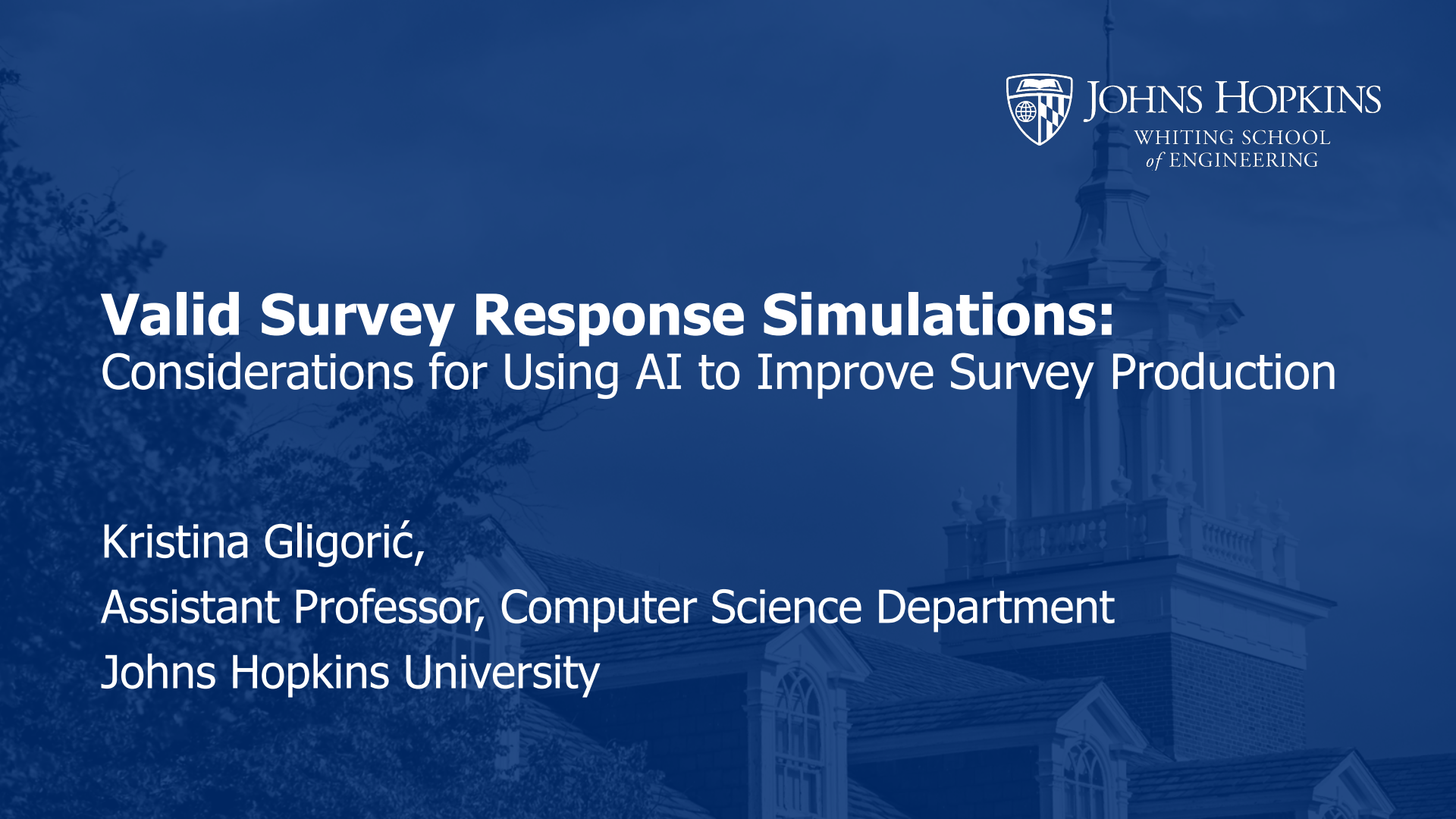


Valid Survey Response Simulations:
Considerations for Using AI to Improve Survey Production

Kristina Gligorić,
Assistant Professor, Computer Science Department
Johns Hopkins University

The problem

- Surveys are invaluable–
  but, **expensive** and **slow**
- $20K for a 4-week survey

- **Eye-tracking studies**
- **Cognitive pretesting**
- **Participant recruitment**
- **Iterative redesign**

What if we had simulated respondents?

Try new questions before sending them into the field.

Iterate on instrument design at the speed of code.

Decide who to ask, in what order, and how.

A growing body of work uses LLMs to simulate respondents

Out of One, Many

Argyle et al. • 2023

"Silicon sampling" — GPT-3 conditioned on socio-demographic backstories matches subpopulations.

Generative Agent Simulations

Park et al. • 2024

Digital twins built from 1,000 in-depth interviews replicate individual-level survey responses.

Scientist & Subject

Manning, Zhu & Horton • 2024

LLMs serve as both researcher and respondent—reproducing classic economics findings.

Why skepticism is warranted!

Yejin Choi

Why AI is incredibly smart and shockingly stupid

Identity flattening

Wang, Morgenstern & Dickerson

LLMs that replace human participants harmfully misportray groups.

Hallucination & prompt sensitivity

Dominguez-Olmedo, Hardt & Mendler-Dünner

Models invent answers and shift with wording and option order.

Social sycophancy

Cheng et al.

LLMs tell us what we want to hear — not an honest respondent.

The questions that we work on

1. Methods

How can we produce reliable synthetic participants whose responses are accurate at the population level?

2. Evaluation

How do we evaluate whether these simulations are valid and useful, and when should practitioners trust them?

Today's case study using NHANES

Valid Survey Simulations with Limited Human Data

The roles of prompting, fine-tuning, and rectification • ACL 2026





EPFL
UC Berkeley
JOHNS HOPKINS UNIVERSITY

Krsteski, S., Russo, G., Chang, S., West, R., & Gligorić, K. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2026*.

NHANES: A national dietary recall survey

8.5K

participants

16K+

full-day recalls

2

longitudinal waves

12

demo. covariates

EXAMPLE: 24-HOUR RECALL

"What did you eat yesterday?"

- Oatmeal — *100 g*
- Rice — *150 g*
- Banana — *45 g*
- Sweet tea, lemon — *77 g*
- Grilled chicken — *150 g*
- Apple — *182 g*

→ *Daily energy intake = 1,766 kcal (target mean)*

Methods

Three places to intervene.

We usually have some human responses at hand that we can leverage to make the AI simulations more accurate.

Fine-tuning

Adapt model parameters using historical (question, demographics, response) triplets.

Prompting

Condition the model on a participant's demographic profile and persona at generation time.

Rectification

Use a small set of human responses to debias model predictions—no retraining required.

Method 1: Supervised fine-tuning

Adapt the model on prior survey waves so its weights encode the population.

WAVE 1 / TRAINING

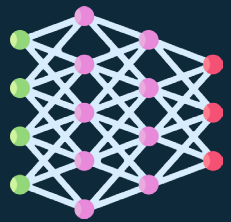
Triplets

(question)

(demographics)

(response)

Cross-entropy loss



FINE-TUNED LLM

weights now reflect population patterns

WAVE 2 / GENERATION

Generate

for new participants from demographics alone; no human responses needed.

Method 2: Persona-based prompting

Condition the LLM on a participant profile — generate a plausible day of recall.

PROMPT

System:

You are participating in a dietary recall survey. Describe what you ate and drank in the past 24 hours, based on your memory. Profile: {demographic_profile}.

User:

List everything you ate and drank in the past 24 hours, in order.

MODEL OUTPUT

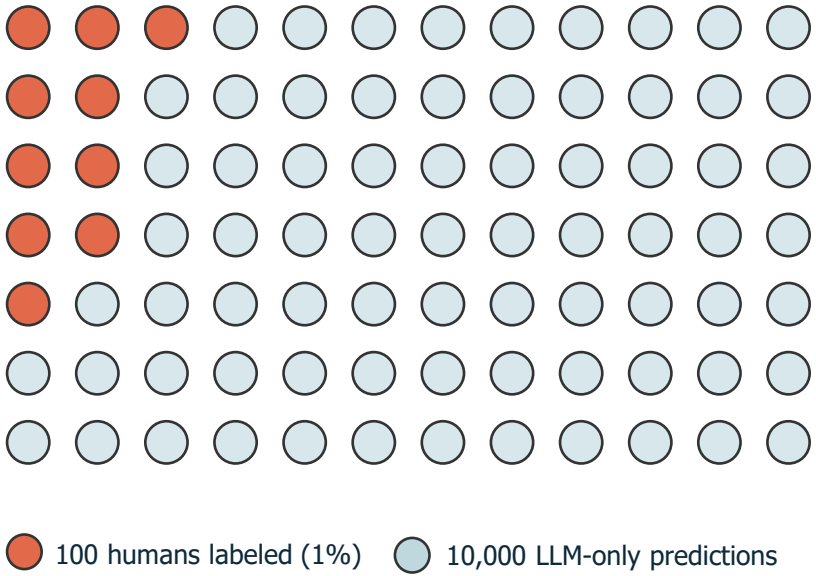
…a full day of recall, conditioned on who that person is.

Sweet tea with lemon	**77 g**
Grilled chicken breast	**150 g**
Apple	**182 g**
Brown rice	**200 g**

Method 3: Statistical correction

Treat the LLM as a powerful but biased predictor. Estimate the bias on a small subset with human answers and correct.

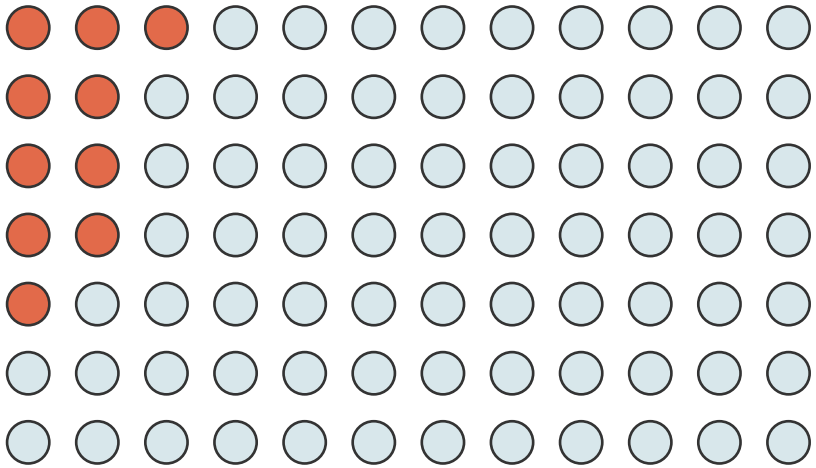
THE INTUITION



Method 3: Statistical correction

Treat the LLM as a powerful but biased predictor. Estimate the bias on a small subset with human answers and correct.

THE INTUITION



100 humans labeled (1%) 10,000 LLM-only predictions

THE ESTIMATOR

synthetic mean + bias correction

from full LLM predictions *from parallel subset*

Builds on: Prediction-Powered Inference
Design-based Supervised Learning

Angelopoulos et al. 2023 • Egami et al. 2023

No retraining: numerical adjustment on top of any synthesis.

What it takes technically

All three methods rely on standard tools. A graduate student or data engineer can stand this up.

Standard libraries

Python

Prompting and FT

API access

E.g., AWS/Azure-hosted

Secure infra

Off-the-shelf packages

Statistics

Metrics: How do we know if this is useful?

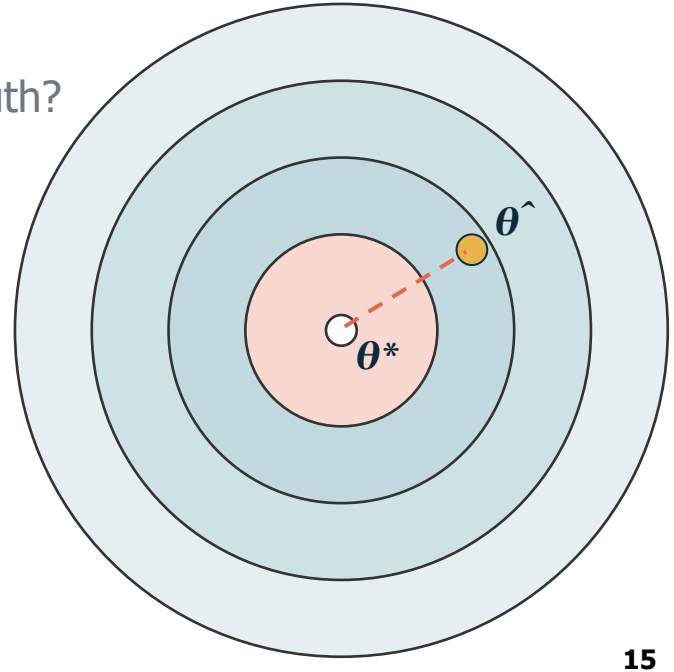
1. Population-level accuracy

How far is the estimated population mean from the truth?

RELATIVE ERROR

< 5%

estimate vs. true population mean



Metrics: How do we know if this is useful?

2. Effective sample size gain

If we add LLM responses on top of human ones, are we actually better off?

Return on investment: does synthetic data add information, or just noise?

wide CI

100 humans

Metrics: How do we know if this is useful?

2. Effective sample size gain

If we add LLM responses on top of human ones, are we actually better off?

Return on investment: does synthetic data add information, or just noise?

ESS GAIN

~ 15%

Added precision from synthetic data

wide CI

100 humans

tighter CI

100 humans + 10,000 LLM

Metrics: How do we know if this is useful?

3. Individual-level accuracy stays hard

People with identical demographic profiles eat very differently, and the models can't distinguish them

ONE PROFILE

Age: 66–70

Sex: female

Weight: 60–70 kg

Non-Hispanic White

No special diet

2,500 kcal

Metrics: How do we know if this is useful?

3. Individual-level accuracy stays hard

People with identical demographic profiles eat very differently, and the models can't distinguish them

ONE PROFILE

Age: 66–70

Sex: female

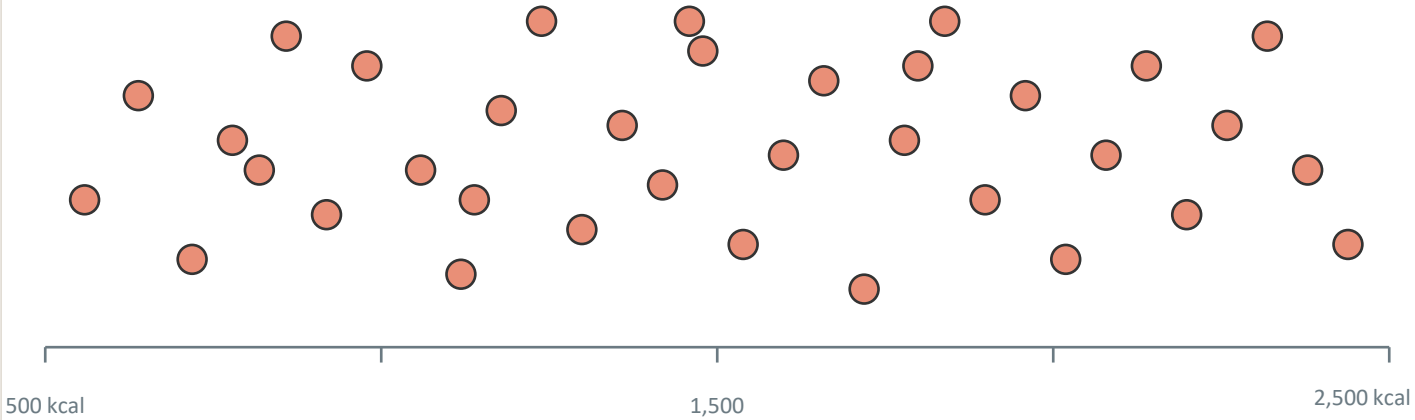
Weight: 60–70 kg

Non-Hispanic White

No special diet

ACTUAL ENERGY INTAKE

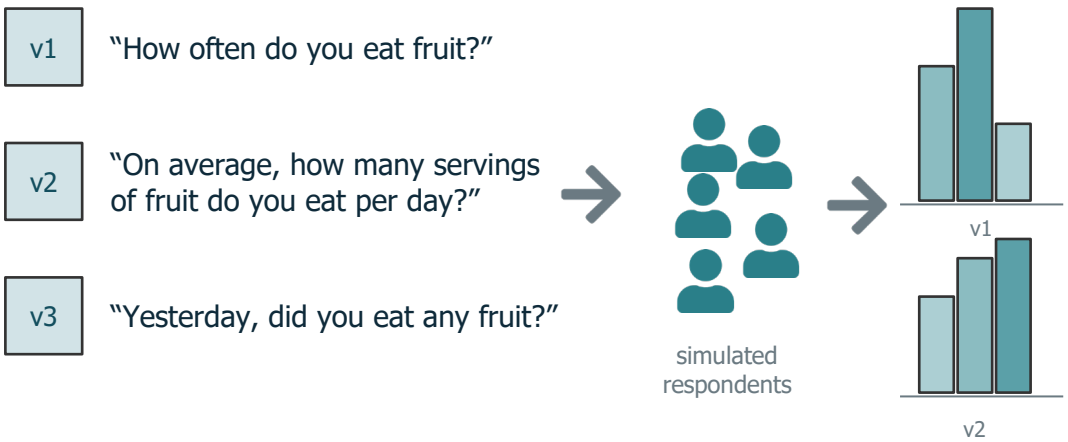
within this single demographic group



How this can support survey production

Pretesting: Use simulated respondents to evaluate new candidate survey questions before fielding them with real participants.

THE WORKFLOW



Compare how each formulation distributes responses; surface ambiguity and order effects before any human is asked; develop questions with high explanatory power.

How this can support survey production

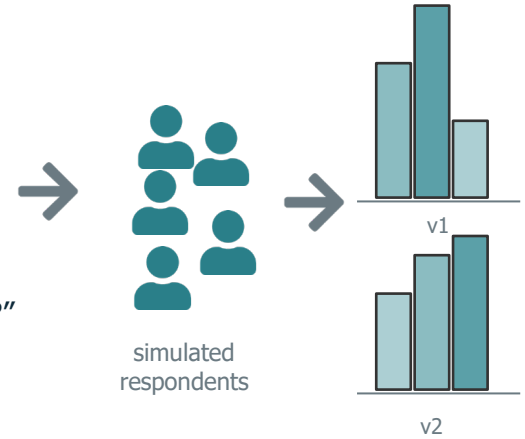
Pretesting: Use simulated respondents to evaluate new candidate survey questions before fielding them with real participants.

THE WORKFLOW

v1 "How often do you eat fruit?"

v2 "On average, how many servings of fruit do you eat per day?"

v3 "Yesterday, did you eat any fruit?"



Compare how each formulation distributes responses; surface ambiguity and order effects before any human is asked; develop questions with high explanatory power.

WHAT IT BUYS US

Surfacing issues

ambiguity, order effects, ceiling effects

Cost reductions

before any human pilot is run

Speeding up discovery

Reducing the number of iteration cycles

How this can support survey production

Adaptive sampling: Use model predictions to identify the most informative respondents

Gligorić, Kristina, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. "Can unconfident LLM annotations be used for confident conclusions?." In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (ACL), 2025.

How this can support survey production

Adaptive sampling: Use model predictions to identify the most informative respondents

LLM simulations guide sampling so that the budget concentrates where the model is least confident

Gligorić, Kristina, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. "Can unconfident LLM annotations be used for confident conclusions?." In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (ACL), 2025.

How this can support survey production

Reducing selection bias: Use simulations to anticipate who responds, and shape outreach to reach the rest.

Frame invitations

Test alternative framings against simulated subgroups before sending. Find phrasings that improve participation.

Set incentives

Estimate the right incentive level to increase participation in hard-to-reach groups, without overpaying easy-to-reach ones.

Predict non-response

Forecast where coverage will be thin and pre-correct the design, or, weight the final estimates accordingly.

The big picture

- AI tools can support efforts to collect high-quality data and statistics at lower costs
- Excellent prototyping tools
- The possibility of lowering costs of survey development via pretesting

...using standard tools and relatively small investments

Thank you!

Questions?

Kristina Gligorić,
Assistant Professor, Computer Science Department
Johns Hopkins University
gligoric@jhu.edu