Human Interaction and Emerging Technologies (IHIET 2025), Vol. 197, 2025, 450–459
https://doi.org/10.54941/ahfe1006738

Evaluating Silicon Sampling: LLM Accuracy in Simulating Public Opinion on Facial Recognition Technology

Charles Ma

Institute of Media and Communications Management, University of St. Gallen (HSG), St. Gallen, 9000, Switzerland

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable capabilities in mimicking human behaviors, leading to growing interest in their potential for survey research through "silicon sampling", a method where LLMs generate responses when prompted with personas. This study evaluates the effectiveness of silicon sampling in emerging technology acceptance by simulating public opinion on facial recognition technology (FRT). I compare LLM-generated responses against actual survey data from 6,076 respondents across China, Germany, the United Kingdom, and the United States. I test three LLMs (GPT-4o, Claude 3.5 Sonnet, and DeepSeek V3) under three prompting conditions: demographic information only, contextual information only, and a combination of both. Performance was assessed using Mean Absolute Error (MAE) and Quadratic Weighted Cohen's Kappa (QWK) metrics. Results demonstrate that demographic-only prompting yields poor simulation accuracy (MAE: 0.90-1.73; QWK: near zero), while incorporating contextual information about FRT experiences and perceptions significantly improves performance. The optimal approach combining demographics with contextual information achieved QWK scores of 0.40-0.45, indicating moderate agreement with human responses. While silicon sampling cannot precisely replicate individual-level survey responses, the findings suggest it holds promise as a complementary tool for survey research, particularly in early research stages. This study provides practical guidance for researchers employing LLMs in surveys and highlights the importance of contextually relevant prompts for effective silicon sampling.

Keywords: Silicon sampling, Large language models, Facial recognition technology, Emerging technology acceptance

INTRODUCTION

Large Language Models (LLMs) are a form of generative artificial intelligence (AI) that can process and generate vast amounts of text. These models are built on the transformer architecture and undergo comprehensive training on vast datasets, including publicly available internet content, utilizing self-supervised learning techniques before being fine-tuned through reinforcement learning from human feedback (Anthropic, 2024; OpenAI, 2024). This sophisticated process allows LLMs to generate intricate content and replicate human-like responses and behaviors (Besta et al., 2024; Bubeck et al., 2023).

The capability of LLMs to mimic human behaviors has sparked research into their use across various disciplines. For example, in health care, Levine et al. (2023) found that GPT-3 could perform diagnoses for common and severe illnesses at accuracy levels close to physicians, though its triage accuracy was significantly lower. Aher, Arriaga and Kalai (2022) conducted four "Turing Experiments" to study how well GPT reproduced human behavior in economic, psycholinguistic, and social psychology experiments. In marketing research, Brand, Israeli and Ngwe (2023) demonstrated that GPT-3.5 responded to survey questions consistent with economic theory and closely matched actual consumer behaviors. LLMs' remarkable ability to generate diverse responses that mirror human patterns of thinking and behavior across different domains presents significant opportunities for social science research (Grossmann et al., 2023). Their utility is especially noticeable as virtual participants when studying specific tasks or simulating targeted samples, which is particularly advantageous in survey-based research (Dillion et al., 2023; Jansen, Jung and Salminen, 2023).

Surveys remain a cornerstone of social science inquiry because they can be tailored to measure attitudes, beliefs, and behaviors across specific contexts (Kriauciunas, Parmigiani and Rivera-Santos, 2011). However, survey research faces growing challenges: respondents demonstrate heightened concerns regarding confidentiality and privacy issues, along with growing distrust of polling organizations and institutions (Berinsky, 2017; Couper, 2017; Jansen, Jung and Salminen, 2023). Consequently, survey response, contact, and participation rates have been gradually declining while costs are increasing to maintain steady response rates (Berinsky, 2017; Couper, 2017). Considering these challenges, emerging research suggests that LLMs may offer new opportunities to identify public opinions. A particular method based on LLMs' replication abilities is "silicon sampling", where LLMs are queried after being prompted with personas usually derived from real demographic data. The concept of silicon sampling was first introduced by Argyle et al. (2023) where they conditioned GPT-3 on thousands of socio-demographic backstories from real human survey participants. The study demonstrated that these silicon samples' responses, which covered tasks describing political partisans and predicting voting behavior, could emulate human response patterns across diverse demographic groups. The idea of silicon sampling is impacting social sciences, leading some researchers to consider if LLMs could even replace human participants in surveys (Bisbee et al., 2024; Dillion et al., 2023). However, Kim and Lee (2023) advised that LLMs are best used to augment, rather than replace, conventional surveys. Several studies also argued that silicon sampling's greatest value lies in the early phase of research design, such as piloting and pre-testing survey items and vignettes (Filippas, Horton and Manning, 2024; Li et al., 2022).

Related Work

Following Argyle et al.'s (2023) initial work, subsequent studies have highlighted concerns and limitations from prompting with demographics information. Cheng, Durmus and Jurafsky (2023) reported that when prompted with demographics, GPT-3.5 and GPT-4 generated responses exhibiting higher rates of racial stereotypes than human-written portrayals using the same prompts. Similarly, Wang, Morgenstern and Dickerson (2024) documented that LLMs are susceptible to misportraying identities and trivializing subgroup distinctions when prompted with demographic identifiers. This tendency of LLMs to create caricatures of demographic groups has also led to poor silicon sampling results when prompts are based solely on demographic characteristics. For example, Sanders, Ulinich and Schneier (2023) found that GPT-3.5 struggled to predict demographic differences across American opinions on issues like abortion and policing when prompted only with demographic information. This notion is corroborated by Lee et al. (2023), whose work on silicon sampling to capture public opinions on global warming demonstrated that both GPT-3.5 and GPT-4 performed poorly when conditioned solely on demographic data. However, the accuracy of both LLMs' responses and predictions significantly improved and aligned closer with actual survey results when additional covariates relevant to global warming were incorporated into the prompts. This highlights the necessity for prompts to go beyond simple demographic data. Gerosa et al. (2023) emphasizes that LLMs must have "situational parameters" that add "interactional context" encoded within the prompts for effective persona-based research. Essentially, LLMs can simulate more accurately when they are prompted with persona variables that are correlated with the specific questions or tasks (Hu and Collier, 2024). Supporting this, Hwang, Majumder and Tandon (2023) showed that GPT-3 produced the most accurate responses when prompted with a persona defined by a combination of demographics and opinions directly relevant to the question. However, they noted that adding excessive opinions and details proved unhelpful and potentially introduced noise, which hindered accuracy. These findings collectively suggest that while demographic characteristics alone are insufficient and potentially problematic for silicon sampling, carefully constructed prompts that combine relevant contextual details can enhance the accuracy and reliability of LLM-generated responses.

Research Objectives

Drawing from these insights, this study examines the effectiveness of silicon sampling in matching survey results on public acceptance towards facial recognition technology (FRT). Specifically, this paper builds upon the work of Kostka, Steinacker and Meckel (2021), who conducted a multinational survey across Germany, China, the United Kingdom (UK), and the United States (US) to analyze public opinion on FRT. Their work collects socio-demographic data and key contextual factors, such as perceived consequences, utility, and reliability of the technology, providing a framework for this paper to address the following two research questions:

RQ 1: Can LLMs simulate an individual's surveyed opinions on FRT when prompted with a persona using only demographic information?

RQ 2: Can LLMs simulate an individual's surveyed opinions on FRT when prompted with a persona using both demographic and relevant contextual information?

To answer these questions, this paper employs three LLMs: GPT-4o, Claude 3.5 Sonnet, and the open-source DeepSeek V3. It compares the LLM-generated responses against the original survey data from Kostka, Steinacker and Meckel (2021) under three prompting conditions: demographic only, contextual information only, and demographic plus contextual information.

This paper makes several contributions. First, it provides empirical evidence on the utility of silicon sampling as a complementary tool for survey research in the domain of emerging technology acceptance, specifically for FRT. Secondly, it provides helpful insights in the impact of different prompt compositions for optimizing the silicon sampling technique. Together, these contributions advance understandings of how LLMs can augment traditional survey methods while highlighting both the potential and limitations of this approach.

METHODS AND DATA

This paper uses data from the original survey from Kostka, Steinacker and Meckel (2021). After removing non-responses from the data, there are a total of 6,076 respondents: 1,629 from China, 1,537 from Germany, 1,512 from the UK, and 1,398 from the US. To generate the silicon samples, I used GPT-4o through the OpenAI API, Claude 3.5 Sonnet through the Anthropic API, and DeepSeek V3 through the OpenRouter API. All models' parameters were set to their defaults except for the temperature, which was set to 0.7 based on prior studies (Argyle et al., 2023; Lee et al., 2023).

Table 1: Variables from the original survey by Kostka, Steinacker and Meckel (2021) split by two categories: Socio-demographics information and experiences and perceptions.

Category	Variables
Socio-demographics Information	Age
	Gender
	Income
	Education
	Ethnic Group
	Living in rural or urban area
Experiences and Perceptions	Exposure to FRT
	Frequency of FRT use
	Consequences of FRT
	Usefulness of FRT
	Reliability of FRT

To create personas for the models, I list the relevant variables in Table 1. These variables are the same main independent variables that make up the conceptual framework of FRT acceptance in the original study by Kostka, Steinacker and Meckel (2021). The persona prompts were entered into the system prompt of the LLMs. The following is an example of a persona prompt with both demographics and contextual information:

You're a 32 years old female who lives in a rural area. Your ethnicity is not a minority group in your country. Financially, you're a high-income earner. Your highest completed level of education is a Bachelor's degree.

You've seen smart devices using facial recognition technology. You've used facial recognition technology everyday. You think you've been unknowingly scanned by facial recognition technology several times a year. You think that facial recognition technology increases efficiency. You think that facial recognition technology could be useful in smart devices, railway or subway stations, and customs or security checks at airports. You think that facial recognition technology is more reliable than other identification methods.

For personas prompted with solely demographics, I only used the first paragraph of the example prompt, which corresponds to the variables categorized as "Socio-demographics Information" in Table 1. For personas with solely contextual information, I only used the second paragraph of the example prompt, which corresponds to the variables categorized as "Experiences and Perceptions" in Table 1. After creating the personas, the LLMs are then asked to answer the three primary questions about FRT acceptance from the survey with the following prompt entered into the user prompt:

Please write a number next to each question to indicate the extent to which you oppose or accept. For example, '(a) 1'. Answer only in the given example format, do not answer in decimals. 1 is Strongly oppose, 2 is Somewhat oppose, 3 is Neither oppose nor accept, 4 is Somewhat accept, 5 is Strongly accept.

(a) In general, do you accept or oppose the use of facial recognition technology?

(b) Do you accept or oppose the use of facial recognition technology in public?

(c) Do you accept or oppose the use of facial recognition technology in the private sphere?

The variables, questions, and answer choices in all the prompts were designed to closely mirror the wording of the original survey. I prompted 6,076 personas, each corresponding to a unique respondent from the survey, three times for the different prompting conditions: demographics only, contextual information only, and demographics plus contextual information. For each condition, I recorded their answers to the three primary questions about FRT acceptance. This procedure was executed using GPT-4o, Claude 3.5 Sonnet, and DeepSeek V3 in February 2025.

Evaluation Metrics

Model performance was evaluated using two primary metrics appropriate for ordinal data: Mean Absolute Error (MAE) and Quadratic Weighted Cohen's Kappa (QWK). MAE quantifies the average magnitude of error as it calculates the mean of the absolute differences between the predicted and the true survey responses. QWK assesses the agreement between predicted and actual responses while also accounting for agreement by chance. Its quadratic weighting penalizes larger misclassifications between categories more severely than smaller ones. I performed bootstrapping with 10,000 samples to obtain the 95% confidence intervals for the MAE and QWK values.

RESULTS AND ANALYSIS

Table 2 shows the MAE and QWK statistics for the three LLMs evaluated under three prompting methods across the three FRT acceptance questions. When prompted with demographics only, all models exhibited substantially high MAE values ranging from 0.90 to 1.73 and QWK values near zero. This indicates that relying solely on demographic information is insufficient for accurate FRT acceptance simulation and results in predictions with large average errors and virtually no agreement with actual survey responses. Incorporating contextual information alone showed substantial improvements: MAE values decreased significantly (ranging from 0.75 to 1.12) and QWK scores markedly increased (ranging from 0.25 to 0.42) across all models. The combination of demographics and contextual information yielded nuanced results. Compared to context only prompting, the MAE values showed mixed changes where they decreased slightly for GPT-4o and DeepSeek V3, but increased for Claude 3.5 Sonnet. However, the QWK values consistently improved or remained stable compared to context-only prompting. Thus, although the average error magnitude slightly increased compared to context only prompting, the combination of demographics and contextual information in the prompts produced the most optimal results because the models achieved more calibrated predictions with fewer severe misclassifications, as reflected by the improved QWK metrics.

Table 2: Mean absolute error (MAE) and quadratic-weighted cohen's kappa (QWK) metrics for the three LLM models across the three FRT acceptance questions by three different prompting methods. 95% confidence interval values are included in the square brackets.

Model	Prompting	MAE General Acceptance	MAE Public Acceptance	MAE Private Acceptance	QWK General Acceptance	QWK Public Acceptance	QWK Private Acceptance
Claude 3.5 Sonnet	Demographics Only	0.93 [0.91, 0.95]	1.10 [1.07, 1.12]	1.73 [1.71, 1.76]	0.06 [0.04, 0.08]	0.03 [0.01, 0.05]	0.02 [0.01, 0.02]
	Context Only	0.75 [0.73, 0.77]	0.86 [0.85, 0.88]	0.91 [0.89, 0.93]	0.39 [0.36, 0.41]	0.38 [0.36, 0.40]	0.36 [0.33, 0.38]
	Demographics and Context	0.79 [0.77, 0.81]	0.88 [0.86, 0.90]	0.95 [0.93, 0.98]	0.44 [0.42, 0.46]	0.40 [0.37, 0.42]	0.38 [0.36, 0.41]
DeepSeek V3	Demographics Only	0.94 [0.92, 0.95]	1.32 [1.30, 1.34]	1.42 [1.39, 1.45]	0.03 [0.02, 0.04]	0.01 [0.00, 0.02]	0.03 [0.01, 0.06]
	Context Only	0.76 [0.74, 0.77]	0.93 [0.91, 0.95]	1.12 [1.10, 1.15]	0.40 [0.38, 0.42]	0.32 [0.30, 0.34]	0.25 [0.22, 0.27]
	Demographics and Context	0.79 [0.77, 0.81]	0.89 [0.87, 0.91]	0.95 [0.93, 0.97]	0.45 [0.43, 0.48]	0.42 [0.40, 0.44]	0.37 [0.35, 0.39]
GPT-4o	Demographics Only	0.90 [0.89, 0.92]	1.13 [1.10, 1.15]	1.14 [1.12, 1.16]	0.04 [0.02, 0.06]	0.03 [0.01, 0.04]	-0.02 [-0.04, 0.00]
	Context Only	0.85 [0.82, 0.87]	0.87 [0.85, 0.89]	0.88 [0.86, 0.91]	0.42 [0.40, 0.44]	0.42 [0.40, 0.44]	0.38 [0.36, 0.40]
	Demographics and Context	0.83 [0.81, 0.85]	0.86 [0.84, 0.88]	0.85 [0.83, 0.87]	0.43 [0.40, 0.45]	0.42 [0.40, 0.44]	0.40 [0.37, 0.42]

With the demographics and context information in the prompts, GPT-4o demonstrated a slight edge, particularly with the lowest MAE for the private and public acceptance of FRT, as well as the highest QWK for private acceptance. However, it should be noted that all three LLMs exhibited largely similar performance across many metrics. The confidence intervals indicate narrow differences across most comparisons, especially for the QWK values clustering around 0.4, which suggest fair to moderate agreement with the actual survey responses (Fleiss, Levin and Paik, 2003).

Overall, these results indicate that silicon sampling with demographic information alone is insufficient and can even be misleading. All models showed substantial reduction in errors and misclassifications when contextual information is included with demographics in the prompts. Despite these significant improvements, the highest QWK scores achieved (around 0.40–0.45) indicate that while LLMs prompted with demographic and relevant contextual information can simulate an individual's surveyed opinions on FRT moderately well, precise simulation remains a challenging task.

CONCLUSION

This study investigated the effectiveness of silicon sampling in simulating survey responses on FRT acceptance, specifically examining the impact of prompt composition on the accuracy of LLM-generated survey responses. Drawing upon the dataset from Kostka, Steinacker and Meckel (2021) and employing three LLMs (GPT-4o, Claude 3.5 Sonnet, and DeepSeek V3), we compared simulations based on prompts with demographic information alone, contextual information alone, and a combination of both. The findings demonstrate that LLMs cannot accurately simulate individual opinions on FRT when prompted solely with demographic information (RQ1), as evidenced by high error rates and minimal agreement with actual survey responses. This result corroborates previous research warning against the risks of demographic-only prompting, which can lead to stereotypical representations and poor simulations (Cheng, Durmus and Jurafsky, 2023; Lee et al., 2023). In addition, the inclusion of contextual information related to FRT experiences and perceptions significantly improved simulation fidelity. While context-only prompts reduced errors compared to demographic-only approaches, the integration of both elements optimized performance by minimizing severe misclassifications, as evidenced by higher QWK scores. Again, this is consistent with past research underscoring the necessity of including relevant contextual information (Gerosa et al., 2023; Hu and Collier, 2024). Across all three LLMs, this combined prompting approach achieved QWK scores in the 0.40–0.45 range, indicating moderate agreement with human survey responses.

Practical Implications

There are several main practical implications of these findings. First, this study offers researchers a template for designing robust persona-based simulations by highlighting the importance of rich, relevant contextual information in prompts for effective silicon sampling. Second, while precise individual-level simulation remains a challenge, silicon sampling shows promise as a complementary tool in survey research. The findings from this paper substantiates the idea from past studies that LLMs can be particularly valuable in early research stages, such as exploring potential response patterns for specific personas or pre-testing survey questions (Filippas, Horton and Manning, 2024). Importantly, while the achieved QWK scores of 0.40-0.45 indicate that LLMs cannot perfectly replicate individual survey responses, they demonstrate sufficient accuracy to capture general patterns of public opinion toward emerging technologies like FRT. This level of agreement suggests that silicon sampling could serve as a useful tool for identifying potential demographic or contextual factors that influence technology acceptance and generating hypotheses about public reactions to new technologies before conducting resource-intensive surveys. Researchers studying other emerging technologies might leverage similar approaches to obtain preliminary insights into public acceptance, though validation with human responses remains essential.

Lastly, by comparing the GPT-4o, Claude 3.5 Sonnet, and DeepSeek V3, the findings showed that performance differences between the models are relatively modest when appropriate prompting methods are used. The choice of LLMs may be less critical than the prompting methodology itself, thus offering researchers greater flexibility in model selection.

Limitations and Future Work

One limitation of this study is that the findings are based on a single dataset concerning FRT acceptance. Generalizability to other emerging technologies or entirely different topics require further investigation. Future work could also include a deeper analysis into the current survey data, such as examining silicon sampling performance by country or correlations between FRT acceptance and the individual variables. Furthermore, Barrie, Palmer and Spirling (2024) have raised critical points and guidelines concerning reproducibility issues of LLMs. Future work could adopt these recommended practices and support reproducible results for silicon sampling. Lastly, the rapid evolution of LLM architectures and capabilities means that future models, especially with reasoning abilities, may exhibit different performance characteristics and present promising avenues for further inquiry.

REFERENCES

Aher, G., Arriaga, R., I. and Kalai, A. T. (2022) *Using large language models to simulate multiple humans and replicate human subject studies*. https://arxiv.org/abs/2208.10264

Anthropic (2024) *The Claude 3 Model Family: Opus, Sonnet, Haiku*, *Anthropic*. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf (Accessed: May 30, 2025).

Argyle, L. P. *et al.* (2023) 'Out of one, many: using language models to simulate human samples,' *Political Analysis*, 31(3), pp. 337–351. https://doi.org/10.1017/pan.2023.2

Barrie, C., Palmer, A. and Spirling, A. (2024) Replication for Language Models. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf (Accessed: May 30, 2025).

Berinsky, A. J. (2017) 'Measuring Public Opinion with Surveys,' Annual Review of Political Science, 20(1), pp. 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724

Besta, M. *et al.* (2024) 'Graph of Thoughts: Solving Elaborate Problems with Large Language Models,' *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), pp. 17682–17690. https://doi.org/10.1609/aaai.v38i16.29720

Bisbee, J. *et al.* (2024) 'Synthetic replacements for human survey data? The perils of large language models,' *Political Analysis*, 32(4), pp. 401–416. https://doi.org/10.1017/pan.2024.5

Brand, J., Israeli, A. and Ngwe, D. (2023) 'Using GPT for market research,' *SSRN Electronic Journal* [Preprint]. https://doi.org/10.2139/ssrn.4395751

Bubeck, S. *et al.* (2023) *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. https://arxiv.org/abs/2303.12712

Cheng, M., Durmus, E. and Jurafsky, D. (2023) *Marked personas: Using natural language prompts to measure stereotypes in language models*. https://arxiv.org/abs/2305.18189

Couper, M. P. (2017) 'New developments in survey data collection,' *Annual Review of Sociology*, 43(1), pp. 121–145. https://doi.org/10.1146/annurev-soc-060116-053613

Dillion, D. *et al.* (2023) 'Can AI language models replace human participants?,' *Trends in Cognitive Sciences*, 27(7), pp. 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Filippas, A., Horton, J. J. and Manning, B. S. (2024) 'Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?,' *EC'24: Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 614–615. https://doi.org/10.1145/3670865.3673513

Fleiss, J. L., Levin, B. and Paik, M. C. (2003) *Statistical methods for rates and proportions*. Wiley-Interscience.

Gerosa, M. A. *et al.* (2023) 'Can AI serve as a substitute for human subjects in software engineering research?,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2311.11081

Grossmann, I. *et al.* (2023) 'AI and the transformation of social science research,' *Science*, 380(6650), pp. 1108–1109. https://doi.org/10.1126/science.adi1778.

Hu, T. and Collier, N. (2024) *Quantifying the persona effect in LLM simulations*. https://arxiv.org/abs/2402.10811

Hwang, E., Majumder, B. P. and Tandon, N. (2023) 'Aligning language models to user opinions,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2305.14929

Jansen, B. J., Jung, S.-G. and Salminen, J. (2023) 'Employing large language models in survey research,' *Natural Language Processing Journal*, 4, p. 100020. https://doi.org/10.1016/j.nlp.2023.100020

Kim, J. and Lee, B. (2023) *AI-Augmented Surveys: Leveraging large language models and surveys for opinion prediction*. https://arxiv.org/abs/2305.09620

Kostka, G., Steinacker, L. and Meckel, M. (2021) 'Between security and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States,' *Public Understanding of Science*, 30(6), pp. 671–690. https://doi.org/10.1177/09636625211001555

Kriauciunas, A., Parmigiani, A. and Rivera-Santos, M. (2011) 'Leaving our comfort zone: Integrating established practices with unique adaptations to conduct survey-based strategy research in nontraditional contexts,' *Strategic Management Journal*, 32(9), pp. 994–1010. https://doi.org/10.1002/smj.921

Lee, S. *et al.* (2023) 'Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2311.00217.

Levine, D. M. *et al.* (2023) 'The diagnostic and triage accuracy of the GPT-3 artificial intelligence model,' *medRxiv (Cold Spring Harbor Laboratory)* [Preprint]. https://doi.org/10.1101/2023.01.30.23285067

Li, P. *et al.* (2022) 'Language Models for Automated Market Research: A new way to generate Perceptual maps,' *SSRN Electronic Journal* [Preprint]. https://doi.org/10.2139/ssrn.4241291

OpenAI (2024) *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774 (Accessed: May 30, 2025)

Sanders, N. E., Ulinich, A. and Schneier, B. (2023) *Demonstrations of the potential of AI-based political issue polling*. https://arxiv.org/abs/2307.04781

Wang, A., Morgenstern, J. and Dickerson, J. P. (2024) *Large language models that replace human participants can harmfully misportray and flatten identity groups*. https://arxiv.org/abs/2402.01908