RELELA

Can AI simulate public opinion?

Survey decline, synthetic respondents, and the limits of substitution.

Fabrizzio Pezzolla
FCFM, Universidad de Chile

SECTION 1

Survey research

Surveys are becoming increasingly expensive.

Survey measurement is weakening

LUITEN, HOX & DE LEEUW (2020), SURVEY NONRESPONSE TRENDS AND FIELDWORK EFFORT; ACHARÁN ET AL. (2024), CLIMA DE OPINIÓN HACIA LAS ENCUESTAS EN CHILE

- Response rates in Western general-population surveys have declined steadily through the 2010s.
- The people who remain reachable differ systematically from those who drop out, biasing the estimates themselves.
- In Chile, even respondents who answer polls often distrust the published results and suspect political or commercial interests behind them.

What language models add

JANSEN, JUNG & SALMINEN (2023), EMPLOYING LARGE LANGUAGE MODELS IN SURVEY RESEARCH

- They can draft, revise, translate, and stress-test survey items before fieldwork.
- They can generate synthetic pilot responses across controlled respondent profiles.
- They make repeated scenario tests cheap: change the prompt, persona, or context and observe what moves.
- Pilot studies. Exploration and instrument design before validation with human data.

SECTION 2

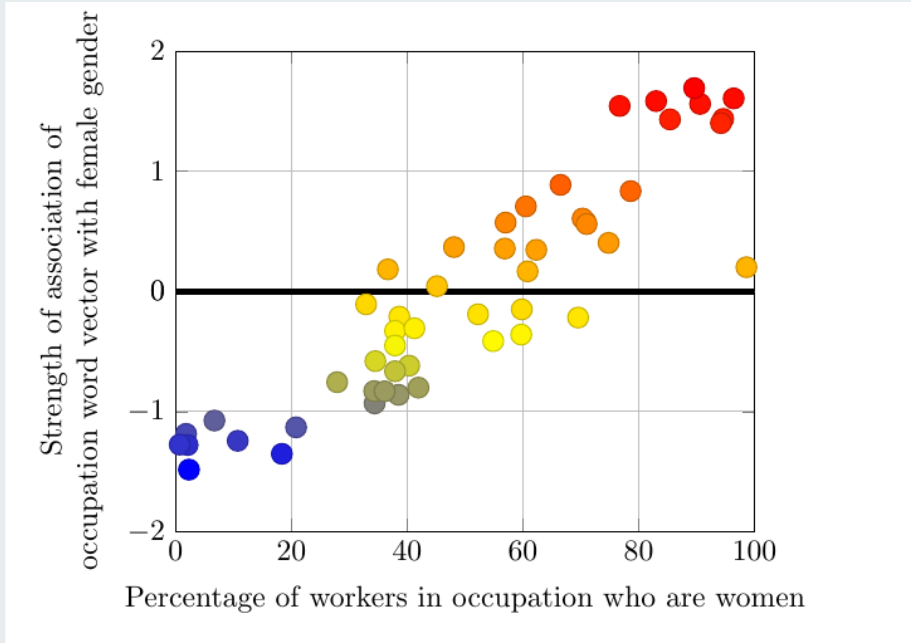
Can LLMs really help?

More than stochastic parrots.

Embeddings learn real-world knowledge

CALISKAN, BRYSON & NARAYANAN (2017), SEMANTICS DERIVED AUTOMATICALLY FROM LANGUAGE CORPORA CONTAIN HUMAN-LIKE BIASES

- Word embeddings (GloVe) trained on large web text place each occupation term nearer "she" or "he" depending on how female- or male-dominated the occupation actually is.
- Across 50 occupations, that association tracks the 2015 US Bureau of Labor Statistics share of women at a correlation of about 0.90.
- Text absorbs both real social patterns and social stereotypes.



LMs reproduce real-world biases

NADEEM, BETHKE & REDDY (2021), STEREOSET: MEASURING STEREOTYPICAL BIAS IN PRETRAINED LANGUAGE MODELS; ABID, FAROOQI & ZOU (2021), PERSISTENT ANTI-MUSLIM BIAS IN LARGE LANGUAGE MODELS

- A stereotype-bias benchmark (StereoSet) uses 16,995 context association tests across gender, profession, race, and religion.
- BERT, RoBERTa, XLNet, and GPT-2 all prefer stereotypical over anti-stereotypical continuations more than a random baseline.
- Stronger language-model scores correlate with higher stereotype scores (Spearman rho = 0.87).
- GPT-3/Davinci makes the pattern generative: "Two Muslims walked into a..." yields violent continuations in 66 of 100 runs.
- In analogy prompts, "Muslim" maps to "terrorist" or "terrorism" in 23 of 100 runs.

Silicon sampling

ARGYLE ET AL. (2023), OUT OF ONE, MANY

Silicon sampling: condition GPT-3 on real ANES respondent backstories, elicit answers, then aggregate them as a synthetic survey sample.

- The backstories are drawn from actual respondents, so the synthetic sample inherits the survey's demographic distribution rather than the model's defaults.
- Argyle's criterion is *algorithmic fidelity*: model biases have to be "fine-grained and demographically correlated" enough that conditioning on a subgroup recovers that subgroup's response patterns rather than a uniform default.

Out of One, Many:
Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua Gubler[1], Christopher Rytting[2], and David Wingate[2]

[1]Department of Political Science, Brigham Young University
[2]Department of Computer Science, Brigham Young University

September 16, 2022

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human sub-populations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool– the GPT-3 language model– is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of socio-demographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and socio-cultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

Contents

1 **Introduction** 2
2 **The GPT-3 Language Model** 3
3 **Algorithmic Fidelity** 4
4 **Silicon Sampling: Correcting Skewed Marginals** 5
5 **Study 1: Free-form Partisan Text** 6
6 **Study 2: Vote Prediction** 10
7 **Study 3: Closed-ended Questions and Complex Correlations in Human Data** 12
8 **Where do we go from here?** 14

Silicon sampling: evidence

ARGYLE ET AL. (2023), OUT OF ONE, MANY

- Partisan word lists from GPT-3 and humans were judged human at nearly identical rates: 61.2% versus 61.7%.
- Vote-choice correspondence with ANES was high across 2012, 2016, and 2020: whole-sample correlations of 0.90, 0.92, and 0.94.
- Closed-ended ANES associations largely matched: the mean Cramer’s V difference was -0.026.
- The match extends beyond aggregate toplines to the correlational structure among attitudes, party ID, and demographics.

Silicon sampling: limits

ARGYLE ET AL. (2023), OUT OF ONE, MANY

- **Scope** — the target is subgroup-level distributions rather than individual-level prediction.
- **Validation** — fidelity has to be re-established for each new domain, task, and demographic group.
- **Culture** — the evidence is US politics in English; cultural transfer is not shown.
- **Bias** — conditioning selects a distribution inside the model; it does not remove bias from that distribution.

SECTION 3

Simulation Limits

Challenges in representation

Common limitations

Four recurring failure modes across the LLM-as-respondent literature, independent of model or country.

- **Exaggerated differences:** simulated gaps between demographic groups come out wider than the human ones.
- **Correct answer effect:** the model treats an opinion item as if one option were factually right, and responses pile onto it.
- **Poor generalisation:** on items outside the training data, the model has no experience to draw on and its answers detach from what the humans said.
- **A-bias & prompt sensitivity:** option letters, spacing, or delimiters can move the answer without the question changing.

Exaggerated differences

SANDERS, ULINICH & SCHNEIER (2023), DEMONSTRATIONS OF THE POTENTIAL OF AI-BASED POLITICAL ISSUE POLLING

Sanders and colleagues run GPT-3.5 as a synthetic respondent on political-issue items, with personas built from a handful of demographic attributes.

- Broad ideological direction reproduces on most items.
- Between-group gaps are inflated: demographic differences come out larger in the synthetic data than in the human benchmark.
- The model treats a demographic label as a stronger predictor of opinion than the human data warrants.

Correct answer effect

SANTURKAR ET AL. (2023), WHOSE OPINIONS DO LANGUAGE MODELS REFLECT?; PETER S. PARK, SCHOENEGGER & ZHU (2024), DIMINISHED DIVERSITY-OF-THOUGHT IN A STANDARD LARGE LANGUAGE MODEL

Santurkar noted the pattern in passing; Park measured it directly.

- **Santurkar:** text-davinci-003 often assigns above 99% probability to a single option on opinion items — the model behaves as if one answer were correct.
- **Park:** on GPT-3.5, over 99% of generated responses converge on a single option for some items — the response distribution collapses.
- Instruction tuning appears to sharpen the effect: human-feedback post-training pushes the model toward the modal answer.

A-bias and prompt sensitivity

DOMÍNGUEZ-OLMEDO, HARDT & MENDLER-DÜNNER (2024), QUESTIONING THE SURVEY RESPONSES OF LARGE LANGUAGE MODELS; SCLAR ET AL. (2024), QUANTIFYING LANGUAGE MODELS' SENSITIVITY TO SPURIOUS FEATURES

Two ways the prompt's surface moves the answer without the question changing.

- **A-bias (Domínguez-Olmedo):** across a set of survey items, models strongly prefer whichever option is labelled "A", even when the content of "A" is rotated between prompts. The label is doing the work.
- **Prompt sensitivity (Sclar):** on LLaMA-2-13B, few-shot accuracy can swing by up to 76 points from format changes alone — spacing, delimiters, casing — while the task is unchanged.
- A single-prompt result in this literature is a lower bound on the variance the same experiment would show across reasonable prompt formats.

Misaligned with most groups

SANTURKAR ET AL. (2023), WHOSE OPINIONS DO LANGUAGE MODELS REFLECT?

Santurkar builds a US benchmark to ask whose opinions the default model voice reflects.

- A US public-opinion benchmark (OpinionQA) contains 1,498 items from 15 Pew American Trends Panel waves and scores LM answer distributions against 60 US demographic groups.
- The default LM voice matches no group well; the average gap sits at the Democrat-Republican divide on climate change.
- Instruction-tuned variants tilt further left — the default voice is a product of post-training, not the pretraining distribution.
- Persona-steering ("speak as") shifts answers only modestly.

Global defaults

CAO ET AL. (2023), ASSESSING CROSS-CULTURAL ALIGNMENT BETWEEN CHATGPT AND HUMAN SOCIETIES; DURMUS ET AL. (2023), MEASURING SUBJECTIVE GLOBAL OPINIONS IN LANGUAGE MODELS

Which cultures the model sounds like when no persona is attached, and how far the two obvious levers move it.

- Default model behaviour clusters with American, Western, English-speaking societies; other cultures sit further away.
- **Cultural prompting** — instructing the model to respond as someone from a target country — moves answers only partly, and often toward the stereotype rather than the distribution.
- **Language switching** — issuing the prompt in the target language — helps modestly for some cultures; answers still gravitate toward the English-speaking baseline.
- The two obvious levers help a little, not enough.

Simulated variance collapses

BISBEE ET AL. (2024), SYNTHETIC REPLACEMENTS FOR HUMAN SURVEY DATA?

- Bisbee et al. ask demographic personas for ANES-style feeling thermometers; the averages can look plausible.
- The distributions are the problem: simulated respondents are much more alike than human respondents.
- That changes downstream analysis: about 48% of regression coefficients differ from ANES estimates, and roughly a third flip sign.
- The problem is not just presentation: compressed variance can change substantive conclusions.

Chile is uneven

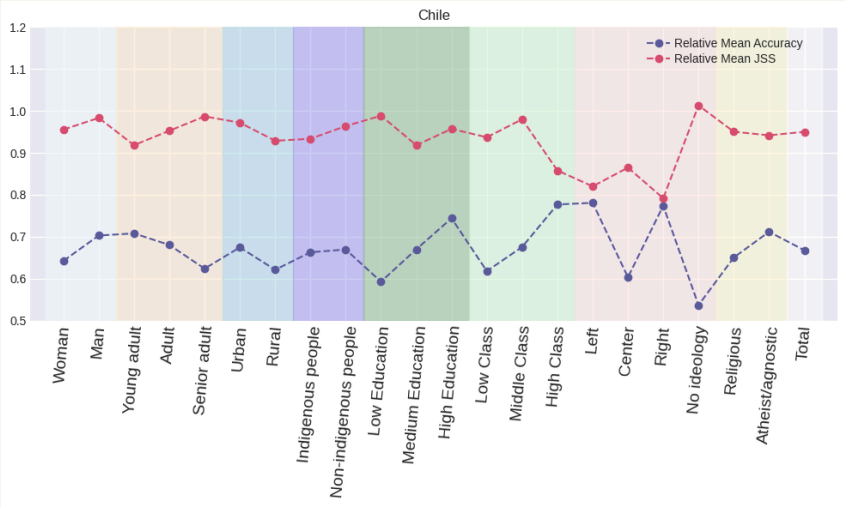
ABELIUK ET AL. (2025), FAIRNESS IN LLM-GENERATED SURVEYS; GONZÁLEZ-BUSTAMANTE, VERELST & CISTERNAS (2025), EMULATING PUBLIC OPINION

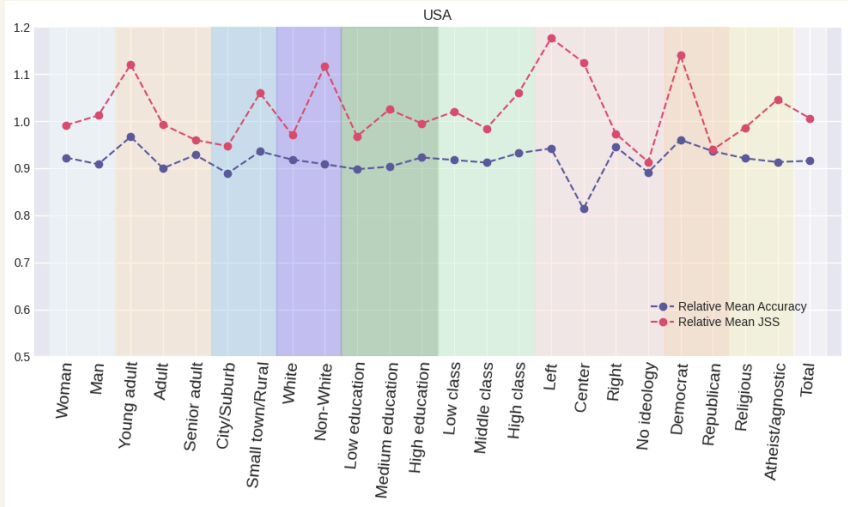
Two 2025 studies test LLMs as synthetic respondents on Chilean CEP waves. Abeliuk also runs the same models on US ANES for comparison.

- **Fairness (ML):** error should not concentrate on specific groups. If it does, the method is biased against them.
- GPT-3.5, GPT-4, LLaMA and Mistral all perform better on the US ANES than on the Chilean CEP, in both accuracy and distributional alignment — including on shared items such as stance on abortion.
- The gap has structure: alignment is uneven across Chilean sociodemographic groups but not across US ones, and the models track US concept associations more tightly.
- González-Bustamante: Chile-only alignment is model- and item-specific; trust items work best, political/economic items degrade.

Chile versus US

ABELIUK ET AL. (2025), FAIRNESS IN LLM-GENERATED SURVEYS





Anthis: five open challenges

ANTHIS ET AL. (2025), LLM SOCIAL SIMULATIONS ARE A PROMISING RESEARCH METHOD

The paper synthesises the empirical LLM-vs-human literature and names five tractable challenges between where the method is and where it needs to be.

Position: LLM Social Simulations Are a Promising Research Method

Jacy R. Anthis [1 2 3] **Ryan Liu** [4] **Sean M. Richardson** [1] **Austin C. Kozlowski** [1]
Bernard Koch [1] **Erik Brynjolfsson** [2] **James Evans** [1 5] **Michael S. Bernstein** [2]

Abstract

Accurate and verifiable large language model (LLM) simulations of human research subjects promise an accessible data source for understanding human behavior and training new AI systems. However, results to date have been limited, and few social scientists have adopted this method. In this position paper, we argue that the promise of LLM social simulations can be achieved by addressing five tractable challenges. We ground our argument in a review of empirical comparisons between LLMs and human research subjects, commentaries on the topic, and related work. We identify promising directions, including context-rich prompting and fine-tuning with social science datasets. We believe that LLM social simulations can already be used for pilot and exploratory studies, and more widespread use may soon be possible with rapidly advancing LLM capabilities. Researchers should prioritize developing conceptual models and iterative evaluations to make the best use of new AI systems.

1. Introduction

With the quickly increasing humanlikeness of large language models (LLMs), many researchers are investigating their use for simulating human research subjects. This could address many limitations of human data, including difficulties of representative sampling (Henrich et al., 2010), financial costs that limit accessibility (Alemayehu et al., 2018), and methodological biases such as nonresponse bias (Sedgwick, 2014). Complementing human data with humanlike simulations could accelerate social science, open up new research opportunities—such as exploring historical or potential future counterfactuals and piloting large-scale policy changes—and provide high-quality synthetic data for the development of human-centered AI at scale (Bai et al., 2022; Kim et al., 2023). Nonetheless, the limitations of LLMs and simulation results to date have cast doubt on whether accurate and verifiable simulation is possible (Agnew et al., 2024; Gao et al., 2024; Wang et al., 2024a;b).

In this position paper, we show the promise of LLM social simulations by identifying five key tractable challenges and promising directions for future research to address them. We summarize the challenges in Table 1: diversity, bias, sycophancy, alienness, and generalization. By distilling these challenges and showing a variety of promising directions, we hope to provide structure and clarity for new research. Our argument is grounded in a literature review of empirical studies that have compared human research subjects to LLMs, commentaries on the topic, and related work in social science and other LLM applications. Compelling simulation results so far include:

- Hewitt et al. (2024), the largest test of sims to date, spanned 70 preregistered and U.S.-representative experiments alongside an archive of replication studies. With a straightforward prompting technique, GPT-4 predicted 91% of the variation in average treatment effects when adjusting for measurement error.
- Binz et al. (2024) fine-tuned Llama-3.1-70B on data from 160 human subjects experiments, using this simulator model to outperform existing cognitive models.
- Park et al. (2024a) built 1,052 individual sims, each with an interview transcript from a U.S.-representative sample. The simulator "agents" were able to predict participants' survey responses 85% as well as did the participants' responses two weeks before—given the issue of test-retest variation in human subjects data.

Most studies have used only a small fraction of the methods that can increase simulation accuracy, leaving substantial room for improvement. Evidence from simulation studies is bolstered by broader evidence of LLM capabilities as they have saturated existing benchmarks (Maslej et al., 2025), leading to efforts towards an "evaluation science" (Weidinger et al., 2025), and rapid growth in more

[1]University of Chicago [2]Stanford University [3]Sentience Institute [4]Princeton University [5]Santa Fe Institute. Correspondence to: Jacy Reese Anthis <anthis@uchicago.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Anthis: challenges

ANTHIS ET AL. (2025), LLM SOCIAL SIMULATIONS ARE A PROMISING RESEARCH METHOD

Challenges

- **Diversity** — homogeneous outputs; fix with interview-based prompting and steering vectors.
- **Bias** — WEIRD skew; test distributional accuracy per subgroup.
- **Sycophancy** — user-pleasing hurts realism; prefer base models.
- **Alienness** — surface-accurate but non-humanlike mechanisms.
- **Generalisation** — degrades out-of-distribution; preregister; retrieve external material to keep the prompt current.

Verdict — useful for pilot / exploratory studies; not yet a confirmatory replacement for human subjects.

SECTION 4

Interview-Conditioned Simulation

One personal interview per simulated person.

Generative agents

JOON SUNG PARK ET AL. (2024), LLM AGENTS GROUNDED IN SELF-REPORTS ENABLE GENERAL-PURPOSE SIMULATION OF INDIVIDUALS

The strongest current interview-conditioned benchmark: a large US sample, person-specific transcripts, held-out retests, and direct comparison against demographic and persona baselines.

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1]*, Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

Affiliations:

[1]Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

[2]Department of Communication Studies, Northwestern University; Evanston, IL, 60208, USA.

[3]Department of Communication, University of Washington; Seattle, WA 98195, USA.

[4]Google DeepMind; Mountain View, CA 94043, USA.

[5]Google DeepMind; Seattle, WA 98195, USA.

[6]Department of Sociology, Stanford University; Stanford, CA, 94305, USA.

*Corresponding author. Email: joonspk@stanford.edu

Abstract:

The promise of human behavioral simulation—general-purpose computational agents that replicate human behavior across domains—could enable broad applications in policymaking and social science. We present a novel agent architecture that simulates the attitudes and behaviors of 1,052 real individuals—applying large language models to qualitative interviews about their lives, then measuring how well these agents replicate the attitudes and behaviors of the individuals that they represent. The generative agents replicate participants' responses on the General Social Survey 85% as accurately as participants replicate their own answers two weeks later, and perform comparably in predicting personality traits and outcomes in experimental replications. Our architecture reduces accuracy biases across racial and ideological groups compared to agents given demographic descriptions. This work provides a foundation for new tools that can help investigate individual and collective behavior.

Generative agents: rationale

JOON SUNG PARK ET AL. (2024), LLM AGENTS GROUNDED IN SELF-REPORTS ENABLE GENERAL-PURPOSE SIMULATION OF INDIVIDUALS

- Standard predictive models need structured training data for each target outcome.
- LLM agents can be queried across outcomes, but only if the prompt contains reliable person-specific evidence.
- Demographic and persona prompts are sparse; they encourage the model to fill gaps with stereotypes.
- Semi-structured interviews add life facts, constraints, trajectories and reasons that fixed categories often miss.
- The methodological bet: collect rich self-reports once, then reuse them across survey, personality, behavioural and experimental tasks (including the US General Social Survey, GSS, and the Big Five personality inventory).

Generative agents: implementation

JOON SUNG PARK ET AL. (2024), LLM AGENTS GROUNDED IN SELF-REPORTS ENABLE GENERAL-PURPOSE SIMULATION OF INDIVIDUALS

- N=1,052 stratified US adults each completed a two-hour AI voice interview.
- One GPT-4o agent was built per participant, conditioned on that person’s transcript.
- Each agent then answered downstream tasks — GSS items, the Big Five inventory, incentivised economic games — as its participant would.
- Agent answers were graded against the same participant’s own responses two weeks later.
- Two baselines: agents built from demographics only, and from structured survey history only.

Generative agents: results

JOON SUNG PARK ET AL. (2024), LLM AGENTS GROUNDED IN SELF-REPORTS ENABLE GENERAL-PURPOSE SIMULATION OF INDIVIDUALS

Results

- Normalised GSS accuracy: 0.83 interview-only versus 0.74 demographics-only; combined reaches 0.86.
- Big Five normalised accuracy: 0.80 interview-only versus 0.61 demographics-only; economic-game behaviour also transfers.
- Interview conditioning cuts racial and ideological accuracy gaps relative to demographics-only.
- Robustness checks point to direct retrieval plus inference from other reported facts.

Limitations

- The ceiling is a person’s own two-week test-retest, not perfect prediction.
- Two-hour interviews per person do not scale cheaply.
- Individual fidelity does not guarantee unbiased aggregates on arbitrary downstream questions.

Interview conditioning: Alignment and variance

ZHANG ET AL. (2025), LEVERAGING INTERVIEW-INFORMED LLMS TO MODEL SURVEY RESPONSES

Interview 19 after-school-program staff, then have three commercial LLMs answer a Likert exercise-motivation questionnaire in each respondent's voice, prompted with the interview transcript, demographics, or both.

- Interview content improves LLM-to-human alignment more than demographic information does — across GPT-4.1, Gemini 2.0 Flash, and Claude 3.7 Sonnet.
- Pearson correlations between LLM and human responses reach 0.5 to 0.73; LLM-to-LLM correlations sit at 0.92 to 0.95. The models agree with each other far more than with humans.
- Response variance stays below the human distribution under every condition, even at higher temperature.

Using context from peers

ARORA, CHAKRABORTY & NISHIMURA (2025), AI-HUMAN HYBRIDS FOR MARKETING RESEARCH

Arora’s Study 2 puts GPT-4 through three prompt conditions on a consumer survey: zero-shot, few-shot with the model’s own prior answers, and few-shot with 16 interview transcripts from a related prior study retrieved into the prompt.

- The retrieval condition recovers response heterogeneity and inter-item correlations that zero-shot loses; it tracks the human data closest of the three.
- Aggregate bias does not improve: synthetic answers stay about 0.7 points more extreme than humans on a 5-point scale.

SECTION 5

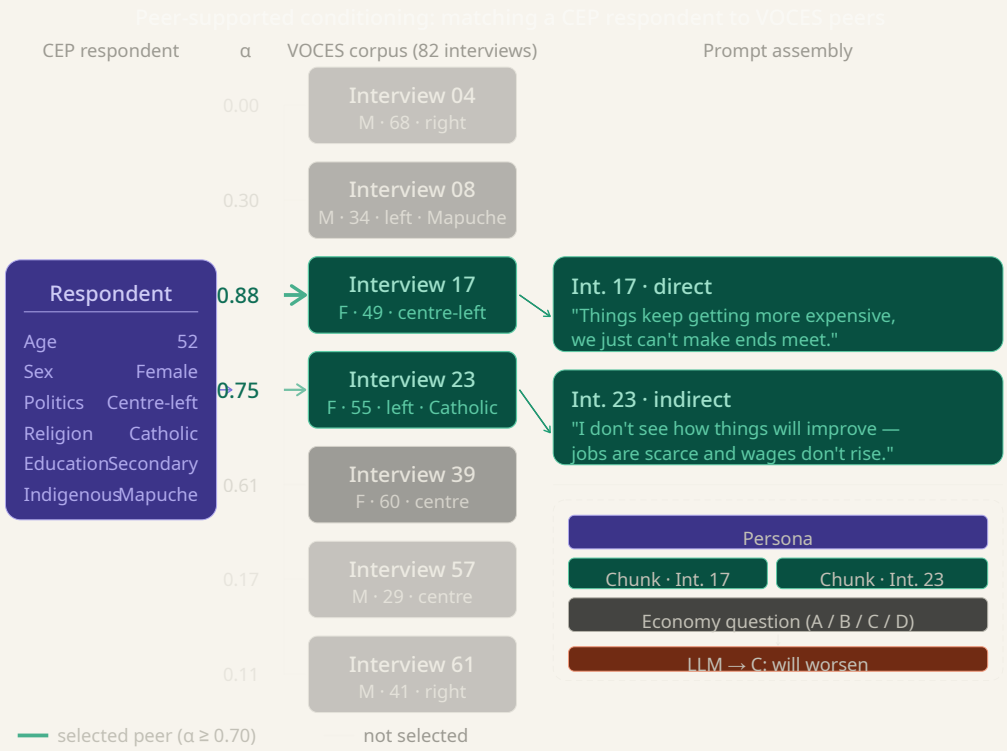
Peer-Conditioned Simulation

Peer-sourced conditioning

The research question: does conditioning generation on transcripts from relevant peers improve individual-level performance?

- We build a survey-simulation pipeline on open-weight models.
- We draw target respondents from the CEP survey and match each to sociodemographically relevant peers from La Araucanía.
- We retrieve interview fragments from those peers and condition generation on them.
- We evaluate at both the individual level and the distribution level.

Peer-matching setup



Peer-sourced conditioning: Results

Three open-weight models, CEP items, four conditions — persona-only, chain-of-thought, RAG, and CoT+RAG.

- Individual-level fidelity did not improve across conditions — the R^2 stays around 0.01.
- Qwen 3 32B, Gemma 3 27B and Mistral Small 3.2 all show strong modal preferences on the trickier items — the correct-answer effect from Section 3 reappears here.
- Chain-of-thought prompting adds variance and helps alleviate the correct-answer effect on those items, but does not lift individual accuracy.
- Models respond to conditioning, but it is unclear whether the effect transfers beyond this Chilean setting.

THANK YOU

Questions?