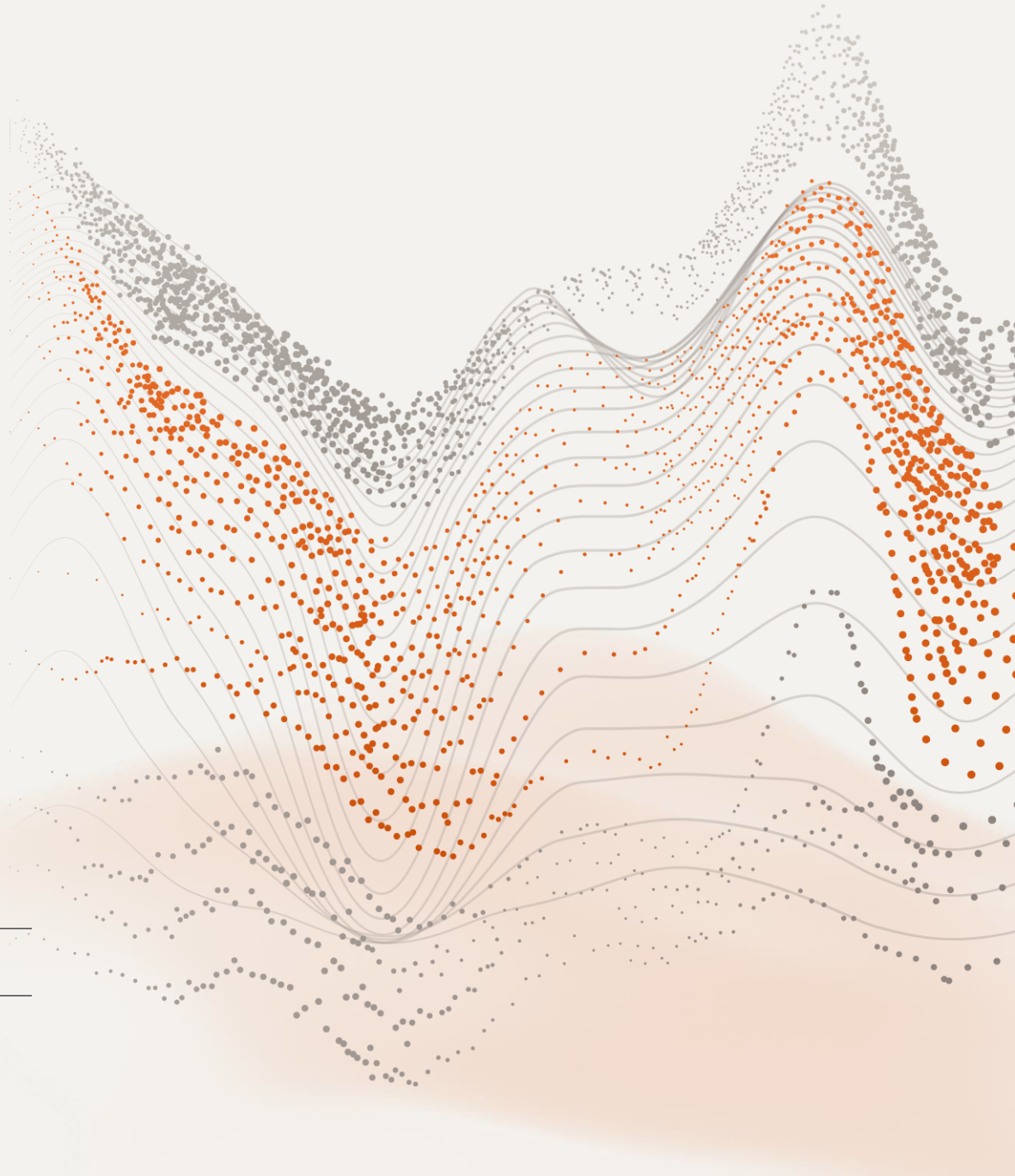
NORC Research Science

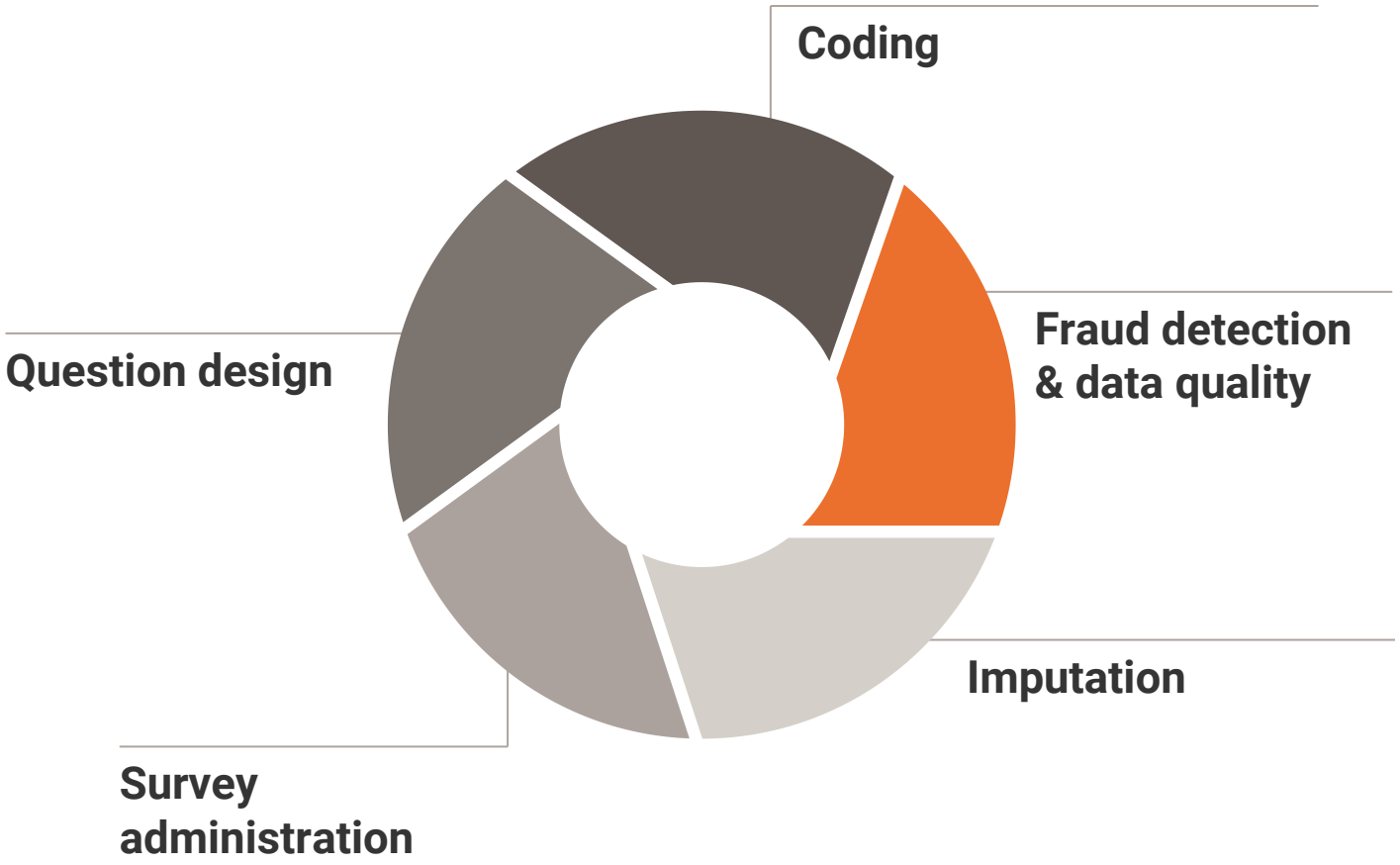


Detecting LLM-Generated Survey Responses

September 24, 2024

Brandon Sepulvado, Joshua Y. Lerner, Lilian Huang, Ipek Bilgen, Leah Christian

AI can help survey methodologists in a range of ways.



However, AI also poses challenges for survey research.

Open-ends entail a sort of balancing act.

- Rich detail but potentially burdensome and error prone (e.g., measurement, processing, nonresponse)
- Too burdensome → use of AI

AI poses a real challenge for survey data quality. (Lebrun et al. 2024)

- Responses from LLMs can be difficult for humans to identify.
- Current AI detection tools perform poorly.
- Existing protocols, e.g., attention checks, are likely to decrease in effectiveness as AI improves.

Lebrun Benjamin , Temtsin Sharon , Vonasch Andrew , Bartneck Christoph. 2024. "Detecting the corruption of online questionnaires by artificial intelligence." *Frontiers in Robotics and AI* 10: 1-15.

However, AI also poses challenges for survey research.

PROBLEM

Open-ended survey responses from LLMs—like ChatGPT—are increasingly common.

SOLUTION

We have piloted a tool that detects AI-generated responses to open-ended questions.

IMPACT

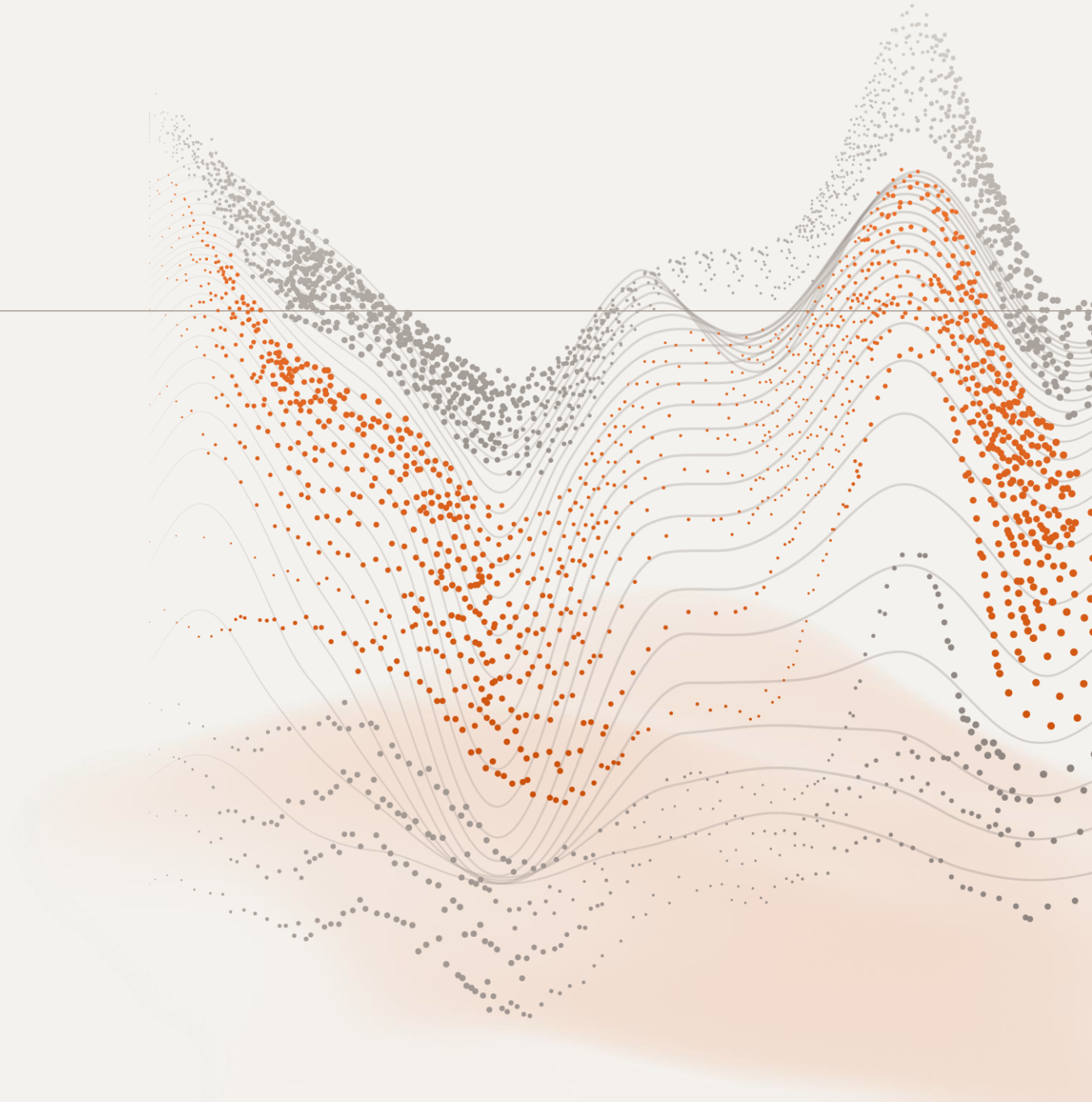
Higher quality data

Increased trust

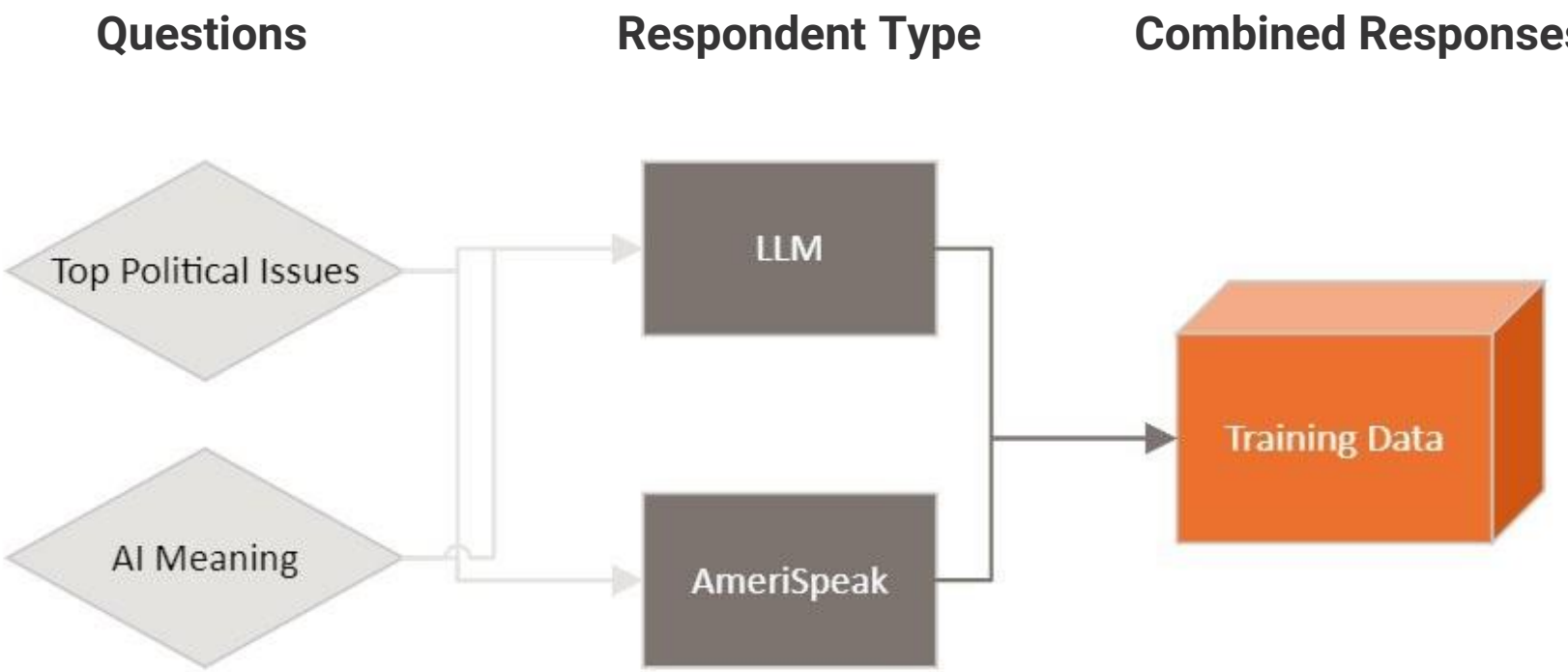
Reduced project costs

Mitigation of other risk(s)

Modeling Process



How did we create training data?



How did we create training data?

Questions

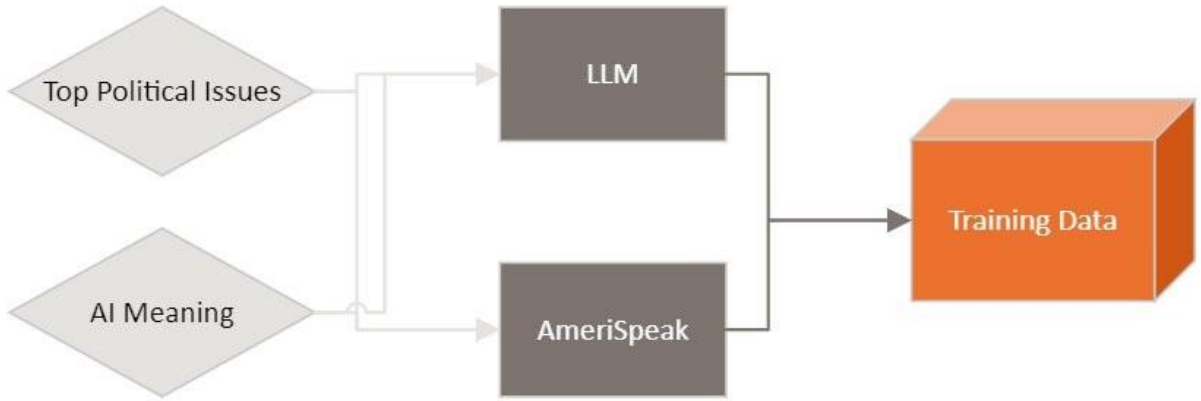
- Understanding of AI
- Most salient policy issues

Sample

- AmeriSpeak Omnibus panel

LLMs

- GPT 3.5
- GPT 4
- Llama 3.1
- Claude Sonnet 3.5



Predicting which responses come from humans

Outcome and outputs

- LLM or person
- Probability that a response comes from an LLM + label

Input

- Open-ended responses

Processing

- “Traditional” text analytic approaches
- No LLMs!

Precision	Recall
0.989	0.999
F1	**Accuracy**
0.994	0.990

Additional baselining

Out of Sample Performance

- Questions from very different survey
  - Medical domain
  - Very technical
- Different population
  - Not general population
- Very good performance
  - Accuracy as high as 94.8%
  - Precision as high as 85.7%
  - Recall as high as 100%

Commercial AI Detectors

- Tested multiple commercially-available AI detector tools
- Their performance tended to be in the 50—75% accuracy range.

What have we learned?

1.

Excellent Performance

- Cross-domain accuracy (topics, respondent type)
- Framework for easy fine tuning

2.

More Efficient Data Collection

- Reduce time to review
- Not computationally demanding

3.

Higher Quality Data

- Increased trust in data product(s)
- Transparent data curation

4.

AI Benefits Survey Methodology and Research!

- AI can help counteract the increasing data quality concerns that AI introduces.

What's next?

Bias

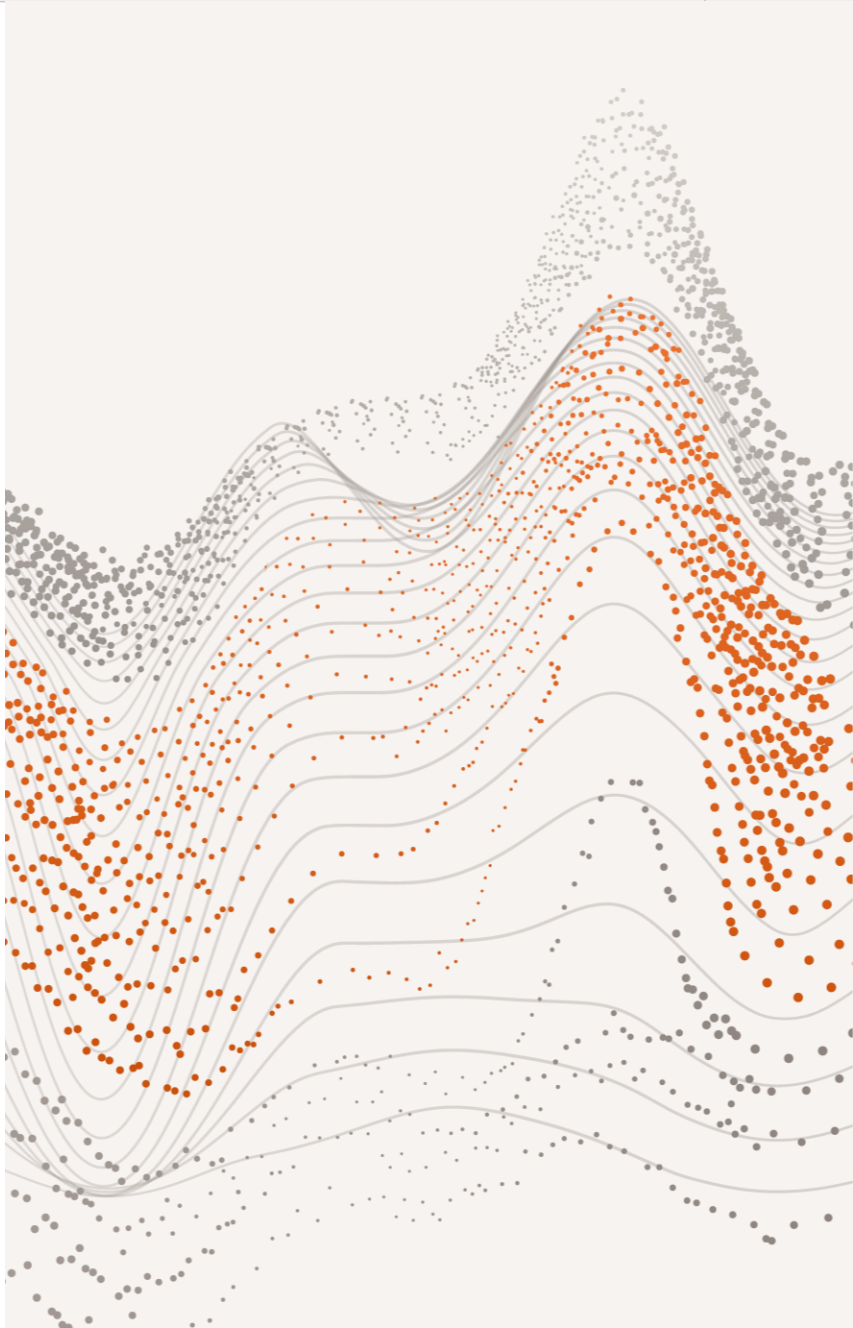
- Formal assessment into whether bias exists in classifications and how we are mitigating it
- When is LLM use permissible?

Deployment

- Application programming interface
- Graphical user interface

Performance Improvement

- Fine tuning, adversarial approaches
- New(er) LLMs



Thank you.

Brandon Sepulvado
Senior Research Methodologist
Huang-Lilian@norc.org

Research You Can Trust™

NORC Research Science