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

Anonymous ACL submission

Abstract

Generative agents powered by large language models (LLMs) can simulate human decision-making in complex settings (e.g., healthcare), yet their alignment with real-world patterns remains unclear. We compare LLM-generated vaccination decisions with responses from the Understanding America Study (UAS), using demographic-grounded prompts across multiple pandemic contexts in 2020. We evaluate five open-source models, Llama3, Gemma2, Ministral, Galactica, and GPT-OSS and assess both internal disparities and alignment with survey data. Results show consistent tendencies: GPT-OSS follows UAS trends more closely than the others but still flattens some group gaps; Llama3 tends to underestimate hesitancy; Galactica, Gemma2, and Ministral often overestimate it. We define bias as misalignment with real-world group patterns: some models flatten differences (e.g., race), while others amplify gaps (e.g., income, education). These findings highlight both the potential and limits of LLMs for behavioural simulation and the need for bias- and context-aware evaluation in public-health research.

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 way to study human behaviour without real-world risks. However, it is unclear if they model complex decision-making accurately. 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 must be carefully evaluated, as LLMs can amplify racial biases in healthcare (Ferrara, 2024).

A key challenge in using LLMs for healthcare decision modelling is determining whether they replicate the complex social, economic, and behavioural factors that shape real-world choices, such as seeking treatment or vaccination, unlike clinical diagnosis, which follows guidelines. While surveys provide structured insights, LLMs offer a scalable alternative for agent-based simulations, though their ability to generate realistic health decisions 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 demographic disparities emerge in LLM-generated decisions, and how do these differ from real-world patterns in survey data?

We hypothesise that LLMs can approximate human decisions under early-stage, low-information conditions, but alignment and bias depend on context and model design. Using a temporal, survey-anchored design, we compare model outputs to UAS responses across four 2020 periods and assess bias as misalignment with real group patterns via DIR and JSD relative to UAS. Focusing on five structural demographics enables clear, comparable analysis of model tendencies and limits in healthcare behaviour modelling.

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 acceptance across racial or Income groups). To understand potential biases, we analyse both internal inconsistencies across demographic groups and compare model outputs to real-world survey data (UAS) where available. This helps identify whether LLMs reflect known behavioural differences or exaggerate group-level patterns. This

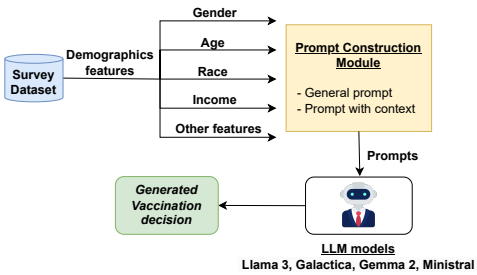

Figure 1: Overview of the experimental setup

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

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, we restrict analysis to five structural demographics (age, gender, race, income, education) and defer proximal attitudinal factors that may mediate hesitancy (e.g., insurance, religion, prior vaccine trust) to future work.

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?"*. We prompted each model with demographic attributes (age, gender, income, race, and education) 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 hos-

pitals), *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 acceptance and whether demographic details mitigate biases. Each model generates 11.5k samples across all demographic profiles and four pandemic phases. We apply majority voting over three generations per prompt to reduce response variance. In designing prompts, we experimented with different formats, including persona-only, location-based, situation-based, and few-shot prompts. However, few-shot prompts often caused the model to repeat or exaggerate specific examples, especially when sensitive attributes such as race or Income were included. To mitigate this risk, we selected a zero-shot prompt with contextual information, which yielded more consistent and balanced results.

2.2.2 Model Selection and Specifications

We evaluate five instruction-tuned open-source LLMs: *Meta Llama-3-8B-Instruct* (Dubey et al., 2024), *Google Gemma-2-9B-IT* (Team et al., 2024), *Mistral AI Ministral-8B-Instruct* (Jiang et al., 2023), *Galactica-6.7B-Evol-Instruct* (Taylor et al., 2022), and *GPT-OSS-20B* (OpenAI et al., 2025). The first three are widely used instruction-tuned models known for their strong performance on general reasoning tasks. Galactica was selected for its domain-specific tuning and prior use in clinical decision-making (Poulain et al., 2024). GPT-OSS provides a high-capacity alternative aligned with recent advancements in LLM development. We selected these models for their architectural diversity, public availability, and suitability for behavioural simulation tasks.

2.2.3 Evaluation Metrics

To assess both internal disparities and alignment with real-world behaviour, we use two metrics: *Disparate Impact Ratio (DIR)* (Feldman et al., 2015) and *Jensen–Shannon Divergence vs. UAS (JSD)* (Lin, 1991). *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 group, the divergence between a model's decision distribution and the corresponding UAS distribution, averaged across demographic groups (e.g., gender, race). Lower values indicate better alignment; higher values indicate misalignment via *amplification* or *flattening* of survey-observed gaps. We also include 95% bootstrap CIs to indicate variability.

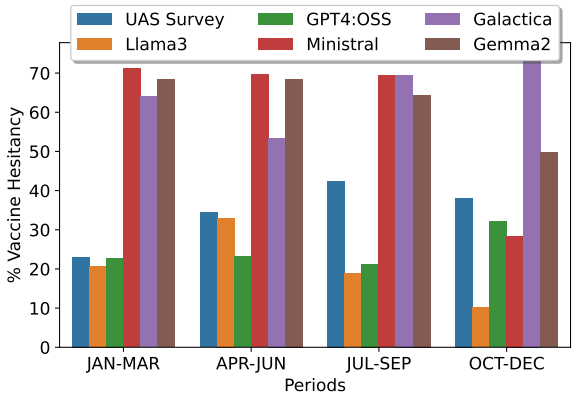

Figure 2: Comparison of survey and LLMs decision outputs 4 different situations

3 Result

3.1 Comparison of LLM predictions with UAS survey data

We compare vaccine hesitancy trends from five LLMs to real-world data from the UAS survey across four COVID-19 phases in 2020 (Fig. 2). Each model received structured demographic profiles with contextual scenarios. Vaccine hesitancy is defined as the percentage of responses where the model (or survey participant) answered that they would not get vaccinated.

Jan-Mar. UAS shows low–moderate hesitancy. GPT-OSS and Llama3 are closest to UAS, while Galactica, Gemma2, and Ministral predict higher hesitancy. **Apr-Jun.** UAS increases. GPT-OSS and Llama3 remain below UAS, whereas Galactica, Gemma2, and Ministral continue to overestimate. The ordering is stable, with Galactica typically highest. **Jul-Sep.** UAS reaches its peak for the year. GPT-OSS and Llama3 underpredict, and Galactica, Gemma2, and Ministral overpredict (remaining above UAS). This period has the largest model spread, indicating stronger model-specific behaviour. **Oct-Dec.** UAS declines. GPT-OSS tracks UAS closely, Llama3 stays low (optimistic), and Galactica, Gemma2, and Ministral stay high (more sceptical), preserving earlier tendencies.

Overall tendencies. GPT-OSS follows UAS most closely across periods. Llama3 tends to be op-

timistic and underestimates hesitancy, while Galactica, Gemma2, and Ministral are more sceptical and often overestimate. These patterns suggest that outputs depend not only on prompts but also on training and alignment. Bias and alignment should therefore be evaluated before using LLMs in public-health simulations.

3.2 Bias analysis in LLM-generated decisions

Table 1: Disparate Impact Ratio (DIR) and Jensen-Shannon Divergence (JSD) with 95% bootstrap confidence intervals (±) across models.

		FEATURE			
MODEL	**METRIC**	GENDER	RACE	INCOME	EDUCATION
Llama3	DIR	0.67 ± 0.03	0.76 ± 0.04	0.15 ± 0.02	0.07 ± 0.01
	JSD	0.14 ± 0.01	0.10 ± 0.01	0.12 ± 0.01	0.22 ± 0.01
Gemma2	DIR	0.90 ± 0.03	0.79 ± 0.03	0.47 ± 0.03	0.30 ± 0.02
	JSD	0.24 ± 0.01	0.21 ± 0.01	0.28 ± 0.01	0.20 ± 0.01
Ministral	DIR	0.95 ± 0.02	0.92 ± 0.04	0.59 ± 0.03	0.51 ± 0.02
	JSD	0.21 ± 0.01	0.22 ± 0.01	0.23 ± 0.01	0.17 ± 0.01
Galactica	DIR	0.99 ± 0.02	0.98 ± 0.03	0.93 ± 0.03	0.95 ± 0.02
	JSD	0.26 ± 0.01	0.26 ± 0.01	0.26 ± 0.01	0.26 ± 0.01
GPT-OSS	DIR	0.92 ± 0.03	0.87 ± 0.03	0.48 ± 0.03	0.76 ± 0.03
	JSD	0.09 ± 0.01	0.08 ± 0.01	0.08 ± 0.01	0.10 ± 0.01
UAS Survey	DIR	0.72 ± 0.03	0.37 ± 0.03	0.42 ± 0.02	0.51 ± 0.03
	JSD	-	-	-	-

We do not assume identical behaviour across groups. We define bias as *misalignment* with real-world group patterns, which may appear as either *amplification* (DIR lower than UAS) or *flattening* (DIR higher than UAS). Table 1 includes UAS as a baseline for comparison. n UAS, disparities are strongest for race (DIR = 0.37), sizable for income (0.42) and education (0.51), and more moderate for gender (0.72). Most models *flatten* race and gender gaps: *Galactica*, *Ministral*, and *GPT-OSS* have DIRs near 0.9–1.0 for these attributes, reducing the disparities seen in UAS (race $= 0.37$, gender $= 0.72$). In contrast, *Llama3 amplifies* inequality in income and education (DIR $= 0.15$ and 0.07 vs. UAS 0.42 and 0.51), and *Gemma2* similarly amplifies education (DIR $= 0.30$). Overall, the open-source models exhibit over-flattening and amplification by feature, with none accurately reproducing UAS disparities.

Figure 3 compares racial patterns in vaccine hesitancy against UAS (leftmost bars), which shows Black > White > Asian. *Llama3* and *GPT-OSS* both underpredict hesitancy for all races, compressing gaps (flattening). *Galactica* and *Ministral* overpredict and produce nearly equal rates across races (flattening at higher levels). *Gemma2* preserves the UAS ordering but amplifies the Black–non-Black gap (Black >70%), indicating disparity amplification.

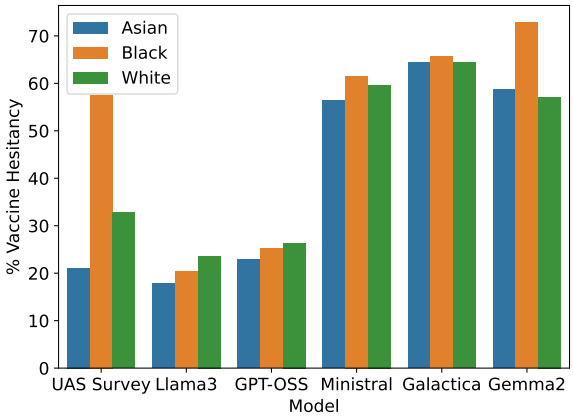

Figure 3: Racial bias in vaccine decisions: LLM outputs vs. survey data

These findings highlight that LLMs can misalign either by amplifying or flattening survey-observed differences, both of which distort behavioural realism and could bias simulation outcomes. Future work should explore mitigation strategies such as refining prompts, improving demographic representation in training data, and adopting fairness-aware modelling techniques.

4 Conclusion

LLMs can simulate decision-making behaviour, but careful evaluation is needed to ensure alignment with real-world patterns. Across periods, models show consistent tendencies: *Llama3* is generally optimistic and underestimates hesitancy, while *Galactica*, *Gemma2*, and *Ministral* tend to overestimate it. *GPT-OSS* follows UAS trends more closely but still flattens group disparities. None of the models reproduces UAS patterns exactly.

We define bias as *misalignment* with survey-observed group patterns, either *amplifying* existing gaps or *flattening* them. Using DIR (within-model disparity) and JSD vs. UAS (divergence from survey distributions), we find that most models flatten race and gender disparities, while some (e.g., *Llama3* for income/education; *Gemma2* for education) amplify them.

Future work will extend to additional datasets and variables, test prompt sensitivity, and include complementary group-level metrics (e.g., Equalised Odds, calibration). We will also evaluate newer/larger LLMs and scenarios less likely to appear in pretraining data to further test generalisation.

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 effects of gender, race, Income, and education in vaccine acceptance surveys quite well. One limitation of this study is that these LLMs may have been exposed to the UAS Survey Data used for evaluation. While the UAS data is not publicly available and should not have been used for model training, it is possible that the dataset or scientific summaries of related sources, such as the ACS, may have been indirectly incorporated during pretraining. Another limitation is the use of a single dataset (UAS) as the primary reference for model evaluation. We acknowledge that this limits generalisability. 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, beginning in early 2020.

In addition, while the UAS panel is large (14,700 respondents in 2024, designed to represent the U.S. population), it is still subject to sampling variance. Another shortcoming of our work is using UAS as the ground truth to evaluate model bias, a necessary simplification in this initial study.Also, Binary decisions reduce nuance relative to Likert scales (e.g., "probably yes/no"). This may compress borderline attitudes. Future work will elicit graded responses and map them to survey Likert scales for finer alignment tests. Finally, we note that LLM-generated decisions may amplify real-world disparities rather than reflect them. This raises 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. A more systematic evaluation of prompt sensitivity and its effect on model outputs is planned for future work, including testing the impact of different prompt styles (e.g., few-shot, simplified persona prompts) across demographic groups.

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*.

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.

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.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*.

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 categorised 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 Prompt variants explored

During early testing, we evaluated several prompt styles to assess how prompt framing affects model behaviour. Below is a list of the main formats considered before selecting the final zero-shot contextual prompt.

A.3.1 Persona-only

> Imagine yourself in the following situation: If you are in the midst of the 2019-2020 COVID-19 outbreak, and there is limited information available. Your background and personal circumstances are as follows: [You are a AGE-year-old GENDER of RACE ethnicity. 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.

A.3.2 Persona-only+ location based prompt

Imagine yourself in the following situation: [If you are in the midst of the 2019-2020 COVID-19 outbreak, and there is limited information available]. Your background and personal circumstances are as follows: [You are a AGE-year-old GENDER of RACE ethnicity. 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]. You live in a diverse country with varying access to healthcare, differing levels of trust in government and medical institutions, and socioeconomic disparities. 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.

A.3.3 Few-shot Prompting

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.

Example 1:
Answer: Yes.
Short reason: I trust the science and want to protect myself and others.

Example 2:
Answer: No.
Short reason: I am concerned about side effects and don't trust the government's messaging right now.

Example 3:
Answer: No.
Short reason: As a 20-year-old living on a limited income of less than $25,000, I'm struggling to make ends meet. The thought of affording the vaccination, even if available to the public, is daunting.

A.4 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.