

Benchmarking AI Models using Surveys

Bolei Ma & Frauke Kreuter

Summer 2026

Course Abstract

This course introduces the concepts of benchmarking AI using surveys as tools. The course provides an overview of understanding the AI-as-respondent interaction and offers practical insights into using AI as proxies for human participants to respond to survey items. The course covers methods, techniques, and real-world examples for designing and conducting benchmark studies to assess AI performance and bias through surveys as a testing tool, as well as its utility in reliably generating and analyzing survey data.

Course Objectives

By the end of the course, students will...

- Understand the "Silicon Sample" concept: how AI models are used as proxies for human respondents.
- Learn modern prompting techniques as a form of "interviewing" the AI.
- Apply survey reliability and validity metrics to AI model outputs.
- Design a benchmark study to test for demographic and cognitive bias within an LLM.
- Explore how to use AI for automated coding of open-ended survey responses.

Course Composition and Teaching Methods

This is a 4-week online course using a flipped classroom design.

- Pre-class: Watch video-recorded lectures and read required literature.
- In-class: Mandatory weekly one-hour online meetings for discussion and live labs on Friday, 12:00-13:00, ET.
- Workload: Approximately 12h/week (3h in-class/videos, 9h independent study and assignments).

Each week, students are expected to complete the recorded lectures and submit the assigned homework by **Thursday at 23:59 ET**. During the weekly Friday meeting, we will review and discuss the course materials and homework assignments.

There will be a total of **five homework assignments**. The first four assignments correspond to the four instructional weeks of the course. The fifth and final assignment will serve as a wrap-up exercise covering the course topics and will be due on the **Thursday following the final class session at 23:59 ET**.

Grading will be based on:

Final grades will be based on the following components:

- Homework assignments (90%)
 - The five homework assignments will be graded on a total of 80 points, which will then be scaled to account for 90% of the final course grade.
 - Homework 1: 15 points
 - Homework 2: 15 points
 - Homework 3: 15 points
 - Homework 4: 15 points
 - Homework 5: 20 points
- Participation (10%)
 - Participation will be evaluated based on attendance and engagement during the weekly online meetings and/or the submission of questions and comments to the weekly discussion forums.

A+ 100 - 97

A 96 - 93

A- 92 - 90

B+ 89 - 87

B 86 - 83

B- 82 - 80

Etc.

Variations for grading on a scale are at the discretion of the instructor.

The final grade will be communicated under the assignment "Final Grade" in the Canvas course. Please note that the letter grade written in parentheses in Canvas is the correct final grade. The point-grade displayed alongside the letter grade is irrelevant and can be ignored.

Dates of when assignment will be due are indicated in the syllabus. There will be a grace period for late assignments, but late assignments will be penalized according to the following rules:

1 day late: 10% off

2 days late: 25% off

3 days late: 50% off

4+ days late: no credit

Readings

- Argyle et al. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis. <https://doi.org/10.1017/pan.2023.2>
- Santurkar et al. (2023). Whose Opinions Do Language Models Reflect? ICML. <https://proceedings.mlr.press/v202/santurkar23a.html>
- Ma et al. (2024). The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. Findings of EMNLP. <https://aclanthology.org/2024.findings-emnlp.513/>

Week 1: Introduction and Background

Focus: Can an AI have an "opinion"? Introduction to LLMs for social scientists; The "Silicon Sample"; A short overview and background of how modern LLMs are trained

In-class session: July 10, 12:00-13:00 ET

Week 2: Prompting and Opinion Generation

Focus: The psychology of the prompt; Social-demographic prompts, Personas; Zero-shot vs. Few-shot prompting; priming effects in AI.

In-class session: July 17, 12:00-13:00 ET

Week 3: Benchmarking and Evaluation

Focus: Testing for demographics; Assigning personas, measuring LLMs performance using human surveys as baseline; identifying "WEIRD" (Western, Educated, Industrialized, Rich, Democratic) bias in benchmarks; Evaluation metrics

In-class session: July 24, 12:00-13:00 ET

Week 4: Extended Challenge on Open-Ended Surveys

Focus: Benchmarking open-ended response generation; Comparing closed-ended and open-ended response generation; Using LLMs for coding open-ended responses.

In-class session: July 31, 12:00-13:00 ET