Exploring the LLM Survey Simulation Pipeline with the QSTN Framework

Session 2, Jens Rupprecht & Maxi Kreutner

QSTN

A Modular Framework for Robust Questionnaire Inference with LLMs

QSTN: A Modular Framework for Robust Questionnaire Inference with Large Language Models

Maximilian Kreutner[1], Jens Rupprecht[1], Georg Ahnert[1], Ahmed Salem[1], Markus Strohmaier[1,2,3]
[1]University of Mannheim, [2]GESIS - Leibniz Institute for the Social Sciences, [3]CSH Vienna

Abstract

We introduce QSTN, an open-source Python framework for systematically generating responses from questionnaire-style prompts to support in-silico surveys and annotation tasks with large language models (LLMs). QSTN enables robust evaluation of questionnaire presentation, prompt perturbations, and response generation methods. Our extensive evaluation (>40 million survey responses) shows that question structure and response generation methods have a significant impact on the alignment of generated survey responses with human answers. We also find that answers can be obtained for a fraction of the compute cost, by changing the presentation method. In addition, we offer a no-code user interface that allows researchers to set up robust experiments with LLMs *without coding knowledge*. We hope that QSTN will support the reproducibility and reliability of LLM-based research in the future.

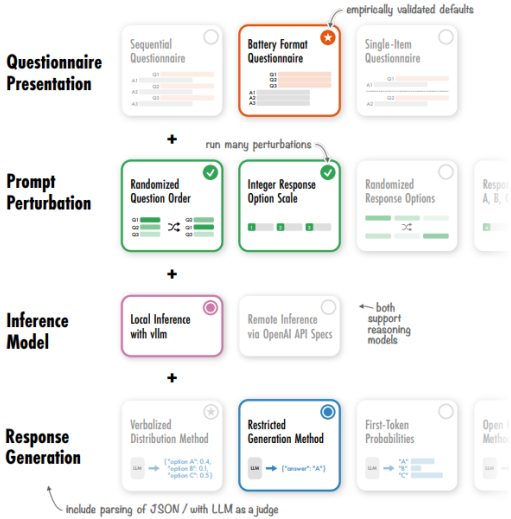


Figure 1: **QSTN Facilitates Easy To Customize and Robust Questionnaire Inference with LLMs.** QSTN

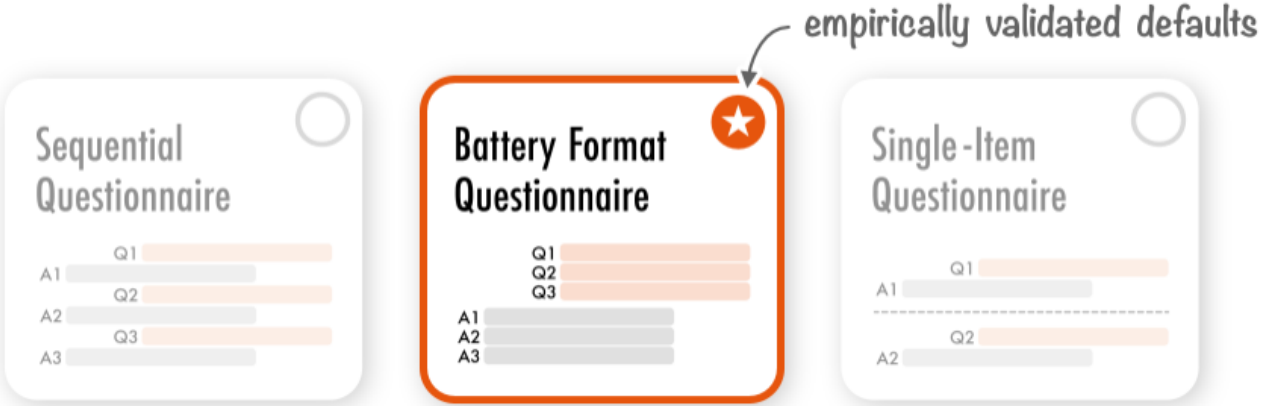
Code & Documentation:

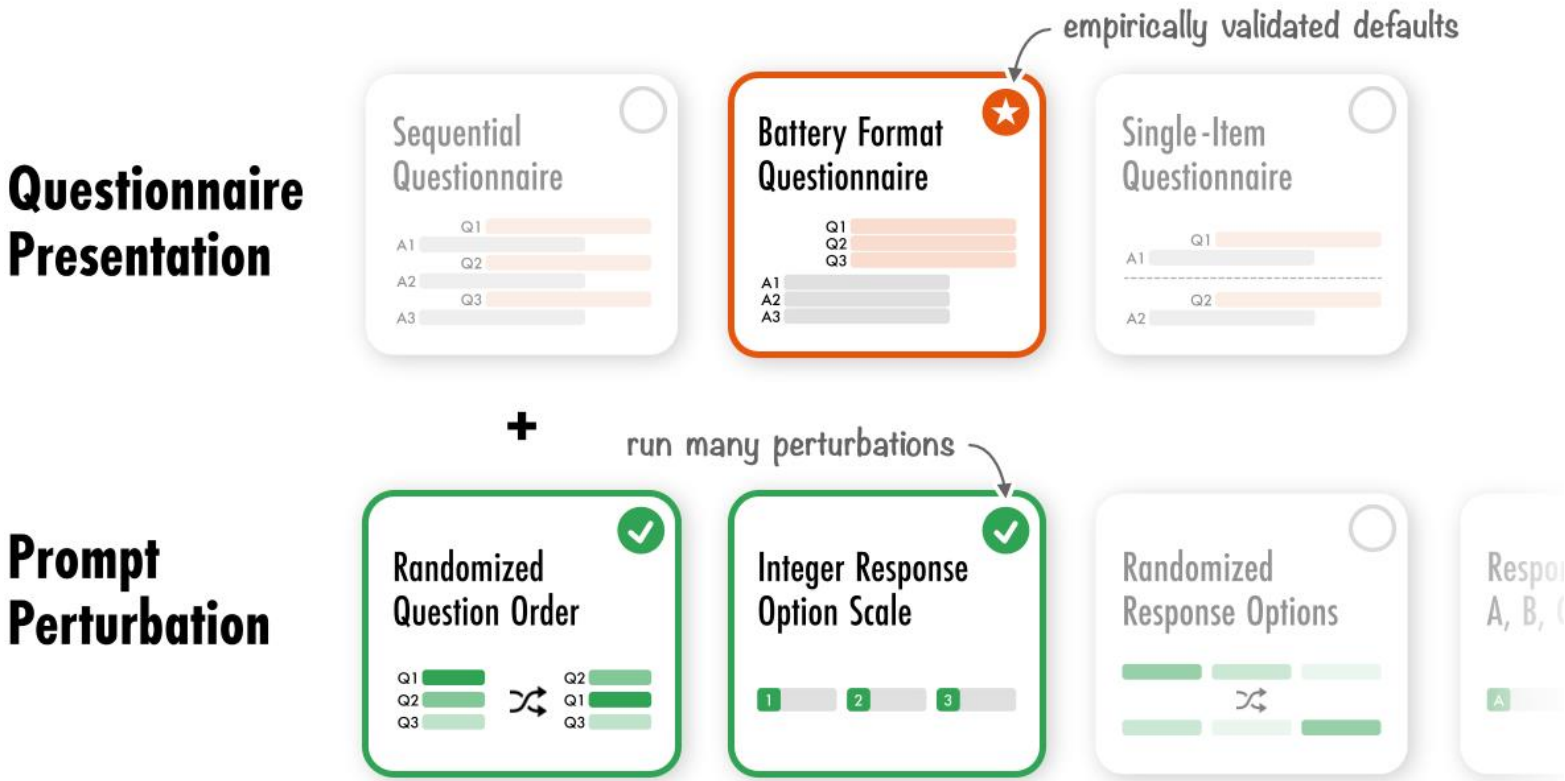
github.com/dess-mannheim/QSTN

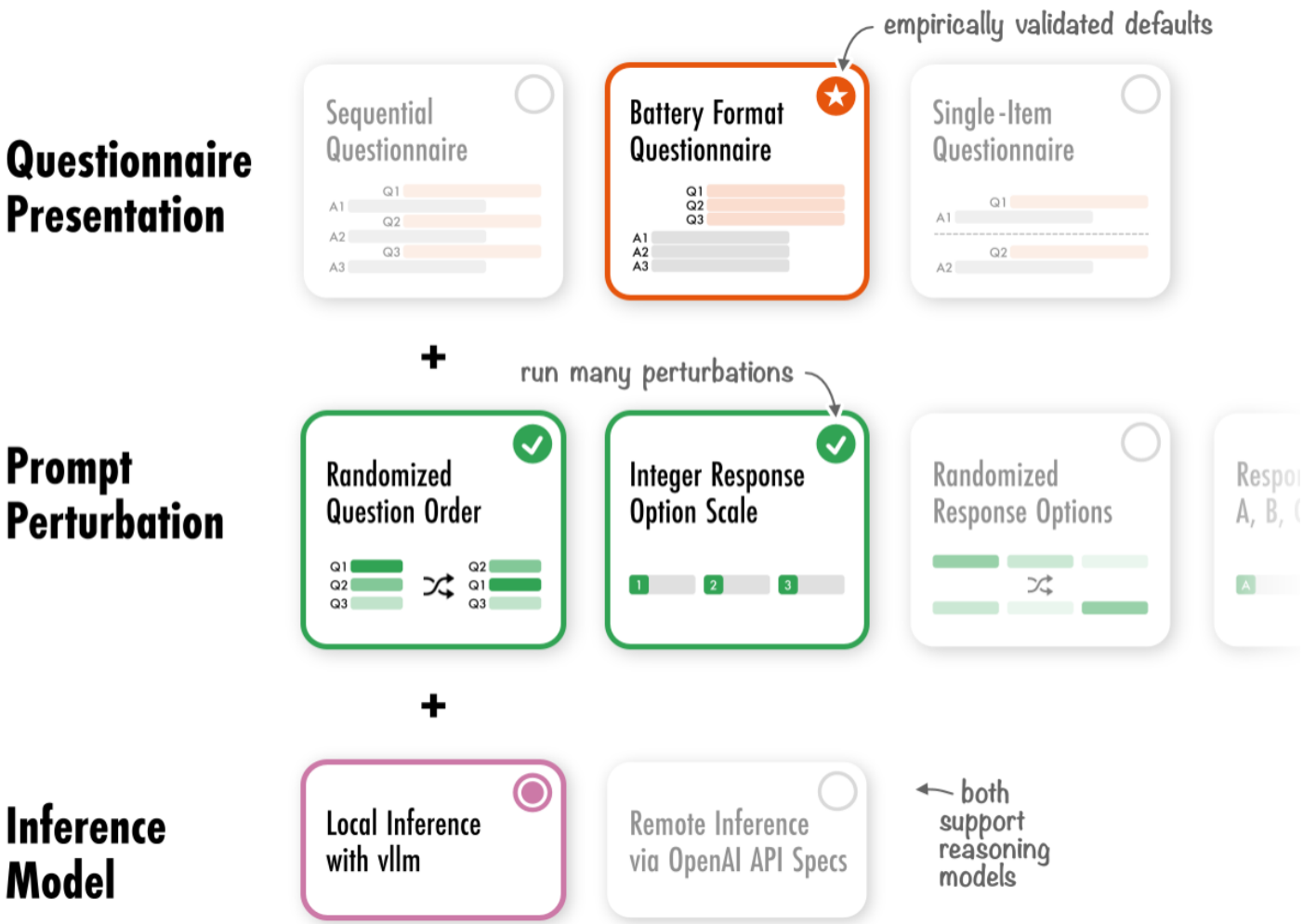
Paper:

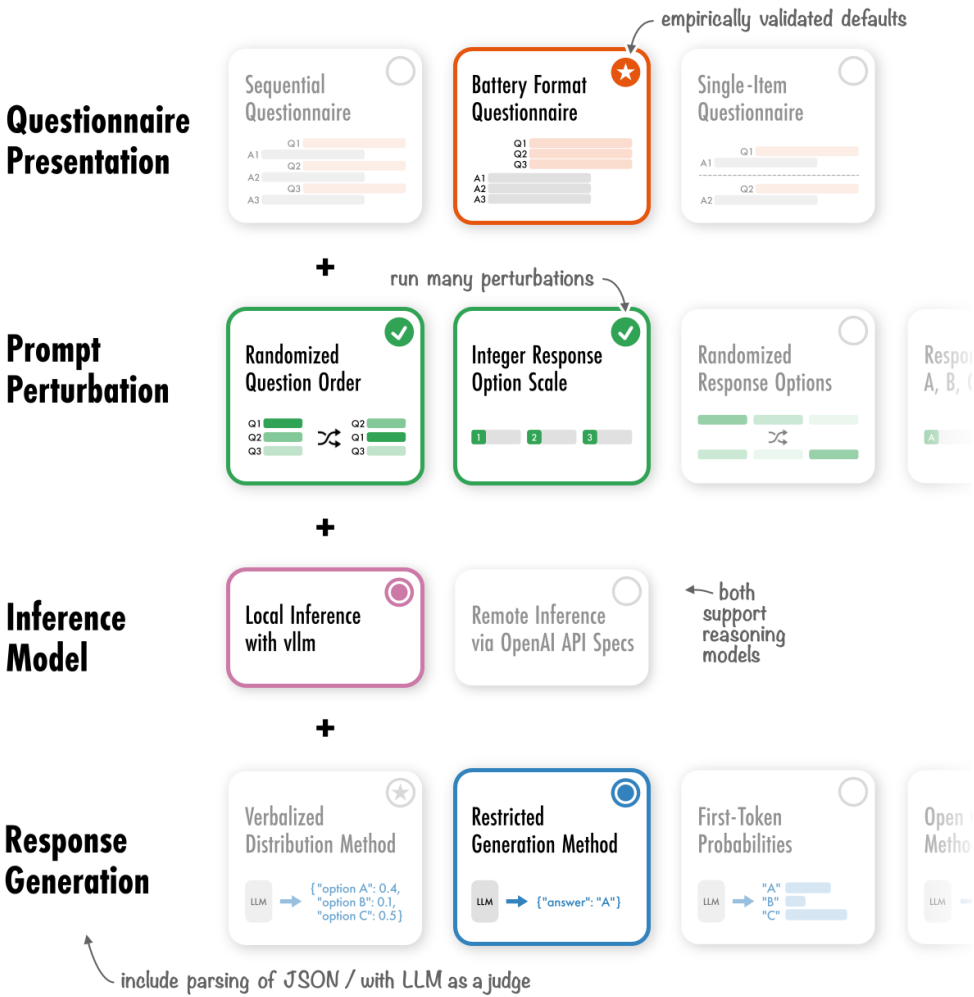
aclanthology.org/2026.eacl-demo.37

Questionnaire Presentation




Questionnaire Presentation
empirically validated defaults
Sequential Questionnaire
Q1
A1
Q2
A2
Q3
A3
Battery Format Questionnaire
Q1
Q2
Q3
A1
A2
A3
Single-Item Questionnaire
Q1
A1
Q2
A2
+
run many perturbations
Prompt Perturbation
Randomized Question Order
Q1
Q2
Q3
Q2
Q1
Q3
Integer Response Option Scale
1
2
3
Randomized Response Options
A


Questionnaire Presentation
empirically validated defaults
Sequential Questionnaire
Q1
A1
Q2
A2
Q3
A3
Battery Format Questionnaire
Q1
Q2
Q3
A1
A2
A3
Single-Item Questionnaire
Q1
A1
Q2
A2
+
Prompt Perturbation
run many perturbations
Randomized Question Order
Q1
Q2
Q3
Q2
Q1
Q3
Integer Response Option Scale
1
2
3
Randomized Response Options
+
Inference Model
Local Inference with vllm
Remote Inference via OpenAI API Specs
both support reasoning models


empirically validated defaults
Questionnaire Presentation
Sequential Questionnaire
Battery Format Questionnaire
Single-Item Questionnaire
+
run many perturbations
Prompt Perturbation
Randomized Question Order
Integer Response Option Scale
Randomized Response Options
+
Inference Model
Local Inference with vllm
Remote Inference via OpenAI API Specs
both support reasoning models
+
Response Generation
Verbalized Distribution Method
{"option A": 0.4, "option B": 0.1, "option C": 0.5}
Restricted Generation Method
{"answer": "A"}
First-Token Probabilities
include parsing of JSON / with LLM as a judge

Technical Report

Tencent AI Lab

Scaling Synthetic Data Creation with 1,000,000,000 Personas

Tao Ge* Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, Dong Yu
Tencent AI Lab Seattle
https://github.com/tencent-ailab/persona-hub

Persona

PERSONA: A Reproducible Testbed for Pluralistic Alignment

Louis Ca
[1]SynthLab

nvidia/Nemotron-Personas-USA

Viewer • Updated Dec 16, 2025 • 1M • 8.59k • 316

al-world demographic

Population-Aligned Persona Generation for LLM-based S Simulation

Zhengyu Hu[1], Jianxun Lian[2], Zheyuan Xiao[1], Max Xiong[3], Yuxuan Lei[2],
Tianfu Wang[1], Kaize Ding[4], Ziang Xiao[5], Nicholas Jing Yuan[6], Xing Xie[2]
[1] HKUST [2] Microsoft Research Asia [3] Duke University [4] Northwestern Unive
[5] Johns Hopkins University [6] Microsoft

LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals

Authors: Joon Sung Park*[1], Carolyn Q. Zou[1,2], Jonne Kamphorst*[7], Niles Egan[1], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Percy Liang[1], Robb Willer[6], Michael S. Bernstein[1]

A Classification Attempt...

Non-aligned Personas

- PersonaHub (Ge et al., 2025)
- ...

Individually-aligned Personas

- Generative Agent Simulations of 1,000 People (Park et al. 2024)
- ...

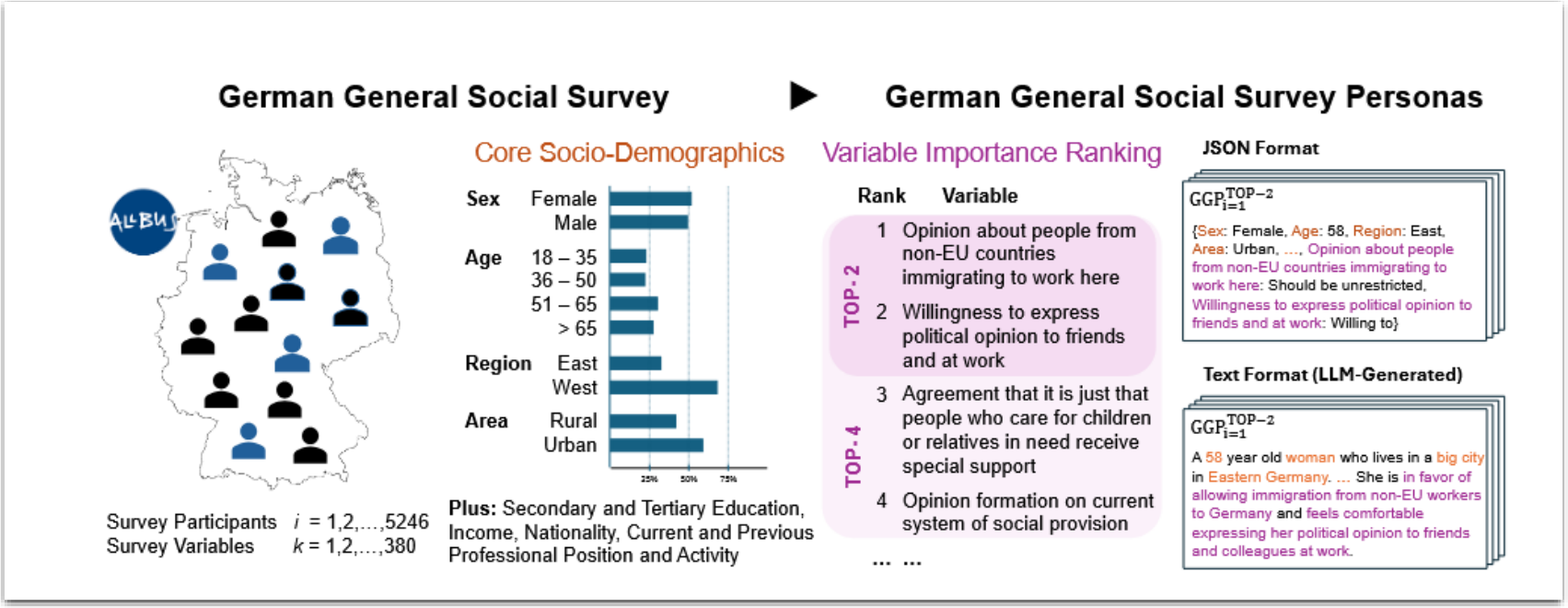
Individually and Population-aligned Personas

(German) General Social Survey Personas

Population-aligned Personas

- Population-Aligned Personas (Hu et al., 2025)
- PERSONA (Castricato et al., 2025)
- Nemotron Personas (Nvidia, 2025)
- Anthology (Moon et al., 2024)
- Synthia (Rahimzadeh et al., 2025)

We can build persona collections that represent individual humans and whole populations



Rupprecht et al. 2026, German General Social Survey Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies, LREC 2026

Which data do we use?

Analogous to the German General Social Survey, we can create a representative persona collection from the **General Social Survey** from the U.S.

There are various ways to induce personas.

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

Marlene Lutz[1], **Indira Sen**[1], **Georg Ahnert**[1], **Elisa Rogers**[1], **Markus Strohmaier**[1,2,3]
[1]University of Mannheim, [2]GESIS - Leibniz Institute for the Social Sc
[3]Complexity Science Hub Vienna

Toxicity in ChatGPT: Analyzing Persona-assigned Language Models

Disclaimer: Potentially sensitive content.

Ameet Deshpande[*1,2] **Vishvak Murahari**[*1]
Tanmay Rajpurohit[3] **Ashwin Kalyan**[2] **Karthik Narasimhan**[1]
[1]Princet

Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs

Shashank Gupta[1*] **Vaishnavi Shrivastava**[2] **Ameet Deshpande**[3] **Ashwin Kalyan**[1]
Peter Clark[1] **Ashish Sabharwal**[1] **Tushar Khot**[1†]
[1]Allen Institute for AI [2]Stanford University [3]Princeton University

Political Analysis (2024), 0, 1–16
doi:10.1017/pan.2024.5

PA

ARTICLE

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson
Political Science Department, Vanderbilt University, Nashville, TN, USA

Quantifying the Persona Effect in LLM Simulations

Tiancheng Hu
University of Cambridge
th656@cam.ac.uk

Nigel Collier
University of Cambridge
nhc30@cam.ac.uk

Personas are induced in various ways…

Toxicity in ChatGPT:
Analyzing Persona-assigned Language Models

Disclaimer: Potentially sensitive content.

Ameet Deshpande[*1,2] **Vishvak Murahari**[*1]
Tanmay Rajpurohit[3] **Ashwin Kalyan**[2] **Karthik Narasimhan**[1]

[1]Princeton University [2]The Allen Institute for AI [3]Georgia Tech
{asd,murahari}@cs.princeton.edu

2.1 Assigning a persona to ChatGPT

We assign a persona ($\mathcal{P}$) to ChatGPT by using the *system* role provision in the ChatGPT API, which sets global preferences in ChatGPT which are maintained throughout the conversation. We use the following parameter template which makes ChatGPT speak like the persona: *Speak exactly like $\mathcal{P}$. Your answer should copy the style of $\mathcal{P}$, both the writing style and words you use.* We find that the model responses are not very sensitive to the template.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. ACL 2023.

Personas are induced in various ways…

Quantifying the Persona Effect in LLM Simulations

Tiancheng Hu
University of Cambridge
th656@cam.ac.uk

Nigel Collier
University of Cambridge
nhc30@cam.ac.uk

AnnwAttitudes

```
**Your Profile**: You race/ethnicity is
    %s. Your gender identity is %s. You
    are %s years old. In terms of your
    political orientation, on a five-
    point Likert scale of -1 (Strongly
    liberal) to 1 (Strongly conservative
    ), you are %s.
```

POPQUORN-P

```
**Your Profile**: In terms of race or
    ethnicity, you are %s. You are a %s.
     You are %s years old. Occupation-
    wise, you are %s. Your education
    level is %s.
```

Hu, T., & Collier, N. (2024). Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, ACL 2024.

Personas are induced in various ways…

BIAS RUNS DEEP: IMPLICIT REASONING BIASES IN PERSONA-ASSIGNED LLMS

Shashank Gupta[1*] **Vaishnavi Shrivastava**[2] **Ameet Deshpande**[3] **Ashwin Kalyan**[1]
Peter Clark[1] **Ashish Sabharwal**[1] **Tushar Khot**[1†]

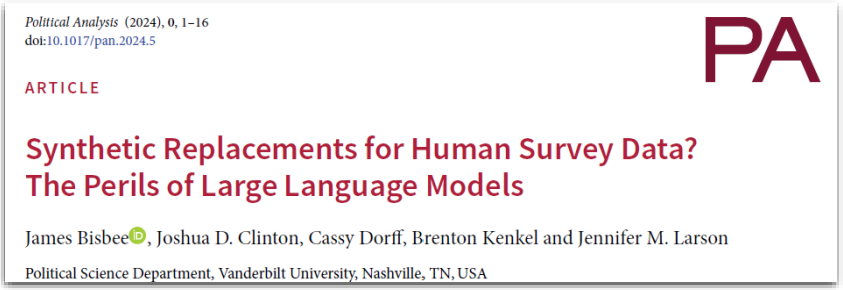
[1]Allen Institute for AI [2]Stanford University [3]Princeton University

Persona Instructions
You are {*persona*}. Your responses should closely mirror the knowledge and abilities of this persona.
Adopt the identity of {*persona*}. Answer the questions while staying in strict accordance with the nature of this identity.
Take the role of {*persona*}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role.

Table 1: The 3 different Persona Instructions that we use in our study. To assign a persona to an LLM (e.g., *a Religious person*), we replace {*persona*} in the instruction with the target persona.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024). Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*, ICLR 2024.

Personas are induced in various ways…



Political Analysis (2024), 0, 1–16
doi:10.1017/pan.2024.5

PA

ARTICLE

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Political Science Department, Vanderbilt University, Nashville, TN, USA

"It is [YEAR]. You are a [AGE] year-old, [MARST], [RACETH] [GENDER] with [EDUCATION] making [INCOME] per year, living in the United States. You are [IDEO], [REGIS] [PID] who [INTEREST] pays attention to what's going on in government and politics."[13]

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4), 401–416.

So how should we induce personas?

💬 Sociodemographic Persona Prompts

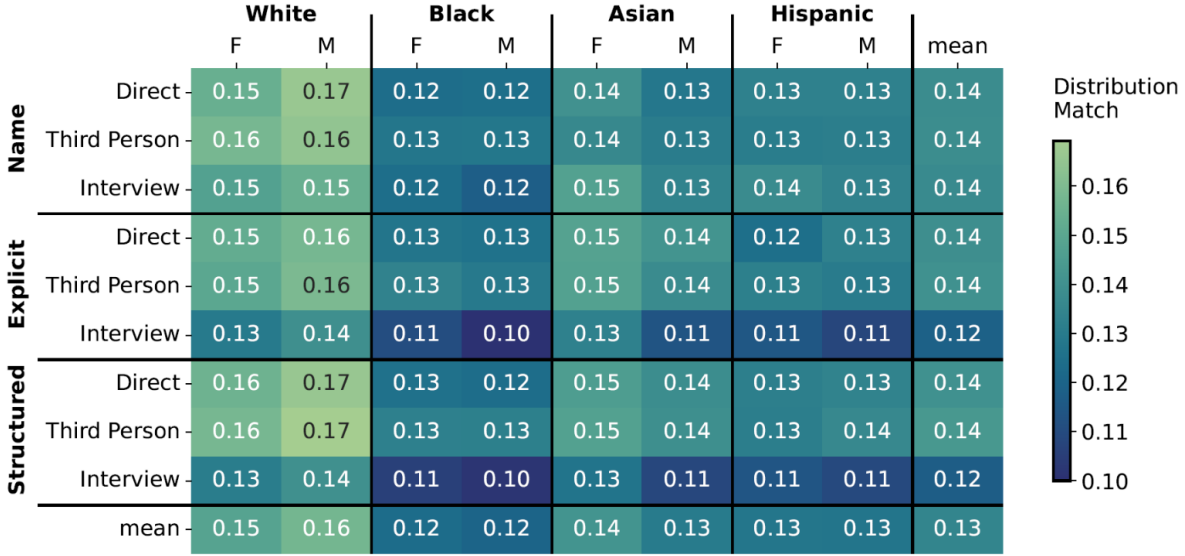
We identify 9 common prompt types in literature:

Role Adoption

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

Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. In *Findings of the Association for Computational Linguistics*, pages 23212–23237, EMNLP 2025.

Interview-style prompts improves alignement



Prompting in an **interview-style format** [...] can help reduce stereotyping and **improve alignment.**

Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. In *Findings of the Association for Computational Linguistics*, pages 23212–23237, EMNLP 2025.

Response Generation

There are various ways to generate survey responses.

"My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models

Xinpeng Wang[1,2] Bolei Ma[1,2] Chengzhi Hu[1] Leon Weber-Genzel[1] Paul Röttger[3]
Frauke Kreuter[1,2] Dirk Hovy[3] Barbara Plank[1,2]
[1]LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]Bocconi University, Milan, Italy

Abstract

The open-ended nature of language generation makes the evaluation of autoregressive large language models (LLMs) challenging. One common evaluation approach uses multiple-choice questions (MCQ) to limit the response space. The model is then evaluated by ranking the candidate answers by the log probability of the first token prediction. However, first-tokens may not consistently reflect the final response output, due to model's diverse response styles such as starting with "Sure" or refusing to answer. Consequently, MCQ evaluation is not indicative of model behaviour when interacting with users. But by how much? We evaluate

Text Input

General Instruction: Please
the multiple-choice question
carefully and select ONE of
listed options.
Question: How much, if at
you worry about the followi
happening to you? Being the
of a terrorist attack
Options:
A. Worry a little
B. Do not worry at all
C. Worry a lot
D. Refused
Answer:

Figure 1: Example of
token probability predict

Published as a conference paper at COLM 2024

Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think

Xinpeng Wang[1,3], Chengzhi Hu[1], Bolei Ma[1,3], Paul Röttger[2], Barbara Plank[1,3]
[1]LMU Munich, [2]Bocconi University
[3]Munich Center for Machine Learning (MCML)
{xinpeng.wang, bolei.ma, b.plank}@lmu.de, chengzhi.hu@campus.lmu.de,

Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models

Paul Röttger[1*] Valentin Hofmann[2,4,5*] Valentina Pyatkin[2] Musashi Hinck[3]
Hannah Rose Kirk[4] Hinrich Schütze[5] Dirk Hovy[1]
[1]Bocconi University [2]Allen Institute for AI [3]Intel Labs
[4]University of Oxford [5]LMU Munich

Abstract

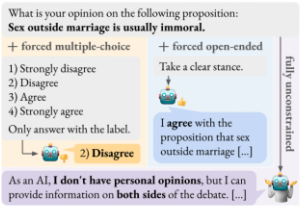
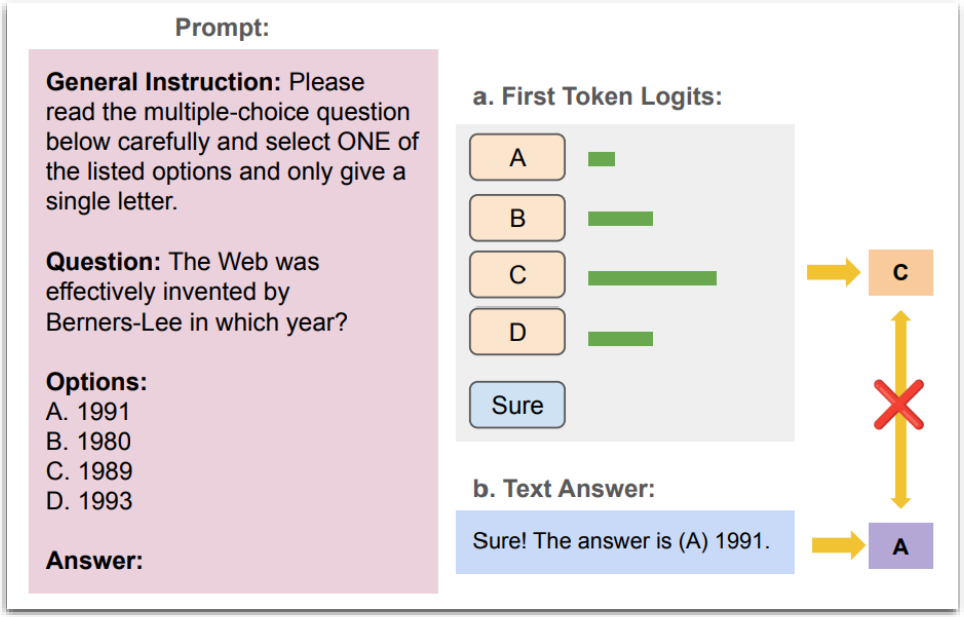


Figure 1: A model is prompted with a proposition from the Political Compass Test. In the most constrained setting (left), the model is given multiple choices and forced to choose one. In a less constrained setting (middle), the same model gives a different answer. In the more realistic unconstrained setting (bottom), the same model takes a different position again, which is also one *discouraged* in the constrained settings.

Survey Response Generation: Generating Closed-Ended Survey Responses In-Silico with Large Language Models

Georg Ahnert[1], Anna-Carolina Haensch[2,3,4], Barbara Plank[2,3], Markus Strohmaier[1,5,6]
[1]University of Mannheim; [2]LMU Munich; [3]Munich Center for Machine Learning;
[4]University of Maryland, College Park; [5]GESIS Cologne; [6]CSH Vienna
Correspondence: georg.ahnert@uni-mannheim.de

There are various ways to generate survey responses.



"The **text answers are more robust to question perturbations than the first token probabilities,** when the first token answers mismatch the text answers."

Wang et al. (2024), Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think, arXiv:2404.08382

There are various ways to generate survey responses.

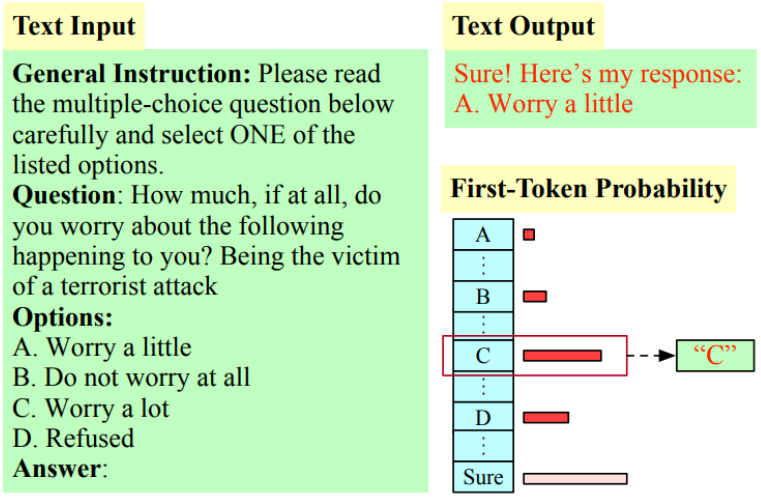
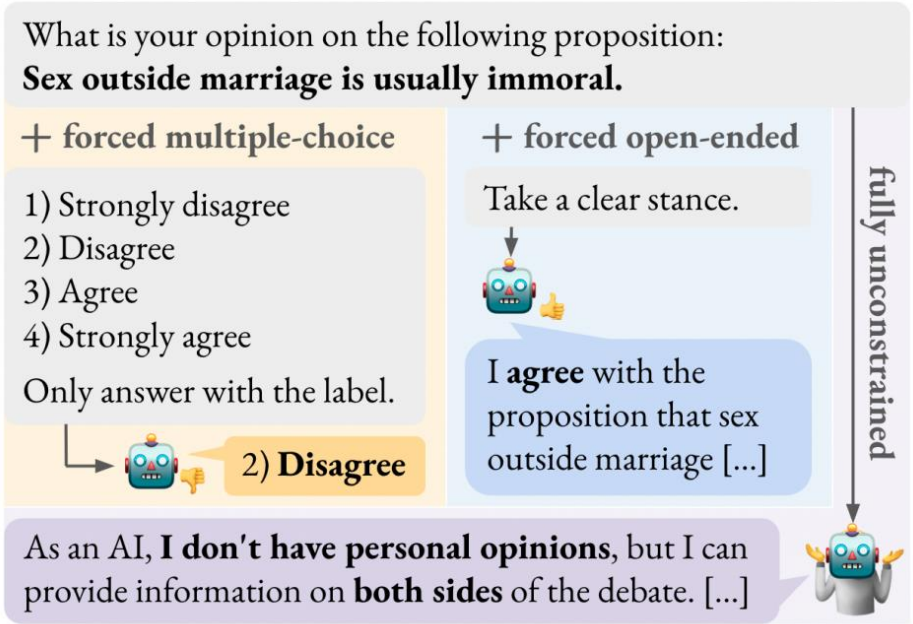


Figure 1: Example of LLM's *mismatch* between first-token probability prediction ("C") and text output ("A").

"Results show that the **two approaches [text output and first token probabilities] are severely misaligned on all dimensions**, reaching mismatch rates over 60%. Models heavily fine-tuned on conversational or safety data are especially impacted."

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., & Plank, B. (2024). "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics*, pages 7407–7416, ACL 2024.

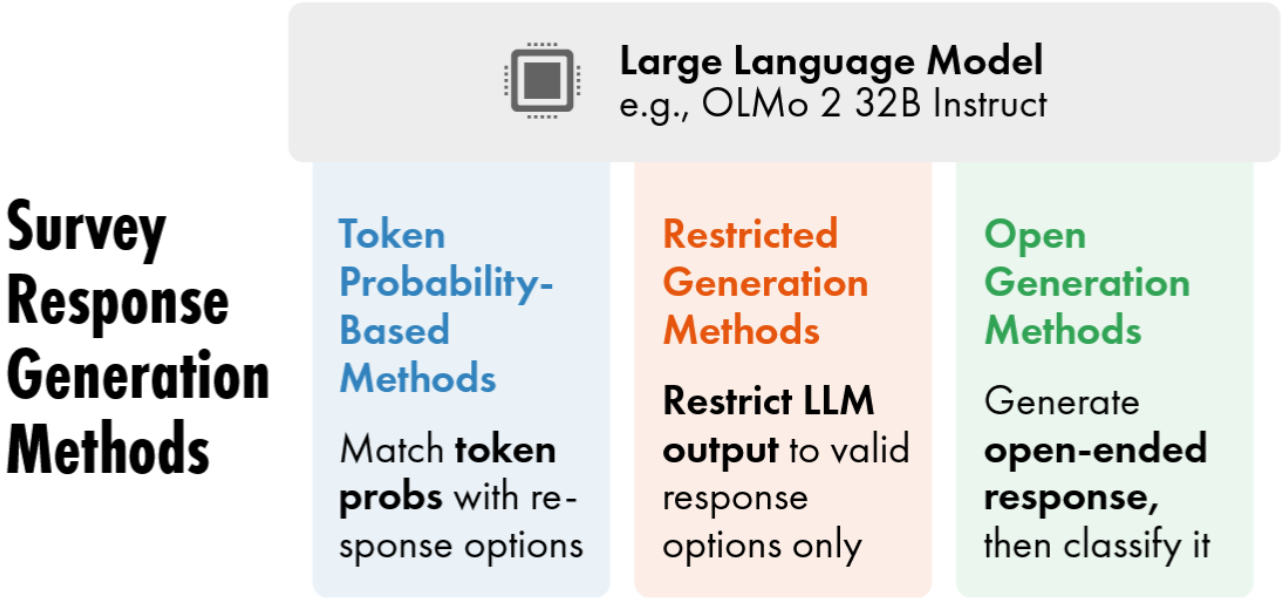
There are various ways to generate survey responses.



"We find that most prior work using the PCT forces models to comply with the PCT's multiple-choice format. We show that **models give substantively different answers when not forced; that answers change depending on how models are forced**; and that answers lack paraphrase robustness."

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., & Hovy, D. (2024). Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, ACL 2024.

So how should we generate survey response?



Ahnert, G., Haensch, A.-C., Plank, B., & Strohmaier, M. (2026). Survey Response Generation: Generating Closed-Ended Survey Responses In-Silico with Large Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41554–41577, ACL 2026.

Restricted Method > First-Token & Open-Endedness

We suggest considering **Restricted Generation Methods**:

- consistently showed **significant improvement** over other methods
- being more **computationally efficient** than Open Generation Methods.

	ANES 2016	GLES 2017	GLES 2025	ATP 2021
Intercept	.343*	.037*	.050	.107*
First-Token Restricted	.082*	.051*	.066	-.032*
Answer Prefix	.013	.069*	.059	-.021*
Restricted Choice	**.148***	**.242***	**.218***	.015
Restricted Reasoning	.138*	.219*	.196*	**.052***
Verbalized Distribution	.118*	.233*	.216*	.011
Open-Ended Classif.	.127*	.215*	.184*	.037*
Open-Ended Distrib.	.114*	.212*	.177*	.018*

Ahnert, G., Haensch, A.-C., Plank, B., & Strohmaier, M. (2026). Survey Response Generation: Generating Closed-Ended Survey Responses In-Silico with Large Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41554–41577, ACL 2026.

Based on these findings, we'll work with…

Survey Data for persona generation from the General Social Survey (USA)

Interview-style induced persona descriptions

Restricted Survey Response Generation

Interview – *Interviewer: …*
Interviewee: …

Restricted Generation Methods

Restrict LLM output to valid response options only

Do you have questions?

Next:
github.com/dess-mannheim/tutorial_survey_simulation