大規模言語モデルは人間のアンケート回答集合を模倣できるか

熊谷 雄介 [1]
[1] 株式会社博報堂 DY ホールディングス
yusuke.kumagae@hakuhodo.co.jp

概要

本研究では従来，人間に対して実施してきたアンケート調査をペルソナを設定した大規模言語モデル（LLM）による仮想個人で代替可能かに取り組む．「仮想個人の回答が個々の人間の回答とどのように類似するか」というミクロな観点と，「仮想個人の回答集合が人間の回答集合とどのように類似し，どの程度カバーできるか」というマクロの観点の両者から検証する．様々な実アンケート，LLM モデル，LLM に対するペルソナ設定による検証の結果，（1）LLM に対するペルソナの設定は，同属性の人間の回答に近付ける効果がない，（2）仮想個人群の回答は人間が持つ多様な回答をカバーしきれない，の 2 点が明らかになった．

1 はじめに

大規模言語モデル（LLM）が持つ自然な対話能力，大量の言語資源による事前学習にもとづく大規模な知識，「陽気な 30 代男性のように振る舞って」といった指示に対する高い追従性能 [1, 2, 3, 4, 5, 6] に着目し，これまで人間に対して行っていた様々な調査を LLM に代替する試みが複数存在する．

このような取り組みはシリコンサンプリング [7] とも呼ばれており，その対象は人間社会の再現 [8, 9]，経済実験 [10, 11, 12, 13, 14, 15, 16]，広告生成 [17, 18]，半構造化インタビュー [19, 20]，世論調査 [21, 7, 22, 23, 24] など多岐にわたる．

本研究では，与えられた質問に対して対応する回答の選択肢を選ぶ多肢選択式のアンケートに着目し，人間に対するアンケート調査を LLM によって代替可能かを検討する．仮に LLM に対するアンケート調査から人間の同程度の精度や多様性を持つ回答が得られるのであれば，時間的[1)]・金銭的[2)]コストや調査対象者への負荷の全てを大きく軽減しうる．しかし，LLM の回答が人間の（個別の回答だけでなく）回答の分布を再現できるかについては既存研究でも十分に検討されていない．

本研究では，LLM に属性や性格などの**ペルソナ** [8] を設定した**仮想個人**に対してアンケート調査を行い，その回答と人間の回答とを比較することで，**仮想個人の回答は人間の回答とどのように類似するか（ミクロの観点）**と**仮想個人の回答集合は人間の回答集合とどのように類似し，どの程度カバーできるか（マクロの観点）**の 2 つの観点で検討する．

複数の実アンケート，LLM モデル，ペルソナ設定による検証の結果，（1）LLM に対するペルソナの設定は，同属性の人間の回答に近付ける効果がない，（2）仮想個人の回答集合は人間が持つ多様な回答をカバーしきれない，ことを示す．

2 関連研究

LLM の出力は人間のフィードバックにもとづく強化学習の結果として均一化すると言われている [25, 26] が，ここでは特に，人間を模倣した LLM の出力における多様性の欠如について述べる．

Argyle らは多肢選択式のアンケートにおいて指示文に応じて回答の多様性が得られると主張し，LLM の回答と人間の回答との高い相関を報告している [7] が，これはあくまで変数間の相関であり，回答全体の多様性までは議論していない．

一方で人間を模倣した LLM から得られる回答が均一的で多様性に欠くという批判は複数存在する．Feuer らは LLM の出力は異なるモデル間であっても類似していると指摘している [27]．Kapania らは LLM に対する半構造化インタビューにおいて，その発話は詳細ではあるものの全く深みがなく，LLM はユーザが指示文で設定した通りにしか返答できないと批判しており [19]，Zhang らによる類似の報告も存在する [21]．

Wang らは自由記述式の回答生成において LLM に

1） 一般に調査には一ヶ月程度の期間が必要になる．
2） 属性やサンプルサイズなどを十分に設計した調査の実施には数十万円から数百万円の費用が必要になる．

特定の属性を指示する時，その振る舞いは対象とした属性そのものではなく，属性を外から観察した時のステレオタイプに非常に近くなり，その際の回答の多様性も失われていると報告している [13]．彼らは同時に自由記述データを離散化したものでも多様性が失われることも報告しており，Argyle らの報告 [7] を間接的に否定している．

本研究は Wang らの取り組みを発展させたものである．Wang らも行った「ペルソナを設定した LLM のアンケート回答がどのように変化するか」というミクロな観点だけでなく，「LLM による回答の集合がどのような多様性を持ち，人間の回答集合をどれだけカバーするか」というマクロな観点も本研究では検証対象とする．

3 設定

3.1 アンケートデータ

本研究では日本人を対象に実施された二種類のアンケートデータを用いる．

生活定点調査 生活定点調査（生活定点）[3] は，博報堂生活総合研究所が 1992 年から隔年で実施する時系列観測調査であり，「ソーシャルメディアの利用状況」や「ココロ重視派 vs モノ重視派」といった質問によって日頃の感情，生活行動や消費態度，社会観など，多角的な設問から，生活者の意識と欲求の推移を分析することを目的としている．

本研究では生活定点の 2022 年度版における 3,488 名，デモグラフィック属性に関する質問を除いた単一選択質問 178 問を対象にした．

世界価値観調査（WVS） 世界価値観調査 [28] は世界中の異なる国や地域，民族の人々の価値観を調査するプロジェクトであり，「次にあげる意見について、あなたはどの程度賛成ですか、それとも反対ですか。 C）一般的に、男性の方が女性より政治の指導者として適している」といった，より政治的・社会的な価値観を問う質問が多く含まれている．

本研究では 2017 年から 2022 年に実施された wave 7 のうち，2019 年に実施された日本人サブセットである 1,353 名，知識を問う問題などを除いた単一選択質問 237 問のうち，5 割以上の人間が有効な回答を行った 188 問を対象にした．

3.2 ペルソナ設定

ペルソナ設定にはデモグラフィック属性と Big Five 因子を用いた．

デモグラフィック属性（Demo） 年齢や性別などの属性から「あなたは {age} 歳の {sex} です.」といった指示文を作成した．その際，アンケートデータの回答者のデモグラフィック属性の組み合わせをサンプリングして用いることで，実データのデモグラフィック属性の分布を反映した．それぞれのアンケートデータに含まれるデモグラフィック属性の詳細については Appendix §Aに示す．

Big Five 因子（Big5） 人間特性を記述する Big Five 因子 [29] について，Jiang ら [6] と同様に各因子の強弱に対応する日本語の単語を連結し，「あなたは話し好きで，陽気な，…，短気，怒りっぽい，…」といった指示文を作成した．それぞれの因子と強弱に対応した単語リストは和田 [30] の結果を用いた．

デモグラフィック属性と Big Five 因子の同時設定（Demo+Big5） デモグラフィック属性と Big Five 因子の同時設定（Demo + Big5）も実施した．アンケートデータの人間には Big Five 因子は付与されていないため，実在のデモグラフィック属性にランダムな Big Five 因子を組み合わせた．

3.3 LLM による回答生成

LLM OpenAI の `gpt-4.1-nano-2025-04-14`（GPT-4n）[31] と Google の `gemini-2.5-flash-lite`（Gemini）[32] を用いた．全実験は 2026 年 1 月に実施した．

生成手続き LLM に「回答は <ans>1</ans> のように選択肢番号を <ans></ans> で囲み，回答以外の情報は含めないでください.」と指示し選択肢番号を得た．ペルソナ設定はシステムプロンプトに入力した．指示文の詳細は Appendix §B に示した．

一人の仮想個人について，全ての質問とその回答を保持したまま回答を生成する方法（stateful）と，個別に質問に対する回答を生成する方法（stateless）の両者を実施したが，傾向に大きな差が見られなかったため，本論文では stateless についてのみ報告する．

仮想個人の人数は Demo および Demo + Big5 は 200 名，Big5 は各因子の強弱の全組み合わせ 32 組を 7 名ずつ合計 224 名を生成した．

3） https://seikatsusoken.jp/teiten/

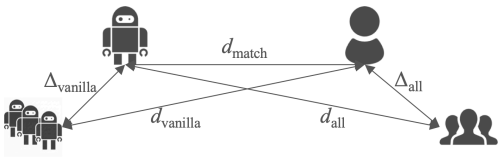

図 1: 仮想個人と同属性の人間との距離の関係.

表 1: 仮想個人，同属性の人間における，人間全体，vanilla LLM それぞれとの距離の比較．対応のある t 検定で Holm 補正を実施したところ $p < .05$ で全て有意．d は効果量（Cohen's d）．

		vs 人間全体		vs Vanilla	
データ	モデル	Δ_{all}	d	Δ_{vanilla}	d
生活定点	GPT-4n	−.028	−.32	−.260	−3.3
	Gemini	−.056	−.74	−1.07	−10.7
WVS	GPT-4n	−.025	−.20	−.424	−2.6
	Gemini	−.031	−.27	−.777	−6.7

4 実験

4.1 分析 1：仮想個人は個別化するか

ミクロの観点である「仮想個人の回答は人間の回答とどの程度近く，遠いか」を，（1）仮想個人の回答が同じデモグラフィック属性を持つ人間[4)]（以降，**同属性の人間**と呼ぶ）の回答とどの程度近く，（2）人間全体や，ペルソナを設定しない LLM（以降，vanilla）の回答とどの程度遠いか，を確認する．

回答間の距離は，（選択肢の多くが順序尺度であることを踏まえて）回答の選択肢番号に対する L1 距離を用いた．距離および類似度の詳細は Appendix §Cに示した．

本分析では距離の差を確認する．仮想個人と同属性の人間の距離を d_{match}，仮想個人と人間の全体との距離 d_{all} についてその差 $\Delta_{\text{all}} = d_{\text{match}} - d_{\text{all}}$ が負であるほど，仮想個人が同属性の人間に近いことを意味する．次に，d_{match} と，同属性の人間と vanilla LLM との距離 d_{vanilla} についてその差 $\Delta_{\text{vanilla}} = d_{\text{match}} - d_{\text{vanilla}}$ が負であるほど，デモグラフィック属性の設定によって LLM が同属性の人間に近付いたことを意味する．距離の関係を図1に図示した．

結果が 表1 である．Δ_{all} については有意差はあるものの $|\Delta_{\text{all}}|$ の値そのものが小さいことから，デモグラフィック属性を設定したとしても LLM の回答は同属性の人間の回答に近付かないことが分かる．

Δ_{vanilla} が負であり有意差があることは，デモグラフィック属性の設定により LLM は（vanilla LLM に比べて）人間全体に近付くが（$\Delta_{\text{vanilla}} < 0$），同属性の人間に近付いていない（$\Delta_{\text{all}} \approx 0$）ことを意味する．この結果は既存研究 [13, 33, 34] が指摘するように，LLM が持つ「人間らしさ」のステレオタイプが現れたものと考えられる．

これらの結果から，LLM に対するデモグラフィック属性の設定は同属性の人間の回答に近付ける効果がないことが分かる．

4) LLM に与えたデモグラフィック属性はアンケートデータに存在する組み合わせをランダムに用いているため，少なくとも一人はマッチングする．

表 2: 人の群内類似度集合と LLM の群内類似度集合の KS 統計量．大きいほど LLM の類似度の集合の分布が人間のものと異なることを意味する．

データ	モデル	Vanilla	Demo	Big5	Demo+Big5
生活定点	GPT-4n	.98	.93	1.00	.26
	Gemini	.72	.51	.32	.41
WVS	GPT-4n	.95	.81	1.00	.43
	Gemini	.63	.69	.40	.11

4.2 分析 2：仮想個人の集合は多様か

次に，マクロの観点である「仮想個人の集合はどのような多様性を持つか」について，（1）仮想個人および人間の**群内類似度**はどのように異なるか，（2）仮想個人の集合は人間の集合をどの程度カバーしているか，の 2 つを確認する．

4.2.1 群内類似度の分布に差はあるか

群内類似度の比較にはコルモゴロフ–スミノフ（KS）統計量 [35] を用いる．KS 統計量は 2 つの累積分布関数 $F_1(x)$, $F_2(x)$ の差 $\sup_x |F_1(x) - F_2(x)|$ で定義される．KS 統計量は 0 から 1 であり，2 つの分布が全く等しい場合には 0，全く異なる場合には最大 1 を取る．

LLM の群内類似度集合と人の群内類似度集合に対する KS 統計量が表2である．ペルソナ設定が行われない LLM では群内類似度の分布が人間の分布と全く異なるが，ペルソナ設定によっては（特に Demo+Big5）KS 統計量が低下し，人間の分布に近付くものがあった．このことから，ペルソナ設定は群内類似度を低下させる効果があることが分かる．興味深いのは GPT-4n において生活定点と WVS 共に Big5 単体での設定は効果がなく，デモグラフィック属性と組み合わせることで類似度分布が多様になる点である．

しかし，KS 統計量は分布が異なることを示すのみであり，LLM が人に対して均一性が高いか低い

表 3: 群内類似度の平均．値が大きいほど多様性が低い．

データ	モデル	人間	LLM			
			Vanilla	Demo	Big5	Demo+Big5
生活定点	GPT-4n	.40	.63	.66	.74	.41
	Gemini		.51	.49	.42	.34
WVS	GPT-4n	.39	.59	.62	.68	.48
	Gemini		.50	.54	.42	.39

かは判断できない．そこで群内類似度の平均値を求めたものが表3である．

ほとんどの LLM 群内類似度の平均は人間の平均よりも大きいことから，仮想個人は人間に比べて似た回答を行っていることが分かる．これは Wang ら [13] の報告と同様である．また，ペルソナ設定は群内類似度の平均を下げる効果があることが分かる．

4.2.2 仮想個人は人間をカバーできるか

§4.2.1では群内類似度集合の分布の違いを確認し，ペルソナ設定によって仮想個人の群内類似度が低下しうることを示した．しかしこれは仮想個人群の間での類似度のばらつきに関する議論であり，仮想個人群が人間にどう近いか，あるいは遠いかを意味しない．よって次に，仮想個人の回答が人間の回答をどの程度カバーするかを確認する．

人間の集合 $\mathcal{H}$ と仮想個人の集合 $\mathcal{L}$ について，人間 $\boldsymbol{h} \in \mathcal{H}$ から最も近い仮想個人 $\boldsymbol{l} \in \mathcal{L}$ までの距離 $d_{\min}(\boldsymbol{h}) = \min_{\boldsymbol{l} \in \mathcal{L}} d(\boldsymbol{h}, \boldsymbol{l})$ を考える．$d_{\min}$ が小さいほど，人間の近くに仮想個人が存在する．$d_{\min}$ の分布の 95%分位点である $q_{95,(\mathcal{H},\mathcal{L})}$ を確認することで「95%の人間における仮想個人までの距離」が分かる．$q_{95,(\mathcal{H},\mathcal{L})}$ が小さいほど，LLM が人間の回答を広くカバーしていることを意味する．

また，ベースラインとして人間の集合 $\mathcal{H}$ 内での $d_{\min}$（それぞれの人間における最も近い他人までの距離）の 95%分位点である $q_{95,(\mathcal{H},\mathcal{H})}$ を求め，その比率（**カバレッジ**）$\frac{q_{95,(\mathcal{H},\mathcal{L})}}{q_{95,(\mathcal{H},\mathcal{H})}}$ を見る．カバレッジが 1 より大きいほど，仮想個人は人間をカバーできていないことを意味する[5)]．

表4がカバレッジの一覧である．ペルソナを設定することによってカバレッジが低下するのはここまでの分析と矛盾しない．最も小さなカバレッジでも 1.5 程度であり，これは仮想個人が人間同士と比較し総じて 1.5 倍ほど人間から離れていることを意味

5) 例えるなら，「人間が存在する空間」を仮想個人の集合でカバーするために，（人間同士でカバーするよりも）各仮想個人がより広い範囲を担当しなければならない状態である．

表 4: カバレッジの比較．この値が 1 より大きいほど仮想個人が人間をカバーできていない事を意味する．

データ	モデル	Vanilla	Demo	Big5	Demo+Big5
生活定点	GPT-4n	2.18	1.70	2.25	1.71
	Gemini	3.27	1.52	2.44	1.62
WVS	GPT-4n	2.25	1.45	1.82	1.46
	Gemini	2.87	1.56	2.28	1.55

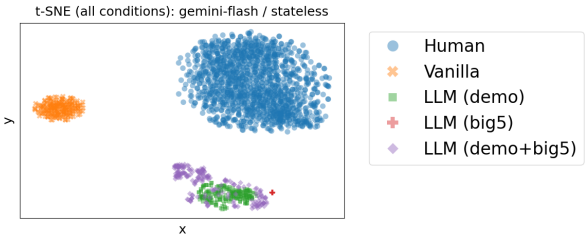

図 2: t-SNE による WVS における人間および Gemini による仮想個人の可視化．Vanilla，ペルソナ設定された仮想個人，人間が大きく異なることが分かる．

している．サンプルサイズに対するカバレッジの感度分析は Appendix §Dにて実施した．

この結果をより明確に示すのが t-SNE[36] による可視化である．図2 は WVS における人間と Gemini による仮想個人の可視化である．この図からは（1）ペルソナの設定により LLM の仮想個人は大きく変化し，わずかに人間に近付くが（2）仮想個人は人間とは依然として大きな隔たりがある，（3）Demo+Big5 のペルソナによって仮想個人の分布が広がる，が分かる．これは，ここまでの分析を裏付ける結果である．

よってこれらの結果から，LLM による仮想個人は人間の多様な回答をカバーしきれないことが明らかになった．

5 結論

本論文では，ペルソナの設定のみによって構築した仮想個人にはミクロおよびマクロの観点において限界があることを示した．

より良い仮想個人の実現のために考えられるのは，適用を想定する領域のデータを用いた追加学習 [15, 37] であるが，追加学習だけでは不十分であるという批判 [38] も存在する．

ステレオタイプの克服のためには「大きいモデルほどバイアスやステレオタイプを持つ」という既存研究 [39] にもとづき，小さなモデルの採用も検討に値するだろう．その場合，ステレオタイプと事前知識とは表裏一体であるため，下流タスクにおけるパフォーマンスとのトレードオフが発生する．

参考文献

[1]Yunfan Shao, et al. Character-LLM: A trainable agent for role-playing. In **EMNLP**, 2023.

[2]Quan Tu, et al. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In **ACL**, 2024.

[3]Xinfeng Yuan, et al. Evaluating character understanding of large language models via character profiling from fictional works. In **EMNLP**, 2024.

[4]Noah Wang, et al. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In **Findings of ACL**, 2024.

[5]Murray Shanahan, et al. Role play with large language models. **Nature**, Vol. 623, No. 7987, pp. 493–498, November 2023.

[6]Hang Jiang, et al. PersonaLLM: Investigating the ability of large language models to express personality traits. In **Findings of NAACL**, 2024.

[7]Lisa P Argyle, et al. Out of one, many: Using language models to simulate human samples. **Polit. Anal.**, Vol. 31, No. 3, pp. 337–351, July 2023.

[8]Joon Sung Park, et al. Generative agents: Interactive simulacra of human behavior. In **UIST**, 2023.

[9]Joon Sung Park, et al. Social simulacra: Creating populated prototypes for social computing systems. In **UIST**, 2022.

[10]Danica Dillion, et al. Can AI language models replace human participants? **Trends Cogn. Sci.**, Vol. 27, No. 7, pp. 597–600, July 2023.

[11]John Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, NBER, 2023.

[12]Benjamin Manning, et al. Automated social science: Language models as scientist and subjects. Technical report, NBER, 2024.

[13]Angelina Wang, et al. Large language models that replace human participants can harmfully misportray and flatten identity groups. **Nat. Mach. Intell.**, Vol. 7, No. 3, pp. 400–411, 2025.

[14]Ayato Kitadai, et al. Can AI with high reasoning ability replicate human-like decision making in economic experiments? **Group Decis. Negot.**, Vol. 34, No. 6, pp. 1303–1326, 2025.

[15]Marcel Binz, et al. A foundation model to predict and capture human cognition. **Nature**, Vol. 644, No. 8078, pp. 1002–1009, 2025.

[16]Joon Sung Park, et al. Generative agent simulations of 1,000 people. **arXiv [cs.AI]**, November 2024.

[17]Daichi Haraguchi, et al. Can GPTs evaluate graphic design based on design principles? In **SIGGRAPH Asia 2024 Technical Communications**, 2024.

[18]Heng Wang, et al. BannerAgency: Advertising banner design with multimodal LLM agents. In **EMNLP**, 2025.

[19]Shivani Kapania, et al. Simulacrum of stories: Examining large language models as qualitative research participants. In **CHI**, 2025.

[20]Perttu Hämäläinen, et al. Evaluating large language models in generating synthetic HCI research data: A case study. In **CHI**, 2023.

[21]Simone Zhang, et al. Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. **Sociol. Methods Res.**, Vol. 54, No. 3, pp. 1197–1242, 2025.

[22]Shibani Santurkar, et al. Whose opinions do language models reflect? In **ICML**, 2023.

[23]Ricardo Dominguez-Olmedo, et al. Questioning the survey responses of large language models. In **NeurIPS**, 2024.

[24]Joseph Suh, et al. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In **ACL**, 2025.

[25]Robert Kirk, et al. Understanding the effects of RLHF on LLM generalisation and diversity. In **ICLR**, 2024.

[26]Jiayi Zhang, et al. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. **arXiv [cs.CL]**, 2025.

[27]Benjamin Feuer and Chinmay Hegde. WildChat-50M: A deep dive into the role of synthetic data in post-training. In **ICML**, 2025.

[28]Christian Haerpfer, et al. World values survey wave 7 (2017-2022) cross-national data-set, 2022.

[29]J M Digman and N K Takemoto-Chock. Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. **Multivariate Behav. Res.**, Vol. 16, No. 2, pp. 149–170, April 1981.

[30]和田さゆり. 性格特性用語を用いた big five 尺度の作成. 心理学研究, Vol. 67, No. 1, pp. 61–67, 1996.

[31]OpenAI, et al. GPT-4 technical report. **arXiv [cs.CL]**, March 2023.

[32]Gheorghe Comanici, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. **arXiv [cs.CL]**, 2025.

[33]Huaman Sun, et al. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In **ACL**, 2025.

[34]Marlene Lutz, et al. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In **Findings of EMNLP**, 2025.

[35]Frank J Massey. The kolmogorov-smirnov test for goodness of fit. **J. Am. Stat. Assoc.**, Vol. 46, No. 253, p. 68, March 1951.

[36]L Maaten and Geoffrey E Hinton. Visualizing data using t-SNE. **Journal of Machine Learning Research**, Vol. 9, No. 86, pp. 2579–2605, 2008.

[37]Akaash Kolluri, et al. Finetuning LLMs for human behavior prediction in social science experiments. In **EMNLP**, 2025.

[38]Yuan Gao, et al. Take caution in using LLMs as human surrogates. **Proc. Natl. Acad. Sci. U. S. A.**, Vol. 122, No. 24, p. e2501660122, June 2025.

[39]Xuechunzi Bai, et al. Explicitly unbiased large language models still form biased associations. **Proc. Natl. Acad. Sci. U. S. A.**, Vol. 122, No. 8, p. e2416228122, February 2025.

A　デモグラフィック属性

生活定点で用いたデモグラフィック属性は一歳刻み年齢，性別，居住地域，婚姻形態，職業，主観的な生活水準である．

WVS で用いたデモグラフィック属性は一歳刻み年齢，性別，最終学歴，職業，世帯収入，婚姻形態，主観的な社会階層である．

B　指示文

WVS における Demo+Big5 のペルソナ設定の指示文例が以下である．この Big Five 因子は開放性が弱く，誠実性が強く，外向性が強く，協調性が弱く，神経症的傾向が強い状態である．

ペルソナ設定指示文

あなたは 43 歳の女性です．婚姻形態は結婚している，最終学歴は 6 年制大学・大学院修士課程，職業はパートタイム・アルバイト (週 30 時間未満)，世帯収入は 800〜900 万円未満です．生活の程度は世間一般から見て「中の上」だと考えています．

あなたは話し好きで，陽気な，外向的，社交的，活動的な，積極的な，短気，怒りっぽい，とげがある，かんしゃくもち，自己中心的，反抗的な，計画性のある，勤勉な，几帳面な，悩みがち，不安になりやすい，心配性，気苦労の多い，弱気になる，傷つきやすい，動揺しやすい，神経質な，悲観的な，緊張しやすい，憂鬱な，平凡な，保守的な，想像力に乏しい，興味の範囲が狭い，新しい変化を好まない，融通が利かない，同調的な，常識にとらわれた性格です

質問の指示文の例が以下である．

質問指示文

以下の質問に答えてください．
回答は次の選択肢番号 (1 2 3 4 5) のいずれかから選んでください．
回答は <ans>1</ans> のように選択肢番号を <ans></ans> で囲み，回答以外の情報は含めないでください．

質問

楽しい生活度

選択肢

- 1: 楽しい方だ
- 2: やや楽しい方だ
- 3: あまり楽しくない方だ
- 4: 楽しくない方だ
- 5: 不明

表 5: 仮想個人をサンプリング（%）した際のカバレッジの変化．

データ	モデル	属性	10%	50%	100%
生活定点	GPT-4n	Vanilla	2.10	2.06	2.05
		Demo	1.70	1.60	1.57
		Big5	2.21	2.15	2.13
		Demo+Big5	1.65	1.58	1.56
	Gemini	Vanilla	2.93	2.86	2.84
		Demo	1.52	1.44	1.41
		Big5	2.45	2.31	2.27
		Demo+Big5	1.58	1.49	1.46
WVS	GPT-4n	Vanilla	1.97	1.93	1.92
		Demo	1.42	1.37	1.36
		Big5	1.78	1.72	1.70
		Demo+Big5	1.42	1.36	1.35
	Gemini	Vanilla	2.49	2.42	2.40
		Demo	1.52	1.46	1.44
		Big5	2.15	2.05	2.01
		Demo+Big5	1.50	1.43	1.40

C　距離および類似度

人間の回答と仮想個人の回答との距離および類似度について述べる．順序尺度を問う質問が多くを占めることから，選択肢番号を連続値と見なし，各質問の選択肢番号の最大値と最小値で [0, 1] にスケーリングした．その際，回答拒否などの欠損値については 0.5 を代入した．その後，回答中の各質問 i について人間の各回答の分布を基準として z-score にもとづいて正規化を行った．

距離関数には L1 距離 $d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{D} \sum_{i=1}^{D} |x_i - y_i|$ を用いた（D は質問数）．

分布の比較においては $\exp(-d(\boldsymbol{x}, \boldsymbol{y}))$ を類似度として用いた．

D　カバレッジの感度分析

§4.2.2で求めたカバレッジはサンプルサイズの影響を受けうる．そこで，仮想個人のサンプルサイズを変化させながらカバレッジを求めたものが表5である．

この結果からは，少なくとも現在のサンプルサイズより少ない状態でもカバレッジは大きく変動しないことが分かる．