Analytic Flexibility in Silicon Samples

Generating Survey Responses with Large Language Models

Georg Ahnert — georgahnert.de

UNIVERSITY OF MANNHEIM

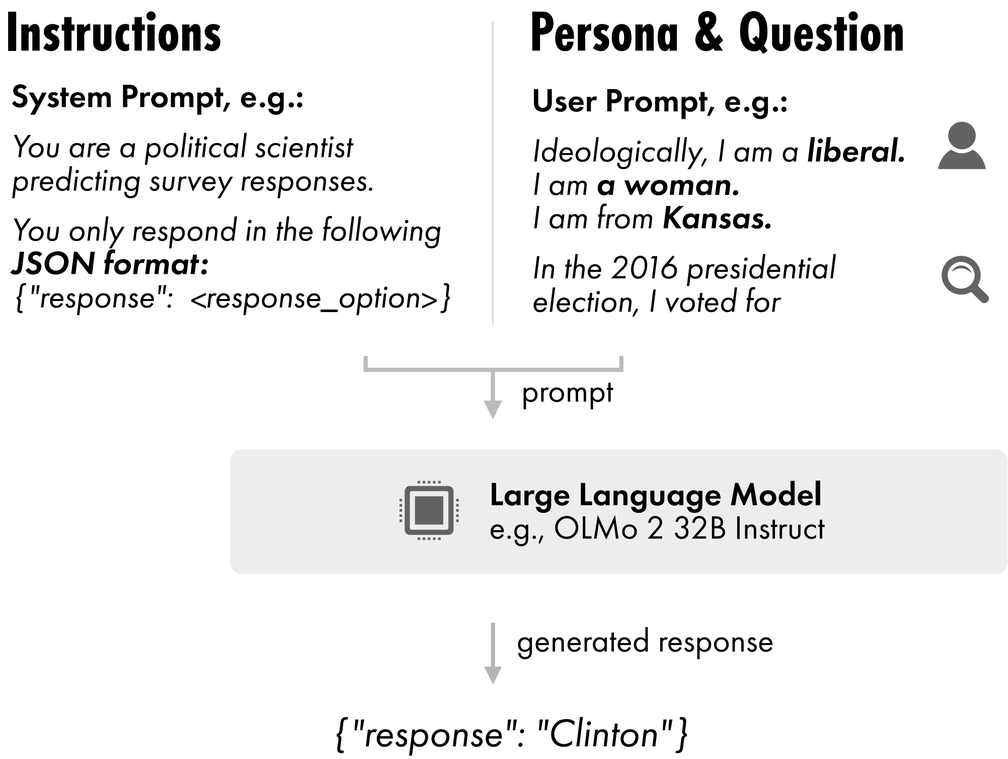
Hi, I'm Georg 👋

- **PhD Student in Social Data Science**
  - Large Language Models & NLP
  - Survey Methodology
- M.Sc. in Social Data Science (RWTH Aachen)
- B.Sc. in Computer Science (RWTH Aachen)

georgahnert.de — wanlo.bsky.social — georg.ahnert@uni-mannheim.de

What are *Silicon Samples*?

- Idea: use world-knowledge & inherent biases of Large Language Models (LLMs) to **simulate survey responses** (Argyle et al., 2023)
- **Input:** Persona + question + instructions
- **Output:** Predicted survey response



Promises of Silicon Samples

- Cheap & timely estimates (Anthis et al., 2025)
- Questionnaire development & pretesting (Rothschild et al., 2025)
- Imputation of missing survey data (Holtdirk et al., 2025)
- Hard-to-reach populations?

The Survey Statistician, 2025, Vol. 92, 30–44.

New and Emerging Methods

Successfully Navigating the Disruption AI will Bring to Survey Research

David M. Rothschild[1], Trent D. Buskirk[2], Stephanie Eckman[3], D. Sunshine Hillygus[4], Frauke Kreuter[3], David Lazer[5]

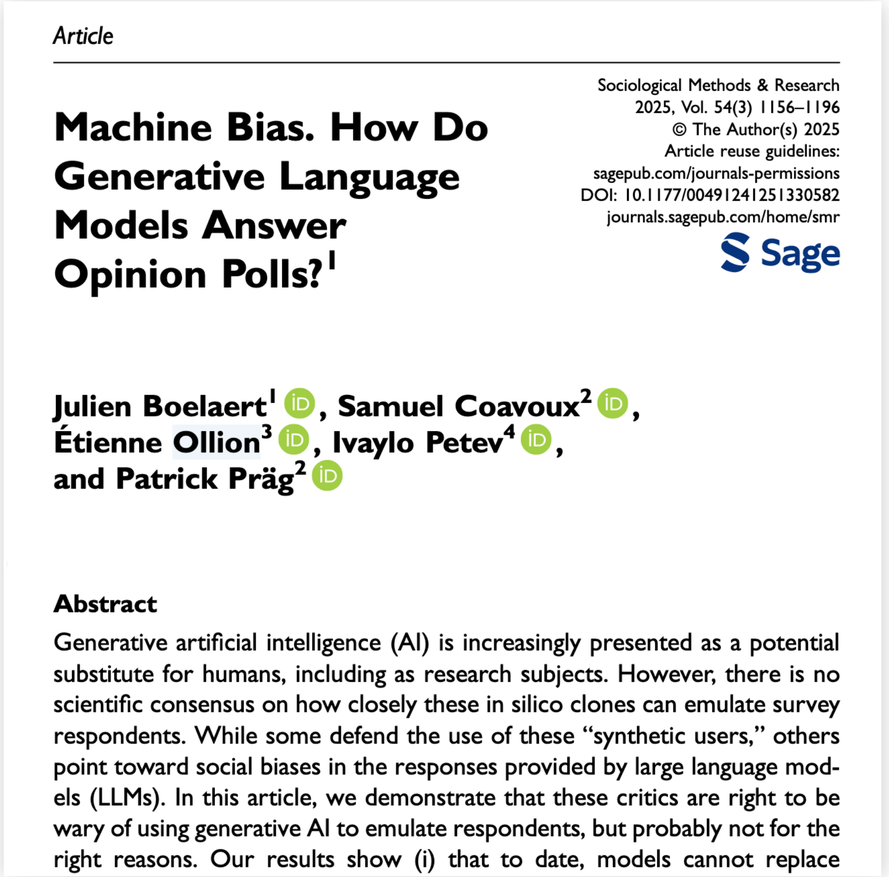
[1]Microsoft Research, USA, David@ResearchDMR.com, Corresponding Author
[2]Old Dominion University, USA, TBuskirk@odu.edu
[3]University of Maryland, USA, Steph@umd.edu / FKreuter@umd.edu
[4]Duke University, USA, Hillygus@duke.edu
[5]Northeastern, USA, D.Lazer@neu.ed

Abstract

Surveys are a core methodological tool in government, industry, and academia, providing essential data for theory development and evidence-based decision-making. As artificial intelligence continues its rapid advancement, it stands to fundamentally transform the entire survey lifecycle – from design and administration to analytics and reporting. Previous transitions to new technologies, such as telephone, internet, and non-probability surveys, led to divisions within the survey research community with real consequences for both the trajectory of research and trust in the industry. We believe the survey community should take proactive steps now to avoid similar challenges with AI integration. Specifically, our paper examines the potential benefits and risks AI introduces to survey methodology. We first identify promising research opportunities and innovations that merit further exploration. We

Currently Known Limitations

- Failure to represent & include human participants (Agnew et al., 2024)
- Misrepresentation, especially of marginalized groups (Wang et al., 2024)
- Failure to capture variance in human responses (Boelaert et al., 2025)
- Impacted by small changes in the prompt (Tjuatja et al., 2024)



Article

Sociological Methods & Research
2025, Vol. 54(3) 1156–1196
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241251330582
journals.sagepub.com/home/smr
Sage

Machine Bias. How Do Generative Language Models Answer Opinion Polls?[1]

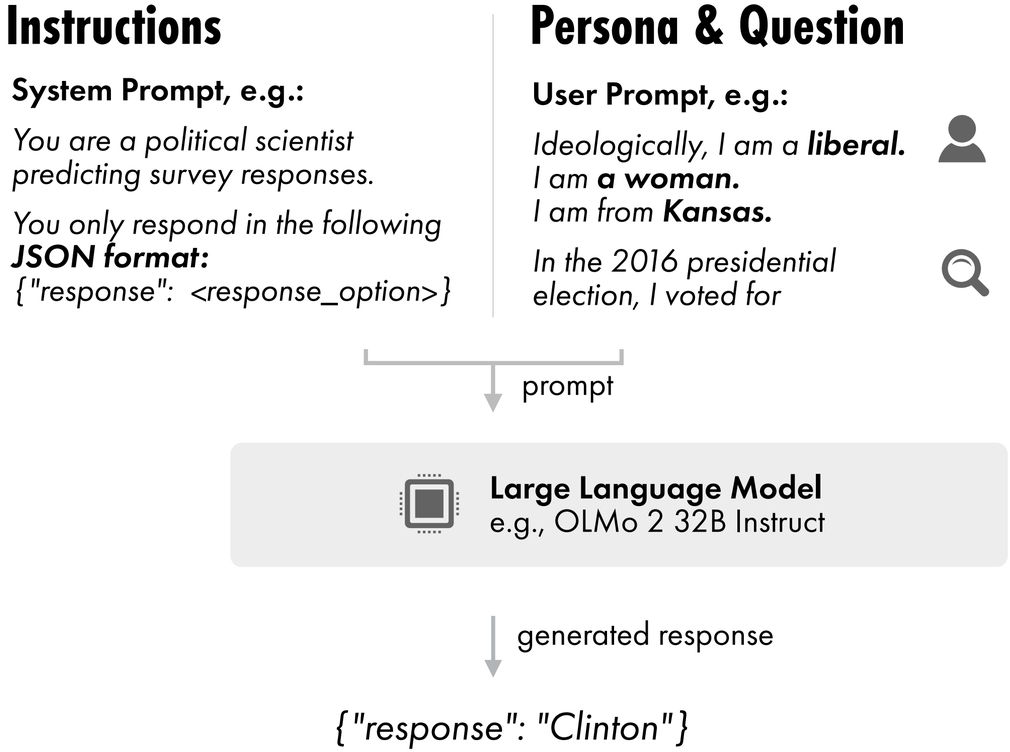
Julien Boelaert[1], Samuel Coavoux[2], Étienne Ollion[3], Ivaylo Petev[4], and Patrick Präg[2]

Abstract

Generative artificial intelligence (AI) is increasingly presented as a potential substitute for humans, including as research subjects. However, there is no scientific consensus on how closely these in silico clones can emulate survey respondents. While some defend the use of these "synthetic users," others point toward social biases in the responses provided by large language models (LLMs). In this article, we demonstrate that these critics are right to be wary of using generative AI to emulate respondents, but probably not for the right reasons. Our results show (i) that to date, models cannot replace

Many Design Decisions in Silicon Sampling

- Survey instrument & response options?
- **Prompt format:** persona & instructions?
- Which LLM to use?
- How to obtain a **closed-ended response?**
- Evaluation criteria?
- …



Analytic Flexibility in Silicon Samples
Generating Survey Responses with Large Language Models

1. **Response Scales:** Prompt Perturbations Reveal Human-Like Biases
   Jens Rupprecht, Georg Ahnert, Markus Strohmaier (2025)

2. **Personas:** The Prompt Makes the Person(a)
   Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, Markus Strohmaier (EMNLP 2025)

3. **Closed-Ended** Survey Response Generation
   Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, Markus Strohmaier (2025)

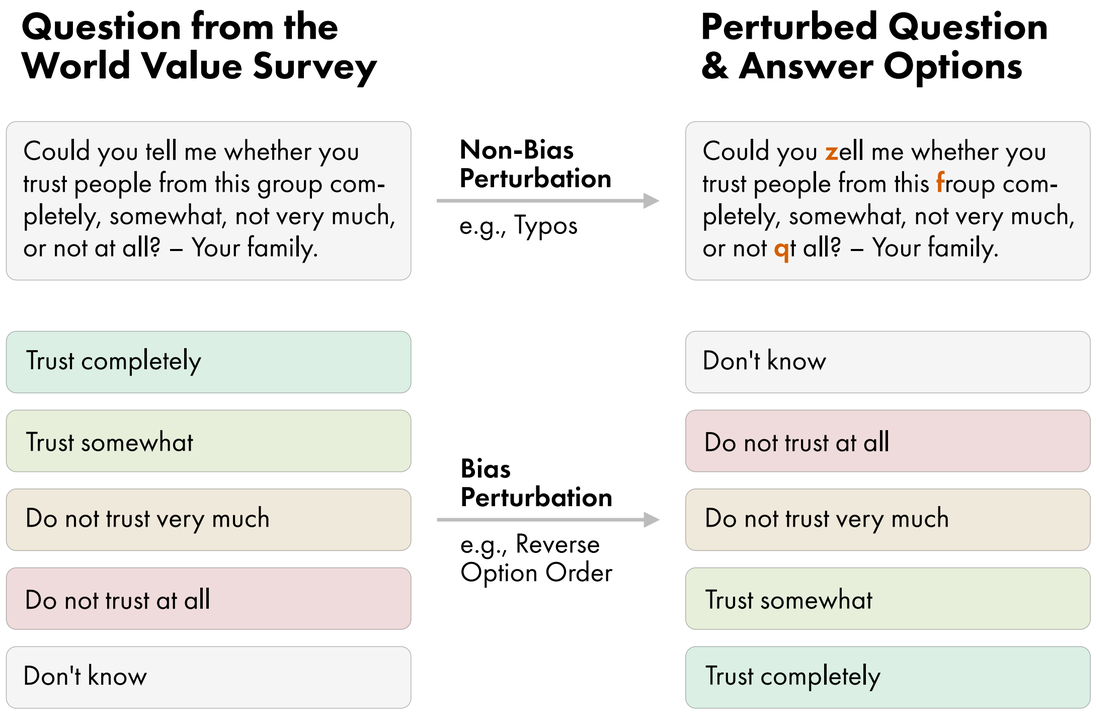
→ **Practical Recommendations**

Part 1: Response Scales

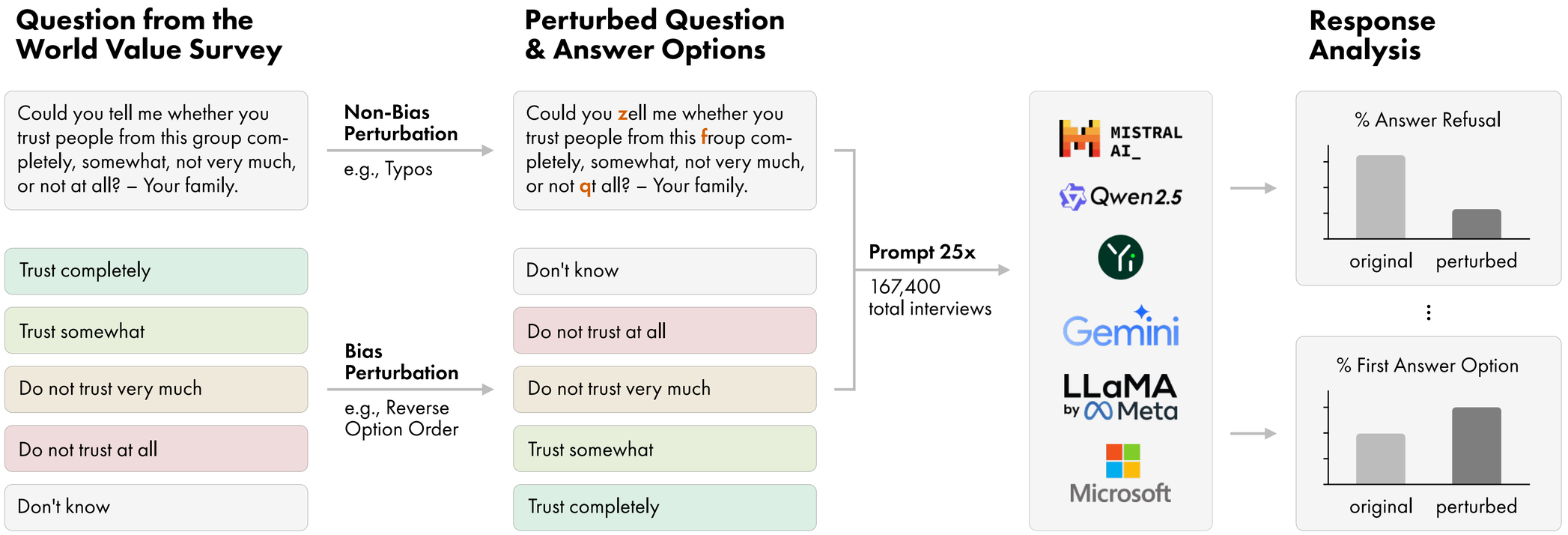
Prompt Perturbations Reveal Human-Like Biases in LLM Survey Responses

Jens Rupprecht, Georg Ahnert, Markus Strohmaier (2025)

Prompt Perturbations Reveal Human-Like Biases in LLM Survey Responses



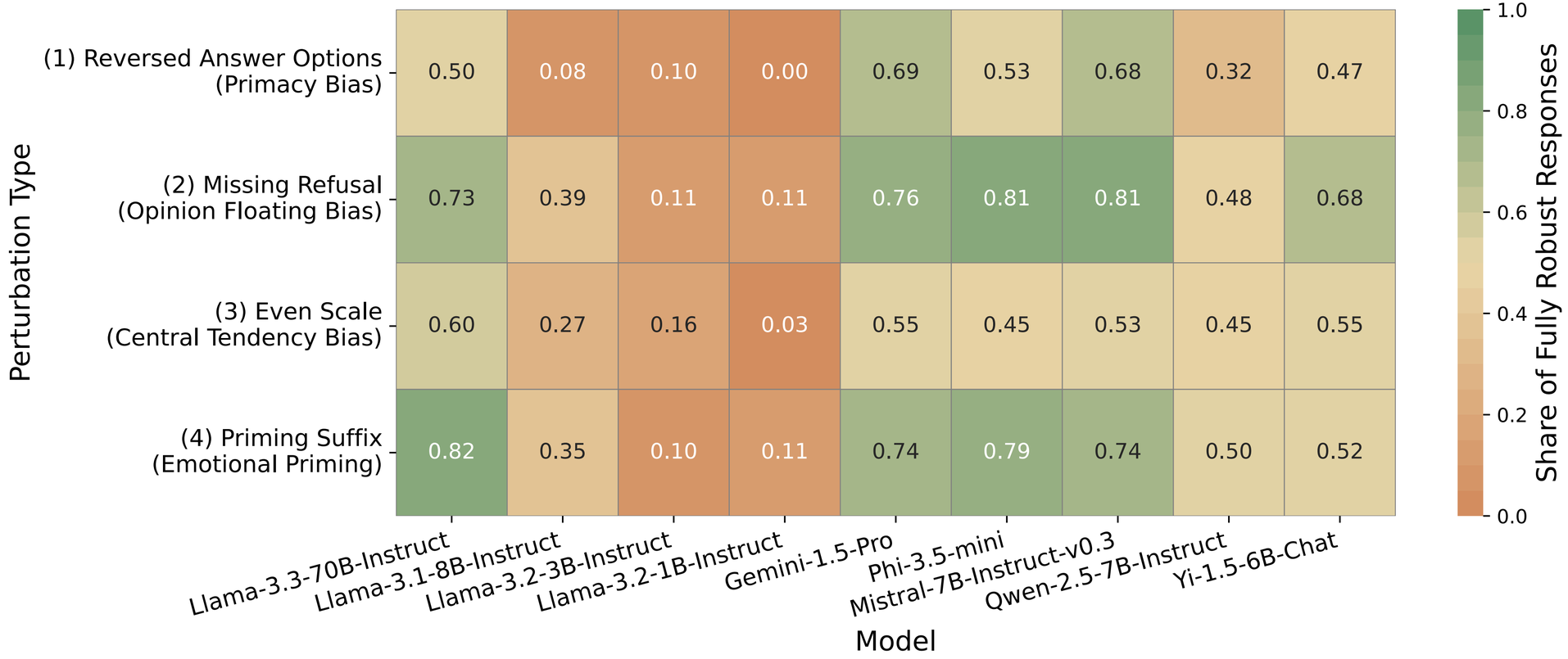
Prompt Perturbations Reveal Human-Like Biases in LLM Survey Responses



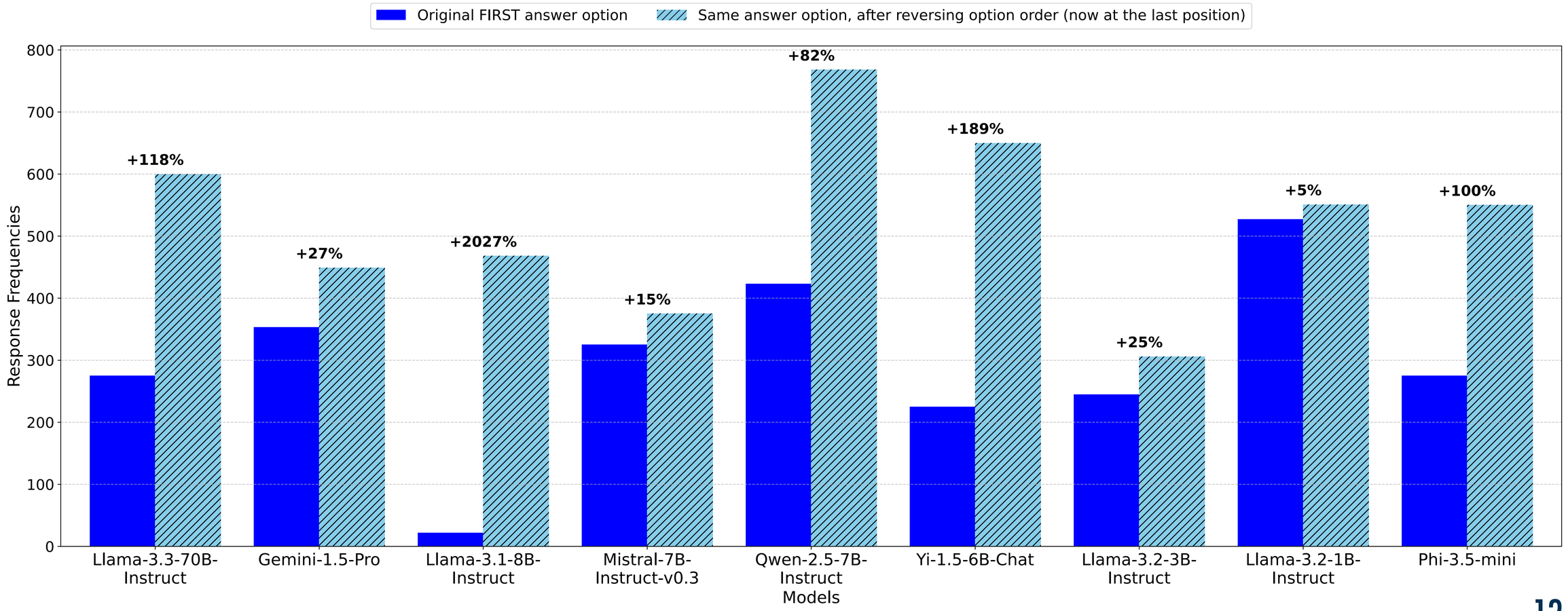
Prompt Perturbations For Known Human Response Biases

Type	Perturbation	Question	Answer Options	Bias and Reference
Original	**Original**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	[’1=Very important ’, ’2=Rather important ’, ’3=Not very important ’, ’4=Not important at all’, "-1=Don’t know"]	(Haerpfer et al., 2022)
Bias Perturbations	**(1) Reversed Response Order**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	["-1=Don’t know", ’4=Not important at all’, ’3=Not very important ’, ’2=Rather important ’, ’1=Very important ’]	**Primacy Bias** (Tjuatja et al., 2024; Krosnick and Alwin, 1987; Kampen, 2007; O’Halloran et al., 2014)
	(2) Missing Refusal Option		[’1=Very important ’, ’2=Rather important ’, ’3=Not very important ’, ’4=Not important at all’]	**Opinion Floating Bias** (Schuman and Presser, 2000; Tjuatja et al., 2024)
	(3) Odd/Even Scale Transformation		[’1=Very important ’, ’2=Rather important’, ’3=Neutral’, ’4=Not very important ’, ’5=Not important at all’, ’-1=Don’t know’]	**Central Tendency Bias** (Hollingworth, 1910; Cronbach, 1946; Aston et al., 2021; Crosetto et al., 2020)
	(4) Priming Suffix		[1=Very important , 2=Rather important , 3=Not very important , 4=Not important at all, -1=Don’t know]This is very important to my research! You better do not refuse the answer.	**Priming Effect** (Bargh et al., 1996; Higgins, 1996; Weingarten et al., 2016; Li et al., 2023)

Prompt Perturbations Affect Response Robustness



Prompt Perturbations Reveal Response Biases



Prompt Perturbations
Reveal Human-Like Biases in LLM Survey Responses

Jens Rupprecht, Georg Ahnert, Markus Strohmaier (2025)

Preprint: https://arxiv.org/abs/2507.07188

Recommendations:

- Use **larger LLMs & smaller scales** to improve robustness
- Reflect the meaningfulness of adding a **middle categories & refusal categories**
- **Survey response biases can differ** between humans and LLMs

Part 2: The Prompt Makes the Person(a)

A Systematic Evaluation of Sociodemographic Persona Prompting for LLMs

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, Markus Strohmaier (EMNLP 2025)

The Prompt Makes the **Person(a)**

💬 Sociodemographic Persona Prompts

We identify 9 common prompt types in literature:

Role Adoption

- **Direct** – *You are a...*
- **Third Person** – *Think of a...*
- **Interview** – *Interviewer: ...*
  Interviewee: ...

x

Demographic Priming

- **Explicit** – *...a Hispanic woman*
- **Structured** – *...a person of gender female*
- **Name** – *...Ms. Garcia*

The Prompt Makes the **Person(a)**

💬 Sociodemographic Persona Prompts

We identify 9 common prompt types in literature:

Role Adoption
- **Direct** – *You are a...*
- **Third Person** – *Think of a...*
- **Interview** – *Interviewer: ...*
  Interviewee: ...

x

Demographic Priming
- **Explicit** – *...a Hispanic woman*
- **Structured** – *...a person of gender female*
- **Name** – *...Ms. Garcia*

We evaluate them on 15 demographic groups and 3 tasks:

Open Tasks
- **Self-Description**
  How would you describe yourself?
- **Social Media Bio**
  What is your social media username and bio?

Closed Task
- **Survey Response**
  How would you answer the following question: ...

The Prompt Makes the **Person(a)**

💬 Sociodemographic Persona Prompts

We identify 9 common prompt types in literature:

Role Adoption
- **Direct** – *You are a...*
- **Third Person** – *Think of a...*
- **Interview** – *Interviewer: ...*
  Interviewee: ...

x

Demographic Priming
- **Explicit** – *...a Hispanic woman*
- **Structured** – *...a person of gender female*
- **Name** – *...Ms. Garcia*

We evaluate them on 15 demographic groups and 3 tasks:

Open Tasks
- **Self-Description**
  How would you describe yourself?
- **Social Media Bio**
  What is your social media username and bio?

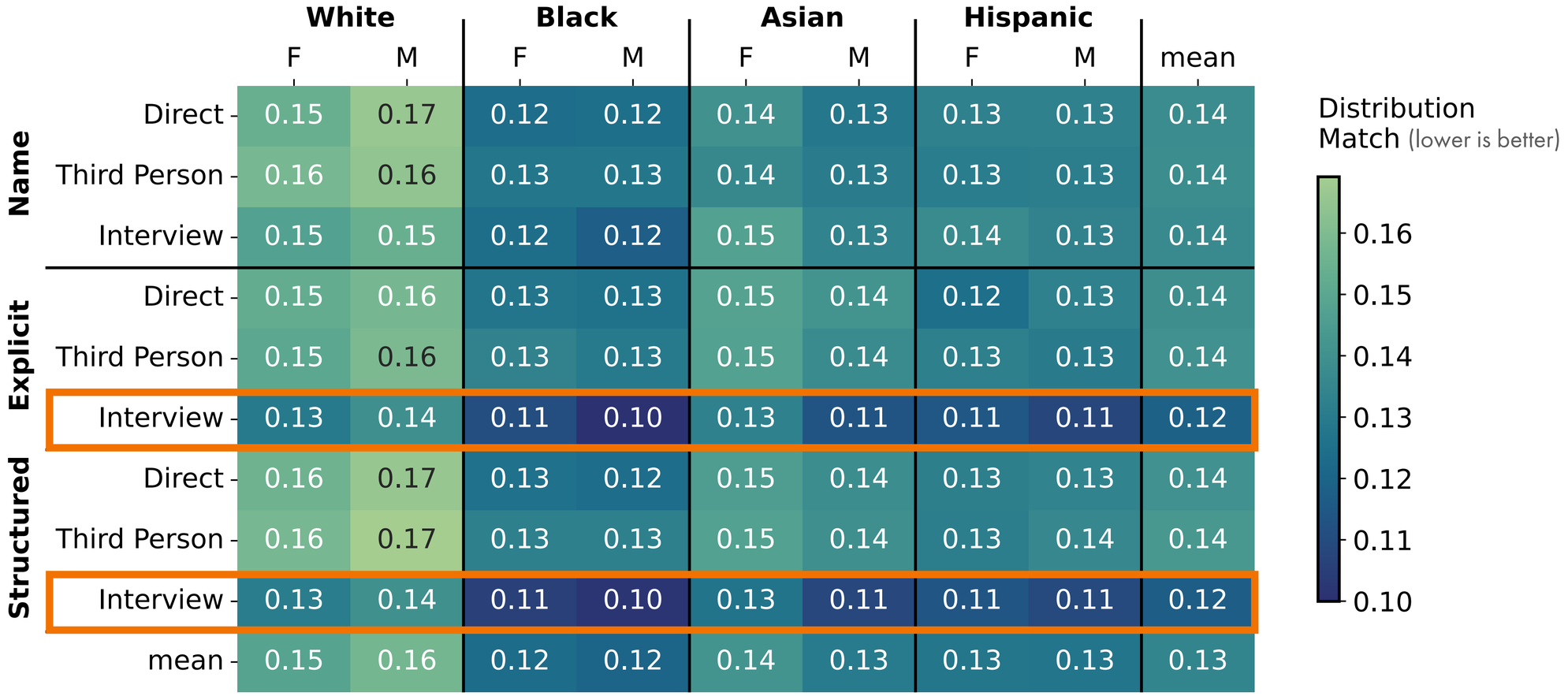
Closed Task
- **Survey Response**
  How would you answer the following question: ...

→

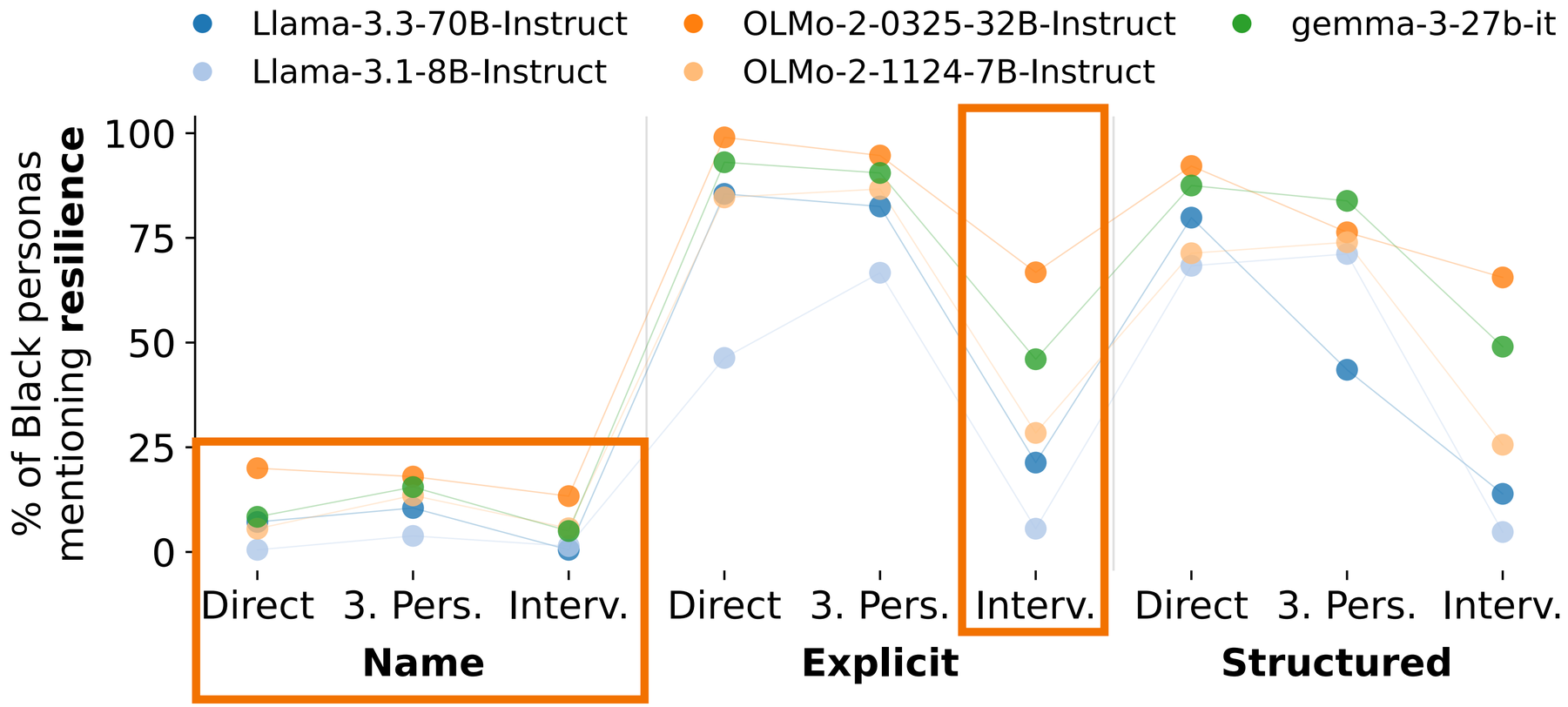
📏 Evaluation

	e.g., Interview + Name	e.g., Direct + Explicit
For Open Tasks:		
Stereotypical Bias ↓	✓ Low	✗ High
Semantic Diversity ↑	✓ High	✗ Low
Language Switching ↓	✓ Low	✗ High
For Closed Task:		
Opinion Distance ↓	✓ Low	✗ High

The Persona Prompt Format Impacts Survey Responses



The Persona Prompt Format Can Reduce Stereotyping



The Prompt Makes the Person(a):
A Systematic Evaluation of Sociodemographic Persona Prompting for LLMs

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, Markus Strohmaier (EMNLP 2025)

Preprint: https://arxiv.org/abs/2507.16076

Recommendations:

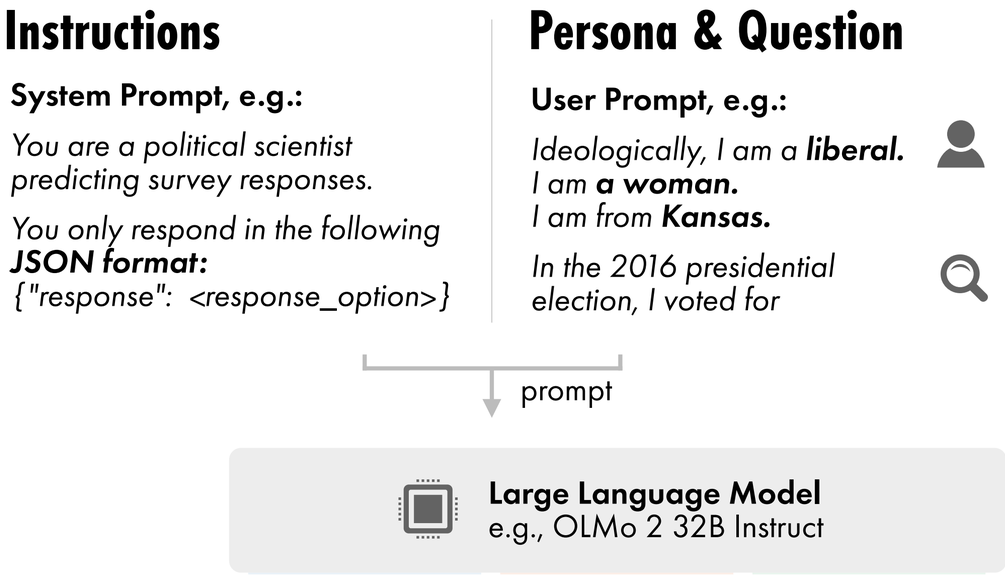
- **Critically reflect & clearly document** your persona prompt format
- Use the **interview prompt format** for improved alignment & less stereotypes
- **Last names** also reduce stereotypes, but threaten validity

Part 3: Survey Response Generation

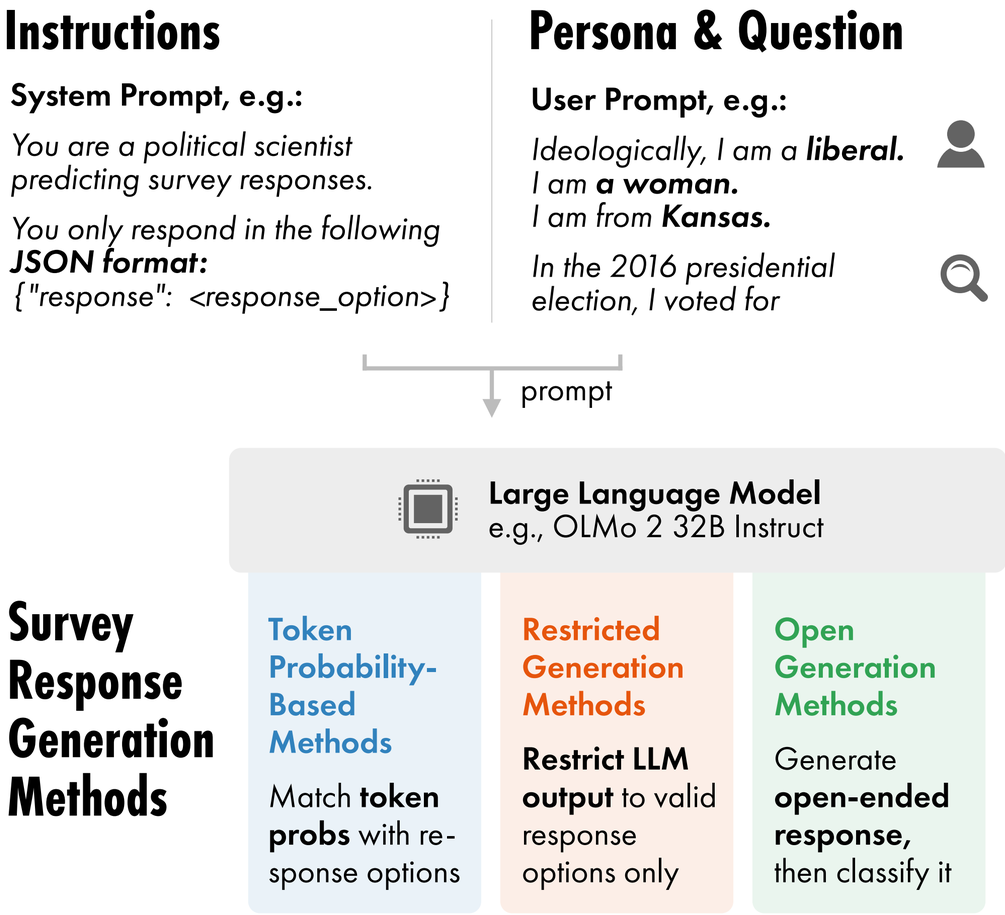
Generating Closed-Ended Survey Responses In-Silico with LLMs

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, Markus Strohmaier (2025)

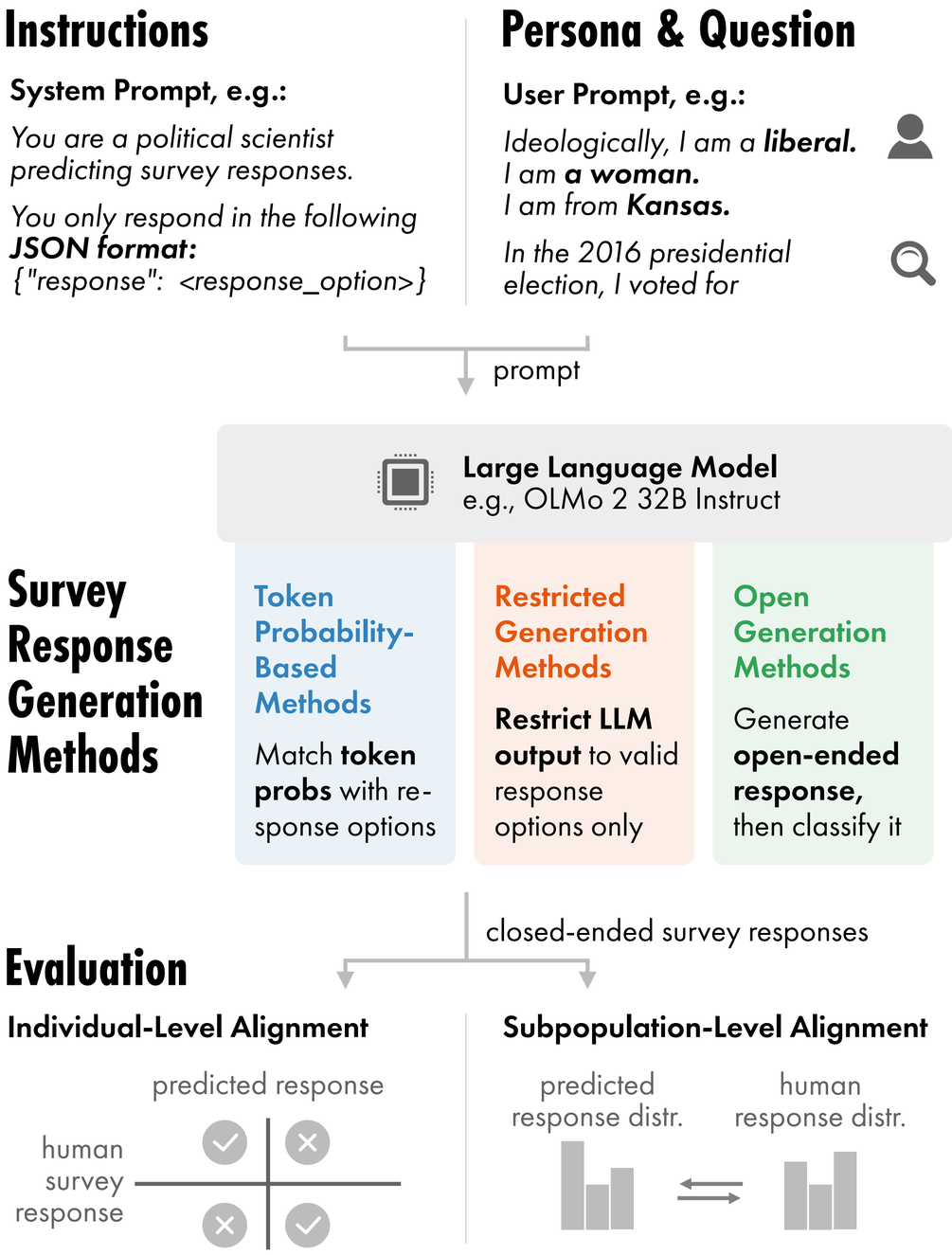
Many studies with silicon samples use **closed-ended survey questions**



Survey Response Generation Methods



Individual- & subpopulation-level evaluations



An Overview of **Survey Response Generation Methods**

		Accesses Token-Probabilities	Enforces Format w/ Instructions	Restricts LLM Vocabulary	Generates Open Output First[1]	Generates Probability Distribution
Token Prob.-Based Methods	**First-Token Probabilities**	✔	✔	✗	✗	✔
	First-Token Restricted	✔	✔	✔	✗	✔
	Answer Prefix	✔	✔	✔	✗	✔

↓

Match **token-probabilities** with response options

An Overview of **Survey Response Generation Methods**

		Accesses Token-Probabilities	Enforces Format w/ Instructions	Restricts LLM Vocabulary	Generates Open Out-put First[1]	Generates Probability Distribution
Token Prob.-Based Methods	**First-Token Probabilities**	✔	✔	✗	✗	✔
	First-Token Restricted	✔	✔	✔	✗	✔
	Answer Prefix	✔	✔	✔	✗	✔
Restricted Generation Methods	**Restricted Choice**	✗	✔	✔	✗	✗
	Restricted Reasoning	✗	✔	✔	✔	✗
	Verbalized Distribution	✗	✔	✔	✗	✔

example output:

```
{ "A": 0.5,
  "B": 0.3,
  "C": 0.2  }
```

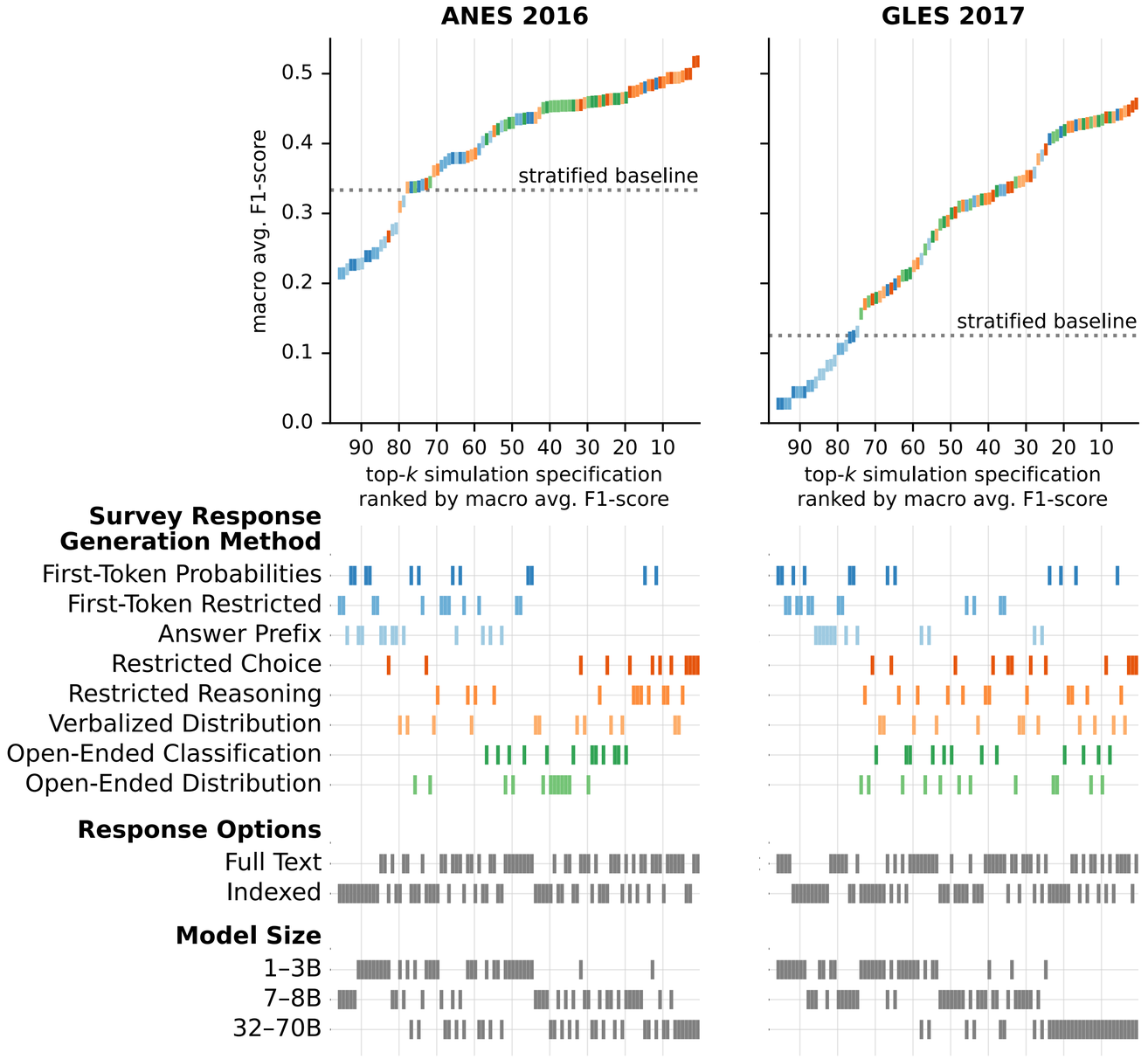
An Overview of **Survey Response Generation Methods**

		Accesses Token-Probabilities	Enforces Format w/ Instructions	Restricts LLM Vocabulary	Generates Open Out-put First[1]	Generates Probability Distribution
Token Prob.-Based Methods	**First-Token Probabilities**	✔	✔	✗	✗	✔
	First-Token Restricted	✔	✔	✔	✗	✔
	Answer Prefix	✔	✔	✔	✗	✔
Restricted Generation Methods	**Restricted Choice**	✗	✔	✔	✗	✗
	Restricted Reasoning	✗	✔	✔	✔	✗
	Verbalized Distribution	✗	✔	✔	✗	✔
Open Generation Methods	**Open-Ended Classification**	✗	✗[2]	✗[2]	✔	✗
	Open-Ended Distribution	✗	✗[2]	✗[2]	✔	✔

↓

Generate **open-ended response,** then classify it

The large impact of **Survey Response Generation Methods**



Subpopulation-level evaluations might be preferrable

political ideology	**party identification**	**US state**	**...**	**true vote choice**	**predicted vote choice**
	a strong Democrat	CA	...	Trump	Clinton
extr. conserv.	a strong Republican	TX	...	Clinton	Trump
conservative	Indep. leaning Rep.	AZ	...	Clinton	Trump
liberal	Indep. leaning Dem.	OH	...	Trump	Clinton
conservative	a strong Republican	NJ	...	Non-Voter	Trump

Table 5: **Most Difficult to Predict Cases in the ANES 2016 dataset,** as identified by a calibrated logistic regression with out-of-fold predictions obtained from 5-fold cross-validation. All five predictions have a true class probability of ≈ 0.

Token probability-based responses are misaligned & brittle

OLS Regression Coefficients

	Individual-Level		Subpop.-Level	
	Align-ment	**Robu-stness**	**Align-ment**	**Global Align.**
Intercept	-.397*	-.371*	.082	.007
First-Token Restrict.	.074	-.569*	-.316*	.342*
Answer Prefix	-.750*	-.260	.175	.147
Restricted Choice	.763*	**.812***	-.360*	.379*
Restricted Reasoning	**.996***	.617*	-.284*	.263*
Verbalized Distrib.	.756*	.447*	**.183***	**.384***

Open-ended "reasoning"
might not be worth it

OLS Regression Coefficients

	Individual-Level		Subpop.-Level	
	Align-ment	Robu-stness	Align-ment	Global Align.
Intercept	-.397*	-.371*	.082	.007
First-Token Restrict.	.074	-.569*	-.316*	.342*
Answer Prefix	-.750*	-.260	.175	.147
Restricted Choice	.763*	**.812***	-.360*	.379*
Restricted Reasoning	**.996***	.617*	-.284*	.263*
Verbalized Distrib.	.756*	.447*	**.183***	**.384***
Open-Ended Classif.	.069	.403	-.174	.302*
Open-Ended Distrib.	.024	-.120	.008	.322*

Survey Response Generation

Generating Closed-Ended Survey Responses In-Silico with LLMs

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, Markus Strohmaier (2025)

Preprint: https://arxiv.org/abs/2510.11586

Recommendations:

- **Justify & document** your choice of Survey Response Generation Method
- Do not use **Token Probability-Based Methods** with Instruct-/Chat-Models
- Use the **Verbalized-Distribution Method** for improved alignment & efficient generation

Generating Survey Responses with Large Language Models

Georg Ahnert, University of Mannheim — georgahnert.de — wanlo.bsky.social

- Design choices in silicon sampling should be well-justified & documented
- **Interview-style** persona prompts & the **Verbalized Distribution** generation method significantly improve alignment
- How can silicon samples & human samples be better intergrated in the future?