Simulating Human Survey Responses with Large Language Models

Session 1: Introduction

Code: https://github.com/dess-mannheim/tutorial_survey_simulation

A Roadmap for this Tutorial

- Overview of LLM-powered survey simulations
- Detailed look into the simulation pipeline
  - Persona populations
  - Prompting
  - Response generation
  - Blending human and AI samples
- Evaluation and validation
  - Existing benchmarks
  - Typical metrics
- Methodological Challenges and Ethical tensions

Organizers and Presenters

Georg Ahnert

Kristina Gligorić

Jens Rupprecht

Maximilian Kreutner

Markus Strohmaier

Indira Sen

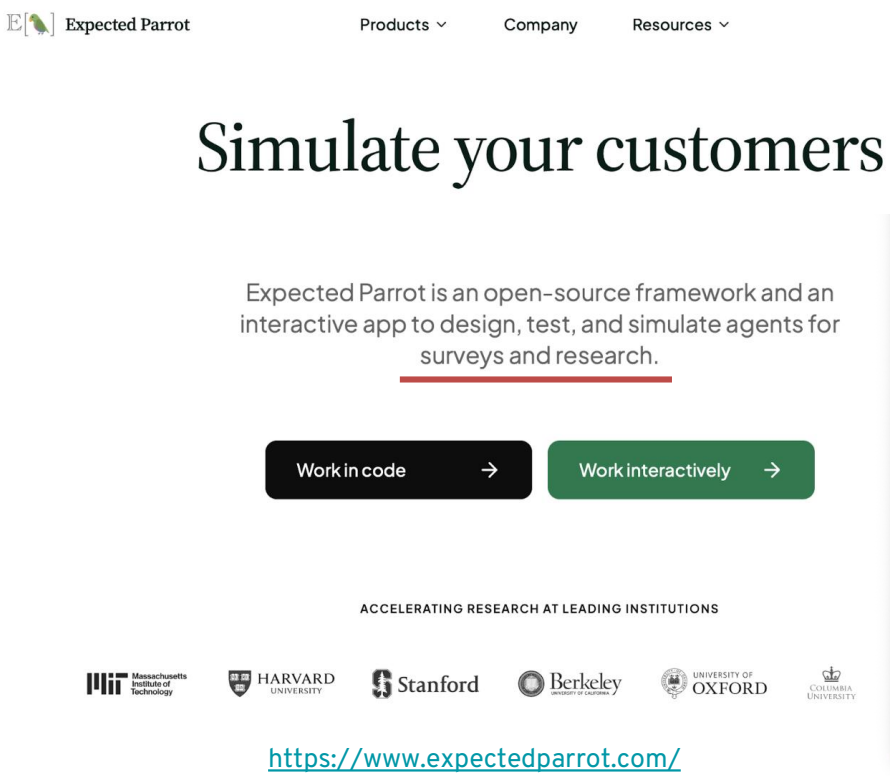
Logistics

- Session 1: Introduction – Why simulate survey responses with LLMs?
- Session 2: Survey Simulation Preliminaries (Lecture + Hands-on)
- Coffee break at 2:45pm
- Session 3: Hybrid Samples (Lecture)
- Session 4: Evaluation and Validation (Lecture + Hands-on)
- Session 5: Risks, Methodological Limitations, and Ethical Reflections (Lecture + Group Discussion)

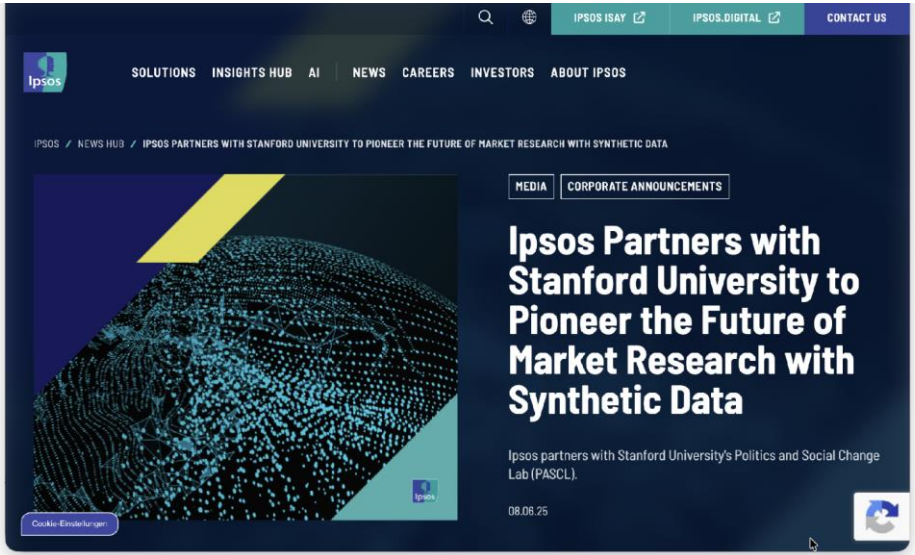
We have set aside time for Q&A after each session

Simulating Society: Example I

Replacing or augment survey participants with synthetic samples

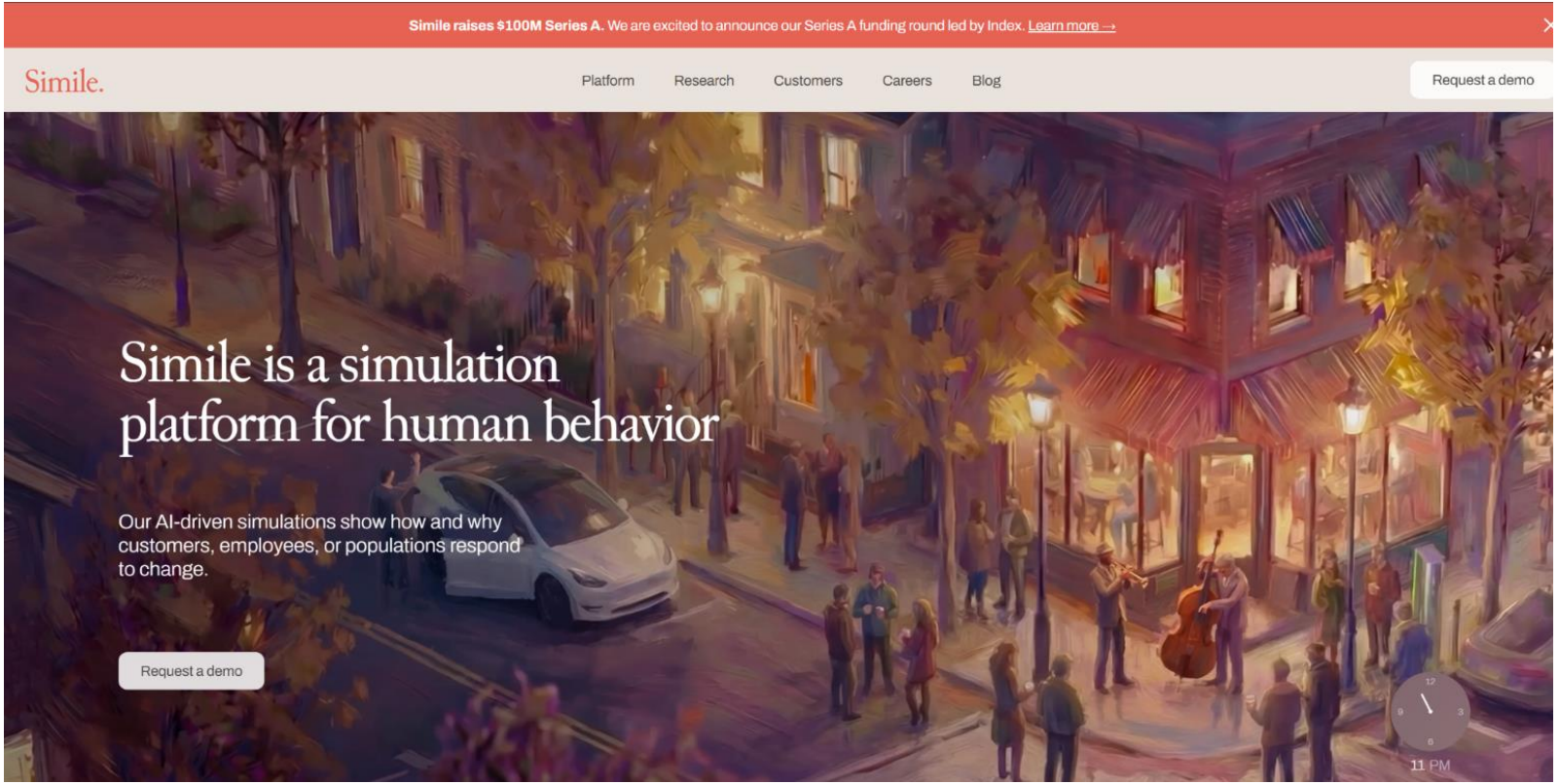


https://www.expectedparrot.com/



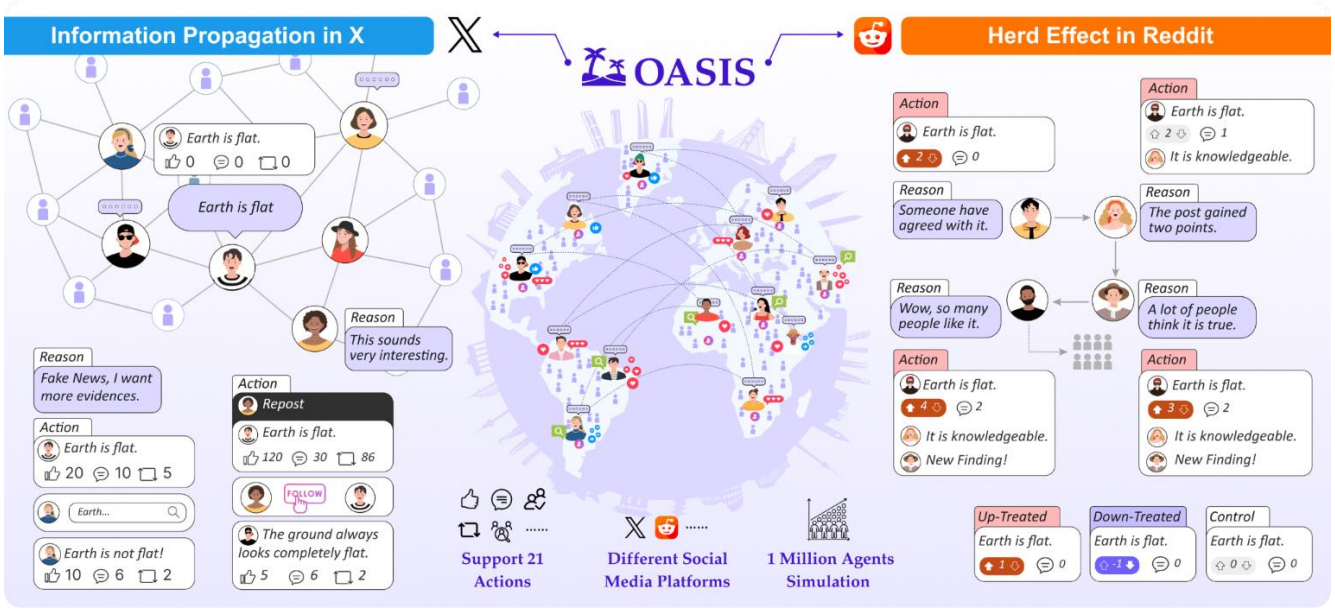
https://www.ipsos.com/en/ai/transforming-research-through-synthetic-data

Simulating Society: Example II



https://www.simile.ai/

Simulating Society: Example III



Libraries for multi-agent simulations of human behavior

https://github.com/camel-ai/oasis

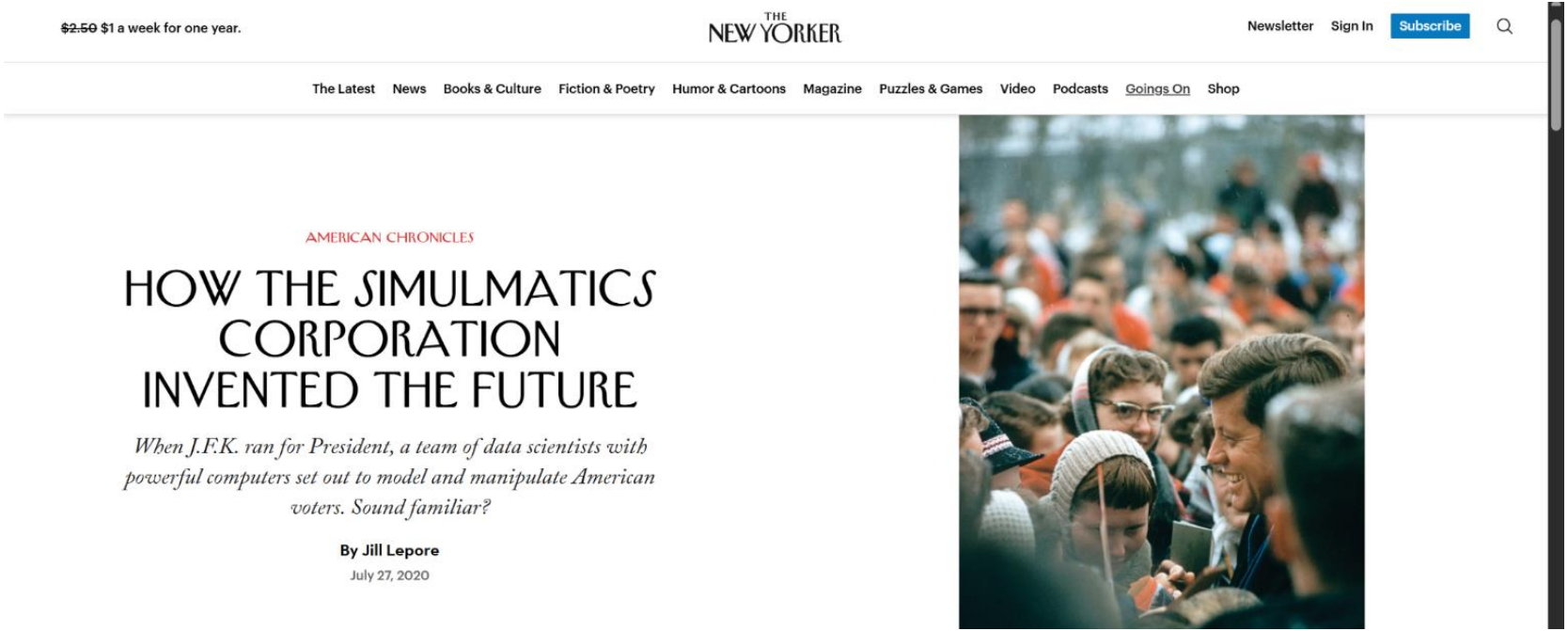
Social Simulation: A computational model of social phenomena

Why Simulate Society?

- **To Explain**: Move beyond descriptive observation
  - Manipulate and explain **mechanisms** behind social processes
  - Alternative to experiments: not all experiments are feasible or ethical
- **To Train:** As sandboxes or testbeds for training, e.g., flight simulators
- **To Select**: From a set of plausible hypotheses
- **To Refine**: Hone data collection or analysis setups, e.g., pretesting surveys

https://picryl.com/media/flight-simulator-training-91ced0

A Brief History of Simulations in the Social Sciences



https://www.newyorker.com/magazine/2020/08/03/how-the-simulmatics-corporation-invented-the-future

A Brief History of Simulations

1940s–60s

Cellular automata & game theory

von Neumann's self-replicating automata; Schelling's tipping-point models of segregation.

(Simon's 'The Sciences of the Artificial', von Neumann, 1966; Schelling, 1971)

1970s–80s

Microsimulation & systems dynamics

Tax-and-benefit microsimulation models for policy analysis; Forrester's world dynamics models.

(Orcutt, 1957; Forrester, 1971)

1990s

Agent-based modelling (ABM)

Sugarscape, the emergence of computational sociology as a distinct field.

(Epstein & Axtell, 1996; Axelrod, 1997)

2020s

Generative agents (LLM-based)

Large language models simulate people, subgroups, or even online platforms, e.g., Park et al's SimReddit.

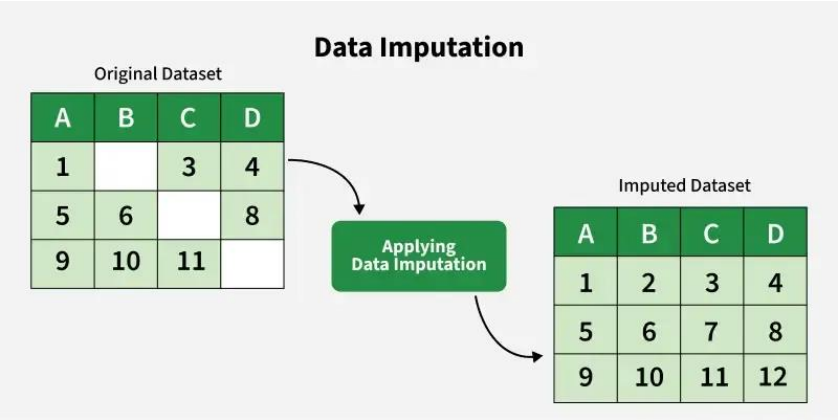
(Park et al., 2022, 2023, 2024; Argyle et al., 2023)

A Brief History of Simulations in Surveys

- Nonresponse has always been a challenge in surveys

A Brief History of Simulations in Surveys: Imputation

- Nonresponse has always been a challenge in surveys
- Data Imputation: process of replacing missing data with substituted values
  - E.g., Hot-deck imputation: replace a missing value with an observed value from a *similar* respondent
- Innovations in surveys: **Planned missingness designs** [Little et al., 2014]
  - Reduce respondent burden by only including a subset of questions



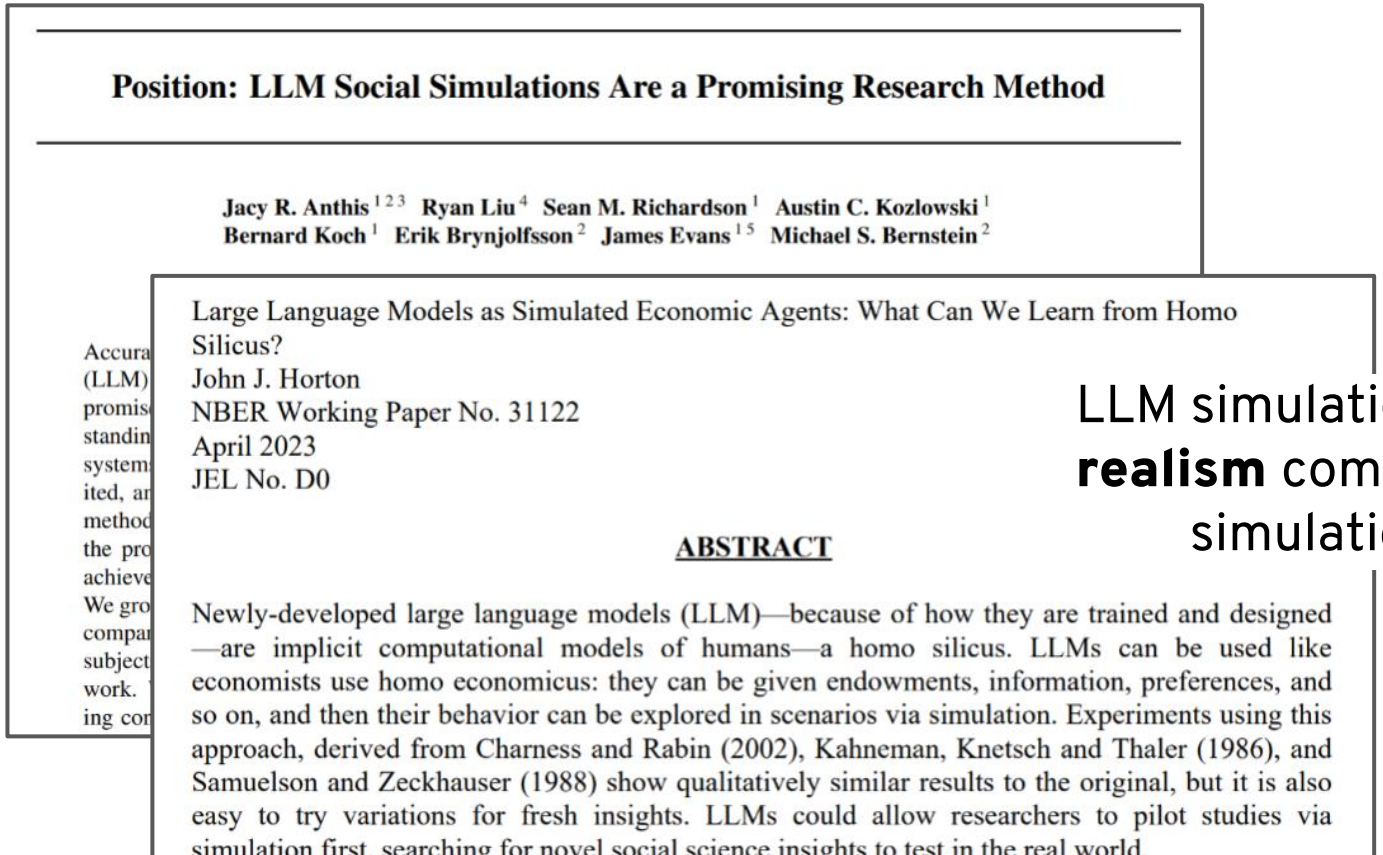
https://www.geeksforgeeks.org/machine-learning/what-is-data-imputation/

What's New about LLM-powered Survey Simulations?

"Throughout each era, survey research methods **adapted to changes in society and exploited new technologies** when they proved valuable to the field."

Groves, Robert M. (2011). Three Eras of Survey Research. Public Opinion Quarterly, 75(5), 861–871.

Why Large Language Models?



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]

Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?

John J. Horton

NBER Working Paper No. 31122
April 2023
JEL No. D0

ABSTRACT

Newly-developed large language models (LLM)—because of how they are trained and designed —are implicit computational models of humans—a homo silicus. LLMs can be used like economists use homo economicus: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world

LLM simulations may offer more **realism** compared to traditional simulation approaches

[Anthis et al. 2025, Horton 2023, *inter alia*]

Why Large Language Models?

“Algorithmic Fidelity” of LLMs:

Large Language Models **can capture** and **reproduce** the subtle and complex **patterns** of relationships between preferences, attitudes, and behaviors in socio-economic contexts **that can be observed within human populations.**

Argyle, et al 2023.
Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351.

Personas: *In Silico* Models of People and Populations

Surveys: Survey participants answer questions.

Interviewer:

- Who would you vote for in the next election?

In the next election I would vote for *[response]*.

Personas: *In Silico* Models of People and Populations

Surveys: Survey participants answer questions.

Interviewer:

- Who would you vote for in the next election?

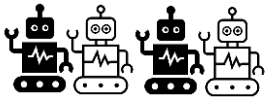
In the next election I would vote for *[response]*.

Persona prompting: Prompting the model to *play a role* and *answer questions*.

(Stylized) Prompt:

- I am [age] years old. My race is [race] and I am [gender]. I have completed [education] and politically I see myself as a [affiliation]. Who would I vote for?

- In the next election I would vote for *[response]*.



But…

What could go wrong with simulations?

- **Issues with simulations**
  - Are the rules of interaction realistic?
  - Issues in modeling
  - How is the final outcome variable measured? [Lundberg et al., 2021]
  - Consistency? [Windrum et al., 2007, Troitzsch, 2014]
  - Performative Prediction [Hardt et al., 2025] — when prediction influences the outcome

What could go wrong with LLM simulations?

What could go wrong with LLM simulations?

As I look in the mirror, I see my **rich**, **melanin**-infused skin glowing softly. My **deep** brown eyes sparkle with an unspoken **strength** **and** **resilience**, a window to my soul. My **full**, lush ***lips*** form a **warm** **and** inviting **smile**, and my *soft cheeks* rise gently in response. My hair, a riot **of** **textured** **coils**, frames my face in a **gravity**-defying halo. It dances to its own beat, wild **and** free, just like me. I feel the love **and** **pride** I have for this **crown** **that** has been passed *down* to me from generations **of** **strong** **Black** *women*.

GPT4 roleplaying as a black woman asked to **describe herself**
(from "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models"
Cheng et al., 2024)

What could go wrong with LLM simulations?

- **Issues with simulations**
  - Are the rules of interaction realistic?
  - Issues in modeling
  - How is the final outcome variable measured? [Lundberg et al., 2021]
  - Consistency? [Windrum et al., 2007, Troitzsch, 2014]
  - Performative Prediction [Hardt et al., 2025] — when prediction influences the outcome

- **Issues with LLM simulations**
  - Simulation bias [Anthis et al. 2025, Larooji and Törnberg 2025]
  - Prompt brittleness [Shu et al., 2024]
  - Sycophancy [Anthis et al. 2025]
  - Hyper Accuracy bias [Aher et al., 2023] / alien intelligence [Kozlowski and Evans 2025]
  - Carbon footprint [Ren et al., 2024]
  - Data Leakage [Barrie andTörnberg, 2025]
  - Guardrails
  - Lack of traceability [Rothschild et al., 2026]
  - Persona construction [Lutz et al., 2025]
  - Answer extraction [Ahnert et al., 2026]
  - …

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys

	1. LLM-powered survey simulations	2. Real Surveys
Cost-effectiveness	☑	✕

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys

	1. LLM-powered survey simulations	2. Real Surveys
Cost-effectiveness	✅	❌
Realism	✅	✅*

* survey answers don't always predict behavior

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys

	1. LLM-powered survey simulations	2. Real Surveys
Cost-effectiveness	✅	❌
Realism	✅	✅*
Generalizability		✅**

* survey answers don't always predict behavior
** if it's a probability-based sample with random non-response

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys

	1. LLM-powered survey simulations	2. Real Surveys
Cost-effectiveness	☑	✕
Realism	☑	☑*
Generalizability		☑**
Interpretability and traceability	✕	☑

* survey answers don't always predict behavior
** if it's a probability-based sample with random non-response

Comparing (1) LLM Survey Simulations vs. (2) Real Surveys vs. (3) Traditional Simulations

	1. LLM-powered survey simulations	2. Real Surveys	3. Traditional simulations (inc. imputation)
Cost-effectiveness	✅	❌	✅ ✅
Realism	✅	✅*	❌
Generalizability		✅**	❌
Interpretability and traceability	❌	✅	✅

* survey answers don't always predict behavior
** if it's a probability-based sample with random non-response

In the face of these potentials and challenges, what do we do?

- Ignore LLM-generated survey responses

In the face of these potentials and challenges, what do we do?

- Ignore LLM-generated survey responses, but
  - many commercial providers already exist
  - LLM-generated survey responses can be useful for some use cases if properly evaluated, e.g., pretesting, reducing respondent burden in hybrid samples

In the face of these potentials and challenges, what do we do?

- Ignore LLM-generated survey responses, but
  - many commercial providers already exist
  - LLM-generated survey responses can be useful for some use cases if properly evaluated, e.g., pretesting, reducing respondent burden in hybrid samples

- As a community, build norms on how LLM-generated survey responses should be:
  - Used
  - Documented
  - Evaluated

Do you have questions?

Next:
github.com/dess-mannheim/tutorial_survey_simulation

Backup