LLM による社会調査再現手法のモデル・データに対する頑健性の検証

田中邦朋 [1]　小林哲郎 [2]　笹野遼平 [1]
[1] 名古屋大学　[2] 早稲田大学

tanaka.kunitomo.z3@s.mail.nagoya-u.ac.jp
tkobayas@waseda.jp　sasano@i.nagoya-u.ac.jp

概要

LLM へのペルソナ付与による社会調査の再現は代替データ収集手法として期待されるが、その頑健性は十分に検証されていない。本研究では、プロンプティングと回答分布の抽出手法の組み合わせが、モデルやデータに対して頑健に機能するかを包括的に検証した。その結果、対象属性の平均的な回答を直接指示し、構造化テキストで出力させる手法が、モデルの違いに対して最も頑健であることが確認された。一方で、調査内容の異なるデータ間では手法の効果に一貫性が見られず、また、LLM への入力言語を対象データの言語に合わせることの効果は限定的であることが示唆された。

1　はじめに

大規模言語モデル (Large Language Model; LLM) は、指示に従ってタスクを遂行するだけではなく、文脈内で与えられたペルソナに沿った出力を生成できることから、特定の属性を持つ人間に代えて、LLM から意見や感情、行動に関わる出力を得る研究が多分野にわたって行われている [1, 2, 3]。その中でも、LLM によって社会調査の結果を再現させる研究は、大規模なデータ収集を代替しうるという社会科学的な側面、そして、LLM が人間の価値観と整合した出力をできるかを検証するという工学的な側面から近年多く取り組まれている [4, 5]。

しかし、プロンプトによって属性情報を LLM に与えることで、実際の社会調査と近い結果が得られるかを調査する研究では、先行研究の一部の設定のみを採用していることが多い。表 1 に主な既存研究での実験設定を示す。多種の社会調査を用いる研究であっても、実験言語が英語でのものが主であり

表 1　社会調査を再現する既存研究における実験設定。#LM は LLM のシリーズ数を表す。分布抽出手法とプロンプトの略記については 3.1 節を参照。

既存研究	#LM	分布抽出	プロンプト	データセット数	入力言語
Meister ら [6]	3	L/V	O	3	英語
Hu ら [7]	7	V	S	20[1)]	英語
Liu ら [8]	5	V	O	1	英/西/中/日/韓/独/露/越/葡
Durmus ら [9]	1	L	O	2	英/中/土/露
Lutz ら [10]	3	L/V	I/O/S	1	英語
Santurkar ら [11]	2	L	I/O/S	1	英語

[6, 7]、多言語にわたる研究であっても、単一のデータセット [8]、あるいは単一のモデル [9] での実験に留まっている。また、属性情報を LLM に与えるためのプロンプトの種類 [10] や、モデルからの分布の抽出手法 [6] といった社会調査の再現手法を検証する既存研究も、英語のみによる物が多く、入力言語や再現対象となる社会調査の内容が異なる場合において、同様の効果が見られるかは明らかではない。

そこで、本研究では、LLM による社会調査の再現において、プロンプティングと分布抽出の手法が、実験設定の差異に対して頑健であるかを検証する。具体的には以下の問い (Research Question: RQ) を設定し、これらに対する答えを得るための包括的な評価を実施する。

RQ1　LLM の差異に対して頑健か
RQ2　社会調査データの差異に対して頑健か
RQ3　LLM への入力言語の差異に対して頑健か

これらの検証のために、各プロンプティング手法により LLM にペルソナを付与した上で、社会調査内の質問への回答予測を取得し、ペルソナを付与しなかった場合の予測と比較する。ペルソナ付与によ

1)　社会調査でないデータセットを含む。

りどれだけ実際の分布と近づいたかを、RQ1 に対しては異なるモデル間、RQ2 に対しては異なる調査データ間、RQ3 に対しては調査で使用された言語と英語で比較することで、プロンプティング手法と分布の抽出手法との組み合わせの頑健性を評価する。

2 検証方法

本研究では、ペルソナ付与のための各種プロンプティング手法 p とモデルからの抽出手法 ϕ との組み合わせの頑健性を検証する。そこで、大規模言語モデル M を用いた社会調査 $Q = \{q_i\}$ に対する回答予測が、ペルソナ付与によって実際の分布に近づいた度合 S_{final} を算出し、評価する。ペルソナの観点 V としては、性別・年齢・イデオロギーの 3 つを考え、分布の近さを測る尺度には、$[0, 1]$ の範囲に正規化された 1-Wasserstein 距離をベースにした類似度 [11] を採用する。

2.1 スコアの算出と集計

最終的なスコア S_{final} は、社会調査 Q に含まれる各質問 q_i に対し、観点 $v_j \in V$ に基づくペルソナを付与した場合に、回答予測が実際の分布に近づいた度合 $S(M, p, \phi, q_i, v_j)$ の平均を取ることで算出する。

$$S_{\text{final}}(M, p, \phi) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{1}{|V|} \sum_{v_j \in V} S(M, p, \phi, q_i, v_j).$$

$S(M, p, \phi, q_i, v_j)$ は次の 2 つの方法により算出する。

1. 属性ごとの集団再現度に基づく算出法 (micro)
LLM からある単一の属性の予測を得るシナリオを考えた算出法である。まず、LLM に観点 v_j 内の属性 $c \in C^{(v_j)}$ のペルソナをプロンプティング手法 p で付与してから、質問 q_i について回答を予測させて分布を得る。これと対象の属性 c に対する実際の回答分布 H_c との類似度を計算し、カテゴリ内での実際の人口割合 w_c ($\sum w_c = 1$) で重み付けして平均を取る。分布間の類似度を $\text{sim}(\cdot \,\|\, \cdot)$ と表すとすると、micro 集計によるスコアは以下で表される。

$$\sum_{c \in C^{(v_j)}} w_c \Big[\text{sim}\big(H_c(q_i) \,\|\, M(q_i | p_c; \phi)\big) - \text{sim}\big(H_c(q_i) \,\|\, M(q_i; \phi)\big) \Big]$$

2. 集団全体の再現度に基づく算出法 (macro)
各人口属性に対する予測を観点内で統合し、回答集団全体の回答を予測するシナリオを考えた算出法である。まず、観点 v_j 内の属性 $c \in C^{(v_j)}$ のペルソナを LLM に付与して予測分布を得る。各属性 c に対する予測分布を実際の人口割合 w_c の重みで統合してから、実際の回答分布と比較する。Micro と同様の記法により、macro 集計でのスコアは以下で表される。

$$\text{sim}\big(H(q_i) \,\|\, \sum_{c \in C^{(v_j)}} w_c \cdot M(q \mid p_c; \phi)\big) - \text{sim}\big(H(q_i) \,\|\, M(q_i; \phi)\big)$$

2.2 RQ に対応するスコア比較

RQ に則した実験設定のもとで最終スコアを計算し評価することで、特定の実験設定の差異に対して各再現手法が頑健であるかを検証する。RQ1 に対しては、LLM が異なる実験設定を考慮して、各再現手法を複数の LLM それぞれに適用し、特定の社会調査での回答を予測することで、1 つの実験設定に対応する S_{final} を求める。RQ2, 3 では、社会調査の内容や入力言語が異なる設定を考慮する。そのため、異なる調査データそれぞれを各再現手法を適用した複数の LLM で予測し、LLM ごとに得られる S_{final} を平均して、1 つの実験設定に対するスコアを得る。

3 実験設定

本研究で検証対象となる再現手法を構成するプロンプティング手法および分布の抽出手法、ならびに変動させる実験条件となる LLM と社会調査データについて説明する。

3.1 再現手法の構成

本研究では、プロンプティング手法 3 種と、LLM からの分布の抽出手法 2 種の組み合わせ、計 6 種の再現手法を検証する。

プロンプティング手法は、Lutz ら [10] や Santurkar ら [11] をもとに、ユーザとの対話中で対象属性を呈示する **Interview** (I)、対象属性の平均的な回答分布を直接指示する **Objective** (O)、LLM が対象の属性を持つことを明示する **Subjective** (S) の 3 種を採用する[2)]。

分布の取得手法については、モデルの出力確率から次単語確率を取得する **LogProb** (L) と、構造化テキストで各選択肢の回答割合を生成させる **Verbalized** (V) の 2 つを採用する。ただし、LogProb での抽出に際しては、温度定数はデフォルト値に設定し、Verbalized では、JSON 形式の出力を促すために指示のプロンプトを追加する[3)]。

2) それぞれの実験で実際に入力されるプロンプトは、A.1 節を参照のこと。
3) 実際のプロンプトは、A.2 節を参照のこと。

図 1　各再現手法によって各モデルで WVS 英語での回答を予測したときのスコア。左が micro、右が macro による算出。

3.2　使用モデルと調査データ

LLM　LLM については、既存研究で用いられていたローカルモデルのシリーズから、Llama-3.1 8B Instruct [12]、OLMo-2-1127 7B Instruct [13]、Qwen3 8B [14]、Gemma-3 12B IT [15] を選定する。

社会調査データ　再現対象とする調査データは、異なる内容のものと、同じ内容で異なる対象国のものを選ぶ。本研究では、66 の国と地域にて価値観に対する質問を行った **World Value Survey** [16] 第 7 波のうち後述する対象国のデータ（以下、WVS）、アメリカの社会調査 **OpinionQA** [11] の Disagreement 500 サブセット（以下、OQA）、およびスマートニュースメディア研究所が日本人 1,902 名を対象に行った無作為抽出世論調査「**スマートニュース・メディア価値観全国調査**（SMPP 調査）」の第 1 回調査（以下、SMPP 調査）を用いる。

なお、各実験では、実用での社会調査の再現を想定して、各調査言語での入力のみによって、対象国の分布予測を誘導する。そのため、特に WVS については、特定の言語の母語話者が人口の大半を占める国として、アメリカ (USA)・ドイツ (DEU)・中国 (CHN)・韓国 (KOR)・日本 (JPN) を選定した。それぞれ言語は、英語・ドイツ語・中国語・韓国語・日本語が対応する。また、本研究で人口属性を規定する観点に含まれる属性について、性別は男性/女性の 2 属性、イデオロギーは左派/中道派/右派の 3 属性、年齢は 30 歳未満/30 歳以上 50 歳未満/50 歳以上 65 歳未満/65 歳以上の 4 属性とする。

この設定のもと、RQ1 に対しては WVS の英語版を用いて各再現手法の異なる LLM に対する頑健性を検証する。RQ2 に対しては、WVS の各国版、OQA、SMPP 調査の回答予測によりスコアを算出して、各再現手法の異なる社会調査に対する頑健性を検証する。RQ3 に対しては、WVS の英語版以外の回答分布を英語での予測分布と比較したスコアを算出し、異なる入力言語に対する各再現手法の頑健性を検証する。

4　検証結果

4.1　LLM の差異に対する頑健性 (RQ1)

異なる LLM に対する頑健性に対応する実験結果を図 1 に示す[4)]。特定の LLM のみに対して顕著に効果が現れる再現手法は見られないが、Objective–Verbalized の組み合わせでの再現が、どの LLM に対しても安定して正のスコアを示している。LogProb よりも Verbalized が優れるという結果は Meister らの研究 [6] と一致する。一方 Lutz らの研究 [10] で述べられていた Interview の優位性は確認できなかった。既存研究でも使用例が少ない Interview–Verbalized は一貫して負のスコアが見られた。

また、micro 集計ではスコアが負になる手法であっても、macro 集計では正に転じる傾向が確認できる。この結果は、特定の属性集団の再現よりも回答集団全体の再現の方が容易であるとした Hu ら [7] の知見とも整合する。

4.2　データの差異に対する頑健性 (RQ2)

異なる社会調査データに対応する実験結果を図 2 に示す。同じ内容で対象国が異なる調査結果について、各国の WVS を再現したときのスコアを見ると、Objective による予測で正のスコアが並ぶものの、そ

4)　以降、図中の p の値は、各スコアの符号を無作為に入れ替えた順列検定（試行回数 10^5）による両側 p 値を表す。

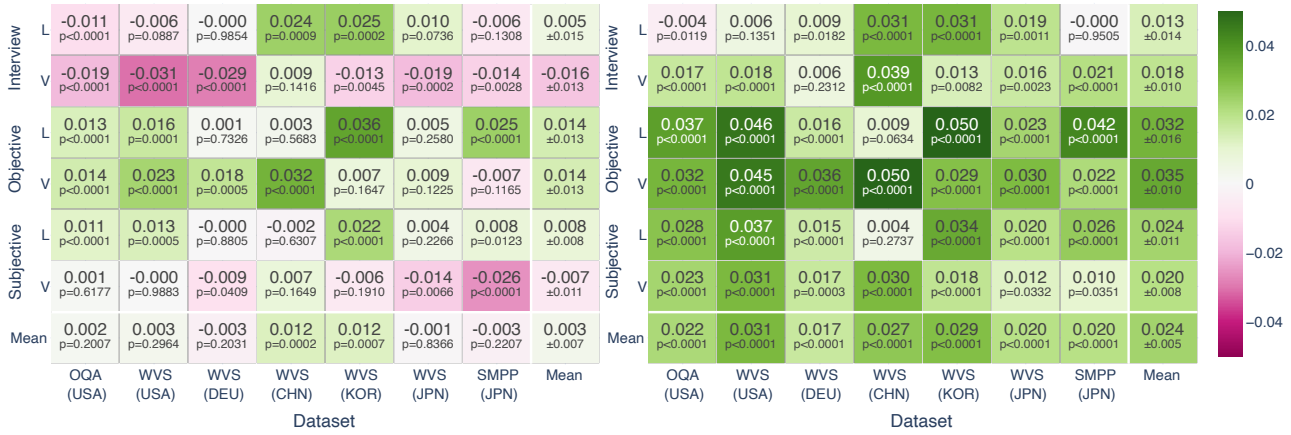

図 2　各再現手法によって各社会調査の回答を予測したときのスコア。 左は micro 集計、右は macro 集計によるもの。

の大きさは再現対象の国ごとに異なっており、他の手法では、国ごとに正負が入れ替わっている。この結果は、各再現手法が異なる対象国に対して十分に頑健でないことを意味する。

異なる内容の社会調査については、まず、OQA と WVS（アメリカ）での micro 集計によるスコアに着目すると、各手法が異なるデータで似た傾向を示している。その一方で、WVS（日本）と SMPP 調査ではスコアの正負が入れ替わる手法が見られる。再現対象の国によって、調査にわたっても一貫したスコアが得られるものと、そうでないものがあり、異なる内容の社会調査に対する頑健性は、対象国によっては限定的であることを示唆する結果となった。また、micro 集計よりも macro 集計の方がスコアが高くなる傾向は、RQ1 に対する結果と同じく、データセットにわたっても見られた。

4.3　入力言語の差異に対する頑健性 (RQ3)

入力言語の違いに対する頑健性を検証するため、アメリカ以外の WVS の回答分布を英語で予測した場合の macro スコアの結果を図 3 に示す[5]。言語ごとの 6 つの再現手法に対する平均スコアの変化[6]はドイツ語、中国語、韓国語、日本語の順にそれぞれ 0.017、0.027、0.029、0.020 から、0.027、0.018、0.023、0.023 への変化にとどまっており、全体的に入力言語の変化による性能の差は確認できない。このことから、入力言語の差異に対しては頑健であると言える。この結果は、ペルソナを付与せずに入力言語のみを変えることでは出力が大きく変化しないという Santurkar ら [11] の知見と整合する。

5）　Micro 集計については、B 節を参照のこと
6）　図 2 の右図の 3 から 6 カラム目が対応する結果である。

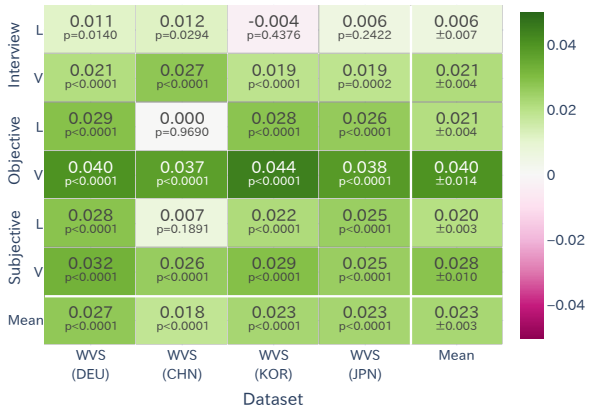

図 3　RQ3: アメリカ以外の WVS の回答分布を英語での予測と比較したときの macro スコア。

5　おわりに

本研究では、LLM を用いた社会調査再現の頑健性を検証するため、プロンプティングおよび分布抽出手法の組み合わせが、モデルやデータ、入力言語の差異に対して頑健であるかを検証した。4 種の LLM と 7 種のデータセットを用いた検証の結果、プロンプティング手法として対象属性の回答予測を指示する Objective、分布抽出手法として構造化テキストで出力させる Verbalized を組み合わせた再現手法は、LLM 間の差異、データ間の差異に対して比較的頑健であったものの、再現手法によっては一部の LLM やデータに対してペルソナ付与の効果がまったく確認できなかった。一方、入力言語の変化によるペルソナ付与の効果に差異は見られなかったことから、入力言語の変化には比較的頑健であることが分かった。今後の課題としては、クローズドモデルを含むサイズの大きい LLM を用いた検証、調査時期の違いの考慮、人口属性の組み合わせの考慮などが挙げられる。

謝辞

本研究では、スマートニュース・メディア価値観全国調査（SMPP 調査）のデータを使用した。調査の実施およびデータ提供にあたり、スマートニュース株式会社ならびにスマートニュースメディア研究所に感謝する。なお、本調査の概要については、https://about.smartnews.com/ja/news/2421.html を参照されたい。また、本研究の一部は JSPS 科研費 24H00727 の助成を受けたものである。

参考文献

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In **Proceedings of the 40th International Conference on Machine Learning (ICML)**, pp. 337–371, 2023.

[2] John J Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Technical report, National Bureau of Economic Research, 2023.

[3] Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: LLM Social Simulations Are a Promising Research Method. In **Forty-second International Conference on Machine Learning (ICML) Position Paper Track**, 2025.

[4] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? **Computational Linguistics**, Vol. 50, No. 1, pp. 237–291, 2024.

[5] Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. In **Findings of the Association for Computational Linguistics (EMNLP)**, 2024.

[6] Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking Distributional Alignment of Large Language Models. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, pp. 24–49, 2025.

[7] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. **arXiv preprint arXiv:2510.17516**, 2025.

[8] Yang Liu, Masahiro Kaneko, and Chenhui Chu. On the Alignment of Large Language Models with Global Human Opinion. **arXiv preprint arXiv:2509.01418**, 2025.

[9] Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In **First Conference on Language Modeling (CoLM)**, 2024.

[10] Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. In **Findings of the Association for Computational Linguistics (EMNLP)**, pp. 23212–23237, 2025.

[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 (ICML)**, pp. 29971–30004, 2023.

[12] 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. Technical report, Llama Team, AI @ Meta, 2024.

[13] Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 Furious. Technical report, OLMo Team, 2024.

[14] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. Technical report, Qwen Team, 2025.

[15] Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. Technical report, Gemma Team, Google DeepMind, 2025.

[16] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0, 2022.

A　プロンプト

本研究で考慮した 3 種のプロンプティングおよび Verbalized による分布取得に際する指示について、実際に LLM へ入力したプロンプトを示す。

A.1　各種ペルソナプロンプティング

実際の社会調査の質問と選択肢を入力する前に、各手法で異なるプロンプトを入力する。以下では、男性のペルソナを付与して日本語の社会調査の回答予測を得る場合のプロンプトを示す。なお、ペルソナプロンプティングに続く質問と選択肢は、Interview の User ロールと同じ形式で入力される。

表 2　Interview

```
User: あなたの性別は？
(1) 男性
(2) 女性
Assistant: 回答: (1)
```

表 3　Objective

```
System: 性別は男性の平均的な回答を出力してください。
```

表 4　Subjective

```
System: あなたは男性です。
```

A.2　Verbalized での指示

Hu ら [7] で LLM に与えた指示を、Gemini 3 で各言語へ翻訳したものを利用した。この指示は System ロールにペルソナプロンプティングを入れる場合、その前に入力される。以下ではその日本語版を示す。

```
回答を JSON 形式で出力してください。
以下のルールを遵守してください。
1. 0 から 100 までの整数を使用してください
2. 各選択肢の合計が 100 になるようにしてください
3. %を含めないでください
4. 以下の JSON 形式を使用してください: {"1": X, "2": X, "3": X, "4": X}
5. 最終回答のみを出力してください。説明文や途中経過は出力しないでください。
X は予測値に置き換えてください。
```

B　RQ3 の micro 集計によるスコア

図 4 に、図 3 の左図に対応する、micro による集計結果を示す。Objective–Verbalized の組み合わせに対しては安定して正のスコアが得られたが、調査対象国の主要言語を使った場合の結果と比較すると、図 3 と同様、全体として性能に大きな差は確認できない。

図 4　RQ3: アメリカ以外の WVS の回答分布を英語での予測と比較したときの micro スコア。“p =” に続く値は、各質問に対するスコアの符号を無作為に入れ替えた順列検定による両側 p 値。“±” に続く値はスコアの標準偏差。