1. Motivation:

What question(s) do you seek to answer and why?

1. Background:

Who has worked on this before? How does this project relate to the content covered in the lectures and tutorials?

1. Data:

Describe the data and any resources used to collect it. Explain how you filtered data, normalized values, computed additional variables, etc.

1. Methods:

Detail the data analysis and statistical methods. Justify your choice.

1. Results:

Show the results of statistical analyses and visualizations that assess the question(s).

1. Discussion:

Evaluate answers to the question and their reliability. Assess limitations and critique. Relate to results from other work or replicated paper

1. **Motivation**

*RQ1: How well do LLMs estimate general public opinion?*

* Analyze how accurately an LLM can predict the results of surveys. Is this answer skewed towards its own stance on the topic?

Check if a better performance can be obtained by inputting comments from social media. *RQ2: Are comments a faithful reflection of society (as measured by the surveys)? When fed into an LLM, does it then yield more polarized estimations of public opinion?*

**2a. Previous results on LLM implementations for in silico survey results**

* Papers that work with ANES – other applications not really this “social census”

I think this is the most related paper: <https://osf.io/5ecfa/download> . **Artificially Precise Extremism: How Internet-trained LLMs Exaggerate Our Differences (Bisbee)**

Introduction

Survey research provides valuable input for the policymakers to make decisions on matters related to economic strategy and public health (Page et al., 1983). In the last decades, despite the money and time investment allocated to surveys (Jansen et al., 2023), these have shown increasingly difficulties regarding their non-response rate and attrition, affecting results accuracy and representativeness (Bisbee et al., 2023; Chu et al. 2023).

Due to their practical consequences and rising problems, finding solutions to the usual pitfalls of surveys with automized pipelines can be of high interest to researchers and governments. Large Language Models (LLMs) have been investigated as a possible answer for some of these shortcomings thanks to their capabilities to reproduce humanlike answers, low costs and fast performance (xxx). However, the new problems that arise with them make still their best implementation unsure, being until now implemented with different methodological approaches.

In this project we test the quality of LLMs results when prompted to create an in silico social census on public opinion matters.

1. **Background**

Previous research on LLM implementations for survey results prediction has leveraged the promising capabilities LLMs have to adopt personas. By stating different sociodemographic group belonging, the LLMs create tailored answers towards public opinion or group affinity questions.

Bisbee et al. (2023), for example, created synthetic personas for ChatGPT 3.5 Turbo and asked it to rate a series of groups using thermometer scores as in the American National Election Survey (ANES). In their results they observed that average synthetic opinions were often substantively and statistically indistinguishable. However, the answers altogether showed a bias towards greater affective polarization by showing greater in-group preference and out-group rejection. Additionally, the in-silico answers were more homogenous, accounting for a smaller standard deviation, which could not be compensated with higher model temperatures.

Open source (OS) models have also been implemented with the aim of creating synthetic survey responses. As compared to black box models, OS models offer the opportunity to be further trained for specific tasks by changing their parameters through fine-tuning. In the research by Chu et al. (2023), a BERT model was fine-tuned with the publications of four major mass media outlets. The fine-tuned model had a greater performance at predicting individuals attitudes towards COVID-19. However, in the field of consumer confidence, questions regarding personal matters such as individuals’ financial situation or housing value had only low or negative correlations with their ground truth counterparts. As explained by the authors, these results align with findings that show that news coverage mostly affect sociocentric and prospective attributes.

Talk about bias in llms, also paper on in silico samples from class.

Implementing LLMs for questions on social matters carries risks related to their training biases. To perform next token prediction and learn insights on correct language forms, LLMs are trained on unsupervised tasks with a big amount of text. Given that social groups do not have equal access to collaborating to these datasets, also called “documentation debt”, models are trained with a bias towards hegemonic viewpoints (Bender et al., 2021). In Besbee (2023) research the model results attributed to Non-Hispanic Whites were

Also maybe ethic problems underlined in KIM paper

Accounting for this background in this project we propose to analyse the accurateness with which black box (GPT-3, GP4 for zero shot?) and OS models can predict the outcome of a survey on social matters from different areas of interest. To do so, it is pursued to further simplify previous approaches by prompting the model to output the frequency of the multiple answer categories. In other words, instead of creating in-silico answers for the different personas, the aim of this research is to assess the model’s ability to perform social census. We are utilizing the 2020 American National Election Studies (ANES) Time Series Study as a benchmark to examine discrepancies in answer frequencies generated by the model. Our objectives include quantifying the extent of these variations and assessing whether the model exhibits any inclination towards specific political attitudes. Our selected questions consist of some of the most regarded topics from the year of the study (2020): Immigration, abortion, gun control, climate change and public health.

As a first step, we are simply asking the model to answer our selected questions from the ANES study by giving an estimate for the frequency distribution of answers (zero-shot).

In addition to zero-shot trials, this project also aims to leverage the capability of LLMs to be sensitive to in-context learning. For this, models will be inputted at prompt time with discussions over the topics of interest in the platform Reddit. Given this additional input, our aim is to evaluate whether Language Models (LLMs) can adjust their initial estimations based on the provided discussion context and potentially yield results closer to the original answers of the ANES survey. In essence, we seek to determine whether the "social sensing" capabilities of LLMs can be calibrated toward a specific representative group, potentially mitigating biases inherent in the training data. Using users’ discussions will lessen the shortcomings of mass media data, by directly collecting citizens opinions on the matter. Nevertheless, polarization bla bla bla.

**2b. Context-learning and context extension**

* Papers that work with extended context-window. We explored SelfExtent blabla

1. **Data**

Reddit data via PRAW.

* Filter for 2020 (ANES years)
* Filter for topics according to key terms; selected submissions according to number of comments/ upvotes threshold
* Manually select submissions and extract comments

ANES Survey

* Topics chosen: something about the possible responses range? Not yes/no questions, or different formats
* Abortion, climate change, immigration, gun control, healthcare: why these topics? More clear inclination between progressive and conservatives?

ANES is a collaboration of [Duke University](https://duke.edu/), [University of Michigan](http://www.umich.edu/), [The University of Texas at Austin](https://www.utexas.edu/), and [Stanford University](http://www.stanford.edu/). The study is conducted among US-American voters every 4 years, before and after the presidential elections. The respondent is asked about multiple demographic details and then about his or her opinion towards multiple societal issues coming from a wide range of topics. We chose topics among the most anticipated issues in the 2020 presidential election (<https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/>). Furthermore, we believe our chosen topics Immigration, abortion, gun control, climate change and public health are object of drastically differing opinions and therefore also likely to be discussed controversially in social media.

The study is conducted every 4 years among American voters, both before and after presidential elections. Respondents are queried about various demographic factors before being asked for their opinions on a wide array of political and societal issues. Our selected topics align with those anticipated to be most significant in the 2020 presidential election (<https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/>). Moreover, we have chosen the topics immigration, abortion, gun control, climate change, and public health, which are known for sparking contentious discussions across social media platforms due to their propensity for evoking drastically differing opinions.

1. **Methods**

We explore in-context learning, using OS fine-tuned models that extend the context window. For GPT we take advantage of the “open session” memory.

Use ~3 models: GPT, Mistral, Codellama

* Zero shot
* Few shots:
  + Top level comments
  + Comment forest
* Several runs (3?)

3\*5\*3 = 45 runs for each model

1. **Results**

* Explore nice measures/visualizations for comparing distributions

1. Discussions

* Bias of Reddit population - also interesting if we show it as a measure of polarization on Reddit though, but without having a non-biased group or exact description of Reddit population we cannot do comparisons.
* Extracted from r/news, which we assume as with no political leaning but probably has a more democrat leaning (maybe we could do a quick exploration of which newspapers are mostly shared there and see the classification). Are we leaving conservatives out? Could explore
* Further applications: parliamentary debates, newspaper articles, other social media.