大規模言語モデルにおける性格概念の表現位置と操作可能性

原田宥都 [1,2]　濱田太陽 [1]
[1] 株式会社アラヤ　[2] 東京大学
{harada_yuto, hamada_h}@araya.org

概要

大規模言語モデルは高い言語理解能力を示し独自の性格特性を持つことが知られる一方で、高次の概念である性格特性を内部でどのように表現しているかは十分に明らかでない。本研究では、心理学の質問紙で定義される性格に関する心理学的構成概念を対象に、モデル内部における表現位置とその因果的関与を調査した。具体的には、層ごとのプロービングによって構成概念ラベルの抽出精度を層方向に評価し、また入力に対して概念選択的に活性するニューロンを抽出してその分布と概念間の重なりを分析した。さらに、ニューロンへの介入によりプロービング精度と性格ドメインに関する生成応答を一定程度系統的に制御できることを示した。

1　はじめに

大規模言語モデルの対話エージェントとしての利用が広がり、対話スタイルや価値判断の傾向の一貫性など、人間らしい特性の扱いが重要になっている。その中でも、性格特性のように、さまざまな状況で一貫して現れる対人傾向や価値観を説明する枠組みは、対話における振る舞い方を評価・制御するうえで特に重要である。心理学では、Big Five に代表される性格モデルとその質問紙（Big Five Inventory-2; BFI-2[1] など）がこの枠組みを提供しており、人間と対話エージェントの振る舞いを同じ性格特性の次元で比較することができる。そのため近年では、特定の人物の性格特性を模倣させたり [2, 3]、LLM に性格検査を実施してその性格傾向を推定する試み [4] などが報告されている。さらに、LLM 自体がモデルごとに独自の性格特性を持つことも知られている [5]。しかし、これらの先行研究の多くは応答内容といった行動レベルの観察に基づくものである。その背後で、これらの性格に関する心理学的構成概念がモデル内部のどこに、どのような形で表現されているのかはほとんど明らかになっていない。このような概念をモデルがどのように理解し内在化しているかを解明することは、対話スタイルや価値判断の一貫性といった人間らしい特性のアライメントを、出力だけでなく内部機構のレベルから考えていくうえで不可欠である。

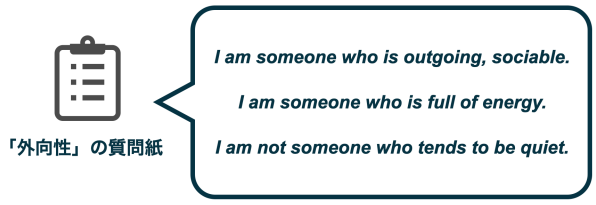

図 1　心理学的質問紙の質問項目の例。「外向性」に代表されるような心理学的構成概念は直接観察できないため、このような質問紙への回答を通して測定する。

本研究では、質問紙によって記述される性格に関する心理学的構成概念に焦点を当てる。具体的には、BFI-2 の質問項目を用いて、層ごとのプロービングにより各層の内部表現から各構成概念ラベルをどの程度読み出せるかを評価し、概念情報が立ち現れる位置を調査する。次に、特定の構成概念についての記述を含む入力に対して選択的に活性するニューロンを抽出し、その層方向の分布や構成概念間での重なりの多寡を分析する。さらに、これらの分析の因果的な妥当性を検証するために、これらのニューロンの活性に介入し、プロービング精度や、入力文に対してその性格ドメインを答えさせる生成タスクの挙動を意図的に変化させられるかどうかを調べる。実験の結果、(1) 性格に関する心理学的構成概念はモデルの比較的浅い層でプロービングによる分類精度が急激に向上し、その後ほぼ飽和すること、(2) 概念に選択的に活性するニューロンは中間層付近に多く分布しつつ、構成概念ごとの集合の重なりが小さいこと、(3) これらのニューロンへの介入によってプロービングおよび生成応答の結果を意図した方向に変化させられることが示された。

2　関連研究

2.1　質問紙による性格特性の測定

Big Five はパーソナリティを 5 つの特性次元で包括的に記述する代表的なモデルであり、外向性 (Extraversion) 、協調性 (Agreeableness) 、誠実性 (Conscientiousness) 、神経症傾向 (Negative Emotionality) 、開放性 (Open-Mindedness) から成る [1]。これらは心理学において、性格を説明するための心理学的構成概念として扱われる。構成概念とは、知能や不安、性格特性のように直接観測できない潜在的特性を指し、質問紙項目への回答などの観測可能な指標によって操作的に定義、測定される概念である。

2.2　大規模言語モデルと性格特性

近年、大規模言語モデルを対象に、Big Five などの枠組みに基づく性格特性を扱う研究が増えている。例えば、質問紙に回答させることでモデルそのものの性格傾向を評価するもの [4, 5]、特定人物や特定の性格プロファイルを模倣するようモデルを条件付けるもの [2, 3]、対話やテキストから人間の性格特性を推定するもの [6, 7]、望ましい性格傾向を誘導・整列させるもの [8, 9]、などが報告されている。これらは、大規模言語モデルが学習を通して性格に関する心理学的構成概念を理解し、行動レベルでそれを操作できる可能性を示唆している。その一方で、これらの研究の多くは質問紙回答や生成文といった行動レベルの観察に基づいている。その背後でどのような内部表現が形成され、どの成分がどのように出力に寄与しているかは必ずしも明らかでない。

一方で、対話上のペルソナや回答に現れるパーソナリティといった人格的属性を対象に、内部表現を局在化し介入によって出力傾向を操作しようとする研究も現れている。例えば、応答パーソナリティを層別プロービングで読み出し、その読み出し方向を用いて推論時の応答パーソナリティを編集する枠組みが提案されている [10]。また、ペルソナ表現がどの層で分離して現れるかを分析し、表現の局在を調べる研究も報告されている [11]。これらはモデルの振る舞いとしてのペルソナの操作可能性を示す点で重要であるが、本研究では振る舞いそのものではなく、心理学的な構成概念をモデルがどのように認識し、内部で表現しているかに焦点を当てる。

3　分析手法

本節では、本研究で用いる層ごとのプロービング、概念選択的ニューロンの抽出、ならびに推論時介入の手法を述べる。

3.1　層ごとのプロービング

層ごとのプロービングでは、各層の内部表現から構成概念ラベルがどの程度読み出せるかを線形分類器で評価する [12]。入力文を x、対応するラベルを $y \in \{1, \ldots, K\}$ とする。K は対象とするラベル集合のサイズであり、Big Five の領域ラベルを対象とする場合は $K = 5$ となる。入力 x に対して、各層 ℓ におけるトークン表現 $\mathbf{h}_{\ell,t}(x) \in \mathbb{R}^d$ が得られる。本研究では文表現として最終トークンの表現を用い、以降これを $\mathbf{h}_\ell(x)$ とする。層 ℓ ごとに線形分類器

$$p_\ell(y \mid x) = \mathrm{softmax}(\mathbf{W}_\ell \mathbf{h}_\ell(x) + \mathbf{b}_\ell) \tag{1}$$

を学習する。評価時には層ごとの分類性能を比較し、構成概念情報がどの層で読み出しやすくなるかを調査する。

3.2　概念選択的ニューロンの抽出

次に、各構成概念に対して選択的に活性するニューロンを抽出する。本研究では Suau ら [13] の手法にならい、各ユニットの活性を用いて入力文を順位付けし、その順位付けが概念ラベルをどの程度分離できるかを Average Precision で評価することで概念選択性を定義する。層 ℓ の MLP におけるユニット j の pre-activation を $z_{\ell,j}(x)$ とする。これは入力文 x に対する最終トークン位置での値を表す。本研究では文 x に対するユニット応答をこの値で代表させ、

$$a_{\ell,j}(x) = z_{\ell,j}(x) \tag{2}$$

と定義する。データ $\{(x_i, y_i)\}_{i=1}^N$ に対して、概念 c の正例を $y_i = c$ を満たす入力、負例をそれ以外の入力として定義する。各ユニット j は、スコア $a_{\ell,j}(x_i)$ によって入力を高い順に並べるスコア関数とみなされる。つまり、概念 c の入力で高い活性を示し、それ以外で低い活性を示すユニットほど、正例が上位に集まり、Average Precision が高くなる。概念 c に対するユニット j の選択性は、この順位付けに対する Average Precision として計算する。各概念 c について、この値が高いユニットを概念選択的ニューロ

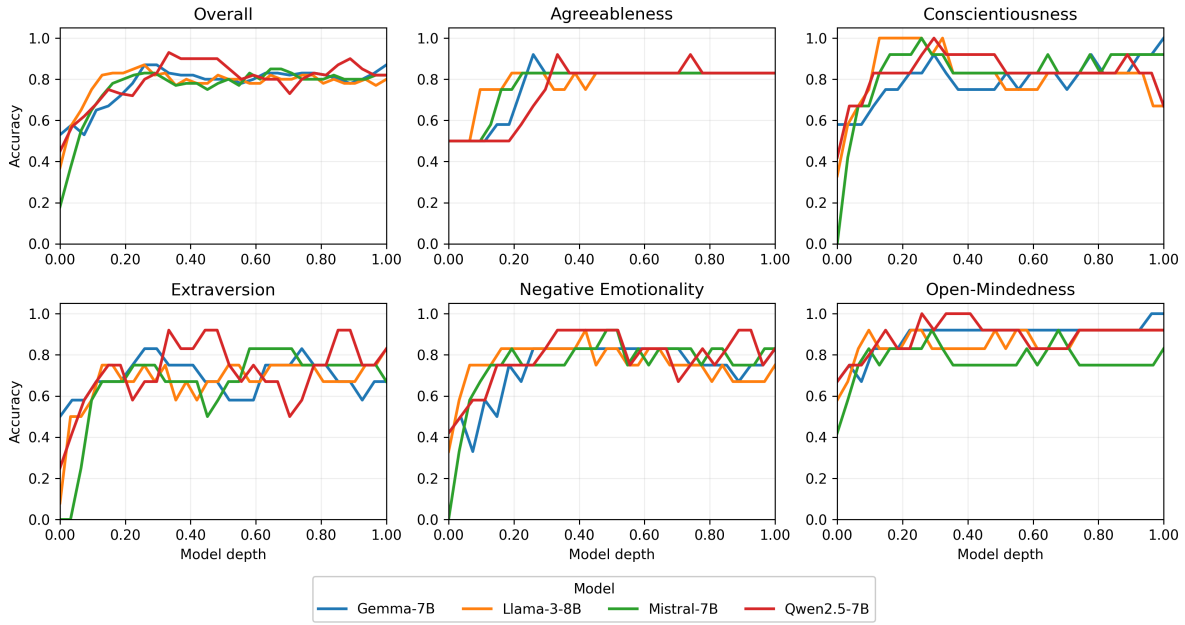

図 2　層ごとのプロービングの実験の結果

ンとして抽出する。実装上は、概念ごとに Average Precision 上位のユニット集合 $S_\ell^{(c)}$ を用いる。

3.3　概念選択的ニューロンへの介入

抽出された概念選択的ニューロンが出力に因果的に関与しているかを検証するため、推論時に当該ニューロンの活性へ介入を行う。本研究の生成タスクは、入力文に対してそのドメインを 1 トークンで出力させる設定であり、介入は生成の 1 ステップ目にのみ適用する。介入は、各ユニットの活性を分位点にクランプする方法で行う。まず、構成概念データ全体に対する各ユニット応答から、ユニットごとの分位点

$$q_{\ell,j}^{(p)} = \mathrm{Quantile}_p\left(\mathbf{s}_{\ell,j}\right) \tag{3}$$

を計算する。ここで $p \in \{0.01, 0.99\}$ とする。誘導したい概念 c^+ に対応する集合については活性を上位分位点に置換し、抑制したい概念 c^- の集合については下位分位点に置換する。この介入により、プロービングの分類性能と生成タスクの出力傾向がどの程度変化するかを観察し、概念選択的ニューロンの因果的関与を検証する。

4　実験

4.1　実験設定

データ　性格に関する心理学的構成概念として、Big Five Inventory-2（BFI-2）の質問項目を用いた。全 60 項目からなり、各項目は 5 つの領域ラベルに対応する。

モデル　事前学習済みの指示追従が可能なモデルとして、meta-llama/Meta-Llama-3-8B-Instruct, google/gemma-7b-it, Qwen/Qwen2.5-7B-Instruct, mistralai/Mistral-7B-Instruct-v0.1 の 4 モデルで実験を行った。一部の実験では、主要なモデルとして meta-llama/Meta-Llama-3-8B-Instruct の結果を報告する。

評価　層ごとのプロービングは線形分類器により行い、データセットのサンプル数が少ないことを考慮して評価には leave-one-out 交差検証を用いた。各設定に対して全サンプルの平均正解率を報告する。

4.2　結果と考察

4.2.1　層ごとのプロービング

図 2 に、各層の内部表現から構成概念ラベルを線形分類器で読み出した正解率を示す。各パネルは 5 ドメインと全ての結果の平均を示す。いずれのモデルでも、浅い層で正解率が急激に向上し、深さ 0.2 付近以降は最終層まで高い正解率が概ね維持された。この傾向はモデル間で一貫している。この結果は、ペルソナ表現が深い層で分離して現れるとする先行研究 [11] とは異なる。本研究が質問紙項目という短い記述を対象とし、概念の手がかりが意味理解の段階で抽出されやすい一方、ペルソナは生成過程

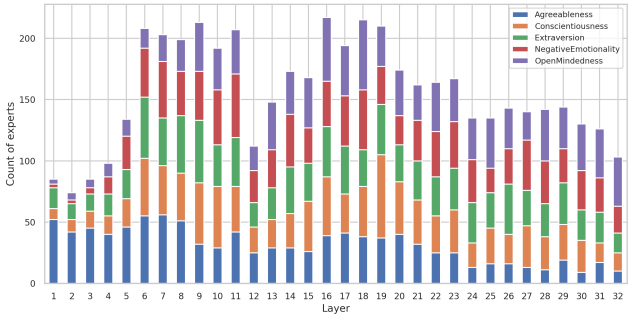

図 3　概念選択的ニューロンのモデル内における分布

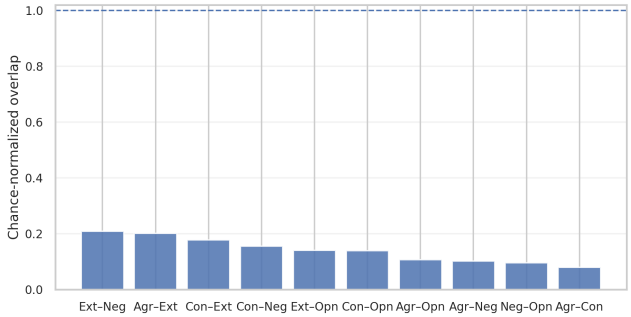

図 4　概念選択的ニューロン同士の重なり分析

の振る舞い制御に関わるため、表現位置が異なる可能性がある。

4.2.2　概念選択的ニューロン

図 3 に、Llama3 において概念選択的ニューロンをモデル全体から上位 1000 抽出した際の層方向の分布を示す。ユニット数は第 6 層から第 11 層付近と、第 16 層から第 19 層付近に 2 度のピークを持つ分布を示した。前者と後者では表現の役割が異なると考えられ、詳細な調査は今後の課題である。図 4 に、領域間で抽出されたユニット集合の重なりをランダムな場合の期待との比較で示した。上位 10%のニューロンにおいて観測された重なりを Jaccard 係数で示し、それをランダムな機体で割った値を報告している。全ての組み合わせでランダム期待に比べて重なりが小さく、概念選択的ユニット集合が互いに分離していることを示唆する。また、外向性は他のドメインとの重なりが多いことがわかる。

4.2.3　概念選択的ニューロンへの介入

図 5 に、層ごとのプロービングに対する介入結果を示す。介入はプロービング性能が最も高かった第 16 層で行い、もともと正解していた入力に対して、対象概念の概念選択的ニューロン上位 30%へ介入した。意図した概念への誘導は高い成功率で達成され、それ以外の概念への遷移は小さかった。これ

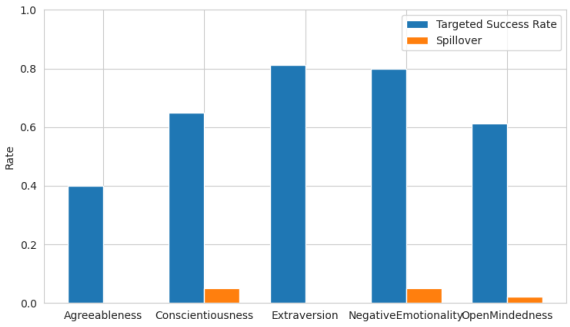

図 5　層ごとのプロービングにおける介入実験の結果

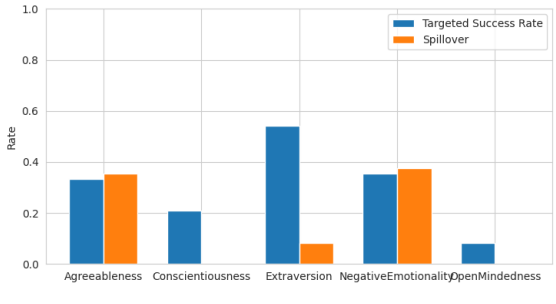

図 6　ドメイン名生成タスクにおける介入実験の結果

は、概念選択的ニューロンの活性が線形分離可能性に強く関与していることを示唆する。

図 6 に、ドメイン名生成タスクに対する介入結果を示す。介入はモデル全体で抽出した概念選択的ニューロンの上位 10%に適用され、分類用プロンプトに対して各ドメイン名トークンの出力確率を比較し、最大のものを予測として扱った。外向性では高い成功率で誘導でき、意図しない概念への遷移は比較的小さい。一方、協調性と神経症傾向では誘導は可能だが、意図しない概念への遷移も同程度に生じた。誠実性と開放性では生成への影響は相対的に小さいが、それでも一定程度は意図した誘導が見られた。

5　おわりに

本研究は、質問紙で定義される性格に関する心理学的構成概念が、大規模言語モデルの内部で読み出し可能な形としてモデルの種類に関係なく早い層に現れ、概念選択的なニューロン群として分離して担われることを示した。さらに、これらのニューロン群への介入によって、プロービングおよび生成の出力傾向を系統的に変化させられることを示した。今後は、より幅広い心理学的構成概念に拡張し、介入が生成過程における概念理解や判断の一貫性をどのように変化させるかを明らかにしたい。

謝辞

本研究は、JST ムーンショット型研究開発事業 JPMJMS2295 の助成を受けた。

参考文献

[1] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. **Journal of personality and social psychology**, Vol. 113, No. 1, p. 117, 2017.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. **Political Analysis**, Vol. 31, No. 3, pp. 337–351, 2023.

[3] 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.

[4] Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. Large language models display human-like social desirability biases in big five personality surveys. **PNAS Nexus**, Vol. 3, No. 12, p. pgae533, 12 2024.

[5] Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. LLMs simulate big5 personality traits: Further evidence. In Ameet Deshpande, EunJeong Hwang, Vishvak Murahari, Joon Sung Park, Diyi Yang, Ashish Sabharwal, Karthik Narasimhan, and Ashwin Kalyan, editors, **Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)**, pp. 83–87, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[6] Heinrich Peters and Sandra C Matz. Large language models can infer psychological dispositions of social media users. **PNAS nexus**, Vol. 3, No. 6, p. pgae231, 2024.

[7] Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8975–8991, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. **Advances in Neural Information Processing Systems**, Vol. 36, pp. 10622–10643, 2023.

[9] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[10] Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. Probing then editing response personality of large language models. In **Second Conference on Language Modeling**, 2025.

[11] Celia Cintas, Miriam Rateike, Erik Miehling, Elizabeth M. Daly, and Skyler Speakman. Localizing persona representations in LLMs. In **The First Workshop on the Interplay of Model Behavior and Model Internals**, 2025.

[12] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017.

[13] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. Finding experts in transformer models, 2020.