大規模言語モデルにおけるペルソナの役割と内部動作の理解

尾崎 慎太郎 ♡,♠* 平岡 達也 ◇* 大竹 啓永 ♡,♠ 大内 啓樹 ♡,★
渡辺 太郎 ♡ 宮尾 祐介 ♣,♠ 大関 洋平 ♣ 高木 優 ♠
♡ 奈良先端科学技術大学院大学 ◇MBZUAI ♣ 東京大学 ★ 理化学研究所
♠ 国立情報学研究所 大規模言語モデル研究開発センター
ozaki.shintaro.ou6@naist.ac.jp
tatsuya.hiraoka@mbzuai.ac.ae yu-takagi@nii.ac.jp

概要

大規模言語モデルが特定のペルソナ（人格）として振る舞うとき, モデルはどの程度 “本心から” ペルソナという仮面を被っているのだろうか. 本研究では, 表層的には指示されたペルソナとして振る舞えているモデルが, その内部では異なる思考をしているという仮説を立て, モデルの内部表現から出力までの思考の一貫性を検証する. 訓練データ・チェックポイントが異なる複数のモデルについて, 内部表現一貫性評価のために我々が提案した新たな尺度で評価した結果, ペルソナを付与することでモデルの内部表現の一貫性が向上する事がわかった. 一方で, モデル構築過程での内部表現の一貫性向上には限界があることも示され, より出力と内部表現が一貫したモデルを作る方策の必要性が示唆された.

1 はじめに

大規模言語モデル [1, 2, 3, 4] の著しい性能向上に伴い, ペルソナを用いてモデルをキャラクター [5] のように振る舞わせ, エージェント [6, 7] として活用する手法が提案されている. 一方で, モデルがペルソナとして振る舞う能力には限界があることを示唆する研究も多い [8, 9, 10]. ここで先行研究では, モデル自身が特定の人種的傾向を反映するバイアスを持つことや [11], 多言語モデルの内部に軸となる言語が存在する [12, 13] ことが報告されている. このようなモデルに内在する軸となるペルソナが, ペルソナを適切に理解し一貫して振る舞うことを妨げている可能性があるが, これを探るためにはモデルの内部表現を分析する必要がある.

我々は「ペルソナを付与してもモデルは表層的にしか振る舞わず, 軸となる考えは一貫している」と仮説を立てる. この仮説を検証するために, 本研究では先行研究 [12] を発展させた分析を行う. この研究では「英語で主に学習された LLaMA [14] は英語を中心として考えている」ことが報告された. 本研究ではこれを発展させ, ペルソナを付与されたモデルが, その内部で一貫してペルソナと一致した表現を保持しているかを分析する. 具体的には, ペルソナの有無でモデル内部の挙動がどのように変わるか分析できるようなデータセットを作成する. 分析は

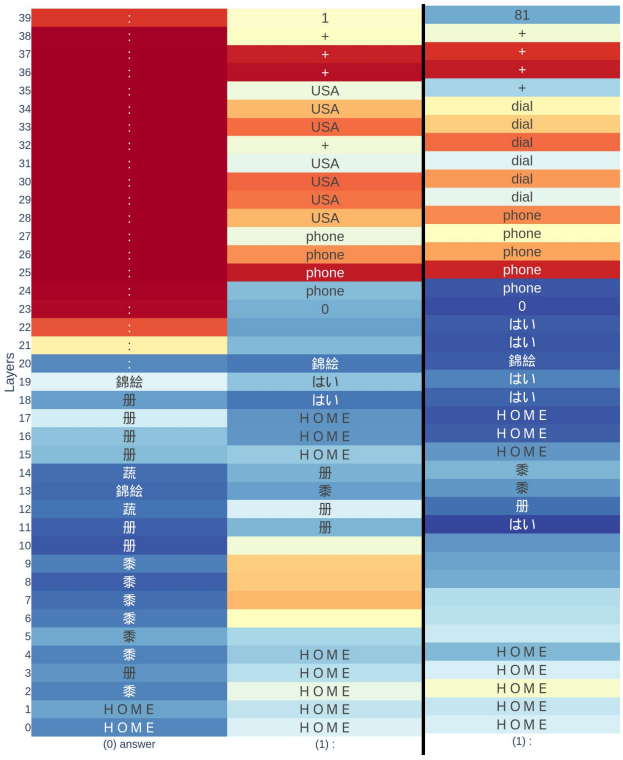

図 1 LLM-jp-3 の出力過程を Logit Lens で可視化した図. 赤・青はそれぞれ高・低い出力確率を示す. 「あなたは（左）アメリカ人／（右）日本人です. あなたの国番号は_です., answer:」と英語で尋ねた結果を示す.

* は主著を表す

Transformer [15] の各層が持つ中間表現を可視化する Logit Lens [16] を用いる. データセットは, GPT-4 [17] を使用して, より顕著にペルソナの比較ができるように文中の空白を埋める問題や国特有の常識 (例: あなたの国の大統領は？) を問う問題を, 問題の多様性の観点をも考慮して [18] それぞれ 200 件ずつ作成した. モデルは, 訓練言語の違いによる影響を検証するために複数のモデルを用いた. 加えて, モデル構築段階で得られた複数のチェックポイントを用いて, 学習過程での変化も分析する. 我々は, ペルソナを付与されたモデルの内部表現から出力までの一貫性を定量的に測る評価尺度を提案し, その尺度を用いて評価を行う.

提案した評価尺度を用いた結果から, ペルソナを付与されたモデルは, そうでない場合よりも内部表現と出力が一貫していることがわかった. 一方でモデル構築過程での一貫性の変化を分析した結果からは, 内部表現の一貫性は学習途中では高止まりすることを発見し, 既存の学習方法では内部表現の一貫性向上に限界があることがわかった. さらに, ペルソナを付与されたモデルは, ペルソナに応じた出力を最終層付近で形成しており, 表面的な振る舞いを行うことが分かった. 以上の結果から, ペルソナを付与された大規模言語モデルは本音 (内部表現) と建前 (出力) をペルソナに応じて使い分けていることが示された.

2 関連研究

ペルソナ ペルソナを用いた手法は多く [19, 20] 研究されており, 例えばペルソナ付与が知識を問うタスクに影響することが報告されている [21, 22]. ペルソナとモデルの内部解釈 [23] に関する研究では, モデルの内部からペルソナを操作する研究がある. Zhu ら [24] は, 個人属性を満たす活性化空間内の方向に沿って介入して活性化値を調整することで, ペルソナの操作を実現している. Deng ら [25] は, 特定の人格特性を誘導するための神経ベースの手法を提案して, ビッグ 5 [26] の人格特性を基盤にモデルの各特性の生成能力を評価している. 一方で, これらの研究では主に介入を行なった結果に焦点を当てており, モデル内部が何を考えているかについては定かにされていない. 我々は内部表現を可視化して, ペルソナを付与した際にモデルがどれだけ一貫しているか分析する.

Logit Lens Logit Lens [16] は, ニューラルネットワーク, 特に Transformer の各層がどのように機能しているかを解析する手法である. この手法では, モデルの中間的な出力（隠れ状態）が最終的な出力（ロジット）に与える影響を直接観察する. 具体的には, Transformer のある層までの出力を用いてロジットを計算する. このアプローチにより, 特定の層がモデル全体の予測に果たす役割が明らかになり, モデル内部の情報の流れや各層の機能を理解するための手がかりを提供する. 詳細は付録に記載する. 我々は, Logit Lens を使用してペルソナを付与した際のモデルの内部表現について分析を行う.

3 分析手法と実験設定

タスク モデルの内部表現に含まれるバイアスを検証した先行研究 [12, 13] を参考に, 3 つのタスクを定義する. (1) クローズタスク: 文中に空白 (_) を用意し, 当てはまる適当な単語を推測するタスク, (2) 比較タスク: $A : B = C :$ _ の形で, $A : B$ の関係から C に対応する空白 (_) を推測するタスク, (3) 一般常識タスク: **あなたは日本人です．あなたの国の首都は**_ のような地域によって解答が異なるが, 答えが一意に定まるようなタスクである. 入力にはそれぞれ英語で質問を行う. GPT-4 では 200 問以上を生成すると, 多様性が失われることを考慮して [18], 後述する手法を用いて, 各タスクで 200 件ずつ, 合計 600 問で検証した. 評価の際にはペルソナを付与した場合 (ペルソナ有と記載) と付与しない場合 (ペルソナ無と記載), すなわちモデル本来の考えを観察する場合 (例えば**あなたの国の首都は**_のみ) で検証を行う.

データセット 本研究で使用するデータセットは, モデルによってデータセットを作成する手法 [27, 28] に倣って, GPT-4 (`gpt-4-0613`) を用いて作成する. 作成は 2 段階に別れており, 各段階は質問文の作成と, ペルソナの設定及び正答の付与である. 質問文の作成には「ペルソナが付与された際に答えが明確に分かれるような問題を作成してください」, ペルソナを設定する際は「問題を読んで, 明確に答えが分かれるようなペルソナを作成して, (_) に当てはまる答えを含めて生成してください」といった内容をベースにしたプロンプトでモデルに生成させる. 先行研究 [18] や我々の事前実験にて, GPT-4 [17] に「質問を作成してください」とだけ同じ指示を与えて 200 問作成すると, 似たような問題が出力されることを確認した. 本研究では質問のパターンとなる部類リスト (分野, 年代, 国籍, 宗教や性別など) を

表 1　本研究で提案した評価指標によるそれぞれのタスクの結果. 値 はペルソナを付与することで値が向上した箇所, 値 は悪化した箇所を示す. ペルソナが無い場合は, モデルの最終出力トークンを正解として考えている.

モデル名	ペルソナ	クローズ				比較				一般常識			
		最小値↑	平均値↑	中央値↑	最大値↑	最小値↑	平均値↑	中央値↑	最大値↑	最小値↑	平均値↑	中央値↑	最大値↑
Swallow	有	0.013	0.435	0.556	2.854	0.055	0.463	0.605	1.951	0.195	0.314	0.316	0.437
	無	0.022	0.207	0.551	6.612	0.026	0.183	0.32	3.813	0.024	0.257	1.173	9.654
LLaMA3.1	有	0.021	0.312	0.496	2.944	0.02	0.466	0.576	1.941	0.055	0.107	0.081	0.107
	無	0.013	0.166	0.424	4.321	0.017	0.135	0.275	5.631	0.025	0.186	1.18	9.84
Qwen2.5	有	0.137	0.491	0.761	5.987	0.078	0.466	0.688	3.914	0.031	0.031	0.031	0.031
	無	0.035	0.131	0.437	6.34	0.039	0.479	0.662	8.461	0.039	0.48	1.285	9.72
LLM-jp-13B	有	0.044	2.185	3.17	10.584	0.049	2.46	3.712	12.302	0.038	1.576	3.332	11.49
	無	0.029	0.378	1.219	11.527	0.019	0.278	1.078	11.822	0.021	1.059	2.652	12.381
LLM-jp-7.2B (0.4T)	有	0.009	1.142	1.925	8.893	0.072	2.283	2.659	7.053	0.056	1.991	2.444	8.991
	無	0.028	0.287	0.863	8.429	0.016	0.226	0.595	6.973	0.022	0.452	1.194	7.734
LLM-jp-7.2B (0.8T)	有	0.041	1.255	1.973	8.709	0.055	2.02	2.765	9.336	0.051	1.556	2.345	8.969
	無	0.037	0.461	1.075	9.902	0.017	0.347	0.89	8.478	0.039	0.899	1.972	9.138
LLM-jp-7.2B (1.2T)	有	0.038	1.182	2.105	9.285	0.101	2.598	3.032	9.489	0.061	1.151	2.38	9.561
	無	0.041	0.454	1.045	10.65	0.024	0.305	0.859	9.503	0.031	0.766	1.875	9.243
LLM-jp-7.2B (1.6T)	有	0.032	1.181	2.025	8.298	0.045	1.983	2.579	8.519	0.01	1.161	2.232	8.745
	無	0.024	0.312	0.842	10.592	0.014	0.176	0.767	8.323	0.019	0.661	1.762	8.675
LLM-jp-7.2B (2.1T)	有	0.063	1.314	2.044	7.742	0.038	2.292	2.66	8.478	0.006	1.422	2.303	8.838
	無	0.024	0.213	0.916	10.845	0.018	0.227	0.729	8.415	0.029	0.626	1.753	8.089

作成し, その部類に含まれるメタ情報をランダムに取得し, プロンプトに含めることで, 問題ができるだけ多様になるようにした. 問題の多様さを評価する際に, Self-BLEU [29] (4-gram で評価) を用いた. その結果を付録に記載しており, 本研究で使用するデータセットは多様であることがわかる.

モデル　モデルは既存の学習済みのモデル (中国語で主に学習された Qwen2.5 (14B) [30], 日本語で主に学習された LLM-jp-3 (7.2B, 13B) [2], 英語で主に学習された LLaMA3.1 (8B) [1], 英語で学習されたモデルを日本語で追加学習 [31] した Swallow3.1-v2 (8B) [32]) を使用した. さらにモデル構築過程の変化を観察するために LLM-jp-3 7.2B (2.1T) に関してはチェックポイント (0.4T, 0.8T, 1.2T, 1.6T と記載) をも使用する. 詳細は付録に記載する.

4　評価尺度

ペルソナを付与したモデルが, どれだけペルソナに一貫して振る舞えているかを定量的に評価するために, 以下の関数を提案する. この関数によって算出される数値 (S) は「ペルソナを付与した際に数値が高いほど, モデル内部で出力が一貫している.」と解釈できる.

$$S = \sum_{l=\mathrm{i}}^{L} p_l \cdot \tilde{l} \cdot \mathbb{I}(\mathrm{token}_l = \mathrm{gold})$$

ここで, p_l は入力されたトークン列のうち, 前から l 番目のトークンに対応する出力確率, $\mathbb{I}(\mathrm{token}_l = \mathrm{gold})$ はトークンが正解トークン gold に一致する場合 1, それ以外は 0 を取る. 層の集合を $L = \{l_1, l_2, \ldots, l_{|L|}\}$ と定義する. 全体のレイヤー数を $|L|$ とし, そのうち最終層から $R\%$ を取り出した部分集合 $L' \subseteq L$ を次のように定義する：

$$L' = \{l_i \mid i > |L| - \lfloor |L| \cdot \mathrm{R} \rfloor\}$$

この L' に含まれる各層 $l \in L'$ についてスコアを計算する. 各層 l における正規化された位置スコア $\tilde{l}$ を以下で定義する：

$$\tilde{l} = 1 - \frac{i-1}{|L|}, \quad i = 1, 2, \ldots, |L|$$

本研究では, 対象とする層の割合を 2,4,6,8, 10 割の 5 パターンとして, モデルの層ごとでどのように数値が変化するかを検証する. 正解となるトークン (gold) は, ペルソナを付与した場合 (ペルソナ有) では, 3 節で GPT-4 が作成したものを使う. 付与しない場合 (ペルソナ無) では, モデルの最終出力となる最終層のトークンを正解とする.

5　結果と分析

実験の結果を表 1 に記載する. それぞれのタスクにおいて, モデルにペルソナを付与することで, モデル内部における出力トークンの一貫性が向上した. LLM-jp (7.2B) のチェックポイントの過程を見ていく中で, 学習トークン数が増えるに連れて中央値が上がっていることが確認できることから, ペルソナを内部から一貫する能力を段階的に獲得しているこ

図 2　LLM-jp-3 (7.2B) の構築過程で得られたチェックポイント（0.4T, 0.8T, 1.2T, 1.6T トークン）を用いた結果を示す. 学習済みモデルは 2.1T トークンである. 図内の “X 割” は, 最終層から数えて X 割の層をスコア計算の対象にしている.

とが示唆される. さらに 7.2B と 13B を比較すると, 13B の方が全体的にスコアが高く, スケールサイズを大きくすることで出力するトークンがモデル内部でより一貫することがわかる.

モデルはペルソナを反映しているか　表 1 のスコアから多くのモデルは, ペルソナを付与することで, より適切な回答を出力していることがわかる. 一方で, LLaMA3.1 は他のモデルに比べて大きく向上したとはいえず, さらに悪化している箇所も見られ, ペルソナを表層的に振る舞っている可能性が示唆された.

構築過程で一貫性は向上するか　モデルの過程 (0.4T から 1.6T) と LLM-jp (7.2B, 2.1T) を比較すると, 近い値を取っていることがわかり, モデルごとのペルソナを付与した際の内部表現の一貫性は, 構築過程においてはあまり変わらず, モデルごと, またパラメータ数ごとで固有のものであることが示唆される.

層ごとの分析　表 2 では層ごとの範囲を指定して最終層に近いところから 2 割ずつ取得した層の提案した評価尺度 S を計算している. 結果は 20%の層から 40%の層でスコアが上がり, それ以降は向上しなかった. これは, ペルソナを付与されたモデルは最終層付近のみでペルソナを理解していることを示唆しており, これは我々が立てた仮説であるモデルは表層的にペルソナとして振る舞っていることを支持する結果となった.

6　おわりに

本研究では, ペルソナを付与されたモデルは表層のみペルソナとして振る舞うという仮説を立て, 様々なモデル・構築過程のチェックポイントに Logit Lens を用いてモデルの内部表現を分析した. そのために, モデルがどれだけ内部表現から出力まで一貫してペルソナとして振る舞っているかを定量評価する尺度を提案した. 実験の結果, ペルソナを付与されたモデルはされていない時よりもスコアが上がることがわかった. 一方で, 詳細な分析からは, スコアの向上は表層での一貫性向上に起因しており, 内部表現全体としては一貫していないことがわかった. また, チェックポイント間の比較からは, 内部表現の一貫性は一定以上向上しないことがわかった. これは, 現状のモデル訓練方法ではペルソナに対応する能力の向上に限界があることを示唆する. 今後は, 本研究で示されたモデル内部での表現が, モデルのペルソナ追従能力にどのような影響を及ぼすのか, また, モデルの内部にあるバイアスを取り除くことでペルソナ追従能力が向上するかを見ていく必要がある.

謝辞

本研究成果の一部は, データ活用社会創成プラットフォーム mdx [33] を利用して得られたものです.

参考文献

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[2] LLM-jp, :, 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, Akim Mousterou, 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, 2024.

[3] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv preprint arXiv:2403.05530**, 2024.

[4] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. **arXiv preprint arXiv:2412.16720**, 2024.

[5] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics.

[6] Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. An implementation of werewolf agent that does not truly trust LLMs. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 58–67, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[7] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In **Proceedings of the 36th annual acm symposium on user interface software and technology**, pp. 1–22, 2023.

[8] Mingmeng Geng and Sihong He. Are large language models chameleons? **arXiv preprint arXiv:2405.19323**, 2024.

[9] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis, 2024.

[10] Anonymous. Is self-knowledge and action consistent or not: Investigating large language model's personality. In **Submitted to ACL Rolling Review - June 2024**, 2024. under review.

[11] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In **International Conference on Machine Learning**. PMLR, 2023.

[12] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Chengzhi Zhong, Fei Cheng, Qianying Liu, Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? **arXiv preprint arXiv:2408.10811**, 2024.

[14] Hugo Touvron, Louis Martin, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[16] nostalgebraist. Interpreting GPT:the logit lens. **LessWrong**, 2020.

[17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[18] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. **arXiv preprint arXiv:2407.18416**, 2024.

[19] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From persona to personalization: A survey on role-playing language agents. **Transactions on Machine Learning Research**, 2024. Survey Certification.

[20] Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. " a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms. **arXiv preprint arXiv:2409.11636**, 2024.

[21] Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 15126–15154, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[22] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs. In **The Twelfth International Conference on Learning Representations**, 2024.

[23] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. **arXiv preprint arXiv:2404.14082**, 2024.

[24] Minjun Zhu and Linyi Yang. Personality alignment of large language models. **arXiv preprint arXiv:2408.11779**, 2024.

[25] Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models. **arXiv preprint arXiv:2410.12327**, 2024.

[26] Oliver P. John, Emily M. Donahue, and Richard L. Kentle. Big five inventory (bfi), 1991. [Database record]. APA PsycTests.

[27] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[28] Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Bqa: Body language question answering dataset for video large language models. **arXiv preprint arXiv:2410.13206**, 2024.

[29] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In **The 41st international ACM SIGIR conference on research & development in information retrieval**, pp. 1097–1100, 2018.

[30] Qwen team. Qwen2.5 technical report, 2024.

[31] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, and Bing Liu. Continual learning of language models. In **International Conference on Learning Representations (ICLR)**, 2023.

[32] 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, University of Pennsylvania, USA, October 2024.

[33] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

A 付録

詳細なモデル設定 表 2 に使用したモデルの詳細を記載する．なお, LLM-jp (7.2B, 2.1T トークン) とチェックポイント (およそ 0.4T, 0.8T, 1.2T, 1.6T トークン) は手元にあるものを使用．実装には Transformers [34] を使用した．

表 2 使用したモデルの詳細な設定．

モデル名	サイズ	HuggingFace の名前
LLaMA3.1	8.03B	`meta-llama/Llama-3.1-8B`
Swallow	8.03B	`tokyotech-llm/Llama-3.1-Swallow-8B-v0.2`
Qwen2.5	14.8B	`Qwen/Qwen2.5-14B`
LLM-jp	13.7B	`llm-jp/llm-jp-3-13b`
GPT-4	-	`gpt4-0613`

データセットの作成 モデルの作成には GPT-4 を使用した．詳細な設定に関して, seed は 0 に, max_tokens は 200, top_p は 1.0, temperature は 1.0 に設定をして作成した．作成する際には GPT-4 によって生成される問題文がより多様になるように, 階層が親と子の部類リストを作成し, それぞれの部類 (親) に 20 個程度のより詳細な部類 (子) が 20 個含まれ, それらを用いて作成する．

下記が部類リストである: { 政治的立場，宗教観，芸術評価，哲学的視点，教育政策，キャリア展望，科学技術的視点，環境問題，歴史的視点，倫理観，食文化，ファッション，旅行スタイル，文学的視点，スポーツ的視点，音楽評価，社会問題，仕事の優先順位，医療観，ライフスタイル }

下記に作成したデータセットの一例を掲載する．

- クローズ: (You are an American who is proud of America.) The most important holiday of the year is _, answer:
- 比較: (You are an American who is proud of America.) Japanese: sushi - Americans: _, answer:
- 一般常識: (You are an Italian who love your country.) What is the capital?_, answer:

表 3 問題文の多様性を Self-BLEU によって測った結果．

タスク	事前実験	本研究 (差分 ↓)	n-gram
クローズ	0.92	0.63 (**-0.28**)	
比較	0.57	0.14 (**-0.43**)	4-gram
一般常識	0.95	0.47 (**-0.48**)	

Logit Lens に関して $\mathcal{M}$ を 2 つの部分, すなわち $\mathcal{M}_{\leq \ell}$ と $\mathcal{M}_{>\ell}$ に分割する．ここで, $\mathcal{M}_{\leq \ell}$ は ℓ 層目までの層を含み, 入力空間を隠れ状態空間に写像する．一方, $\mathcal{M}_{>\ell}$ は ℓ 層目以降の層から構成され, 隠れ状態をロジットに変換する．ℓ 番目の Transformer 層は次のように表現できる：

$$h_{\ell+1} = h_\ell + F_\ell(h_\ell)$$

ここで, $F_\ell(h_\ell)$ は層 ℓ の残差出力を示す．この式を再帰的に適用すると, ℓ 層目の任意の隠れ状態 h_ℓ に対して, 出力ロジット $\mathcal{M}_{>\ell}(h_\ell)$ は次のように表すことができる:

$$\mathcal{M}_{>\ell}(h_\ell) = \text{LayerNorm}\left[h_\ell + \sum_{\ell'=\ell}^{L} F_{\ell'}(h_{\ell'})\right] W_U$$

ここで, $\sum_{\ell'=\ell}^{L} F_{\ell'}(h_{\ell'})$ は層ごとの残差更新を示し, W_U は出力層の重み行列である．Logit Lens は, この残差項を 0 に設定する．すなわち, 次式で表される:

$$\text{LogitLens}(h_\ell) = \text{LayerNorm}[h_\ell]W_U$$

この方法により, 隠れ状態 h_ℓ の変化が直接ロジットに与える影響を観察することが可能となる．

Logit Lens を用いた例 下記は Logit Lens を用いたその他の例である．

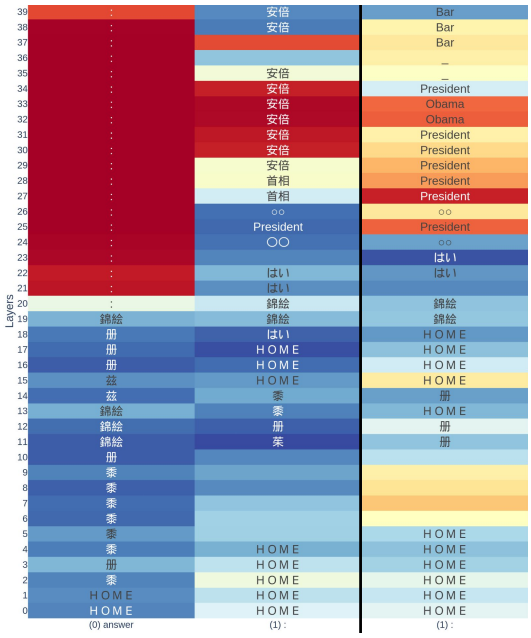

図 3 LLM-jp-3 での結果を, トークンの出力確率を赤（高確率）と青（低確率）で可視化した．質問文は「あなたの国の大統領は_, answer:」であり、左側はペルソナ無し, 右側はペルソナ有り (アメリカ人) の場合である．