LLM と分布表現

Arseny Tolmachev　児玉壮平
株式会社博報堂テクノロジーズ
{arseny.teramachi,sohei.kodama}@hakuhodo-technologies.co.jp

概要

LLM に確率分布を学習させる際、適切な表現方法は明らかでない。人工データを用いて直接的・間接的な 5 つの表現手法を比較した結果、分布を明示的に表現する手法は、頻度ベースの暗黙的な学習より高い推定精度を達成した。学習データへの適度なノイズ付与は未知条件への汎化性能向上に寄与する。

1　はじめに

広告代理店では、マーケティング戦略立案に調査データを活用している。なかでも、生活者を対象としたアンケート調査は、生活者の態度や意識を高い信頼性で把握できる手法として広く用いられているが、実施には多大なコストと時間を要する。

近年、この課題に対するアプローチとして、大規模言語モデル（LLM）を用いて仮想的なアンケート回答を生成する研究が注目を集めている。もちろん、LLM によるシミュレーションが人間による実調査を完全に代替することはありえない。しかし、実調査に先立つ仮説構築の支援や、調査票設計の予備検証に活用することで、業務を効率化できる。

LLM をアンケートシミュレーションに活用する研究は、大きくプロンプトベースとファインチューニングベースに分類できる。プロンプトベースでは、属性情報をプロンプトに含めることで LLM の回答を誘導するが、実際の回答分布との乖離やプロンプト設計への感度が課題として指摘されている [1, 2, 3]。一方、近年ではアンケートデータを用いたファインチューニングにより分布予測精度を向上させる手法が提案されており、プロンプトベースの手法を大幅に上回る性能が報告されている [4, 5, 6, 7]。

しかし、設問への回答のモデリングにおいて、回答データをどのように表現し、LLM に学習させるべきかという点は未解決の課題として残されている。先行研究の多くは、個々の回答者の具体的な回答内容を直接生成・再現するように LLM をファインチューニングするアプローチを採用している。本研究では果たしてこの個別の回答生成というアプローチが、調査結果の全体像や傾向を捉える上で最適な手法であるかについて再考する。

そこで本研究では、設問への回答を確率分布として捉え、その分布を直接推定するよう LLM をファインチューニングする手法を検討する。また、分布を LLM に学習させる際の表現形式について、複数の手法を実装し比較する。

本研究では、具体的に以下の点を明らかにする。

- 設問への回答の分布表現として、どのような形式が有効であるかを比較検討する。
- データに含まれるノイズがモデルの学習挙動や汎化機能に与える影響を分析する。

実験設定として、回答者のデモグラフィック情報を入力とし、設問への回答分布を出力するモデルを構築する。検証には、あらかじめ定義された確率分布に基づいて生成された人工データを用いる。これにより、LLM が背後にある分布構造をどの程度正確に捉えることができるかを定量的に評価し、理想的な条件下での LLM の分布表現力の上限を確認する。

2　実験の設計

LLM には回答者のデモグラフィック情報を入力として与え、対応する設問への回答分布を予測させるタスクを課す。

2.1　データセット

本実験の検証には、あらかじめ定義した確率分布に基づいて生成した人工データを用いる。評価は、学習データに含まれる設問への適合度（既存設問）と、学習データに含まれない設問への汎化性能（未知設問）の 2 軸で行う。加えて、学習データへのノイズ付与がモデルの学習挙動に与える影響についても検証する。

アンケートデータ　本実験では人工的なアンケートデータを用いる。データ作成にあたり、まずペルソナ（仮想的な回答者）を定義する。次に、設問を用意する。最後に、確率分布に基づき各ペルソナの回答を生成する。

アンケートの回答者として、nvidia/Nemotron-Personas-Japan から 10,000 人分のペルソナを抽出した。さらに、国税庁の申告所得税統計における所得階級別人員を参照し、各ペルソナの居住地域に基づいて年収を付与した。各ペルソナは、性別・年代・婚姻状況・子供の有無・居住地域・年収をデモグラフィック情報として持つ。

設問は、生活者の嗜好に関する 7 つのカテゴリ（お金・仕事・価値観・情報・政治・消費・生活）に分類し、各カテゴリに回答形式の異なる 8 つの設問を設定した。内訳は 2 値選択が 3 問、3 値選択が 3 問、5 値選択が 2 問であり、合計 56 設問となる。2 値選択の選択肢は「はい」「いいえ」、3 値選択は「はい」「いいえ」「どちらとも言えない」である。5 値選択は 1 週間あたりの頻度に関する設問であり、「全くない」「1 日程度」「2–3 日程度」「4–5 日程度」「ほぼ毎日」の 5 段階で回答する。設問の例を表 1 に示す。

本実験では、各回答者の回答を生成する際に、デモグラフィック属性によって回答傾向が変化するような確率分布を設計した。具体的には、各設問に対して基準となる回答確率（基本分布）を定義し、これに回答者の属性に応じた効果を加算することで最終的な回答確率を算出する。基本分布はロジットスケールで表現され、属性効果は各属性値に対する重みとして定義される。最終的な回答確率は、基本分布のロジットと該当する属性効果の和をソフトマックス関数で正規化することで得られる。例えば、ある 2 値選択の設問では、基本分布として「はい」75%・「いいえ」25%を設定し、「既婚」「子供あり」を「はい」の確率を高める属性、「若年層」「高齢層」を「いいえ」の確率を高める属性として定義している。この生成ルールは設問番号（Q1–Q8）に対して一意に定まり、カテゴリには依存しないため、異なるカテゴリであっても同じ設問番号であれば、同一の属性を持つ回答者からは統計的に等価な回答分布となる。このデータセット構造を、汎化性能の検証に利用する（2.3）。

2.2　分布の表現

各形式の具体的な出力例は付録に示す。

個票 [1, 2]　回答テキストをそのまま出力ターゲットとする最も単純な形式である（図 1）。学習データ内の出現頻度によって確率分布が暗黙的に表現されるが、最尤推定 (MLE) では最頻出回答への収束が生じやすい。推論時はサンプリングを行い、得られた回答群から分布を復元する。

個票+ID　デモグラフィック情報に加え、ランダムな ID（UUID 等）を入力することでモデルにランダム性の源泉を与える。単純な個票学習で生じやすい「最頻出回答への収束」の緩和を意図した手法である。推論方法は**個票**と同様である。

個票+回答確率 [4, 5, 7]　学習時の損失関数において、回答の先頭トークンの生成確率分布が正解の回答分布と合致するような項を追加する。それ以外のトークン生成には通常のクロスエントロピー損失を用いる。推論方法は**個票**と同様である。

個票+複数回答 [8]　1 つの学習事例に複数の回答を含めることで、回答間のばらつきや分布形状の効率的な学習を意図する（図 2）。推論時は、生成された複数の回答から分布を計算する。

カテゴリカル　設問への回答をカテゴリカル分布として捉え、その分布を直接モデル化する（図 3）。学習データにおける回答の偏り（インバランス）の影響を受けにくい特徴がある。推論時は出力文字列を解析して確率分布を得る。

積率　回答分布をパラメータ付き分布と仮定し、その積率（モーメント）を推定する。「はい」「いいえ」等の選択肢を持つ設問には正規分布を仮定し、正の値を「はい」、負の値を「いいえ」、0 付近を「どちらとも言えない」に対応付ける。頻度に関する設問にはポアソン分布を適用する。モデルは分布パラメータ（平均・分散等）を含む文字列を直接予測するため（図 4）、損失関数の変更を必要としない。不確実性を伴う回答の曖昧さを適切に表現できる可能性があり、推論時は出力パラメータから分布を再構築する。

2.3　評価設定

評価は、学習データに含まれる設問への適合度（既存設問）と、学習データに含まれない設問への汎化性能（未知設問）の 2 軸で行う。加えて、学習データへのノイズ付与がモデルの学習挙動に与える

表 1: アンケート設問の例

カテゴリ	設問	選択肢
お金	毎月、決まった額の貯金をしている	はい, いいえ
お金	経済的な余裕があるか	はい, どちらとも言えない, いいえ
仕事	好きなことを仕事にできているか	はい, どちらとも言えない, いいえ
価値観	多少大変でも成長できる環境に身を置きたい	はい, どちらとも言えない, いいえ
情報	生成 AI 系のサービス・アプリを利用している	はい, いいえ
消費	買う前に値段を比較する	はい, いいえ
生活	1 週間のうち、運動をする日	全くない, 1 日程度, ..., ほぼ毎日

表 2: データ分割（●: 学習, ○: 未知設問）

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
お金	○	●	●	●	●	●	●	●
仕事	●	○	●	●	●	●	●	●
価値観	●	●	○	●	●	●	●	●
情報	●	●	●	○	●	●	●	●
政治	●	●	●	●	○	●	●	●
消費	●	●	●	●	●	○	●	●
生活	●	●	●	●	●	●	○	●

影響についても検証する。

データ分割　各カテゴリから 1 つの設問を未知設問として除外し、残りを学習データとする分割を行う。除外する設問番号がカテゴリ間で重複しないよう、カテゴリごとに異なる設問を割り当てた（表 2）。

回答分布の生成ルールは設問番号に対して一意に定まるため、カテゴリが異なっても同一設問番号であれば分布構造は共通である。未知設問の分布予測では、汎化の精度ではなく、汎化自体が生じうるか否かを検証する。

ノイズ付与　学習データにノイズを付与することは、モデルの汎化性能を向上させる手法として広く知られている。画像認識におけるデータ拡張 [9] や、ニューラルネットワークにおけるドロップアウト [10] はその代表例である。本実験では、学習データへのノイズ付与が分布推定タスクにおいても汎化に寄与するかを検証する。

個票+回答確率・カテゴリカル・積率では、学習例を作成する際に回答者をサブサンプリングして集計することでノイズを導入する。同一のデモグラ・設問の組み合わせであっても、サンプルされる回答者が異なることで学習例ごとに回答分布が変動する。サンプル率を 1.0（無）、0.8（小）、0.5（中）、0.2（大）と変化させることでノイズの強度を制御する。加えて、デモグラや設問内容とは無関係に Dirichlet 分布から回答分布を直接サンプルした条件（全）を設け、背後に学習可能な分布構造が存在しない場合のベースラインとする。

個票・個票+ID では、各回答者の回答を生成する際に、算出された回答分布にノイズを加えることで実現する。ノイズ最大条件（全）では同様に、Dirichlet$(0.8, \ldots, 0.8)$ からサンプルした回答分布に基づいて回答を生成する。

2.4　実験の手順

ベースモデルとして Qwen2.5-7B-Instruct[11] を使用し、LoRA[12] によるファインチューニングを行った。学習時の各事例には、学習データから無作為に選択された 1〜20 問の設問とその回答を含めた。個票および個票+ID では約 500 万件、その他の手法では約 100 万件の学習の用例を用意した。学習は全データを読み切る前に終了し、通常約 0.6 エポック程度を消化した。これにより、1 回の推論で複数の設問に対する回答分布を同時に予測できるようモデルを学習させた。詳細な学習パラメータは付録に記載する。

評価は、学習データに含まれる設問を対象とした分布内評価（既存設問）と、学習データに含まれない設問を対象とした分布外評価（未知設問）の 2 つの設定で行った。既存設問では、モデルが学習データで与えられた分布にどの程度適合できるかを測定する。未知設問では、データ分割で述べた設問間の構造的な重複を利用して、モデルが未知の設問に対しても分布構造を学習・汎化できるかを検証する。

3　分析

評価結果を表 3 に示す。L2 距離は、推定された分布が本実験のために定義した生成元の確率分布にどの程度近いかを測定する指標である。平均生成エントロピーは、推論された分布の非一様性を測定する指標である。一様分布のエントロピーが 1 となるよう正規化しており、値が 1 に近いほど分布が一様で

表 3: 評価結果：L2 距離とモデル出力のエントロピー（L2 距離が最小のチェックポイントを使用）

手法	指標	既存設問					未知設問				
		無	小	中	大	全	無	小	中	大	全
個票	L2	0.232	0.244	0.237	0.238	0.340	0.373	0.376	0.345	0.371	0.329
	Ent	0.774	0.770	0.801	0.792	0.915	0.865	0.863	0.888	0.864	0.930
個票+ID	L2	0.231	0.235	0.238	0.233	0.345	0.382	0.375	0.395	0.370	0.340
	Ent	0.792	0.770	0.781	0.797	0.908	0.883	0.856	0.878	0.844	0.933
個票+回答確率	L2	0.196	0.197	0.191	0.204	0.318	0.358	0.383	0.383	0.373	0.339
	Ent	0.706	0.745	0.763	0.835	0.947	0.873	0.888	0.865	0.907	0.947
個票+複数回答	L2	0.174	0.178	0.175	0.177	0.292	0.365	0.367	0.354	0.360	0.326
	Ent	0.824	0.808	0.829	0.825	0.984	0.914	0.907	0.924	0.914	0.986
カテゴリカル	L2	0.152	0.156	0.175	0.219	0.457	0.354	0.326	0.322	0.353	0.435
	Ent	0.894	0.895	0.907	0.914	0.753	0.916	0.938	0.943	0.940	0.755
積率	L2	0.172	0.177	0.194	0.220	0.433	0.353	0.366	0.331	0.360	0.476
	Ent	0.886	0.880	0.897	0.899	0.705	0.889	0.899	0.913	0.905	0.714

あることを示す。

既存設問において、ノイズが増加するにつれて全ての手法で予測精度が低下した。興味深い点として、個票・個票+ID・個票+回答確率はノイズ増加に伴い一様分布に近づく傾向を示したのに対し、カテゴリカル・積率は非一様な分布を維持した。これは、ノイズ条件（全）では Dirichlet$(0.8, \ldots, 0.8)$ からサンプルした分布の期待値が一様分布となるためと考えられる。個票ベースの手法では各学習事例が単一の回答であるため、大量の事例を通じて一様分布への収束が学習される。一方、カテゴリカル・積率では各学習事例が非一様な分布を明示的に含むため、モデルは非一様な出力を維持する。

既存設問において、カテゴリカルと積率は個票ベースの手法よりも正解分布に近い予測を示した。カテゴリカルは積率よりもわずかに良好な結果を得た。個票+回答確率のように損失関数を直接最適化するアプローチも一定の効果を示したが、LLM が理解しやすい形式で分布を明示的に記述する手法ほど効果的ではなかった。

個票・個票+ID は最頻出回答への収束という失敗モードには陥らなかったが、これは学習データが比較的大規模であったためと考えられる。一方で、個票+ID はランダム ID によってこの収束を緩和することを意図した手法であるが、個票と比較して明確な改善は見られなかった。

未知設問では、ノイズ最大の設定において既存設問とほぼ同等のスコアを示した。これは、モデルが背後にある分布構造を学習し、常に何らかの修正を加えていることを示唆している。ノイズが小さい設定ではより良好なスコアが得られたが、最良の汎化性能はノイズがゼロの設定ではなく、適度なノイズを含む設定で達成された。この結果は、適度なノイズが汎化に有益であることを示している。

また、未知設問においてカテゴリカル・積率は個票ベースの手法よりも一様分布に近い出力を示した。これは、LLM が未知の設問に対して一様分布を事前分布として学習している可能性を示唆しており、今後の検証課題である。

4 おわりに

人工データを用いて LLM によるアンケート回答分布推定における 5 つの分布表現手法を比較検討した。カテゴリカルおよび積率は個票ベースの手法を上回る性能を示し、LLM が解釈しやすい明示的な分布表現が、暗黙的な頻度ベースの学習よりも効果的であることが確認された。また、適度なノイズが汎化性能の向上に寄与することが示された。

カテゴリカル形式は分布推定において実用的な選択肢であるが、この優位性が実際の調査データでも成立するかは検証を要する。本実験は理想的な人工データ条件下での分布推定能力の上限を確認するものであり、実際の調査データはより複雑なパターンを含む可能性がある。今後の課題として、実データによる検証および未知設問に対する LLM の事前分布に関する調査が挙げられる。

生成 AI の利用

本論文では、執筆支援および実装補助に Claude と Gemini を使用した。なお、実験データの生成には生成 AI を使用していない。

参考文献

[1] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 29971–30004. PMLR, 23–29 Jul 2023.

[2] Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. Employing large language models in survey research. **Natural Language Processing Journal**, Vol. 4, p. 100020, 2023.

[3] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. In **Proceedings of the 42nd International Conference on Machine Learning (ICML)**, 2025.

[4] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 3141–3154, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[5] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. 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. 21147–21170, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] Ji Huang, Mengfei Li, and Shuai Shao. Distribution shift alignment helps llms simulate survey response distributions, 2025.

[7] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 29971–30004. PMLR, 23–29 Jul 2023.

[8] 並河進, 山本覚, 木幡容子, アグチバヤルアマルサナー, ツェレンサンブーバサンドルジ. 生成 ai を活用した大規模生活者予測モデルの研究. 人工知能学会全国大会論文集, Vol. JSAI2025, pp. 3I4GS1105–3I4GS1105, 2025.

[9] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. **Journal of Big Data**, Vol. 6, No. 1, pp. 1–48, 2019.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, Vol. 15, No. 56, pp. 1929–1958, 2014.

[11] Qwen Team. Qwen2.5 Technical Report. **arXiv preprint arXiv:2412.15115**, 2024.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

表 4: 各設問の生成設定例

	選択肢と基本分布	追加効果の概要
Q1	はい: 0.75 いいえ: 0.25	**はい** ↑: 既婚, 子有, 関東; **いいえ** ↑: 中年層
Q2	はい: 0.30 いいえ: 0.70	**はい** ↑: 女性, 北海道・東北, 高所得; **いいえ** ↑: 既婚, 子有, 高齢層
Q3	はい: 0.50 いいえ: 0.50	**はい** ↑: 男性, 中年層, 低所得, 中部・近畿; **いいえ** ↑: 子有
Q4	はい: 0.25 中立: 0.20 いいえ: 0.55	**はい** ↑: 男性, 未婚, 若年層, 中国・四国; **中立** ↑: 子無
Q5	はい: 0.30 中立: 0.40 いいえ: 0.30	**はい** ↑: 女性, 中年層, 子有, 低所得; **中立** ↑: 関東
Q6	はい: 0.55 中立: 0.20 いいえ: 0.25	**はい** ↑: 高齢層, 子有; **中立** ↑: 既婚, 九州
Q7	全くない: 0.30 1 日: 0.25 2–3 日: 0.20 4–5 日: 0.15 ほぼ毎日: 0.10	**高頻度** ↑: 男性, 既婚, 子有; **中頻度** ↑: 中部・近畿; **低頻度** ↑: 若年層
Q8	全くない: 0.15 1 日: 0.20 2–3 日: 0.30 4–5 日: 0.20 ほぼ毎日: 0.15	**高頻度** ↑: 高齢層, 高所得; **中頻度** ↑: 女性, 関東; **低頻度** ↑: 未婚

A　人工データ生成の詳細設定

本実験で使用した人工データの生成ルールを詳述する。各設問は、基準となる回答確率（基本分布）に対し、回答者のデモグラフィック属性に応じた効果（属性効果）を加算することで生成される。属性効果はあらかじめテンプレートとして定義されており、年齢が上がるほど特定の選択肢を選びやすくなる傾向や、既婚者・特定の居住地域に対するバイアスなどが含まれる。

実験に使用した設問ごとの設定を表 4 に示す。各設問（Q1–Q8）はカテゴリ（お金・仕事等）に依存せず共通の構造を持つ。表中の「追加効果」は、基本分布に対して正の影響（確率上昇）または負の影響を与える条件を記述している。なお、表中の記述は主要な効果を示したものであり、ある属性で特定の選択肢（例：「はい」）の確率が上昇する場合、その逆の属性では対立する選択肢（例：「いいえ」）の確率が上昇する関係にある。

B　表現の回答例

```
## 回答 1
1 日程度
## 回答 2
いいえ
```

図 1: 回答表現例：個表

```
## 回答 1
1 日程度
2〜3 日
ほぼ毎日
1 日程度
## 回答 2
いいえ
はい
いいえ
いいえ
```

図 2: 回答表現例：複数回答

```
## 回答 1
全くない: 55%
1 日程度: 15%
2〜3 日: 10%
4〜5 日: 5%
ほぼ毎日: 15%
## 回答 2
はい: 30%
いいえ: 70%
```

図 3: 回答表現例：カテゴリカル

```
## 回答 1
Poisson(0.5)
## 回答 2
N(-0.3, 0.5)
```

図 4: 回答表現例：積率

C　学習のパラメーター

設定	値
Model	Qwen2.5-7B-Instruct
Optimizer	AdamW
LR	$3 \cdot 10^{-5}$
LoRA r	8
LoRA α	16
GPU	A100@80Gx8
Batch Size	平均 5 (x8) (可変長)
Gradient Accumulation	なし
モデル更新数	12000