Simulating Human Survey Responses with Large Language Models

Session 3, Kristina Gligoric: Combining human & simulated survey responses: fine-tuning, prompting, PPI, adaptive sampling

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.

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.

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)*

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.* https://aclanthology.org/2026.acl-long.498/

Method 1: Supervised fine-tuning

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

WAVE 1 / TRAINING

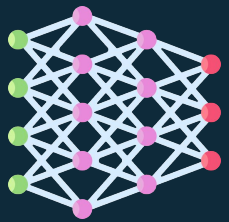
Triplets

(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

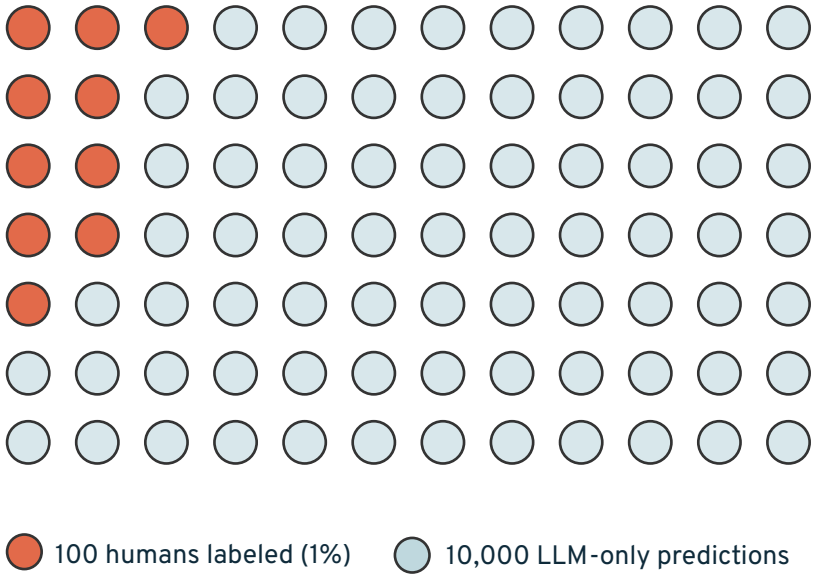
…a full day of recall, conditioned on who that person is.

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.

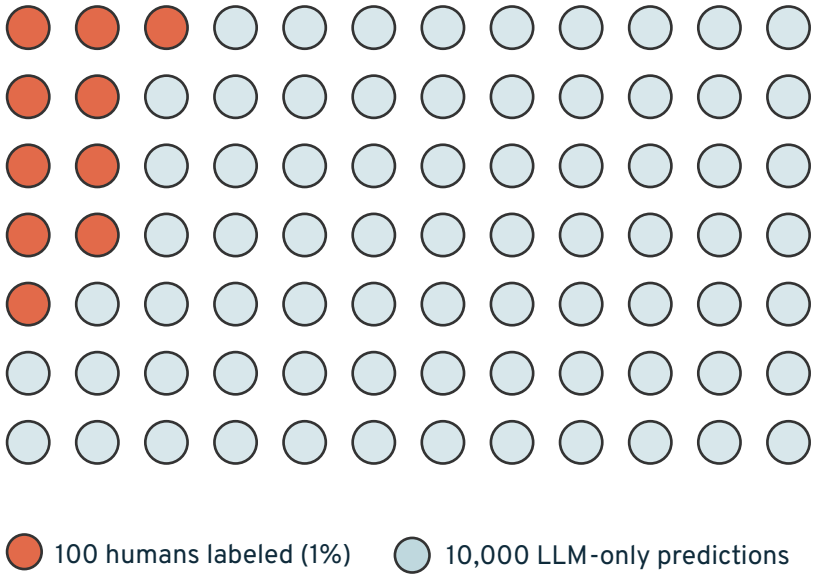
THE INTUITION



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



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. Science
https://www.science.org/doi/full/10.1126/science.adi6000
Egami et al. 2023. Neurips https://arxiv.org/abs/2306.04746

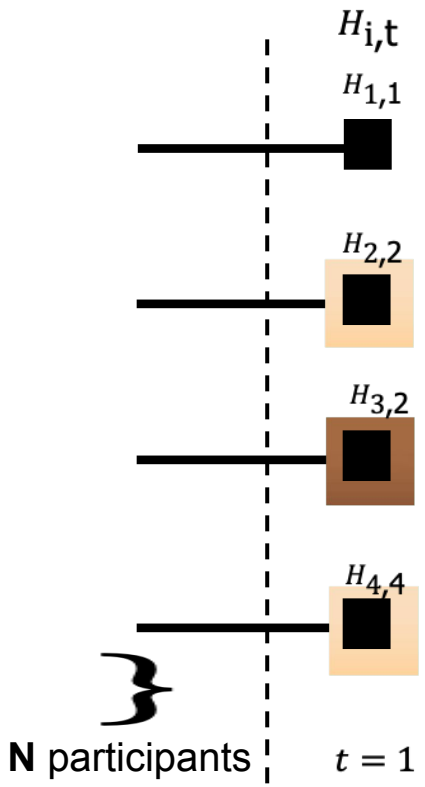
No retraining: numerical adjustment on top of any synthesis

A deeper dive into statistical correction

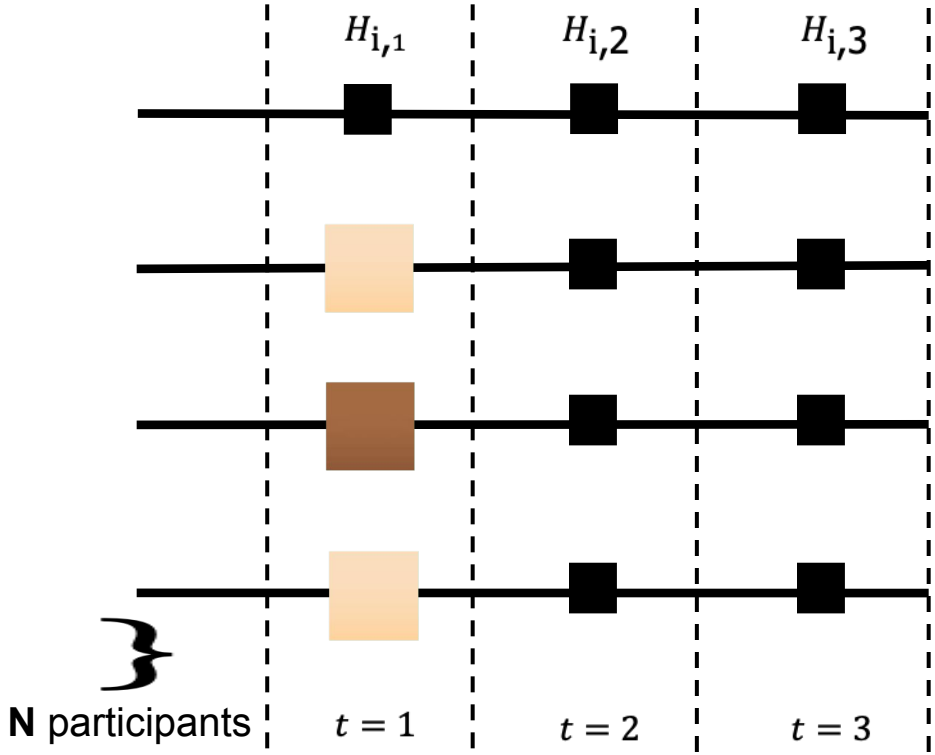
The setup: Let’s define a survey

N participants

The setup: Let's define a survey

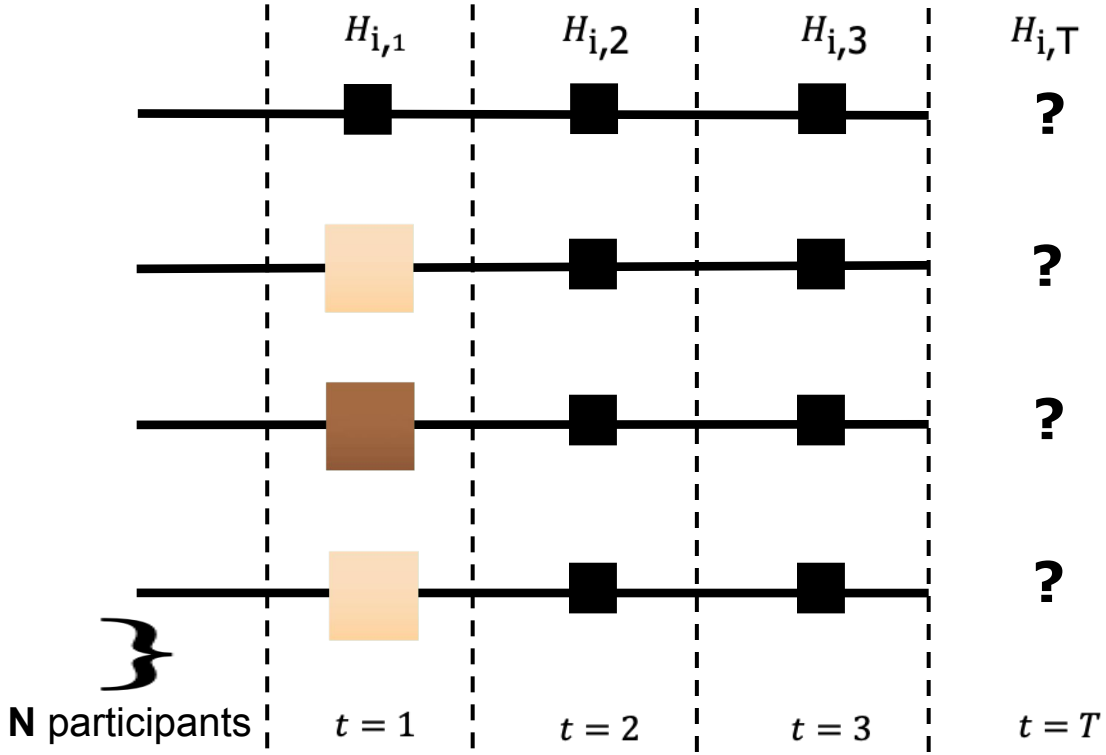


The setup: Let's define a survey



N participants

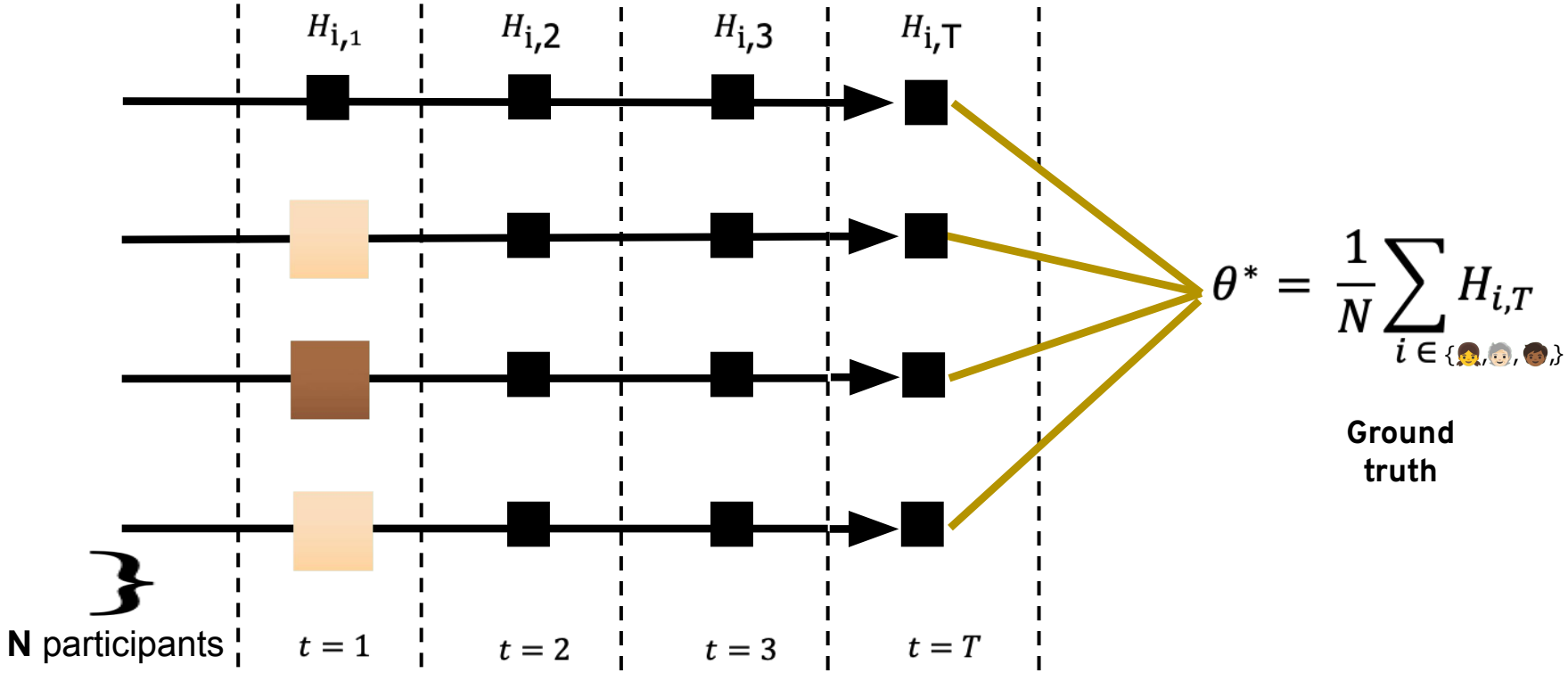
The setup: Let's define a survey



N participants

missing survey answers $H_{i,T}$

The setup: Let's define a survey



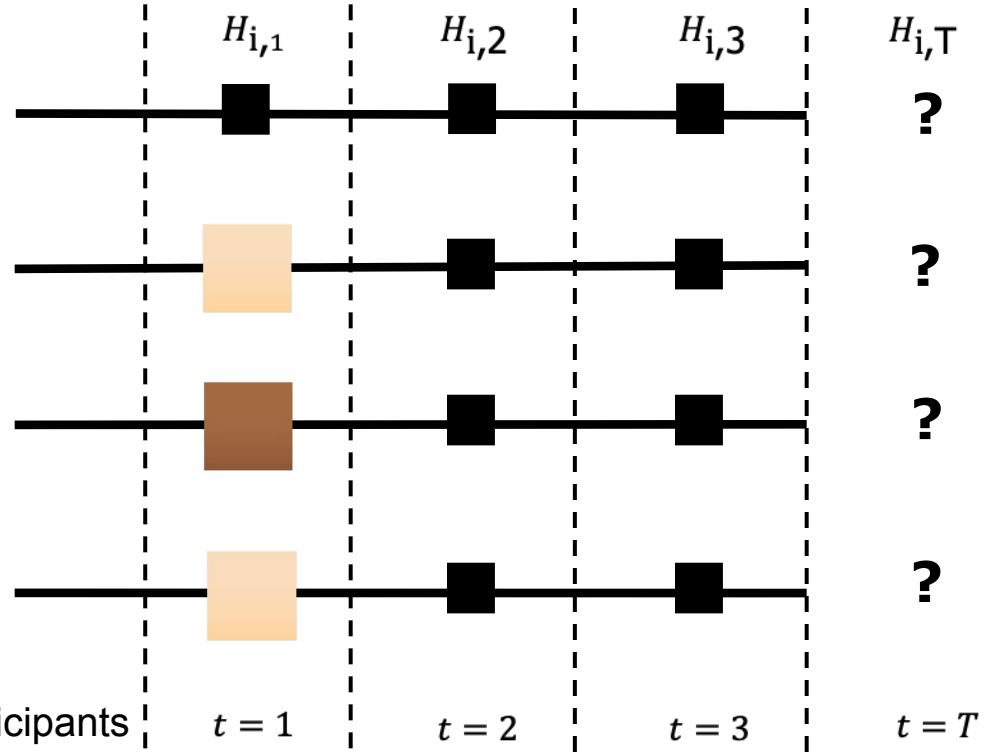
The setup: Let's define a survey

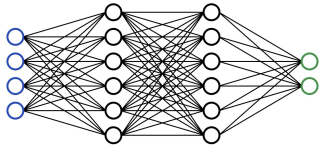
goal: estimate quantity of interest θ^*

examples of θ^*:

- Average number of calories
- Percentage of people who ate fruit
- Whatever we care about learning once we have survey answers!

The setup: Let's define a survey

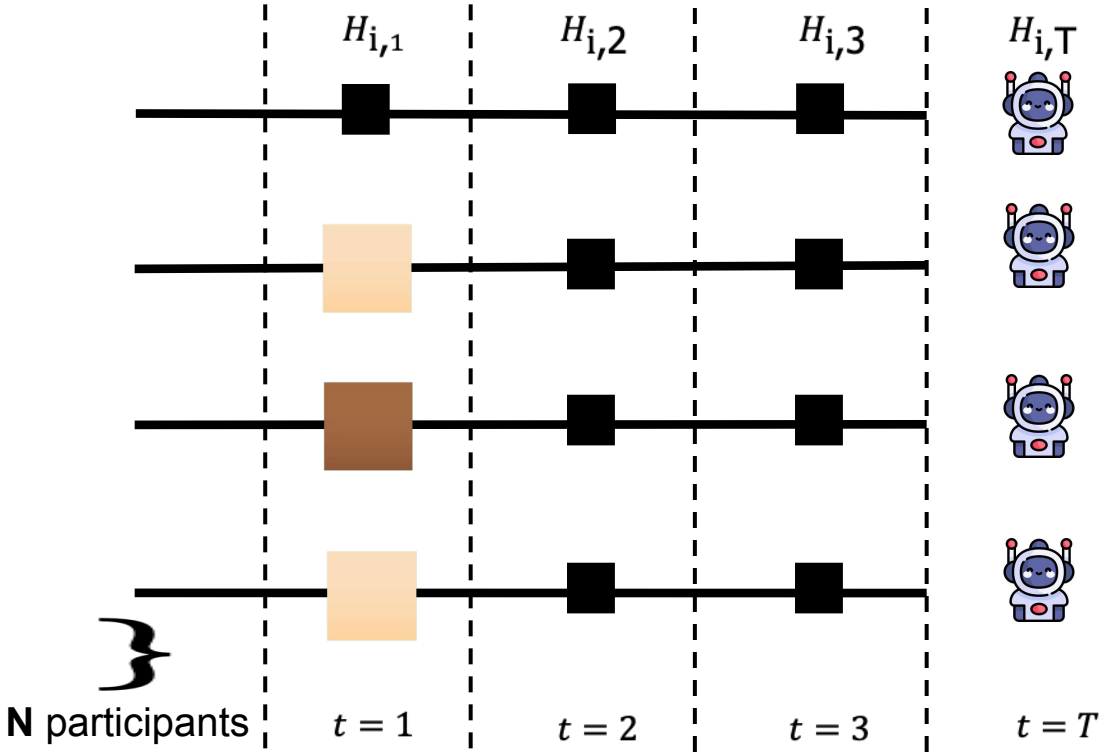


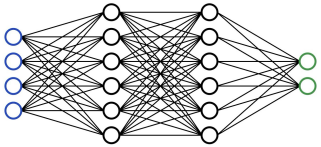


large language model

can be used to produce the estimate

The setup: Let's define a survey



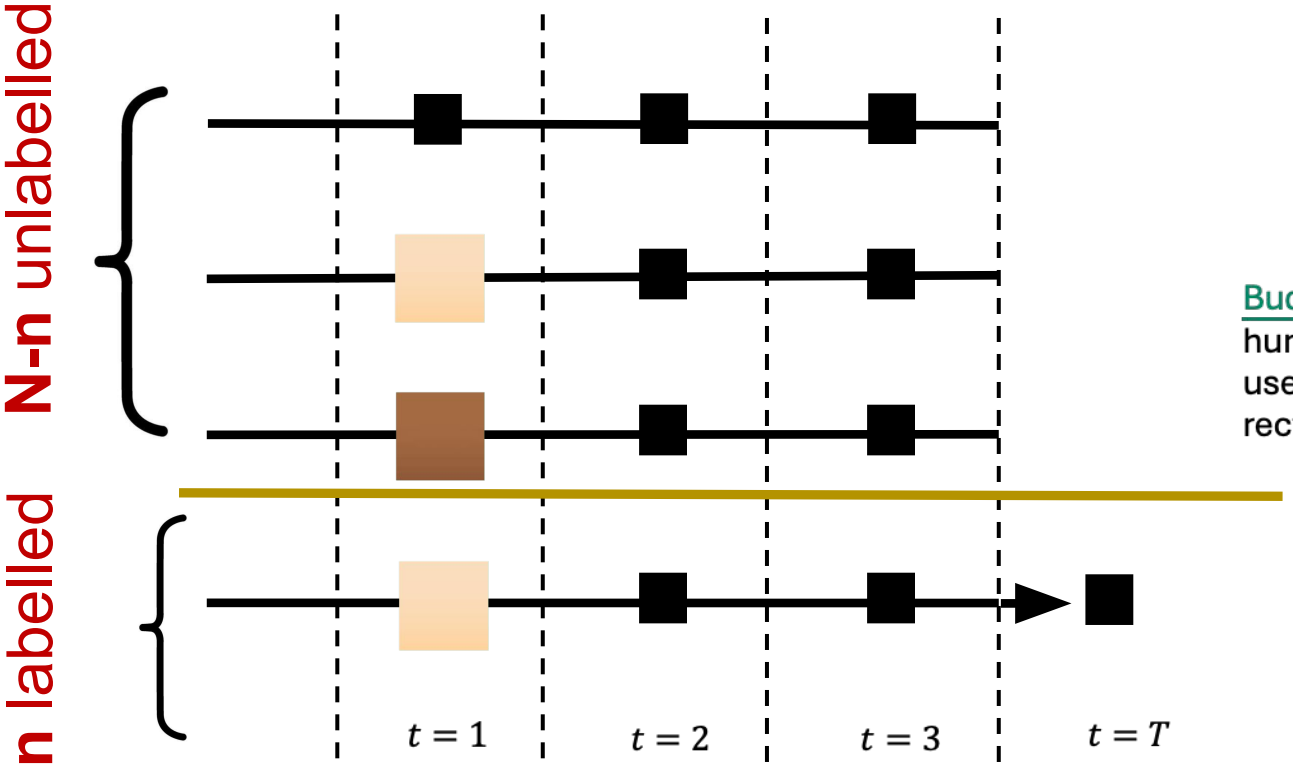


large language model

Issue: these are are potentially inaccurate answers!

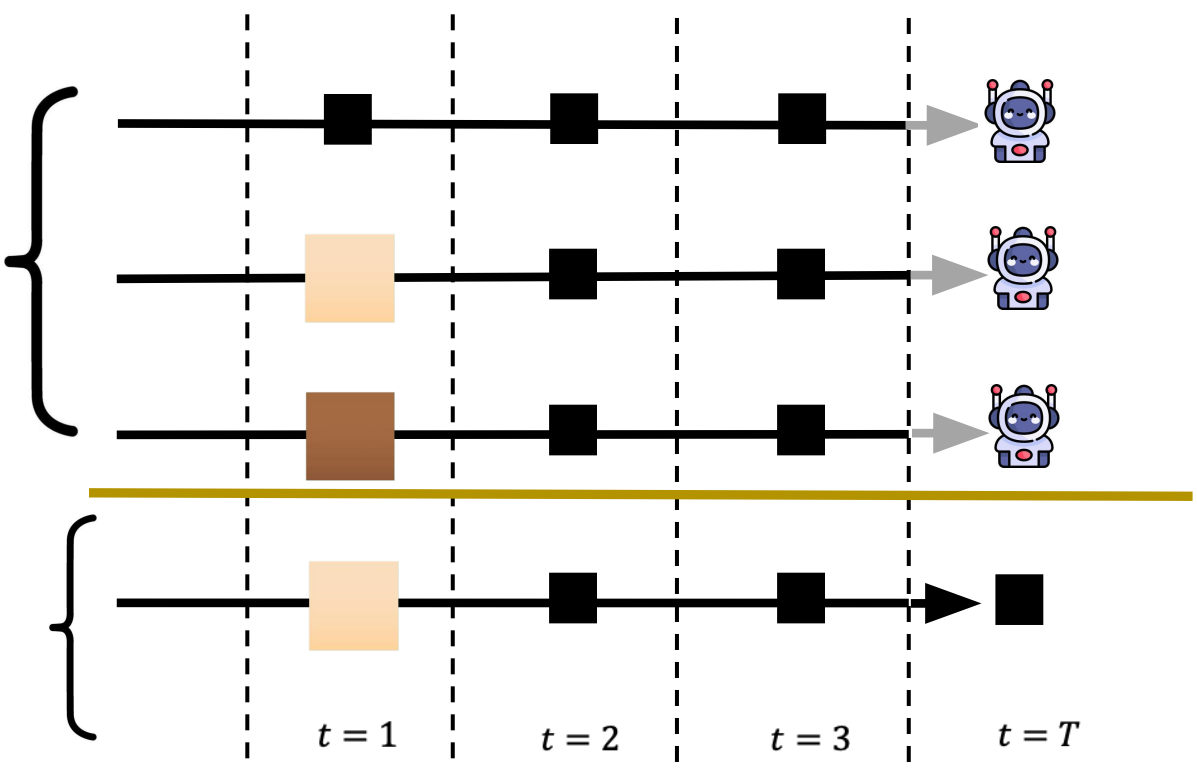
Unless we are willing to assume that the LLM is accurate, there is no hope of reaching valid conclusions without any human answers!

The setup: Let's define a survey

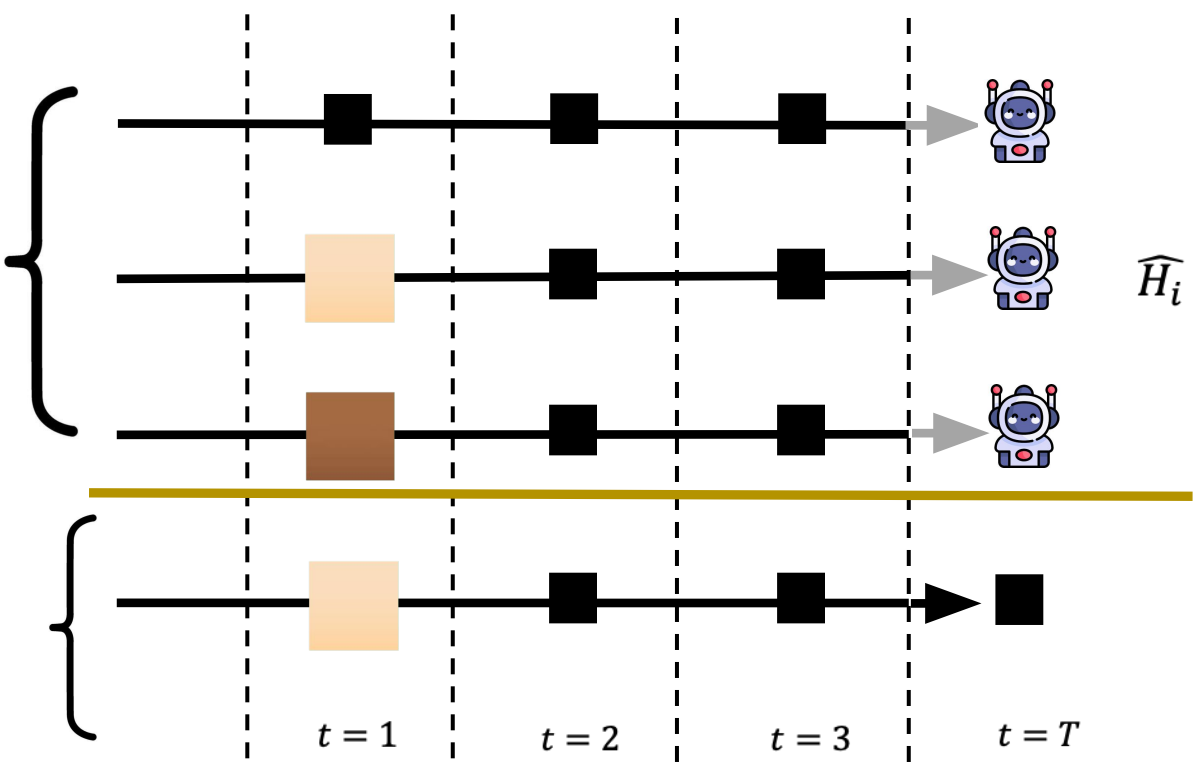


Budget: can collect at most $n \ll N$ human survey answers that can be used for synthesis and/or rectification

The setup: Let's define a survey



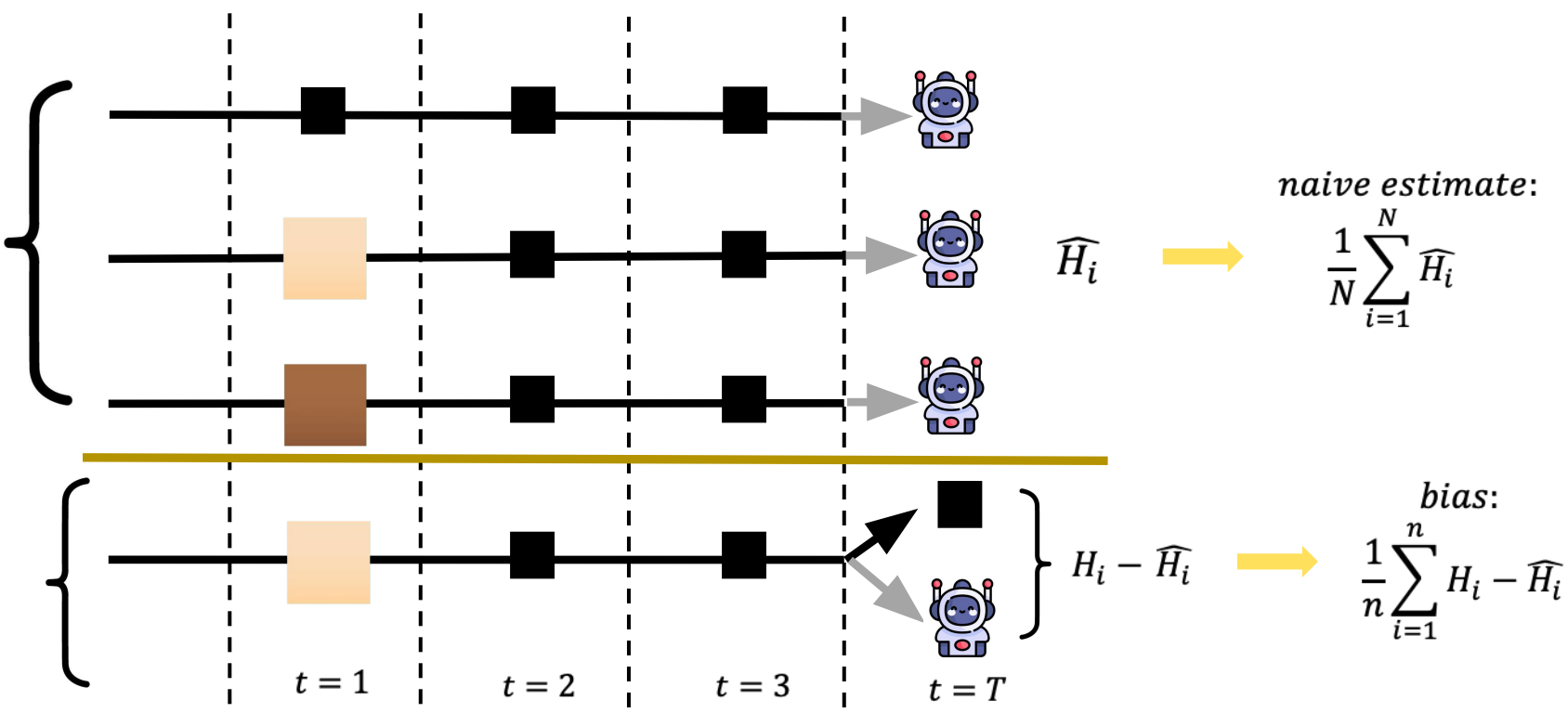
The setup: Let's define a survey



naive estimate:

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{H_i}$$

Prediction-Powered Inference



Prediction-Powered Inference

$$\hat{\theta}^{\mathrm{PPI}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\widehat{H_i}}_{\text{naive estimate}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}H_i - \widehat{H_i}}_{\text{bias}}$$

Theorem. For any data, $\hat{\theta}^{\mathrm{PPI}}$ is:

- accurate: $\hat{\theta}^{\mathrm{PPI}} \to \theta^*$ as the data size grows
- well-behaved: $\hat{\theta}^{\mathrm{PPI}} \approx N(\theta^*, \sigma^2)$

$\Rightarrow$ can form a confidence interval $(\hat{\theta}^{\mathrm{PPI}} \pm r)$

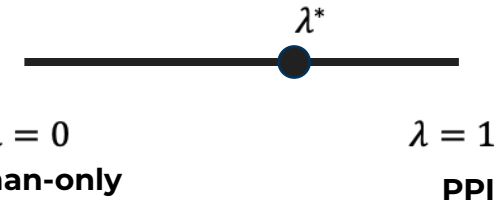
Angelopoulos et al. 2023. Science https://www.science.org/doi/full/10.1126/science.adi6000
Egami et al. 2023. Neurips https://arxiv.org/abs/2306.04746

Safeguard Against Poor LLM Answers

If the LLM outputs very noisy responses, we might be worse off than with the human-only approach!

Power tuning interpolates between using and not using LLMs

$$\hat{\theta}^{\text{PPI}} = \underbrace{\text{mean}\left(\lambda \cdot \widehat{H}_{n+1}, \ldots, \lambda \cdot \widehat{H}_N\right)}_{} - \underbrace{\text{mean}\left(\lambda \cdot \widehat{H}_1 - H_1, \ldots, \lambda \cdot \widehat{H}_n - H_n\right)}_{}$$



Optimal tuning λ^* is proportional to how well H and $\widehat{H}$ correlate and can be computed explicitly

What it takes technically

All three methods rely on standard tools. A graduate student can set 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 augmented approach is useful?

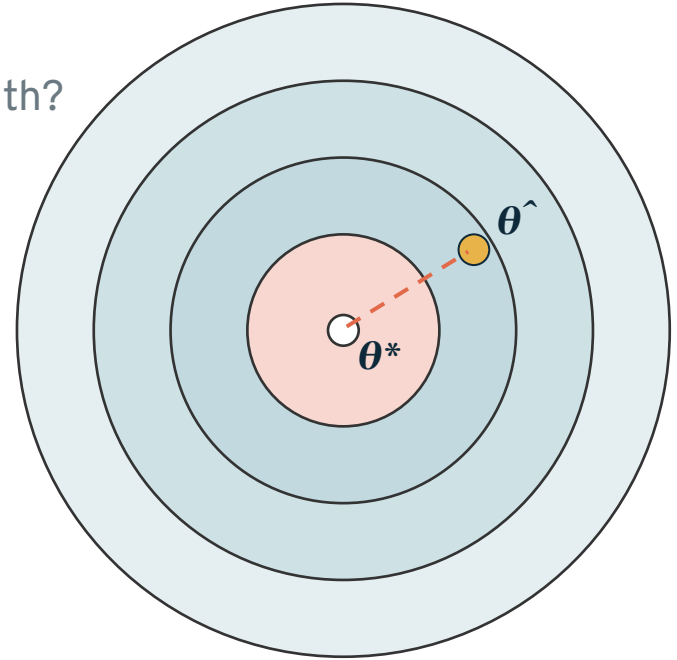
1. Population-level accuracy

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 augmented approach 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 augmented approach 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 augmented approach 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 augmented approach 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

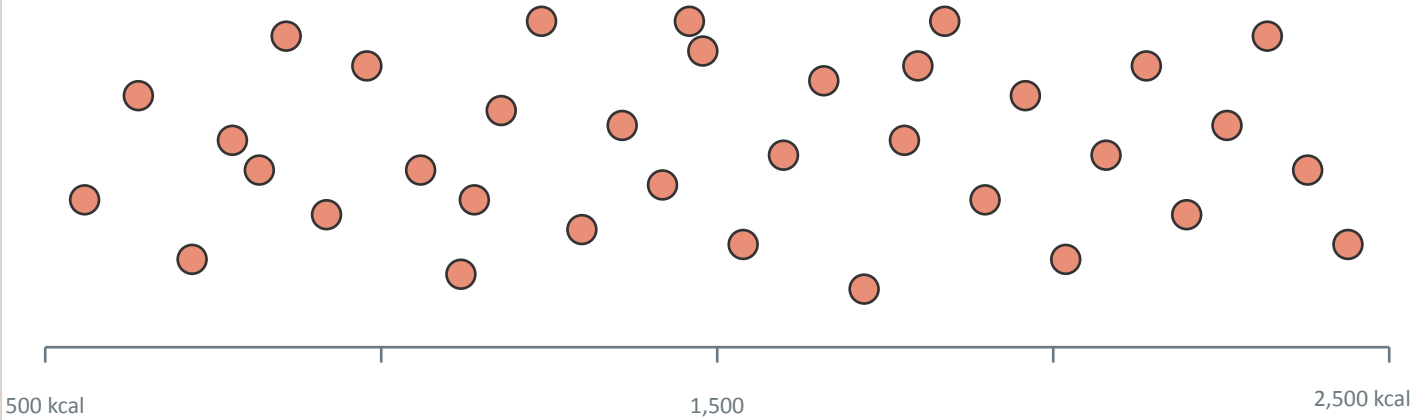
Weight: 60–70 kg

Non-Hispanic White

No special diet

ACTUAL ENERGY INTAKE

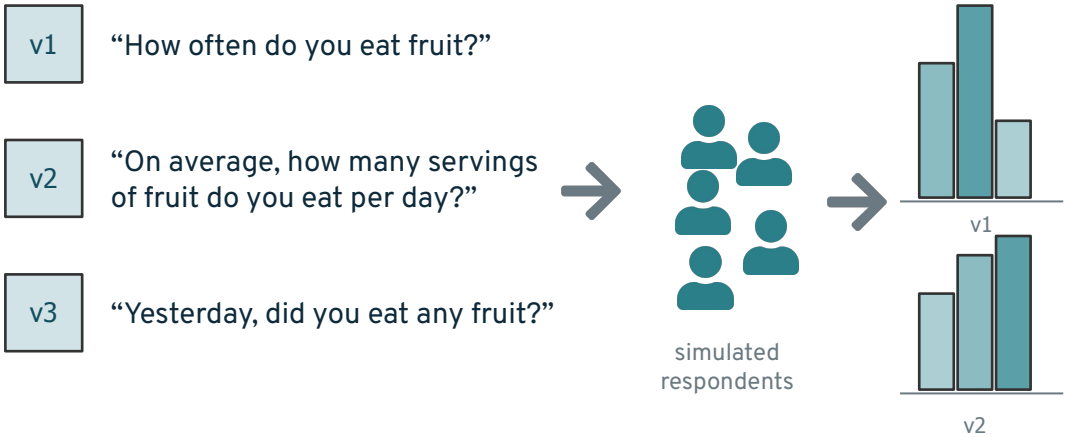
within this single demographic group



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

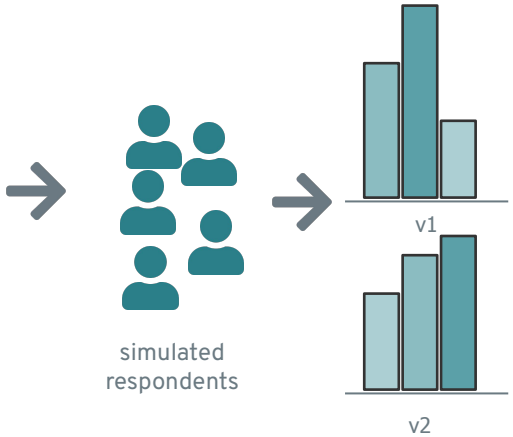
Pretesting: Use simulated respondents to evaluate new candidate survey questions before fielding them with real participants.

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

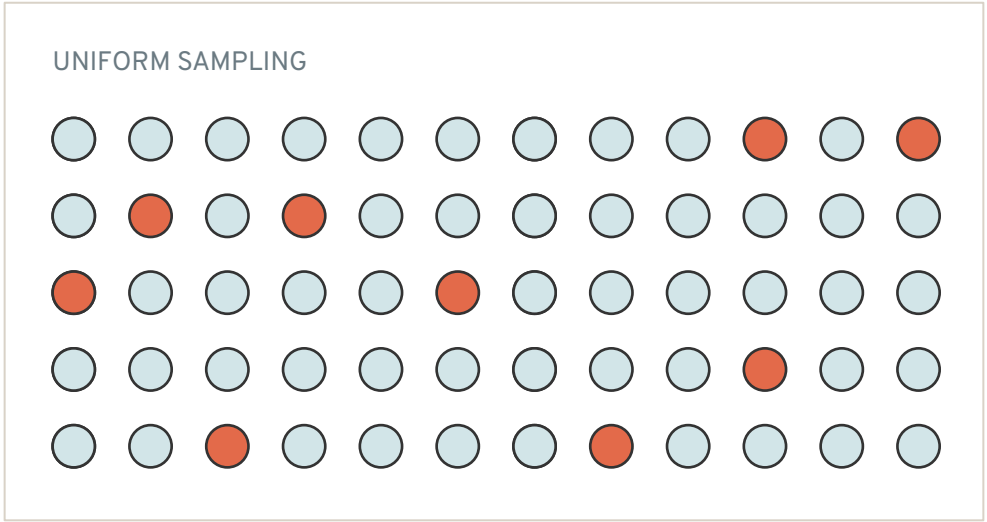
before any human pilot is run

Speeding up discovery

Reducing the number of iteration cycles

How this can support survey production

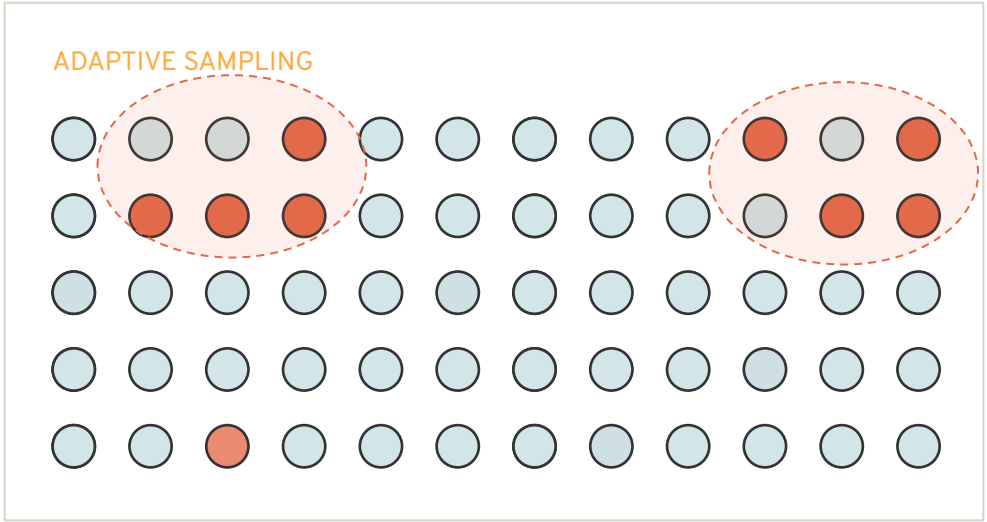
Adaptive sampling: Use model predictions to identify the most informative respondents



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. https://aclanthology.org/2025.naacl-long.179/#

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. https://aclanthology.org/2025.naacl-long.179/#

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
- Hybrid approaches bring the best of both worlds
- Excellent prototyping tools
- The possibility of lowering costs of survey development via pretesting

  …using standard tools and relatively small investments

Q&A

Questions?