Evaluating the Bias in LLMs for Surveying Opinion and Decision Making in Healthcare

Anonymous ACL submission

Abstract

Large language models (LLMs) can simulate human decision-making without labelled training data, yet their alignment with real-world behavioural The patterns remain unclear. We compare five open-source LLMs (GPT-OSS, Qwen3, Llama3.1, Gemma3, Ministral-3) against supervised baselines (logistic regression, random forest, XGBoost) on COVID-19 vaccine intent prediction across four pandemic periods in 2020, evaluated against the Understanding America Study (UAS) survey responses. Overall, traditional models outperform LLMs in mean accuracy and F1. However, LLMs require no prior labelled data, giving them a unique advantage at pandemic onset, when supervised models cannot be deployed; GPT-OSS:20b achieved the highest single-period accuracy in the study. Bias analysis shows that most LLMs misrepresent group-level hesitancy, amplifying disparities by race, education, and income while flattening gender differences, so we treat them as zero-data early-phase simulators and defer to supervised models once labelled data becomes available.

1 Introduction

The rise of large language models has enabled agentic AI systems that simulate human behavior at scale through "simulacra" populations (Park et al., 2023; Shanahan et al., 2023). These generative agents show promise for studying social dynamics without real-world risk. However, their ability to model *decision-making* under demographic and contextual constraints remains largely unexplored and has not been systematically evaluated against empirical group-wise data, which is precisely what public-health applications require.

Vaccine acceptance during COVID-19 provides a critical testbed: hesitancy varied sharply by race, income, and education, shaped by structural and social factors rather than medical facts alone (Kapteyn et al., 2024). If LLM-based simulacra are to inform policy analysis or intervention design, they must reproduce, rather than distort, these observed group-level patterns.

Existing work leaves this question open. Clinical LLM studies primarily target guideline-based diagnostic or treatment tasks (Hager et al., 2024), while most generative-agent studies emphasise qualitative plausibility (e.g., believable routines, personality fidelity) rather than quantitative alignment with large-scale behavioural datasets (Park et al., 2023; Wang et al., 2024). Fairness and bias research in LLMs has documented demographic disparities, yet comparatively little work benchmarks model outputs against *empirical* population-level behaviour from large surveys in healthcare decision-making (Ferrara, 2024; Kim et al., 2025).

We address this gap by comparing LLM-generated vaccination decisions to the Understanding America Study (UAS) (Kapteyn et al., 2024), a nationally representative survey covering early 2020 before vaccines existed. We prompt five instruction-tuned open-source LLMs with structured demographic personas across four pandemic periods and compare their outputs to UAS responses, also including classical statistical baselines trained directly on UAS tabular data. We define *bias* as misalignment with UAS group patterns, either *amplification* (exaggerating disparities) or *flattening* (erasing them), measured via Disparate Impact Ratio (DIR) and Jensen–Shannon Divergence (JSD) (Feldman et al., 2015; Lin, 1991).

Contributions:

- We explore a survey-anchored framework for evaluating LLM-based simulacra against empirical health decisions, using UAS vaccine-intention data across four early-pandemic periods.
- We investigate how instruction-tuned LLMs reproduce or distort vaccine hesitancy patterns

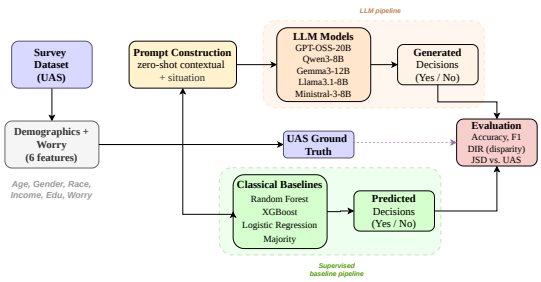

Figure 1: Overview of the experimental setup

across demographic groups, finding consistent tendencies to amplify or flatten disparities relative to UAS.

- We examine how design choices in prompting, temporal context, and demographic specification are reflected in LLM-generated vaccine decisions, offering initial guidance for using LLMs as behavioural components in public-health simulations.

2 Method

To analyse how LLMs approximate human decision-making in healthcare, we design a study that integrates demographic attributes, contextual prompts, and LLM-generated decisions (Figure 1). Model predictions are compared to UAS survey data across four pandemic periods to assess alignment and detect biases.

2.1 Dataset

We use the Understanding America Study (UAS) Coronavirus in America survey (Kapteyn et al., 2024) and analyse waves from March 2020 to January 2021, when vaccine information was still limited. UAS uniquely covers this early period; other surveys such as HPS (Beleche et al., 2021) began vaccine questions in 2021, after broader availability. UAS also provides rich demographic and behavioural fields relevant to our goals (see Appendix A.1).We align model outputs to UAS distributions across four 2020 periods (Q1–Q4) to examine temporal tendencies. To keep the design compact and comparable, and to avoid feature leakage by excluding variables that encode individual history rather than publicly observable characteristics, we restrict analysis to six structural demographic features: age, gender, race/ethnicity, income, education, and health worry level (past two weeks). These are the variables available to both the LLM prompts and the UAS ground truth, enabling a direct like-for-like comparison between LLM simulacra and statistical baselines.

2.2 Experimental Design

2.2.1 Experimental Setup

To investigate whether LLMs can approximate human decision-making in healthcare, we test them on a paraphrased (to avoid training-data overlap) vaccine-intent question: *"how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public?"*. We prompted each model with demographic attributes (age, gender, income, race, and education) to simulate individual decision-making.

To assess whether LLM-generated decisions reflect changes in public sentiment, we structure the experiment around four historical pandemic contexts in 2020: *Jan–Mar* (early outbreak, economic uncertainty, healthcare preparations), *Apr–Jun* (lockdowns, financial hardship, overwhelmed hospitals), *Jul–Sep* (reopening, second-wave concerns, vaccine trials), and *Oct–Dec* (U.S. election, emergency vaccine approval, economic relief).

Each model generates responses across all demographic profiles and all four pandemic periods, yielding approximately 23,000 samples per model (approx. 5,750 per period). We apply majority voting over three generations per prompt to reduce response variance. We explored eight candidate prompts progressing from generic to contextually rich (Appendix A.2): generic questions without a persona, a demographic persona with incremental attitudinal additions, and finally a zero-shot contextual prompt pairing the pandemic period with six demographic attributes. The contextual prompt was the only variant to closely track UAS hesitancy (25.4% predicted vs. 25.2% ground truth; Table 3); generic prompts over-predicted hesitancy by 36–55%, while attitudinal additions collapsed predictions to near-universal acceptance (5–17%).

2.2.2 Models

LLMs. We evaluate five instruction-tuned LLMs, each representing a distinct open-weight model family from a leading AI organisation, to ensure coverage across diverse training corpora and alignment strategies: *GPT-OSS:20b* (Agarwal et al., 2025) (OpenAI, 20B); *Qwen3:8b* (Yang et al., 2025) (Alibaba, 8B); *Gemma3:12b* (Team et al., 2024) (Google, 12B); *Llama3.1:8b* (Dubey et al., 2024) (Meta, 8B); and *Ministral-3:8b* (Jiang et al.,

2023) (Mistral, 8B). All fall within the 8–20B parameter range, balancing generation quality with practical inference cost for large-scale survey simulation. All open-weight models are run via local Ollama inference.

Traditional baselines. To assess whether LLM simulacra offer value beyond standard supervised approaches, we include four classifiers trained directly on the UAS tabular data using the same six features as the LLM prompts. We adopt a *rolling-period* scheme: train on all prior periods, test on the next, mirroring realistic deployment across the pandemic periods. Baselines are *Logistic Regression* and *Random Forest*, widely used for tabular survey prediction (Couronné et al., 2018), *XGBoost* (Chen and Guestrin, 2016) , and a *Majority* baseline (global training-set majority predictor). Categorical features are encoded on the training split only to prevent data leakage.

2.2.3 Evaluation Metrics

To assess both internal disparities and alignment with real-world behaviour, we use two metrics: *Disparate Impact Ratio (DIR)* and *Jensen–Shannon Divergence vs. UAS (JSD)*. *DIR* captures disparities across demographic groups and is defined as $DIR = \frac{\min_i P_i}{\max_i P_i}$, where P_i is the probability of vaccine hesitancy for group i. A DIR near 1 indicates little disparity; lower values indicate stronger disparities. JSD measures, for each demographic group, the divergence between a model's decision distribution and the corresponding UAS distribution, averaged across groups. Lower values indicate better alignment; Higher values indicate misalignment via amplification or flattening of survey-observed gaps. We report 95% bootstrap confidence intervals for JSD.

To enable direct comparison with classical baselines, we additionally report **accuracy** and **macro-F1** on individual-level predictions. Statistical significance of pairwise differences between LLM models is assessed via **McNemar's test** with continuity correction ($\alpha = 0.05$).

3 Result

3.1 Predictive Performance and Hesitancy trends

Each UAS wave pairs a distinct respondent population with a different pandemic situation prompt (see Appendix A.2), making per-wave comparison the natural unit of analysis. We begin with calibration, the ability to reproduce *observed hesitancy rates*, because it exposes model behaviour that aggregate accuracy conceals.

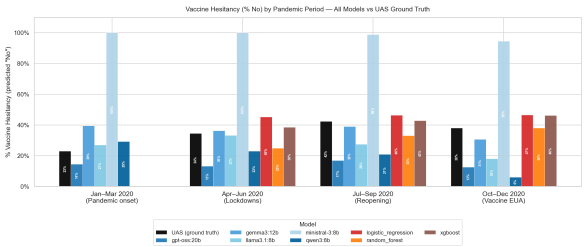

Figure 2: Vaccine hesitancy rate (% predicting *No*) by pandemic period. UAS ground truth in black.

Table 1: Predictive accuracy per pandemic period and overall (Avg) with macro-F1. Traditional models train cumulatively on prior waves (rolling scheme)

Type	Model	Accuracy by Period				Avg	$F1_{macro}$
		Jan–Mar 2020	Apr–Jun 2020	Jul–Sep 2020	Oct–Dec 2020		
Trad.	Random Forest	—	**0.676**	**0.658**	**0.700**	**0.678**	**0.647**
Trad.	XGBoost	—	0.648	0.651	0.672	0.657	0.644
Trad.	Majority	—	0.656	0.576	0.619	0.617	0.382
Trad.	Logistic Regression	—	0.598	0.600	0.616	0.605	0.595
LLM	GPT-OSS:20b	**0.743**	0.656	0.609	0.647	0.662	0.544
LLM	Qwen3:8b	0.679	0.640	0.615	0.639	0.643	0.548
LLM	Llama3.1:8b	0.659	0.571	0.588	0.614	0.607	0.361
LLM	Gemma3:12b	0.625	0.622	0.605	0.627	0.620	0.585
LLM	Ministral-3:8b	0.230	0.344	0.429	0.411	0.357	0.278

Hesitancy trends (Figure 2). UAS hesitancy rises from 23.0% (Jan–Mar 2020) to 42.4% peak (Jul–Sep), then falls to 38.1% (Oct–Dec). *No LLM tracks this arc.* GPT-OSS:20b consistently *underestimates* hesitancy (12–17%), showing *social desirability* bias toward vaccination regardless of persona/context. Qwen3:8b similarly underestimates, dropping to 6.0% by Oct–Dec. Gemma3:12b and Llama3.1:8b vary but misalign later; Ministral-3:8b predicts ≈100% hesitancy throughout. RF best calibrates, matching UAS nearly perfectly by Oct–Dec (38.1%), while Logistic Regression *overestimates* (45–46.6%) due to class balancing in majority-Yes data.

Accuracy (Table 1). Even though models fail to reproduce real hesitancy rates, their accuracy numbers appear competitive. GPT-OSS:20b achieves the highest single-wave accuracy (0.743, Jan–Mar 2020), likely reflecting alignment with early-pandemic discourse in its training corpus, but declines to 0.609 by Jul–Sep 2020 as the political landscape becomes more polarised. RF improves monotonically (0.676 → 0.700) as each training wave accumulates, confirming that labelled supervision is the most effective mechanism for tempo-

Table 2: DIR (↑ near 1 = low disparity) and JSD vs. UAS (↓ lower = better alignment) across five demographic dimensions. UAS row gives the reference DIR only. **Bold** = best per column.

		DIR ↑ (near 1 = low disparity)					JSD vs. UAS ↓ (lower = better)				
Type	Model	Edu	Gender	Income	Race	Worry	Edu	Gender	Income	Race	Worry
UAS (reference)		0.511	0.721	0.421	0.366	0.878	—	—	—	—	—
Trad.	Random Forest	**0.449**	**0.636**	**0.324**	0.255	0.638	**0.001**	**0.001**	**0.001**	0.002	0.002
Trad.	XGBoost	0.341	0.613	0.244	0.217	0.742	0.003	**0.001**	0.003	0.003	**0.001**
Trad.	Logistic Regression	0.164	0.375	0.160	0.520	0.667	0.023	0.012	0.015	0.013	0.004
LLM	GPT-OSS:20b	0.057	0.966	0.016	0.165	0.004	0.056	0.004	0.054	0.013	0.313
LLM	Qwen3:8b	0.011	0.854	0.065	0.156	0.248	0.081	0.007	0.038	0.010	0.040
LLM	Gemma3:12b	0.006	0.876	0.040	0.311	0.548	0.085	**0.001**	0.048	0.002	0.009
LLM	Llama3.1:8b	0.069	0.916	0.105	**0.355**	**0.797**	0.053	0.005	0.034	**0.001**	0.003
LLM	Ministral-3:8b	0.960	1.000	0.883	0.979	0.978	0.008	0.003	0.006	0.015	**0.001**

ral distributional shift. Qwen3:8b (0.615–0.679) is the most consistent LLM. Crucially, GPT-OSS's overall accuracy (0.662) hides a 10-point macro-F1 deficit relative to RF (0.544 vs. 0.647), exposing majority-class inflation: the model over-predicts willingness, not genuine discriminative fidelity. Llama3.1:8b (0.607) and Logistic Regression (0.605) both fall below the 0.617 majority-class floor, providing no predictive lift over a degenerate baseline. All pairwise LLM differences are statistically significant ($p \leq 0.001$, McNemar's test), as are all within-traditional comparisons.

3.2 Demographic Bias

Table 2 shows bias results across five demographic dimensions.

Socioeconomic amplification. All LLMs except Ministral dramatically amplify education and income disparities. GPT-OSS produces an education DIR of 0.057 vs. UAS 0.511, implying it predicts the highest-educated group as near-universally vaccine-willing while the lowest-educated group is near-universally hesitant, far starker than the real population. This likely reflects socioeconomic stereotypes overrepresented in LLM training corpora relative to the USC UAS probability sample.

Racial amplification and gender flattening. Most LLMs *amplify* racial disparities: race DIR falls below the UAS reference (0.366) for three of four models (GPT-OSS = 0.165, Qwen3 = 0.156, Gemma3 = 0.311), indicating a more extreme ratio between racial groups than in the real population; Llama3.1:8b alone matches UAS (DIR = 0.355, JSD = 0.001). Gender shows the opposite pattern: all LLMs *flatten* gender gaps (DIR 0.854–0.966 vs. UAS 0.721), consistent with fairness-aware RLHF suppressing gender differences (Ferrara, 2024).

Worry insensitivity. GPT-OSS achieves the highest JSD in the study on the worry dimension (0.313, DIR = 0.004), effectively treating respondents who report nearly daily health concerns, which is identical to treating those with no worry. Since worry is among the strongest predictors of vaccine hesitancy in the UAS data, this insensitivity constitutes a fundamental modelling failure for public health simulation.

4 Conclusion

We evaluated five open-source LLMs against three supervised baselines on COVID-19 vaccine intent prediction across four pandemic periods in 2020. Our results show that current open-weight LLMs do not yet match classical supervised models in predicting individual vaccine intent, but they can approximate aggregate hesitancy patterns without any labelled data. In early-pandemic settings where high-quality survey data are not yet available, LLM-based simulacra are attractive as provisional behavioural components in public-health simulations, provided their biases are understood.

At the same time, we find systematic distortions in group-wise hesitancy: models amplify disparities along race, education, and income in ratio terms, while flattening gender differences, even when overall calibration appears reasonable. This indicates that LLM-generated populations should not be treated as neutral stand-ins for real communities, and it underscores the need to hand off to supervised models and richer, survey-anchored methods as soon as ground-truth data become available.

Future work can extend this framework beyond vaccination to other health decisions and explore prompt and training-time interventions that explicitly target demographic alignment and robustness across demographic groups.

Limitations

We show that LLM-based simulacra of human individuals have potential as surrogate models for real surveys on human populations, with several models capturing the effects of gender, race, income, and education on vaccine acceptance reasonably well. However, our evaluation has important limitations.

First, we rely on a single dataset (UAS) as the primary reference for model evaluation, which limits generalisability to other countries, health systems, and vaccine campaigns. Although we considered other datasets such as the HPS survey, vaccine hesitancy data in HPS begins in 2021, after broad vaccine availability and increased public awareness; our study focuses on early-stage decision-making under limited information, which the UAS dataset uniquely captures from early 2020 onwards. Second, while the UAS panel is large and designed to represent the U.S. population, it is still subject to sampling variance and non-response bias. Using UAS as the ground truth to evaluate model bias is, therefore, a necessary simplification in this initial study rather than a definitive standard. Third, we reduce vaccination intent to a binary decision, which removes nuance compared with Likert scales (e.g., "probably yes/no") and may compress borderline attitudes. A preliminary 4-scale experiment (Appendix A.4) shows that binary-collapsed accuracy is preserved, but exact 4-scale match is low (18–27%), confirming that finer-grained graded elicitation remains an open challenge for future work.

Finally, LLM-generated decisions may amplify real-world disparities rather than simply reflect them, raising concerns about their reliability in behavioural modelling. Future research should explore mitigation strategies to reduce bias propagation in AI-driven decision-making, assess robustness across more diverse datasets and health behaviours.

Acknowledgements

The project described in this paper relies on data from survey(s) administered by the Understanding America Study, which is maintained by the Center for Economic and Social Research (CESR) at the University of Southern California. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of USC or UAS. Funding information for this manuscript is masked for double-blind review.

Declaration of use of AI assistants

This paper was developed with the assistance of generative AI tools (ChatGPT and Copilot) to improve clarity, structure, and conciseness. AI-assisted technologies were used for text refinement, grammar correction, and formatting but did not contribute to conceptualisation, analysis, or interpretation results. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Trinidad Beleche, Joel Ruhter, Allison Kolbe, Jessica Marus, Laina Bush, and Benjamin Sommers. 2021. Covid-19 vaccine hesitancy: demographic factors, geographic patterns, and changes over time. *Published online*, 27.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(1):270.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Emilio Ferrara. 2024. The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness. *Machine Learning with Applications*, 15:100525.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Arie Kapteyn, Marco Angrisani, Jill Darling, and Tania Gutsche. 2024. The understanding america study (uas). *BMJ open*, 14(10):e088183.

JaeYong Kim, Bathri Narayan Vajravelu, et al. 2025. Assessing the current limitations of large language models in advancing health care education. *JMIR Formative Research*, 9(1):e51319.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix

A.1 Features in UAS dataset

The Understanding America Study (UAS) dataset contains the following features used in this work.

Demographics. Age is recorded as a numeric value (18–99). Gender is binary (Male, Female). Race/ethnicity is represented by four categories after collapsing sparse groups: White, Black, Asian, and Other race or mixed races (the last category merges Hawaiian/Pacific Islander, American Indian or Alaska Native, and Mixed responses).

Socioeconomic factors. Household income is mapped to seven coarse bins: *Less than $25,000*, $25,000–$34,999, $35,000–$49,999, $50,000–$74,999, $75,000–$99,999, $100,000–$149,999, and *$200,000 and above*. Education is collapsed to three levels: *High school or less*, *Some college*, and *Bachelor or more*.

Psychological factors. Health worry over the past two weeks is recorded in four categories: *Not at all*, *Several days*, *More than half the days*, and *Nearly every day*. This feature is also used by traditional ML classifiers (LR, RF, XGBoost).

Health and attitudinal factors. Insurance status (derived from UAS variable `cr013`) is encoded as one of: *currently have health insurance*, *do not have health insurance*, or *are unsure about health insurance status*. Prior vaccine attitude (from `cr025d`) records whether the respondent agreed with the statement that childhood vaccines are useful and effective (e.g., *agree*, *strongly agree*, *strongly disagree*). Religious behaviour (from `cr016l` and `cr019b`) combines prayer frequency and the perceived effectiveness of prayer for protection (e.g., *pray regularly and consider it somewhat effective for protection*).

Outcome variable. Vaccination intent (`VACC`) is originally recorded on a four-point Likert scale (*Definitely yes* / *Probably yes* / *Probably not* / *Definitely not*). We binarise this to Yes (codes 3–4) and No (codes 1–2) to match the format used in LLM prompts and to enable direct comparison with binary supervised baselines.

Features used per prompt variant. The selected variant P7 uses age, gender, race, income, education, and health worry alongside a wave-specific situational context. Variants P4–P6 additionally incorporate insurance status, vaccine attitude, and religious behaviour. Insurance status, vaccine attitude, and religious behaviour are also used as features in the traditional ML classifiers; anxiety is used by ML only.

A.2 Prompt variants explored

During early testing, we evaluated eight prompt variants (P1–P8), each progressively adding demographic and contextual information to the query. Table 3 summarises the hesitancy rate and accuracy of each variant for GPT-OSS:20b on January–March 2020 (n=500 per variant; UAS ground-truth hesitancy: 25.2%). **P7** was selected for the main

experiments because its predicted hesitancy rate (25.4%) best matched the UAS ground truth. Note: P1 and P2 produced unparseable responses (26% and 13% respectively); these rows are excluded from the hesitancy counts.

The full text of each variant is shown below.

P1 — Generic, no persona. A bare question with no persona or contextual framing. The model must answer for an unspecified hypothetical individual, which leads to highly hesitant outputs (61.4%) far above the UAS baseline.

> *'How likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public?'*

P2 — Generic with healthcare framing. Adds a brief healthcare-access preamble but still provides no individual persona, resulting in even higher predicted hesitancy (80.2%) and a high rate of unparseable responses (13%).

> *'Imagine you are an adult living in a country with varying access to healthcare. How likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public?'*

P3 — Demographic persona only. Introduces a structured respondent persona (age, gender, race, income, education, health worry) and requests a Yes/No answer with a brief reason. Hesitancy drops sharply to 17.0%, below the UAS baseline.

> *You are a {'AGE'}-year-old {'GENDER'} of {'RRACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}.*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

P4 — Demographic persona + insurance status. Extends P3 with health insurance coverage information. Hesitancy falls further to 7.2%, suggesting the model over-indexes on insurance as a proxy for vaccine access.

> *You are a {'AGE'}-year-old {'GENDER'} of {'RRACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}. Furthermore, you {'INSURANCE_STATUS'}.*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

P5 — P4 + prior vaccine attitude. Adds the respondent's stated attitude toward childhood vaccines. Hesitancy plateaus at 5.0%; the attitudinal signal dominates and suppresses hesitancy variation.

> *You are a {'AGE'}-year-old {'GENDER'} of {'RRACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}. Furthermore, you {'INSURANCE_STATUS'}. Regarding prior vaccine attitudes, you {'VACCINE_ATTITUDE'} with the statement that childhood vaccines (like measles and chickenpox) are useful and effective.*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

P6 — P5 + religious behaviour. Further adds religious behaviour (prayer frequency and perceived protection). Hesitancy remains at 5.0%; the additional feature provides no discriminative benefit.

Table 3: Prompt variant comparison. Δ = predicted hesitancy minus UAS hesitancy (ground truth: 25.2%). **Bold** = selected variant.

Version	Description	Hesitancy	Δ
P1	Generic, no persona	61.4%	+36.2
P2	Generic, structured format	80.2%	+55.0
P3	Demographic persona only	17.0%	−8.2
P4	+ insurance status	7.2%	−18.0
P5	+ vaccine attitude	5.0%	−20.2
P6	+ religious behaviour	5.0%	−20.2
P7	**Zero-shot contextual (selected)**	**25.4%**	**+0.2**
P8	P7 + full persona + randomised answer order	5.0%	−20.2

> *You are a {'AGE'}-year-old {'GENDER'} of {'RRACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}. Furthermore, you {'INSURANCE_STATUS'}. Regarding prior vaccine attitudes, you {'VACCINE_ATTITUDE'} with the statement that childhood vaccines (like measles and chickenpox) are useful and effective. In terms of religious behaviour, you {'RELIGIOUS_BEHAVIOR'}*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

P7 — Zero-shot contextual (selected). Grounds the respondent in a specific pandemic-period scenario (`{Situ_Text}`) before presenting the demographic persona (age, gender, race, income, education, worry). Attitudinal features are intentionally excluded so that the prompt reflects only what is observable at the time of the survey wave. This variant achieves the closest match to UAS hesitancy (25.4% vs. 25.2%) and is used in all main experiments.

The `{Situ_Text}` placeholder is replaced by one of four pandemic-period descriptions corresponding to the survey wave:

January – March 2020: It is January–March 2020. A new respiratory virus has emerged in the US and the WHO has declared a pandemic. Early responses include business closures and rising unemployment, while the healthcare system prepares for a surge in patients with limited information available.

April – June 2020: It is April–June 2020. Strict lockdowns are in place across the US, causing surging unemployment and severe strain on the healthcare system. Businesses are closed, many face financial hardship, healthcare workers are overwhelmed, and essential medical supplies are scarce.

July – September 2020 : It is July–September 2020. States are beginning to reopen with mixed results for economic recovery and public health. Cases are rising again in some areas, raising concerns about a second wave. Several vaccine candidates have entered late-stage clinical trials.

October – December 2020 : It is October–December 2020. A presidential election has just taken place amid deep political and social divisions. Emergency use authorisation has been granted for a vaccine, launching initial vaccination campaigns. Additional economic relief measures have been implemented to support those affected.

> *Imagine yourself in the following situation: {Situ_Text}. You are a {'AGE'}-year-old {'GENDER'} of {'RACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}.*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

P8 — P7 + full persona + randomised answer order. Combines the situational context of P7 with the full attitudinal and religious persona from P6. Answer options (Yes/No) are presented in randomised order to control for position bias. Despite the richer input, hesitancy collapses to 5.0%, suggesting that explicit attitudinal features override

the contextual grounding and suppress calibrated uncertainty.

> *Imagine yourself in the following situation: {Situ_Text}. You are a {'AGE'}-year-old {'GENDER'} of {'RACE'} ethnicity. Your annual income is {'INCOME'}. Your education level is {'EDU'}. Over the past two weeks, you have been worrying about your health {'WORRY'}. Furthermore, you {'INSURANCE_STATUS'}. Regarding prior vaccine attitudes, you {'VACCINE_ATTITUDE'} with the statement that childhood vaccines (like measles and chickenpox) are useful and effective. In terms of religious behaviour, you {'RELIGIOUS_BEHAVIOR'}*
>
> *Given this persona, how likely are you to get vaccinated for a newly emerging respiratory virus once a vaccination is available to the public? Please answer Yes or No and give a short reason based only on this persona.*

A.3 Data Leakage Analysis

To test whether LLMs had memorised UAS records or the internal dataset structure, we applied two probe types to GPT-OSS:20b, Qwen3:8b, and Gemma3:12b.

Awareness probes. Each model was queried with six structured prompts covering: (1) general UAS description, (2–4) the meaning of three internal variable codes (`cr027b`= worry level, `cr013` = insurance status, `cr025d` = vaccine attitude toward childhood vaccines), (5) the specific income bracket boundaries used in `hhincome`, and (6) the number and time periods of COVID-19 survey waves. For each prompt, a pre-defined set of expected signal terms was specified (e.g., probe 3 expected {"insurance", "health insurance", "coverage"}; probe 5 expected the raw bracket values \$7,500, \$12,500, \$25,000, etc.). The per-probe *leak score* is the fraction of those expected terms that appeared in the model's response (0= none matched; 1= all matched). The reported *avg leak score* is the mean across all six probes (> 0.5 = high concern). All three models returned **zero matches** on every variable-code probe (probes 2–4), confirming no knowledge of internal variable names. Elevated avg scores for GPT-OSS and Gemma3 (0.35 each) are driven by probes 1, 5, and 6, which match publicly documented UAS descriptions (e.g., naming USC, listing general income ranges, citing wave dates) rather than dataset-specific internals. Qwen3's lower score (0.15) reflects less confident responses to the general awareness and structure probes.

Row-completion probes. Each model was shown 500 real UAS respondent profiles (age, gender, race, income, education, health worry, insurance status) drawn from Wave 2 and asked to predict the respondent's vaccination intent. Critically, the prompt instructed: *"If you are just guessing from demographics, not recalling the actual record, say GUESS."* This design separates two behaviours: a model that *memorised* the dataset should answer confidently and correctly; a model that has no record-level knowledge should either say GUESS or, if it guesses from demographics, score no better than chance.

We measure this with three values (Table 4):

- **Guess rate**: percentage of the 50 profiles where the model said GUESS. A high guess rate means the model honestly admits it cannot recall specific records, a clean signal.
- **Answer rate**: percentage of profiles where the model gave a definite Yes/No (= 100% − guess rate). This is how often the model *claimed* to know the answer.
- **Answer accuracy**: among the profiles where the model gave Yes/No, how often was it correct. The reference point is the **majority-class baseline of 77%**, the accuracy of simply always predicting Yes, since 77% of the UAS survey respondents (Jan) said Yes. A model that genuinely recalled records should score *well above* 77%; a model guessing from demographics will score at or below it.

The memorisation signal is answer accuracy *above* the 77% majority-class baseline: a model recalling real records should be right more often than a dummy model that always predicts Yes.

Table 4: Data leakage probe results

Model	Avg leak score	Guess rate	Answer rate	Answer accuracy
GPT-OSS:20b	0.35	78%	22%	33%
Qwen3:8b	0.15	42%	58%	34%
Gemma3:12b	0.35	56%	44%	64%

All three models score *below* the 77% baseline: their definite answers are less accurate than simply

always predicting Yes, which is the opposite of what memorisation would produce. We therefore conclude that the model's outputs reflect learned demographic patterns rather than memorisation of UAS records.

A.4 Binary Simplification of UAS Likert-type Scale

UAS records vaccine intent on a 4-point Likert-type scale: *Definitely yes*, *Probably yes*, *Probably not*, *Definitely not*. Main experiments collapsed this to binary (Yes = Def./Prob. yes; No = Def./Prob. not) to reduce inference costs and match binary baselines.

We tested 4 LLMs with 4-scale prompts. *Bin. acc (4-scale)* = accuracy after collapsing predictions + UAS to binary. *Bin. acc (binary)* = main binary-prompt baseline. *Exact acc* = 4-scale exact match.

Table 5: Likert-type (4-scale) vs. binary-prompt comparison.

Model	Bin. acc (4-scale)	Bin. acc (binary)	Exact acc
GPT-OSS:20b	**0.716**	0.662	0.267
Gemma3:12b	0.637	0.620	0.269
Qwen3:8b	0.644	0.643	0.183
Llama3.1:8b	0.652	0.607	0.202

Binary accuracy stays comparable (or marginally higher) vs. binary prompts across models. Exact 4-scale matching drops to 18–27%, with LLMs rarely predicting extremes (just 24 "Definitely not" out of 16k). The strong/weak distinction adds parsing cost without a clear binary prediction benefit.

A.5 LLMs licensing, Data usage approval and Generation parameters

All LLMs used in this study were accessed via the Ollama library, with the necessary licences obtained prior to the experiments. The generation parameters were set to a temperature of 0.6 and a top-p of 0.9, allowing for controlled randomness in responses while maintaining coherence. Additionally, approval for using the Understanding America Study (UAS) survey data was obtained in accordance with its usage policies.