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

Anonymous ACL submission

Abstract

Generative agents have been increasingly used to simulate human behaviour in silico, driven by large language models (LLMs). These simulacra serve as sandboxes for studying human behaviour without compromising privacy or safety. However, it remains unclear whether such agents can truly represent real individuals. In this work, we compare survey data from the Understanding America Study (UAS) on healthcare decision-making with simulated responses from generative agents. Using demographic-based prompt engineering, we create digital twins of survey respondents and analyse how well different LLMs reproduce real-world behaviours. Our findings show that some LLMs fail to reflect realistic decision-making, such as predicting universal vaccine acceptance. However, Llama 3 captures variations across race and income more accurately but also introduces biases not present in the UAS data. This study highlights the potential of generative agents for behavioural research while underscoring the risks of bias from both LLMs and prompting strategies.

1 Introduction

The rise of large language models (LLMs) has enabled advances in agentic artificial intelligence (AI), where AI systems can make independent choices and act autonomously (Acharya et al., 2025; Park et al., 2023; Xi et al., 2025; Shanahan et al., 2023). Generative agents, in particular, have been shown to create realistic synthetic human populations, or simulacra, where individual agents follow daily life patterns and interact with each other (Park et al., 2023; Shanahan et al., 2023; Han et al., 2024; Wang et al., 2024). These simulacra offer a promising approach to studying human behaviour in silico, raising the question of whether they can effectively model complex decision-making in real-world scenarios. In healthcare, where decisions are shaped by personal, social, and policy factors, the ability of simulacra to approximate human choices has significant implications. If LLMs can reliably simulate decision-making, they could serve as valuable tools for policy analysis, health behaviour prediction, and intervention design. However, their accuracy and potential biases when applied to real-world data require careful evaluation, particularly as LLMs have been shown to amplify racial biases in healthcare applications (Ferrara, 2024).

A key challenge in using LLMs for healthcare decision modelling is determining whether they effectively replicate factors shaping real-world health decisions. Unlike clinical diagnosis, which follows medical guidelines, social, economic, and behavioural influences shape choices such as seeking treatment or vaccination. While surveys provide structured insights into human intentions, LLMs offer a scalable alternative for modelling decisions in agent-based simulations. However, their ability to generate realistic health choices remains uncertain.

To investigate this, we compare health decision-making in a disease simulation framework, focusing on vaccination as a case study. Individuals make choices based on varying levels of contextual information, including personal risk perception, demographics, and external messaging. We compare LLM-generated vaccine decisions to survey data from the Understanding America Study (UAS) (Kapteyn et al., 2024), which includes socioeconomic, risk perception, and personal belief data. This enables the assessment of LLMs' alignment with human decision patterns and potential biases diverging from real-world behaviours.

Despite this potential, several challenges remain. First, while LLMs generate human-like responses, it is unclear whether they truly capture the reasoning behind health-related decisions. Prior research suggests LLMs can retrieve medical knowledge, but their ability to simulate human decision processes is still in question (Hager et al., 2024). Since

vaccination intentions are shaped by social and psychological factors, it is critical to assess whether LLMs accurately model these influences or merely reflect statistical patterns from their training data.

LLM-generated decisions may exhibit biases that diverge from human decision-making, raising concerns about their reliability in public health modelling. Biases in LLM training data can create demographic disparities (Kim et al., 2025), making assessing them against actual human decisions essential. This study explores: **RQ1:** Can LLMs effectively model healthcare decisions, such as vaccination intentions?; **RQ2:** What biases emerge in LLM-generated decisions across demographic groups, and how do models distribute decisions among populations?

We hypothesise that LLMs can approximate human decision-making, but their effectiveness depends on the amount and type of contextual information provided (H1). Additionally, pretraining data and prompt formulation may cause LLMs to exhibit biases that differ from human biases (H2).

This study tests these hypotheses through a structured experiment. We compare LLM-generated vaccination decisions with survey responses to assess alignment and examine biases by analysing disparities across demographic groups. Our findings enhance an understanding of LLMs' strengths and limitations in modelling healthcare behaviours and decision-making.

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). LLMs are prompted with structured demographic profiles under various pandemic scenarios, and their responses are analysed to assess decision patterns and potential biases.

We evaluate vaccination decisions by testing models across four historical pandemic contexts from 2020. Each model is presented with a standardised decision-making prompt, incorporating demographic details and situational factors. Model predictions are then compared to UAS survey data to assess alignment with real-world trends across different pandemic phases. To analyse biases, we examine disparities in LLM-generated decisions within each demographic category (e.g., variations in vaccine hesitancy across racial or income groups). Instead of benchmarking these disparities against survey data, we assess internal inconsistencies, identifying whether models treat similar demographic profiles differently.

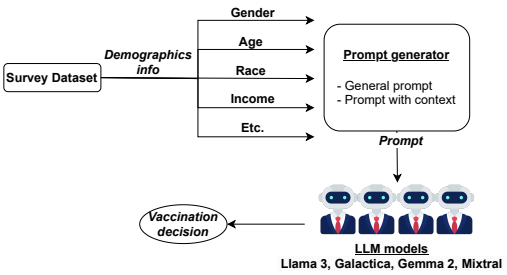

Figure 1: Overview of the experimental setup

This approach assesses LLMs' reliability in healthcare decisions and highlights potential biases from demographic variations in responses.

2.1 Dataset

Our study utilises data from the Understanding America Study's Coronavirus in America survey (Kapteyn et al., 2024), which tracks U.S. attitudes, health behaviours, and policy responses to COVID-19. We analyse data from the national long-form questionnaire, focusing on survey waves from March 2020 to January 2021. The initial survey (Wave 1) launched on March 10, 2020, followed by bi-weekly tracking surveys (Waves 2–21) to capture shifting public sentiment. By restricting our analysis to this period, we examine decision-making patterns before and during early vaccine distribution. This dataset helps assess how demographics influenced vaccination intentions and preventive behaviours in the pandemic's initial stages (see Appendix A.1).

2.2 Experimental Design

2.2.1 Experimental Setup

To investigate whether LLMs can approximate human decision-making in healthcare, we test LLMs on the question: *"How likely are you to get vaccinated for coronavirus once a vaccination is available to the public?"*. Each model is prompted with demographic attributes (age, gender, income, race, education, and worry level) to simulate individual decision-making. The responses are compared to real-world survey data from the UAS to evaluate alignment and detect biases in predictions.

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).

We assess each LLM to see how contextual variations influence decision-making. We also analyse bias to identify disparities in LLM responses regarding vaccine hesitancy and whether demographic details mitigate biases. Each model generates 11.5k samples covering all demographic profiles through four pandemic phases. To enhance robustness and minimise variability, we run each sample three times and utilise majority voting for the final outcome, ensuring stable predictions. By comparing LLM predictions with survey data, we assess generative models' strengths and limitations while exploring potential biases that may impact their use in policy and healthcare.

2.2.2 Model Selection and Specifications

We evaluate four instruction-tuned LLMs with diverse architectures to compare their decision-making in healthcare contexts: *Meta Llama-3-8B-Instruct* (Dubey et al., 2024), optimized for instruction-following with reinforcement learning from human feedback (RLHF); *Google Gemma-2-9B-IT* (Team et al., 2024), designed for improved generalization and contextual understanding; *Galactica-6.7B-Evol-Instruct* (Taylor et al., 2022), fine-tuned for structured instruction-following and domain-specific knowledge; and *Mistralai Mistral-8B-Instruct* (Jiang et al., 2023), known for balancing efficiency and reasoning performance.

These models were selected based on their diverse architectures and training methodologies, allowing us to examine how different LLM families handle structured prompts and demographic attributes when predicting vaccination decisions.

2.2.3 Evaluation Metrics

We employ two key metrics, the Disparate Impact Ratio and Jensen-Shannon Divergence, to assess bias and alignment in LLM-generated decisions.

Disparate Impact Ratio (DIR) measures disparities in decision distributions across demographic groups (Feldman et al., 2015). It is defined as: $DIR = \frac{\min(P_i)}{\max(P_i)}$, where P_i represents the probability of vaccine acceptance for each demographic category. When multiple categories exist, we compute the ratio of the best to worst outcomes to identify the largest bias. A DIR near 1 suggests fair treatment, while lower values indicate significant disparities.

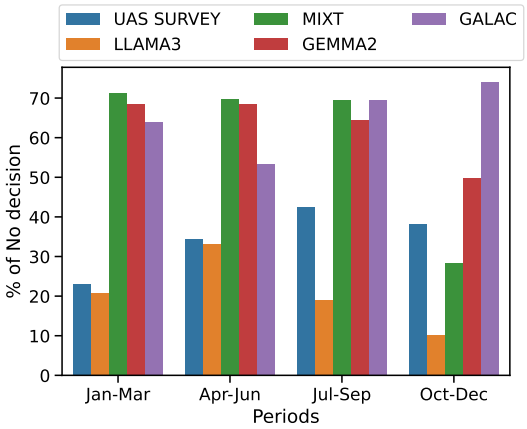

Figure 2: Comparison of survey and LLMs decision outputs 4 different situations

Jensen-Shannon Divergence (JSD) quantifies differences in decision distributions (Lin, 1991) within LLM-generated outputs across demographic groups (e.g., Male vs. Female, White vs. Black vs. Asian). A higher JSD value indicates greater inconsistencies in decision patterns, suggesting potential demographic biases in the model's decision-making.

3 Result

3.1 Comparison of LLM predictions with UAS survey data

Our study examines vaccination intentions using four LLMs, prompting them with pandemic-phase-specific contexts and demographic profiles. The structured prompts cover four COVID-19 phases in the U.S. (Jan–Dec 2020), which allows us to assess how LLMs simulate decision-making trends compared to UAS survey data.

In Jan–Mar 2020, early uncertainty led to a moderate hesitancy of (≈25%) in the UAS survey. Llama3 closely matched this, while Mistral and Galactica overestimated hesitancy nearly threefold. By Apr-Jun 2020, as the lockdowns and economic strain intensified, hesitancy increased slightly. Llama3 remained the closest match, while Gemma2 and Galactica still overestimated hesitancy but adjusted slightly.

During Jul–Sep 2020, vaccine trials progressed, but concerns over a second wave grew. Hesitancy exceeded 40% in the UAS data. Gemma2 aligned well, while Llama3 underestimated hesitancy, suggesting it assumed vaccine acceptance earlier than

observed. Mistral and Galactica continued to overestimate scepticism. In Oct-Dec 2020, as vaccines gained emergency approval, hesitancy remained just under 40% in the UAS data. Mistral closely matched, while Galactica and Gemma2 overestimated hesitancy and Llama3 again underestimated, indicating a bias toward optimism.

These findings highlight LLM tendencies: Llama3 assumes early acceptance, while Mistral and Galactica persistently overestimate scepticism, even as vaccine availability improves. This suggests LLMs interpret decision-making differently, with some over-representing early fears and others assuming a more rational acceptance curve. Understanding these biases is essential for evaluating LLMs' reliability in public health modelling.

3.2 Bias analysis in LLM-generated decisions

Table 1: Disparate Impact Ratio (DIR) and Jensen-Shannon Divergence (JSD) across models. Values are shown as DIR / JSD

	Llama3		Gemma2		Mistral		Galactica	
Feature Set	DIR	JSD	DIR	JSD	DIR	JSD	DIR	JSD
Gender	0.918	0.004	0.853	0.002	0.933	0.000	0.974	0.0001
Race	0.942	0.001	0.602	0.009	0.864	0.001	0.973	0.0000
Income	0.624	0.019	0.236	0.035	0.404	0.013	0.889	0.0001
Education	0.614	0.066	0.061	0.140	0.349	0.045	0.941	0.0020

Our results reveal significant biases in LLM-generated vaccine decisions across income, education, and race, as shown in Table 1. Gemma2 and Mistral exhibit the most pronounced disparities, with low DIR for income (0.236 and 0.404) and education (0.061 and 0.349). Their high JSD scores indicate substantial deviations from real-world trends. Prior studies confirm that income and education strongly influence vaccine hesitancy, with lower-income and less-educated individuals being more reluctant (Aw et al., 2021; Allen et al., 2021). Galactica maintains the most balanced predictions across demographic groups, while Llama3 performs moderately well but tends to underestimate hesitancy.

To illustrate these disparities, Figure 3 presents the racial distribution of "No" decisions across models. Gemma2 shows the highest hesitancy rate for Black respondents (above 70%), deviating significantly from UAS data. Llama3, in contrast, exhibits the lowest hesitancy rates, aligning more closely with UAS trends. These findings suggest that some models may amplify existing biases, overestimating vaccine scepticism among certain populations.

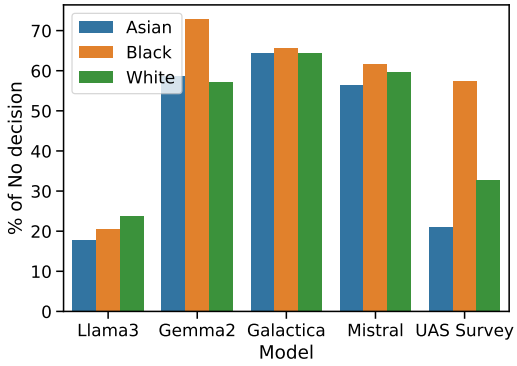

Figure 3: Racial bias in LLM-generated vaccine decisions

These results highlight the need for bias-aware evaluation in LLM-driven decision modelling, as disparities in model-generated outcomes may reinforce real-world inequities. Future work should explore mitigation strategies, including refining prompts, improving training data representation, and integrating fairness-aware techniques.

4 Conclusion

This study evaluates how LLMs simulate vaccine decision-making across different phases of the COVID-19 pandemic, examining biases in model-generated responses. Our findings reveal distinct disparities: Llama3 aligns well with early trends but underestimates scepticism in later phases, while Mistral and Galactica consistently overestimate hesitancy. Gemma2 exhibits the most significant demographic disparities, particularly across income and education, where lower-income and less-educated groups show higher hesitancy—trends also observed in real-world survey data. By analysing bias through DIR and JSD, we show how LLMs reflect and potentially reinforce demographic disparities rather than model decision-making equally.

This work contributes by quantifying LLM biases in vaccine decision modelling and demonstrating how disparities vary across demographic groups. Using DIR and JSD, we provide a structured approach to assessing bias in AI-generated decisions. Our findings highlight the importance of evaluating and mitigating demographic biases in LLM-based public health applications. Future work should explore how adjustments in training data, prompt design, and bias-mitigation techniques can improve fairness and reliability in behavioural modelling.

Limitations

We show that LLM-based simulacra of human individuals show the potential to be used as a surrogate model for real surveys on human populations, specifically with LLama3 and Galactica capturing the effect of gender, racial, income, and education in vaccine acceptance surveys quite well. One limitation of this study is that these LLMs may have seen the UAS Survey Data that we used for our evaluation. The UAS Data is not publicly available and should not have been used for training of these LLMs, but it is possible that this dataset may have been directly used in training, or that scientific results summarising the ACS data may have been used to expose this data indirectly. Another shortcoming of our work is using the UAS data as a ground truth to evaluate the bias of models. While the UAS household panel is quite large, having 14,700 respondents in 2024, representing the entire United States, even such a large sample is still a sample and subject to sampling variance. Additionally, LLM-generated decisions may amplify biases rather than merely 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 and assess model robustness across more diverse datasets.

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

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*.

Jennifer D Allen, Nadia N Abuelezam, Rebecca Rose, and Holly B Fontenot. 2021. Factors associated with the intention to obtain a covid-19 vaccine among a racially/ethnically diverse sample of women in the usa. *Translational behavioral medicine*, 11(3):785–792.

Junjie Aw, Jun Jie Benjamin Seng, Sharna Si Ying Seah, and Lian Leng Low. 2021. Covid-19 vaccine hesitancy—a scoping review of literature in high-income countries. *Vaccines*, 9(8):900.

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.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation. *arXiv preprint arXiv:2407.01093*.

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.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

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.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

A Appendix

A.1 Features in UAS dataset

The Understanding America Study (UAS) dataset features include demographic attributes and behavioural indicators relevant to vaccine decision-making. The demographic attributes consist of gender (Male/Female), age (15–99), and race (White, Black, Asian). Socioeconomic factors include household income, categorised into eight bins from 'Less than $25,000' to '$200,000 and above', and education level, classified as "High school or less", "Some college", and "Bachelor or more". Psychological factors, such as worry levels over the past two weeks, are grouped into four categories: 'Not at all', 'Several days', 'More than half the days', and 'Nearly every day'. The survey also includes vaccination intent, which is recorded as a binary outcome (Yes/No).

For prompt generation, we selected representative values for each attribute: ages (18-99), genders (Male, Female), races (White, Black, Asian), and income levels spanning eight bins. Education was categorized into three levels, and worry levels followed the original survey classification. These structured inputs allowed us to systematically analyse how LLMs simulate vaccine decision-making across different demographic groups.

A.2 Prompt template

Imagine yourself in the following situation: [SITU PROMPT]. Your background and personal circumstances are as follows: [You are a AGE-year-old GENDER of RACE ethnicity, living in a diverse country with varying access to healthcare, differing levels of trust in government and medical institutions, and socioeconomic disparities. Your annual income is INCOME. Your education level is EDU_LEVEL. Over the past two weeks, you have been worrying about your health WORRY_LEVEL]. Please use this persona to answer the question below:

'How likely are you to get vaccinated for coronavirus once a vaccination is available to the public?'

In this context, please answer based on your persona. Answer: [Yes/ No] Short reason: [FILL IN] based on your persona

The [SITU PROMPT] will be replaced by the following contextual prompt from a different period:

January - March 2020: From January to March 2020, COVID-19 emerged in the US, leading to the first reported cases and the declaration of a pandemic by the WHO. The early economic impact included business closures and rising unemployment while the healthcare system began preparing for an influx of patients. Consider the initial response to the virus, the economic impact, and healthcare system preparations.

April - June 2020: During April to June 2020, the US experienced strict lockdown measures, a surge in unemployment, and significant strain on the healthcare system due to COVID-19. Businesses were closed, and many people faced financial hardships. Healthcare workers were overwhelmed, and there were shortages of essential medical supplies. Considering these challenges and public health measures

July - September 2020: From July to September 2020, states in the US began to reopen, leading to mixed responses in terms of economic recovery and public health. Concerns about a second wave of COVID-19 emerged as cases began to rise

again in some areas. Progress was made in vaccine development, with several candidates entering late-stage trials. Include considerations of reopening efforts, second-wave concerns, and progress in vaccine development.

October - December 2020: In the period from October to December 2020, the US presidential election took place, creating significant political and social implications. COVID-19 vaccines received emergency use authorization in December, leading to the beginning of vaccination campaigns. Additional economic relief measures were implemented to support individuals and businesses affected by the pandemic.

A.3 LLMs licensing, Data usage approval and Generation parameters

All LLMs used in this study were accessed through Hugging Face, with the necessary licences acquired before the experiments. The generation parameters were configured with a temperature of 0.6 and a top-p of 0.9, which allowed 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.