

Evaluation & Validation

Session 4, Georg Ahnert

The “Validate-then-Simulate” Approach

(Hullman et al., 2026 – more in the session *Understanding Large Language Models (1)* on Fri at 10:45)

- **Compares generated responses to:**
 - human survey responses (treated as ground truth)
 - external data (e.g., election results)
 - theoretical expectations
- **Aims for generalization** to different personas/questions/topics/points in time

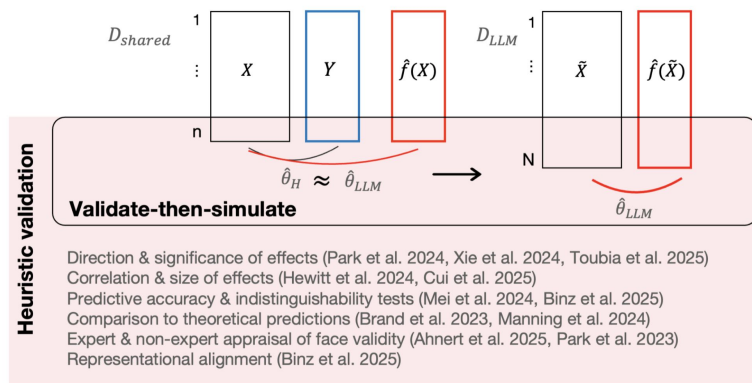
This human study did not involve human subjects:
Validating LLM simulations as behavioral evidence

Jessica Hullman
Northwestern University

David Broska
Stanford University

Huaman Sun
Northwestern University

Aaron Shaw
Northwestern University



Existing Benchmark Datasets

- OpinionQA ([Santurkar et al., 2023](#))
 - GlobalOpinionQA ([Durmus et al., 2024](#))
 - Psych-101 ([Binz et al., 2024](#))
 - Psych-201 ([Binz et al., 2026](#))
 - SimBench ([Hu et al., 2025](#))
 - GGSS Personas ([Rupprecht et al., 2026](#))
- Is your target construct covered?
 - Generalization beyond specific countries and points in time?
 - Training data contamination

SIMBENCH: BENCHMARKING THE ABILITY OF LARGE LANGUAGE MODELS TO SIMULATE HUMAN BEHAVIORS

Tiancheng Hu¹, Joachim Baumann², Lorenzo Lupo³,
Nigel Collier¹, Dirk Hovy³, and Paul Röttger³

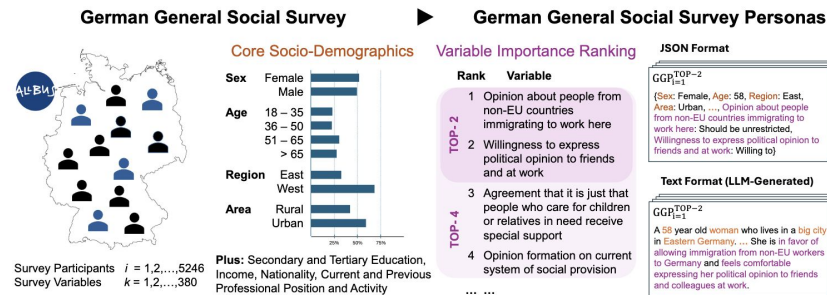
¹University of Cambridge, ²University of Zurich, ³Bocconi University,

German General Social Survey Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies

Jens Rupprecht^{1,*}, Leon Fröhling^{2,*}, Claudia Wagner^{2,3,4}, Markus Strohmaier^{1,2,4}

¹University of Mannheim, ²GESIS – Leibniz Institute for the Social Sciences,

³RWTH Aachen University, ⁴Complexity Science Hub



Common Approaches to Heuristic Validation

- **Face validity** (e.g., [Park et al., 2023](#))
- Easy to test
- Known limitations apply

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu



Common Approaches to Heuristic Validation

- Face validity
- Individual-level alignment
(e.g., [Argyle et al., 2023](#))
- Do we include all categories?
(e.g., presidential candidates)
- How to treat missing or invalid responses in human & LLM data?
- Do we account for survey weights and group sizes?
- How relevant are individual responses for social science research?

PA

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

Lisa P. Argyle¹, Ethan C. Busby¹, Nancy Fulda²,
Joshua R. Gubler¹, Christopher Rytting² and David Wingate²

Variable	2012	2012	2016	2016	2020	2020
	Tetra.	Prop. Agree	Tetra.	Prop. Agree	Tetra.	Prop. Agree
Whole sample	0.90	0.85	0.92	0.87	0.94	0.89
Men	0.90	0.85	0.93	0.88	0.95	0.88
Women	0.91	0.86	0.92	0.86	0.94	0.90
Strong partisans	0.99	0.97	1.00	0.97	1.00	0.97
Weak partisans	0.73	0.74	0.71	0.74	0.84	0.82
Leaners	0.90	0.85	0.93	0.87	0.95	0.89
Independents	0.31	0.59	0.41	0.62	0.02	0.53
Conservatives	0.84	0.84	0.88	0.86	0.91	0.89
Moderates	0.65	0.77	0.76	0.78	0.71	0.77

Common Approaches to Heuristic Validation

- Face validity
- Individual-level alignment
- **(Sub-)population-level distributions**
(e.g., [Santurkar et al., 2023](#))
- Mostly unaffected by individual-level noise
- Do we simulate individuals or groups?
- Do we account for survey weights, group sizes?
- Which subpopulations do we include in the evaluation?

303.17548v1 [cs.CL] 30 Mar 2023

Whose Opinions Do Language Models Reflect?

Shibani Santurkar
Stanford
shibani@stanford.edu

Esin Durmus
Stanford
esindurmus@cs.stanford.edu

Faisal Ladhak
Columbia University
faisal@cs.columbia.edu

Cinoo Lee
Stanford
cinoolee@stanford.edu

Percy Liang
Stanford
pliang@cs.stanford.edu

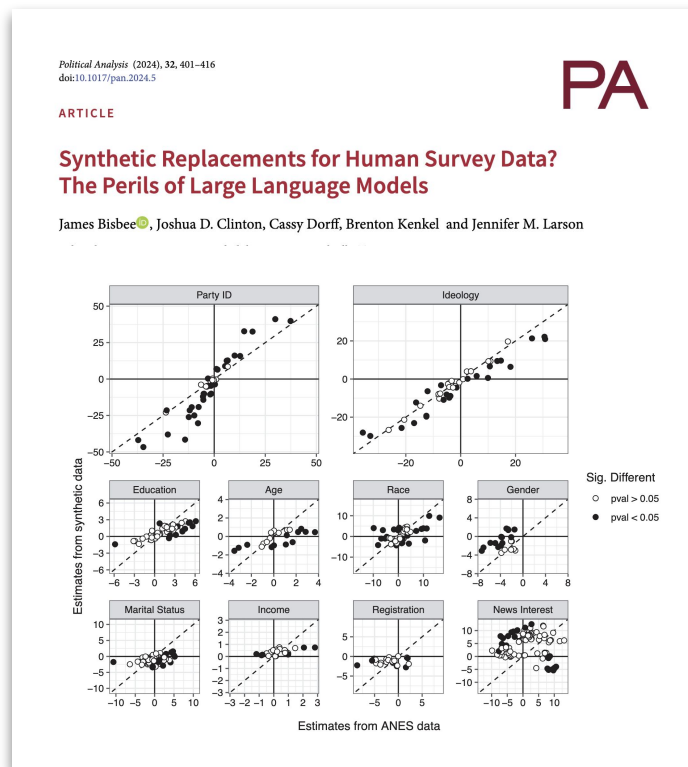
Tatsunori Hashimoto
Stanford
thashim@stanford.edu

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions reflected by LMs in response to subjective queries can have a profound impact, both on user satisfaction, as well as shaping the views of society at large. In this work, we put forth a quantitative framework to investigate the opinions reflected by LMs – by leveraging high-quality public opinion polls and their associated human responses. Using this framework, we create OpinionQA, a new dataset for evaluating the alignment of LM opinions with those of 60 US demographic groups over topics ranging from abortion to automation. Across topics, we find substantial misalignment between the views reflected by current LMs and those of US demographic groups: on par with the Democrat-Republican divide on climate change. Notably, this misalignment persists even after explicitly steering the LMs towards particular demographic groups. Our analysis not only confirms prior observations about the left-leaning tendencies of some human feedback-tuned LMs, but also surfaces groups whose opinions are poorly reflected by current LMs (e.g., 65+ and widowed individuals). Our code and data are available at https://github.com/tatsu-lab/opinions_qa.

Common Approaches to Heuristic Validation

- Face validity
- Individual-level alignment
- (Sub-)population-level distributions
- **Selected downstream regr. coef.**
(e.g., [Bisbee et al., 2024](#))
- Typically the most relevant for substantive social science research
- How are hypotheses converted into statistical models? (specification multiverse)
- Selection of estimands vs. unknown effects and secondary use of survey data



Survey Data Imputation (Holtdirk et al., 2026)

- Survey data is often partially missing
- Data imputation is an extensively studied issue (Rubin, 1987; Little and Rubin, 2019)
- Takes imputation uncertainty into account
- Assumptions about missingness mechanisms
- Computationally expensive, especially for LLM simulations

In-Context Learning for the Imputation of Public Opinion Data with Large Language Models

Tobias Holtdirk^{1,2}, Georg Ahnert³, Joseph W Sakshaug^{4,1,3}, Anna-Carolina Haensch^{1,2,5}

¹LMU Munich; ²Munich Center for Machine Learning; ³University of Mannheim;

⁴Institute for Employment Research (IAB); ⁵University of Maryland, College Park

Abstract

Large language models have been widely evaluated as simulators of individual survey responses. In practice, however, fully unobserved responses are rare; the dominant problem is partial non-response. *Imputation* aims to restore the overall structure of a survey dataset by filling in these missing values. It has its own well-defined evaluation criteria and differs fundamentally from prediction. We propose to impute missing survey data through in-context learning (ICL). We systematically evaluate ICL design choices across different missingness mechanisms (MCAR, MAR, MNAR) on 150 opinion variables spanning 15 waves of the American Trends Panel. Compared to well-established statistical methods for data imputation like MICE PMM, our ICL approach consistently reduces absolute error across all missingness mechanisms, with the largest gains under non-random missingness (MNAR). Notably, the best-performing specification (gpt-oss-120b with 100 in-context examples) achieves near-nominal aggregate coverage (approaching the 95% level) with confidence intervals two to five times narrower than MICE PMM. We publish a Python package with an sklearn-like API to enable easy deployment of our method using local and proprietary LLMs.

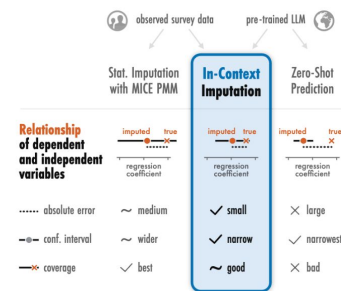


Figure 1: **In-context imputation combines the model's prior knowledge with evidence from the observed data.** In-context imputation (middle) produces confidence intervals substantially narrower than MICE PMM (left) at acceptable coverage, and is less biased than the zero-shot LLM (right).

Imputation aims to restore the overall data structure by filling in (*imputing*) the missing values (Little and Rubin, 2019). If left untreated or treated incorrectly, missing data can lead to biased estimates and reduced efficiency in data collection (Hoxen-

Evaluation Pipeline for Data Imputation

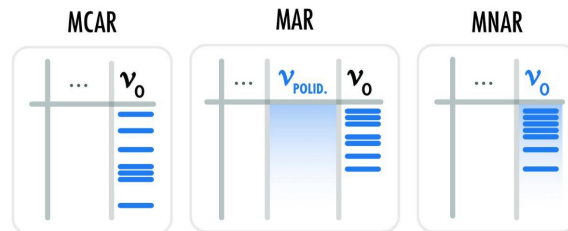
Public opinion data

Separate datasets $\mathcal{D}_{\text{true}}$ for each of the 150 selected variables v_o

	$v_{A1} \dots v_{An}$	$v_{\text{POLID.}}$	v_o
...	4 1	0	2
...	...	...	...

1. Draw 500 complete cases from the dataset $\mathcal{D}_{\text{true}}$

2. Simulate missingness pattern with 50% missing values



3. Impute missing values

ICL Imputation



5 samples

→ $\mathcal{D}_{\text{imputed}}$

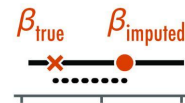
4. Fit OLS regressions

separately on $\mathcal{D}_{\text{true}}$ and $\mathcal{D}_{\text{imputed}}$

$$v_{\text{POLIDEOLGY}} = \alpha + \beta v_o + \varepsilon$$

5. Evaluate regression coefficients

..... absolute error
 —●— conf. interval
 —×— coverage



Holtdirk, T., Ahnert, G., Sakshaug, J. W., & Haensch, A. C. (2026). In-Context Learning for the Imputation of Public Opinion Data with Large Language Models. *arXiv preprint [arXiv:2606.09351](https://arxiv.org/abs/2606.09351)*.

Results

Across Imputation Methods

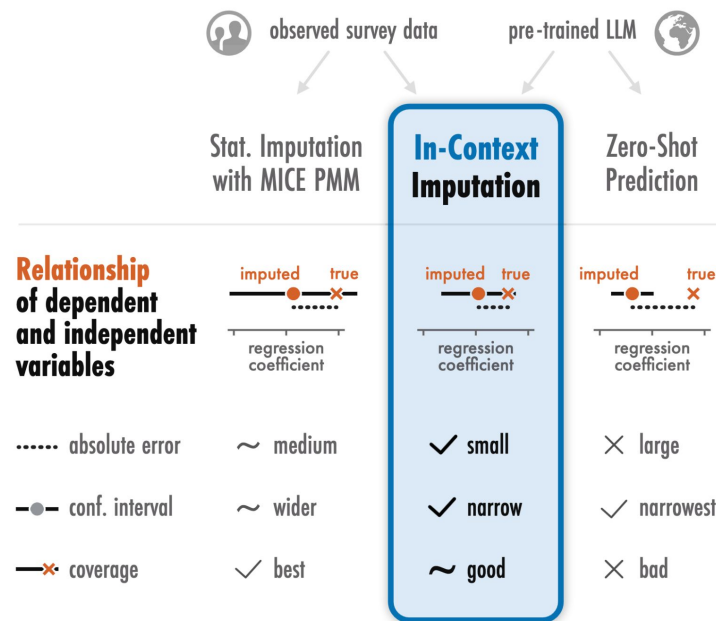


Figure 1: **In-context imputation combines the model’s prior knowledge with evidence from the observed data.** In-context imputation (middle) produces confidence intervals substantially narrower than MICE PMM (left) at acceptable coverage, and is less biased than the zero-shot LLM (right).

Fit for Purpose Evaluation

Imputation $\neq$ Prediction

Holtdirk, T., Ahnert, G., Sakshaug, J. W., & Haensch, A. C. (2026).
In-Context Learning for the Imputation of Public Opinion Data with
Large Language Models. *arXiv preprint [arXiv:2606.09351](https://arxiv.org/abs/2606.09351)*.

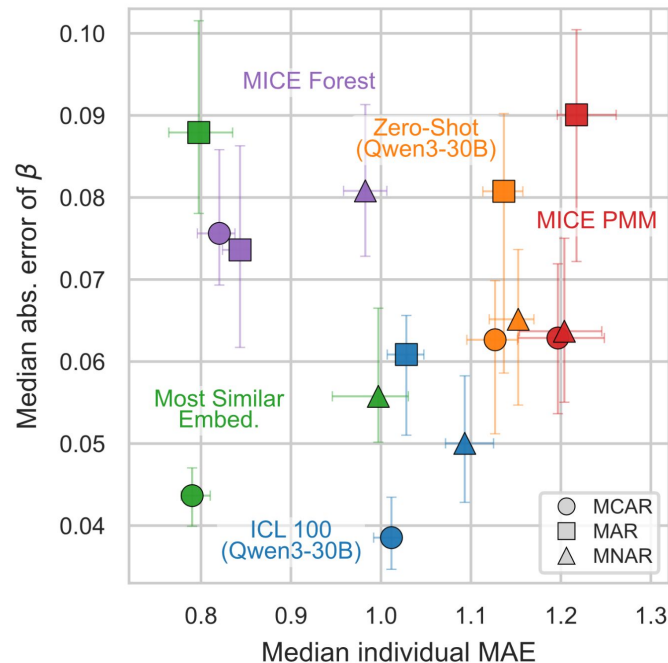


Figure 4: **Individual-level prediction (x-axis) and coefficient-level inference (y-axis) are only loosely related.** Each marker is one (method, mechanism) pair from Study 2; both axes ($\downarrow$) are variable-medians with 95% bootstrap CIs on the standardized v_o scale (Section 2.2). The x-axis is $|\hat{v}_o - v_o^{\text{true}}|$ averaged over imputed cells and the m draws.

Threats to Heuristic Evaluation

- **Model Training**

- data leakage
- memorization
- lack of generalization
- “potemkin understanding”

- **Human survey responses**

- subject to measurement & representation errors

- **Scientific Process**

- Specification multiverse encourages *LLM hacking* ([Baumann et al., 2025](#))
- Underspecified applications

Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation

Joachim Baumann¹, Paul Röttger¹, Aleksandra Urman², Albert Wendsjö³,
Flor Miriam Plaza-del-Arco⁴, Johannes B. Gruber⁵, and Dirk Hovy¹

We replicate 37 social science
data annotation tasks.

ideology	topic	emotion
frames	stance	politeness
tone	issue	hate speech
misinformation	...	
2,361 realistic hypotheses		

LLMs take over the role of
data annotators.

18 large language models
13 million annotations 

We test the **validity of
scientific conclusions.**

1.4 million regressions

LLM annotations are unreliable
for downstream statistical analyses!

Task: Is manifesto <i>left</i> or <i>right</i> ?			
Ground truth: <i>no effect</i> ($p > 0.05$) 			
 LLM 1, config 1	$p = 0.242$		✓
 LLM 1, config 2	$p = 0.001$		✗
 LLM 2, config 1	$p = 0.424$		✓
...	...		...
LLM hacking risk = 31% – 50%			

Mitigating the Threats (Hullman et al., 2026)

- “Simulate-then-validate”

- Pre-testing
- Effect size estimation
- ...

- Rectification with Prediction Powered Inference (PPI)

- Statistical guarantees through power-tuning
- Also addresses LLM hacking

This human study did not involve human subjects:
Validating LLM simulations as behavioral evidence

Jessica Hullman
Northwestern University

David Broska
Stanford University

Huaman Sun
Northwestern University

Aaron Shaw
Northwestern University

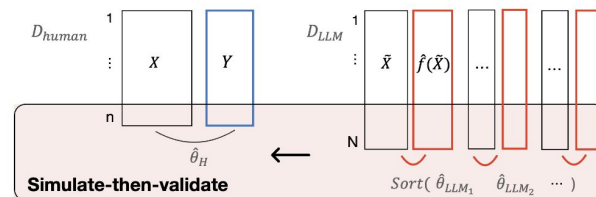


Figure 3: In a simulate-then-validate approach, the researcher uses LLMs exclusively for exploratory piloting to find hypotheses with support, then conducts a human study on those that appear most promising.

Do you have questions?

Next:

**[github.com/dess-mannheim
/tutorial_survey_simulation](https://github.com/dess-mannheim/tutorial_survey_simulation)**

