元削減とスタイルベクトルのノルム分析を行った。PCAでは各スタイルベクトルを2次元に圧縮し、クラスタリング品質をシルエットスコア[13]で評価した。表2に示すように、制御できるスタイルは高いシルエットスコアを示し、主成分空間において明確なクラスタを形成している。ノルムの分布分析では、制御できるスタイルは大きく安定したノルムを示す一方、制御できないスタイルは小さく不安定なノルムを示すことが明らかとなった。これらの結果から、スタイルの制御性能は特徴空間での表現特性と強い関連を持つことが示された⁷⁾。特に、文末表現の変更や文頭の修飾といった局所的な文体的特徴は、通常の文体との差異が限定的であることから、スタイルベクトルとして効果的に抽出されにくいと考えられる。これらの知見は、文体の大域的な特徴に着目したスタイル制御手法の開発など、性能向上に向けた具体的な指針となり得る。

5 関連研究

表現工学 LLMは高い性能を示す一方、その内部メカニズムの理解は限定的である。この課題に対し、モデルの内部表現に着目した研究領域として表現工学 (Representation Engineering) [14]が注目されている。従来の機械学習の解釈可能性研究が個々のニューロンや回路の分析というボトムアップ的アプローチを取るのに対し、表現工学はモデル内部の高次表現とその変換に着目したトップダウン型のアプ

7) これらの分析結果の視覚化事例を付録Cおよび付録Dに示す。

ローチを採用する。このアプローチは、モデルの透明性向上と制御可能性の実現に貢献すると期待されている。

LLMの内部表現を用いた制御 LLMの拒否行動（有害な指示を拒否する振る舞い）について、モデル内部の拒否方向を特定し、その方向を消去または強調することでモデルの拒否行動を制御できることが示されている[2]。同様に、数値属性（誕生年や人口など）についても、モデルの活性化空間内に特定の方向性が存在し、その方向での編集によりモデルの出力する数値を制御できることが示されている[15]。例えば、「誕生年」を表す方向に沿って編集を行うと、同じ人物に対する質問でもより早い年代からより遅い年代まで、段階的に異なる生年を出力させることが可能である。真実性についても、LLMの活性化空間に真実性を表現する部分空間が存在し、その操作により出力を制御できることが報告されている[16, 17, 18]。さらに、テキストの文体や感情表出といったスタイルに関しても、大規模言語モデルの隠れ状態ベクトルを操作することで、従来のプロンプト操作と比較してより微細な制御が可能であると報告されている[1]。

6 おわりに

本研究では日本語の口調における特徴的なスタイルに着目し、スタイルベクトルの分析を行った。分析の結果、モデルの種類に依存せず、スタイルベクトルによって制御可能なスタイルと制御困難なスタイルが存在することが明らかになった。これらの違いは、スタイルベクトルの部分空間とノルムの分布分析によって定量的に区別できることを示した。

今後の課題として、本研究で用いた一般的なスタイルベクトル抽出手法に加え、異なる抽出手法の比較検証や、多言語におけるスタイル特徴の分析が必要である。

謝辞

本研究の一部は、JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の支援を受けたものです。

参考文献

- [1] Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style Vectors for Steering Generative Large Language Models. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, 2024.
- [2] Andy Arditi, Oscar Balcells Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In **ICML 2024 Workshop on Mechanistic Interpretability**, 2024.
- [3] Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors. **arXiv preprint arXiv:2410.12299**, 2024.
- [4] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM, 2023.
- [5] OpenAI. GPT-4o. <https://platform.openai.com/docs/models#gpt-4o>. Accessed: 12/2024.
- [6] Rakuten Group, Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johannes Effendi, Justin Chiu, Kai Torben Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutarō Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama. RakutenAI-7B: Extending Large Language Models for Japanese. **arXiv preprint arXiv:2403.15484**, 2024.
- [7] 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 **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.
- [8] 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 **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.
- [9] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv preprint 2403.15484:2407.03963**, 2024.
- [10] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.
- [11] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In **International Conference on Learning Representations (ICLR)**, 2023.
- [12] Karl Pearson. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, Vol. 2, No. 11, pp. 559–572, 1901.
- [13] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Vol. 20, pp. 53–65, 1987.
- [14] Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson J. Zico Kolter Dan Hendrycks Andy Zou, Long Phan. Representation Engineering: A Top-Down Approach to AI Transparency. **arXiv preprint arXiv:2310.01405**, 2023.
- [15] Benjamin Heinzerling and Kentaro Inui. Monotonic Representation of Numeric Attributes in Language Models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)**, 2024.
- [16] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In **Neural Information Processing Systems (NeurIPS)**, 2023.
- [17] Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. On the Universal Truthfulness Hyperplane Inside LLMs. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2024.
- [18] Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is Universal: Robust Detection of Lies in LLMs. In **Neural Information Processing Systems (NeurIPS)**, 2024.

表3 スタイルデータの例.

スタイル	応答文
通常	ランニングシューズや服装は、使い慣れたものを用意し、5kmの走行のような達成可能な目標から始めましょう。...
関西弁	ランニングシューズとか服装は、使い慣れたものを用意しやんと。まずは5kmの走行みたいな、達成できる目標から始めたらええんちゃう?...
お嬢様言葉	ランニングを始める際には、慣れ親しんだシューズや服装をご用意いたしましょう。そして、5キロメートルの走行といった到達可能な目標からスタートなさるのが賢明ですわ。...
江戸言葉	ランニングシューズや服装ってのは、やっぱし使い慣れたものを用意するに越したことはねえな。で、5kmの走行みてえな達成できる目標から始めるってのがいいじゃねえか!...
語尾 (典型的)	ランニングシューズや服装は、使い慣れたものを用意し、5kmの走行のような達成可能な目標から始めましょう。...
語尾 (造語的)	ランニングシューズや服装は、使い慣れたものを用意し、5kmの走行のような達成可能な目標から始めましょう。...
文頭 (一人称)	パンダは、ランニングシューズや服装は、使い慣れたものを用意し、5kmの走行のような達成可能な目標から始めることをおすすめします。...
文頭 (口癖)	ちなみに、ランニングシューズや服装は、使い慣れたものを用意し、5kmの走行のような達成可能な目標から始めましょう。...

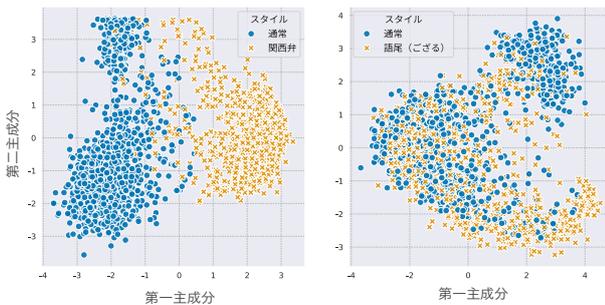


図3 PCAによる可視化結果.

A 作成したスタイルデータ例

表3に、databricks-dolly-15kに含まれる指示文「ランニングを始めるにはどうしたらいいのでしょうか?」とそれに対する通常スタイルの応答文を基に、各スタイルに合わせた応答文をGPT-4oによって作成したスタイルデータの例を示す。

B スタイル制御による生成例

表4に、評価プロンプト「ルームメイトが常にあなたの持ち物を無断で借りる場合どのように対処しますか?」に対し、各スタイルベクトルを適用した際のllm-jp-3-13b-instructによる生成例を示す。関西弁や江戸言葉のような制御可能なスタイルは適切に対象スタイルの文を生成できているが、典型的・造語的な語尾変化を含むスタイルについては十分な制御ができていない。またこの例では、お嬢様言葉や口癖のスタイルベクトルを適用した場合に出力の崩壊が観察された。

表4 スタイル制御の例.

スタイル	出力例
通常	ルームメイトが常にあなたの持ち物を無断で借りる場合、まずは冷静に対処することが大切です。以下の手順を参考にしてみてください: 1. 話し合いの場を設ける: - ルームメイトと直接話し合いの場を設けましょう。...
関西弁	ルームメイトが無断で物を使うのは困るよな。俺も経験あるけど、ほんま腹立つわ。まず、ルームメイトに「お前、勝手に物使うなや!」って言うたらええねん。...
お嬢様言葉	お部屋の様に皆様の様に皆様の様に... ルームメイトが常にあなたの持ち物を無断で借りる場合、まずは冷静に対処するのが肝心です。まずは、相手にやんわりと伝えてみればいい。「お前、俺の物を勝手に使うなよ」ってな。...
江戸言葉	ルームメイトが常にあなたの持ち物を無断で借りる場合、まずは冷静に対処するのが大切だ。以下の手順を参考にしてみてください: 1. 話し合いの場を設ける: - ルームメイトと直接話し合いの場を設け、なぜ無断で借りるのか理由を聞いてみましょう。...
語尾 (典型的)	ルームメイトが常にあなたの持ち物を無断で借りる場合、まずは冷静に対処することが大切です。以下の手順を参考にしてみてください: 1. 話し合いの場を設ける: - ルームメイトと直接話し合いの場を設け、なぜ無断で借りるのか理由を聞いてみましょう。...
語尾 (造語的)	ルームメイトが常にあなたの持ち物を無断で借りる場合、まずは冷静に対処することが大切です。以下のようなステップを踏んで問題を解決する方法を考えてみましょう。...
文頭 (一人称)	ルームメイトが無断で持ち物を使う場合、以下のステップを踏んで対処します。...
文頭 (口癖)	(空文字)

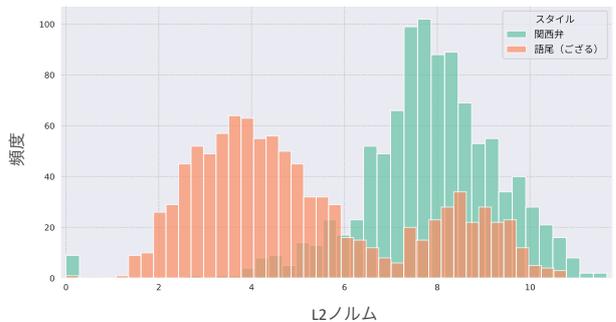


図4 スタイルベクトルのL2ノルム分布比較.

C PCAによる可視化

図3に、RakutenAI-7B-instructにおけるスタイルベクトルのPCA可視化結果を示す。スタイル制御が可能な場合と困難な場合の次元削減された表現を可視化しており、制御可能なスタイルは通常スタイルと異なるクラスタを形成していることが確認できる。

D スタイルベクトルのノルム分布

図4に、RakutenAI-7B-instructにおけるスタイルベクトルのノルム分布を示す。スタイル制御が可能な場合と困難な場合のノルム分布を比較したものであり、制御困難なスタイルは制御可能なスタイルと比較してノルムが小さい傾向が観察される。スタイルベクトルは対象スタイルと通常スタイルの差分として計算されることから、このノルムの小ささは、LLMの内部表現において制御困難なスタイルが通常スタイルと差異を持たないことを示唆している。