LLM を用いたアンケート回答生成における精度の要因分析

尾城 奈緒子 [1] アフィカ アディラ [1] 小田 虎之介 [1]
[1] 株式会社インテージ
{oshiro.44969,afiqah-a,oda.44970}@intage.com

概要

消費者パネル調査は広く利用されているが，追跡調査やデータ結合に伴う離脱や欠損が課題となっている．本研究では欠損データ補完への応用を見据え，LLM を用いたアンケート回答生成の精度に影響する要因を明らかにする．対象は車に関する定点調査 4 問である．入力データとして，定点調査以外の個人特性を表すアンケート回答を 5 つの要素に分類し，要素ごとの推論精度への影響を評価した．さらに，車に関する定点調査の 1 年前の過去回答の入力有無による精度差を検証した．結果，過去回答がない場合には属性情報が精度に最も寄与する一方，過去回答がある場合には全要素で正解率 0.7 以上を維持し，要素間の差は小さいことが示された．

1 はじめに

近年，社会調査やマーケティング・リサーチの分野におけるアンケート調査では，コスト効率や調査速度の観点から消費者パネルが広く利用されている [1, 2]．こうしたパネル調査は調査実務の効率化に寄与する一方で，回答率の低下や若年層の離脱といった課題が指摘されている [3, 4]．特に，同一対象者に複数回の回答を求める追跡調査や，複数の調査データを回答者単位で結合する場合には，調査回答者の離脱やデータ欠損により，期待するサンプルサイズが確保できないことがある．

欠損データの補完手法としては，従来の多重代入法に加え，深層モデルを応用する研究が進んでいる．特に，GAN を用いた GAIN [5] や，VAE を基盤とする MIWAE [6]，および任意条件付き VAE である VAEAC [7] などは，表形式データの欠損補完において高い性能を示している．

しかし，これらの手法は観測された回答分布に基づいて欠損値を補完することを前提としている．そのため，個人 ID をキーとして複数のアンケート回答データや購買データなどを回答者単位で結合する際に欠落するデータの補完に必要な学習データに存在しない未知カテゴリや未観測属性値に対しては適切な推定が困難である．一方で，大規模言語モデル（LLM）は学習データに含まれない表現やカテゴリを生成できることから，このような欠損データを補完できる可能性がある．そこで本研究では，プロンプトに含める要素を分類し，LLM を用いたアンケート回答生成の精度およびその精度に影響を与える要因を，回答者（個票）単位で検証する．

2 関連研究

LLM を用いてアンケート回答を生成する手法については，近年さまざまな研究が提案されている．Lutz ら [8] は，LLM に与えるペルソナの書き方を体系的に比較し，生成される回答のバイアスや意見の一貫性が大きく変化することを示した．特に，インタビュー形式でデータを入力すると，最も整合性の高い回答を生成しやすく，属性を入れずに人物名のみを与える場合はステレオタイプ的な回答傾向を低減できる可能性が示唆されている．

Zhao ら [9] は，LLM を「仮想的な調査回答者」として用い，部分属性シミュレーション（PAS）と全属性シミュレーション（FAS）という 2 つのタスクを通じて LLM が生成した回答の評価を行った．PAS では，回答者プロフィールの一部情報に基づいて欠損属性を予測し，FAS ではゼロコンテキスト条件と，社会・公的統計データを付与したコンテキスト強化条件の双方において，完全な合成アンケートデータセットの生成を行っている．その結果，プロンプト設計や付与される文脈情報がシミュレーション精度に大きな影響を及ぼすこと，またアンケート項目全体を LLM によって仮想生成することには一定の限界があることが示された．一方で，文脈要素が推論精度に寄与する影響度については明確には示されていない．

これらの研究はいずれも，人口レベルでの分布や傾向の再現性を主な評価対象としている．個票レベ

ルの既存回答情報を条件として，新たな回答項目を生成する研究としては，Kim・Lee[10] が挙げられる．この研究では，未観測項目の生成精度について AUC ＝ 0.73 という結果が報告され，一定の有効性が示されている．ただし，同研究は似たようなテーマを扱う調査票の未回答項目を生成する設定に限定されており，例えば文化的志向や宗教観といった属性情報を入力として，同性婚に対する賛否を生成するなど，既存項目間の関係性を利用した生成にとどまっている．そのため，全く異なるトピックやテーマに関する新規カテゴリの回答生成については十分に検証されていない．

3 実験

3.1 問題設定

本研究では，個人 ID をキーとして紐づけられた複数の調査データを入力として用い，特定の個人が定点調査における将来のアンケートにどのように回答するかを生成するタスクを扱う．具体的には，購買行動，性格特性，過去の定点調査回答などの情報を条件として，新たなアンケート設問に対する回答を LLM により生成する．

個人 i に対して与えられる入力情報は以下の 3 種類である：

- **定点調査以外のアンケート回答**: 年齢や性別などの属性情報，性格特性，購買履歴，行動傾向など，過去に回答された各種アンケート項目
- **生成対象となる定点調査の過去回答**: 同一個人による過年度の定点調査回答
- **外部データ**: アンケート設問に関連する，政策動向や社会情勢などの外部情報

出力は，個人 i に対して，ターゲットとなる設問集合 Q に含まれる各設問 $q \in Q$ の回答である．

$$y_{i,q}$$

$y_{i,q}$ の回答形式は，q に依存し，選択式，数値回答，自由記述などが考えられるが，本研究では単一の回答形式のみを対象とする．

3.2 データセット

アンケート回答生成の精度にどのような個人特性の情報が影響を与えるかを検証するため，入力データとして与える定点調査以外のアンケート回答を，5 つの要素に分類した．アンケート回答データは，インテージが保有するデータベース[1)]から取得した．各要素の概要を表 1 に示す．

表 1 定点調査以外のアンケート回答の分類

要素	代表的な項目
属性（Attributes）	性別／年齢／独身／社会人
性格特性（Personality traits）	やさしい／しっかりした
志向性（Orientations）	食に対する考え方／健康志向
嗜好性（Preferences）	好きな色／ペット／興味関心
行動（Behaviors）	購入経験／情報収集／ SNS 発信

また，アンケート回答の生成に用いる外部情報として，近年の政策動向や社会情勢に関するデータを，GPT-5 により作成した．

モデルが生成するアンケート回答は，インテージが保有する業界サブパネルデータ Car-kit[2)] に含まれる設問 4 問を対象とする．概要を表 2 に示す．

表 2 生成するアンケートの調査項目および選択肢の概要

調査項目	選択肢例	選択肢数
車の保有有無	家に車はあるが，自分は運転しない	4
前月までの車購入有無	“あなたが運転する車” を新たに買った	4
現有車新車中古区分	新車で購入	4
今後の車購入予定	1 ヶ月以内	11

評価には 2025 年 6 月時点の Car-kit データを用い，同一個人の 1 年前にあたる 2024 年 6 月の回答を過去情報として入力に含めた．

個人 ID をキーとして複数のアンケート回答データを結合した結果，合計 3,694 件のサンプルが得られた．このうちランダムに 1,000 件を抽出し，テストデータとして用いた．

3.3 実験方法

本研究では，以下の 2 点について検証を行う．

- アンケート回答生成の精度に影響を与える要素
- 過去回答の有無が生成精度に及ぼす影響

アンケート回答生成の精度に影響を与える要素 表 1 に示した 5 つの要素である属性，性格特性 (性格)，志向性 (志向)，嗜好性 (嗜好)，行動が，LLM によるアンケート回答生成の精度にどの程度寄与している

1） https://www.intage.co.jp/service/services/database/
2） https://www.intage.co.jp/service/services/database/car-kit/

表 3　車保有有無および車購入意向における各モデルの精度（Acc/F1）

調査項目	Methods	-属性		-性格		-志向		-嗜好		-行動		ALL		+外部	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
過去回答なし															
車保有有無	GPT5	0.54	0.18	0.54	0.26	0.54	0.25	0.54	0.25	0.54	0.26	**0.56**	0.28	0.55	0.26
	GPT-5 mini	0.54	0.18	0.54	0.22	0.54	0.21	0.54	0.22	0.54	0.21	0.55	0.22	0.55	0.19
	Gemini 2.5 Flash-Lite	0.54	0.18	0.51	0.30	0.51	0.30	0.51	**0.31**	0.52	0.28	0.50	0.30	0.53	0.20
車購入予定	GPT5	0.05	0.03	0.13	0.07	0.12	0.05	0.13	0.07	0.14	0.08	0.13	0.07	0.11	0.07
	GPT-5 mini	0.25	**0.11**	0.34	0.10	0.28	0.07	0.34	0.08	0.34	**0.11**	**0.37**	**0.11**	0.24	0.07
	Gemini 2.5 Flash-Lite	0.07	0.05	0.19	0.10	0.15	0.07	0.22	**0.11**	0.18	0.09	0.23	0.10	0.06	0.04
4 調査項目	**Average**	0.43	0.18	0.46	0.20	0.44	0.20	0.46	0.20	0.46	0.19	**0.47**	**0.21**	0.44	0.19
過去回答あり															
車保有有無	GPT5	0.96	0.94	0.96	0.94	0.96	0.94	0.96	0.94	0.96	0.94	0.96	0.71	0.96	0.70
	GPT-5 mini	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95
	Gemini 2.5 Flash-Lite	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.95
車購入予定	GPT5	0.70	0.32	0.71	0.32	0.71	0.34	0.71	0.32	0.71	0.32	0.71	0.32	0.70	0.32
	GPT-5 mini	0.73	0.33	0.73	0.32	0.73	0.32	0.73	0.32	0.73	0.32	0.73	0.33	0.73	0.32
	Gemini 2.5 Flash-Lite	0.74	0.32	0.74	0.32	0.74	0.32	0.74	0.32	0.74	0.33	0.74	0.32	0.73	0.32
4 調査項目	**Average**	0.88	0.66	0.87	0.65	0.88	0.66	0.87	0.65	0.87	0.65	0.88	0.64	0.87	0.63

かを検証するため，要素ごとに入力データから除外する条件を設定した比較実験を行った．

具体的には，すべての要素を入力に含めた条件（**All-features**）を基準とし，各要素を 1 つずつ入力から除外した条件でモデルに回答を生成させた．これらの生成結果を比較することで，要素全体としての効果と，各要素が生成精度に及ぼす寄与を評価した．定点調査以外のアンケート回答のみを用いた場合と外部データを追加した場合との差を明確にするため，本研究では All-features に外部データを含めていない．

過去回答の有無が生成精度に及ぼす影響　LLM による未知カテゴリのアンケート回答生成において，過去の定点調査回答が生成精度に与える影響を検証するため，過去回答を入力に **含める場合** と **含めない場合** の 2 条件で比較実験を行った．

以上の 2 種類の条件をすべて組み合わせ，Zhao ら（2025）[9] を参考に構築したアンケート回答生成プロンプトを用いて実験を実施した．使用した生成モデルは，GPT-5，GPT-5 mini，および Gemini 2.5 Flash-Lite の 3 種類である．

各条件で生成された回答について，Car-kit の正解データと照合し，設問単位での正解率 (Accuracy) および F1(F1-score) を評価指標として用いた．評価では，テストデータに対して実験を 5 回繰り返し，その平均値を最終的な評価結果とした．

3.4　実験結果

各条件に対する実験結果を表 3 および付録 B の表 4 に示す．表 3 より，アンケートの過去回答がない場合，データ生成の正解率の平均は最大でも 0.47，F1-score は 0.21 と全体的に低いことがわかる．特に車の次期購入意向に関しては，他の調査項目と比べて精度が低い結果となった．このことから，属性や性格といった個人特性の情報のみを用いて，異なるトピックのアンケート回答を生成することは困難であり，とりわけ将来の行動意向を予測することは難しいことが示唆される．また，モデル別の精度を確認すると，車の次期購入意向に関しては GPT-5 よりも GPT-5 mini と Gemini 2.5 Flash-Lite の方が精度が高い結果となった．

一方，過去回答がある場合には，入力からどの要素を除外しても正解率は概ね 0.70 以上を維持しており，F1-score も調査項目による差はあるものの平均で 0.60 以上と高い値を示した．また，モデルの違いや，属性・性格などの要素の有無による精度差はほとんど見られず，個人の過去回答の有無が生成精度に最も大きな影響を与えていることがわかる．

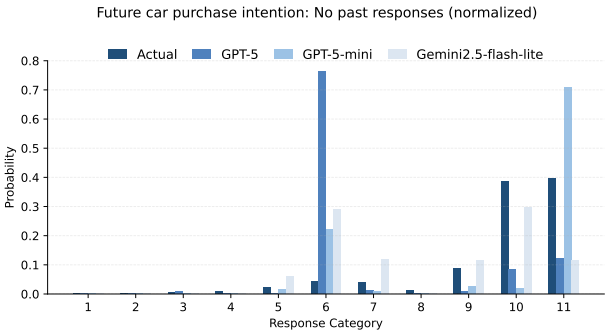

図 1　過去回答なしの回答分布と LLM での生成分布

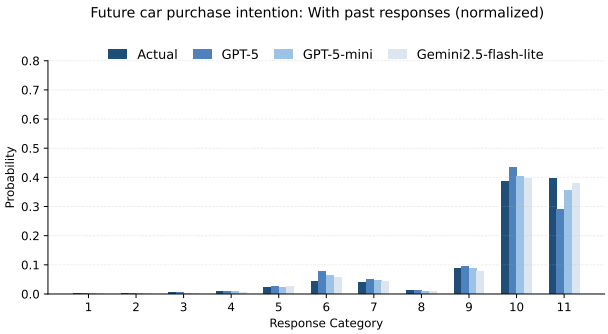

図 2　過去回答ありの実回答分布と LLM での生成分布

さらに，要素ごとの精度を比較すると，過去回答がない条件では，属性を入力から除外した際に精度が低下しており，属性情報が生成精度に与える影響が相対的に大きいことが確認できる．車の次期購入意向ではすべての要素を含めた条件と比較し，正解率が 0.16 ポイント下がっている．また，社会情勢のような外部データを要素として追加した場合には All-features と比較し精度が低下する傾向が見られた．これらの結果から，アンケート回答生成においては，関連性の高い情報のみを適切に選択して入力として与えることが重要であると考えられる．

4　考察

LLM で生成した回答分布について　LLM が生成するアンケート回答の分布が実際の回答分布とどの程度一致しているかを確認するため，各モデルについて All-features 条件で生成された回答分布と実回答分布を棒グラフで比較した．車の次期購入意向の結果の過去回答なしを図 1 に，過去回答ありを図 2 に示す．図 1 より，GPT-5 は過去回答を入力しない条件において，他のモデルと比べて選択肢「6」を生成する傾向がある．アンケート回答生成において，大規模な GPT-5 モデルが GPT-5 mini や Gemini 2.5 Flash-Lite よりも低い精度を示した要因の一つとして，このような中心化が考えられる．一方で図 2 より，過去回答を入力した場合には，いずれのモデルにおいても実データに近似した回答分布が得られた．車の購入頻度は食品や雑貨と比較し高くないため，過去回答が生成に大きく影響されていると考えられる．

5　結論

本研究では，LLM を用いたアンケート回答生成において，生成精度に影響する要因を回答者単位で検証した．その結果，過去回答がある場合，比較的軽量なモデルでも一定の精度で回答生成が可能であるが，過去回答がない場合はどの調査項目でも正解率と F1-score が低い．このことから，個人に結び付く過去に回答された各種アンケート情報のみを用いて，異なるトピックの新たなアンケート回答を生成することは困難であるといえる．プロンプトに含める情報としては属性情報が生成精度に与える影響が相対的に大きいが，過去回答がある場合は過去回答の影響が大きく要素間の差は小さいことが示された．本研究で用いた過去回答は直近 1 年分に限定されており，それ以前の回答でも同様に実分布と近い回答分布を生成できるかどうかは未検証である．過去回答参照期間やアンケート回答の対象とする調査カテゴリの範囲を拡張し，より包括的な評価を行うことを今後の課題とする．

謝辞

本研究は株式会社インテージホールディングスグループ R&D センターの助成で行われた．

参考文献

[1] Mario Callegaro and Charles DiSogra. Computing response metrics for online panels. **Public opinion quarterly**, Vol. 72, No. 5, pp. 1008–1032, 2008.

[2] Reg Baker, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau. Summary report of the aapor task force on non-probability sampling. **Journal of survey statistics and methodology**, Vol. 1, No. 2, pp. 90–143, 2013.

[3] Paul Malschinger, Susanne Vogl, and Brigitte Schels. Drop in, drop out, or stay on: patterns and predictors of panel attrition among young people. **Österreichische Zeitschrift für Soziologie**, Vol. 48, No. 3, pp. 427–450, 2023.

[4] Nicole Rübsamen, Manas K Akmatov, Stefanie Castell, André Karch, and Rafael T Mikolajczyk. Factors associated with attrition in a longitudinal online study: results from the habids panel. **BMC medical research methodology**, Vol. 17, No. 1, p. 132, 2017.

[5] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In **International conference on machine learning**, pp. 5689–5698. PMLR, 2018.

[6] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In **International conference on machine learning**, pp. 4413–4423. PMLR, 2019.

[7] Oleg Ivanov, Michael Figurnov, and Dmitry P. Vetrov. Variational autoencoder with arbitrary conditioning. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[8] 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 2025**, pp. 23212–23237, Suzhou, China, November 2025. Association for Computational Linguistics.

[9] Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. Large language models as virtual survey respondents: Evaluating sociodemographic response generation. **CoRR**, Vol. abs/2509.06337, , 2025.

[10] Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. **CoRR**, Vol. abs/2305.09620, , 2023.

```
#システム
出力フォーマットに従い、2025年6月の調査結果のみを出力してください

#背景情報
{{survey_content_information}}

#分析の目的
あなたはデータサイエンティストでありマーケターです。 調査を回答する人物の情報とともに自身の社会的常識やモデル構
築力を用いて、リアルな現有車購入パターンを仮定し現有車や次期意向など車に関する調査結果をシミュレーション生成
してください。

#調査を回答する人物の情報
{%- if include_attributes %}
### 属性情報
{{attribute_data}}
{%- endif %}

（省略）

{%- if include_current_car_info %}
# 2024年6月の車に関する情報
{{Current_car_info_data}}
{%- endif %}

#予測タスク
回答は以下の四つの質問に回答します。回答は数値をシングルアンサーで答えます。
{{question}}

#出力フォーマット
次のフォーマットの辞書型で出力してください(valueは数値形式です。唯一の正解例):
{ 'QS1':value, 'QS2':value, 'QS3':value, 'QS4':value }
```

図 3　アンケート回答生成に用いたプロンプト

A　プロンプト例

アンケート回答生成に用いたプロンプトを，図 3 に示す.

B　設問別の詳細結果

3.4 節で示した実験結果に含まれない設問について，その詳細な評価結果を，表 4 に示す.

表 4　車購入有無および車中古区分における各モデルの精度（Acc/F1）

調査項目	Methods	-属性		-性格		-志向		-嗜好		-行動		ALL		+外部	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
過去回答なし															
車購入有無	GPT5	0.65	0.29	0.68	**0.32**	0.68	0.31	0.69	0.31	**0.70**	0.24	0.69	0.31	0.68	**0.32**
	GPT-5 mini	0.66	0.25	0.65	0.25	0.65	0.25	0.64	0.25	0.66	0.20	0.66	0.25	0.65	0.26
	Gemini 2.5 Flash-Lite	0.67	0.20	0.67	0.24	0.67	0.23	0.68	0.24	0.67	0.22	0.68	0.24	0.66	0.20
車中古区分	GPT5	0.44	0.23	0.42	0.20	0.39	0.21	0.41	0.20	0.41	0.18	0.41	0.19	0.44	0.18
	GPT-5 mini	0.45	0.20	0.45	0.20	0.45	0.21	0.45	0.20	0.45	0.20	**0.46**	0.20	0.44	0.20
	Gemini 2.5 Flash-Lite	0.29	0.23	0.35	0.22	0.34	0.23	0.35	0.24	0.32	0.22	0.33	0.23	0.38	0.25
過去回答あり															
車購入有無	GPT5	0.96	0.52	0.95	0.53	0.95	0.54	0.95	0.54	0.96	0.49	0.95	0.53	0.95	0.53
	GPT-5 mini	0.96	0.59	0.95	0.55	0.95	0.54	0.95	0.55	0.95	0.55	0.96	0.59	0.95	0.53
	Gemini 2.5 Flash-Lite	0.94	0.57	0.93	0.57	0.94	0.56	0.92	0.58	0.92	0.57	0.93	0.58	0.92	0.58
車中古区分	GPT5	0.86	0.70	0.85	0.69	0.86	0.73	0.87	0.69	0.85	0.73	0.85	0.69	0.84	0.73
	GPT-5 mini	0.93	0.87	0.91	0.87	0.92	0.87	0.91	0.87	0.91	0.84	0.92	0.87	0.93	0.84
	Gemini 2.5 Flash-Lite	0.86	0.84	0.87	0.86	0.87	0.85	0.87	0.85	0.88	0.83	0.87	0.87	0.83	0.85