Prompted for a Discussion About Prompt Engineering

Trent D. Buskirk, Old Dominion University

SoDa Symposium: Prompt Engineering to Support AI Enables Research

Presentation with discussant and Q&A

Extrinsic Motivation…

- **Large Language Models (LLMs) that power ChatBots are set to converge to so called "general purpose tools" in the future**
  - **capable of completing many tasks currently performed by humans.**
- **Eloundou and colleagues (2023) recently published research that explored the impact of GPTs on several different types of jobs and occupations and in their work they estimated that Survey Research was:**

the **second-most** exposed field for work that could be completed by GPTs

the **highest** exposed occupation relating to GPT-powered software.

The exposure provides an estimate of how GPTs or GPT-software can be leveraged to save workers in these fields a significant portion of time for completing a majority of their tasks.

Levels of Interaction with LLMs…

Prompting

Please provide the reading level of the following question: “Do you love this talk so far?”

The question **“Do you love this talk so far?”** is written at a **basic reading level**, typically suitable for **Grade 2-4** or **ages 7-9**.

Few Shot Learning

The reading level of the following question is 4rd grade:
Do you think this meeting is fun?
The reading level of the following question is 8th grade:
“Do you like coming to MAPOR?”
Please provide the reading level of the following question:
“Do you love this talk so far?”

Based on the comparisons provided:

- The question **“Do you think this meeting is fun?”** (4th grade) uses simple vocabulary and a straightforward structure.
- The question **“Do you like coming to MAPOR?”** (8th grade) introduces a proper noun ("MAPOR"), which increases the complexity slightly due to its need for specific context or familiarity.

Estimated Reading Level: 5th-6th grade

Fine Tuning

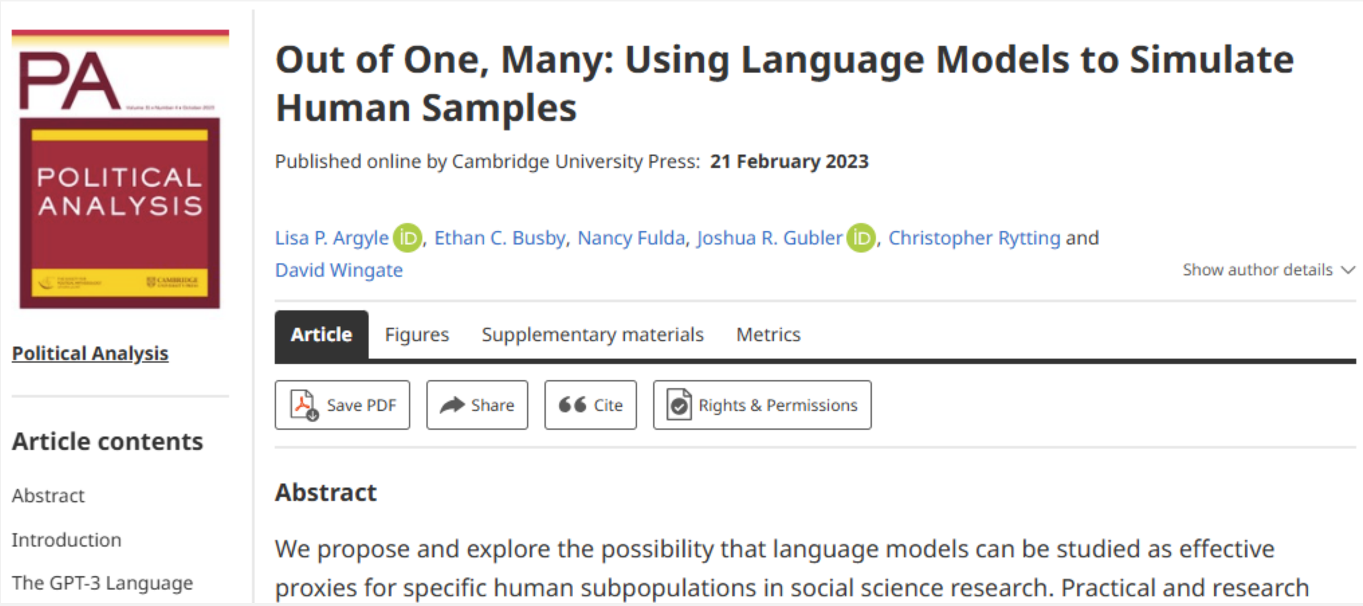
A data set with at least 10 to more than 100 examples are provided and input into the LLM which then creates a modified model that can be used to answer questions or respond to prompts.

LLMs/AI as a tool for Survey Responses: Silicon or Synthetic Samples

- One area in Survey Research where we have also seen some experimentation in prompting is around generating Synthetic Samples.
- Prompts may vary in their complexity and amount of detail they include about an actual respondent whose answers are to be "simulated" using and LLM.

Argyle and colleagues (2023): Introduced the concept of "silicon samples" and criteria for assessing "algorithmic fidelity" for LLMs and report nuanced similarities between human and AI generated responses.

https://bit.ly/ArgyleEtAl2023


PA
POLITICAL ANALYSIS

Political Analysis

Article contents

Abstract

Introduction

The GPT-3 Language

Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: **21 February 2023**

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate

Show author details

Article Figures Supplementary materials Metrics

Save PDF Share Cite Rights & Permissions

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research

LLMs/AI as a tool for Survey Responses: Silicon or Synthetic Samples

Bisbee and colleagues (2023) report contrary findings that suggest that silicon samples generate responses that are <u>far less variable compared to actual survey respondents' responses</u>. They also remark that results can be highly dependent on prompt and LLM version being used.

https://bit.ly/BisbeeEtAl2024

PA
POLITICAL ANALYSIS

Political Analysis

Article contents

Abstract

Research Design and Data

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

Published online by Cambridge University Press: 17 May 2024

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Show author details

Article Figures Supplementary materials Metrics

Save PDF Share Cite Rights & Permissions

Abstract

Large language models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular LLM ChatGPT can recover public opinion, prompting the LLM to adopt

Sun and colleagues (2023): Improved upon the concept of silicon samples by introducing so called "random silicon sampling" and showed it performed as well or better than silicon sampling for many tasks.

https://arxiv.org/pdf/2402.18144

Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information

Seungjong Sun[1], Eungu Lee[1], Dongyan Nan[2], Xiangying Zhao[2], Wonbyung Lee[1], Bernard J. Jansen[3], Jang Hyun Kim[1,2]

[1]Department of Applied Artificial Intelligence, [2]Department of Interaction Science, Sungkyunkwan University

[3]Qatar Computing Research Institute, Hamad Bin Khalifa University

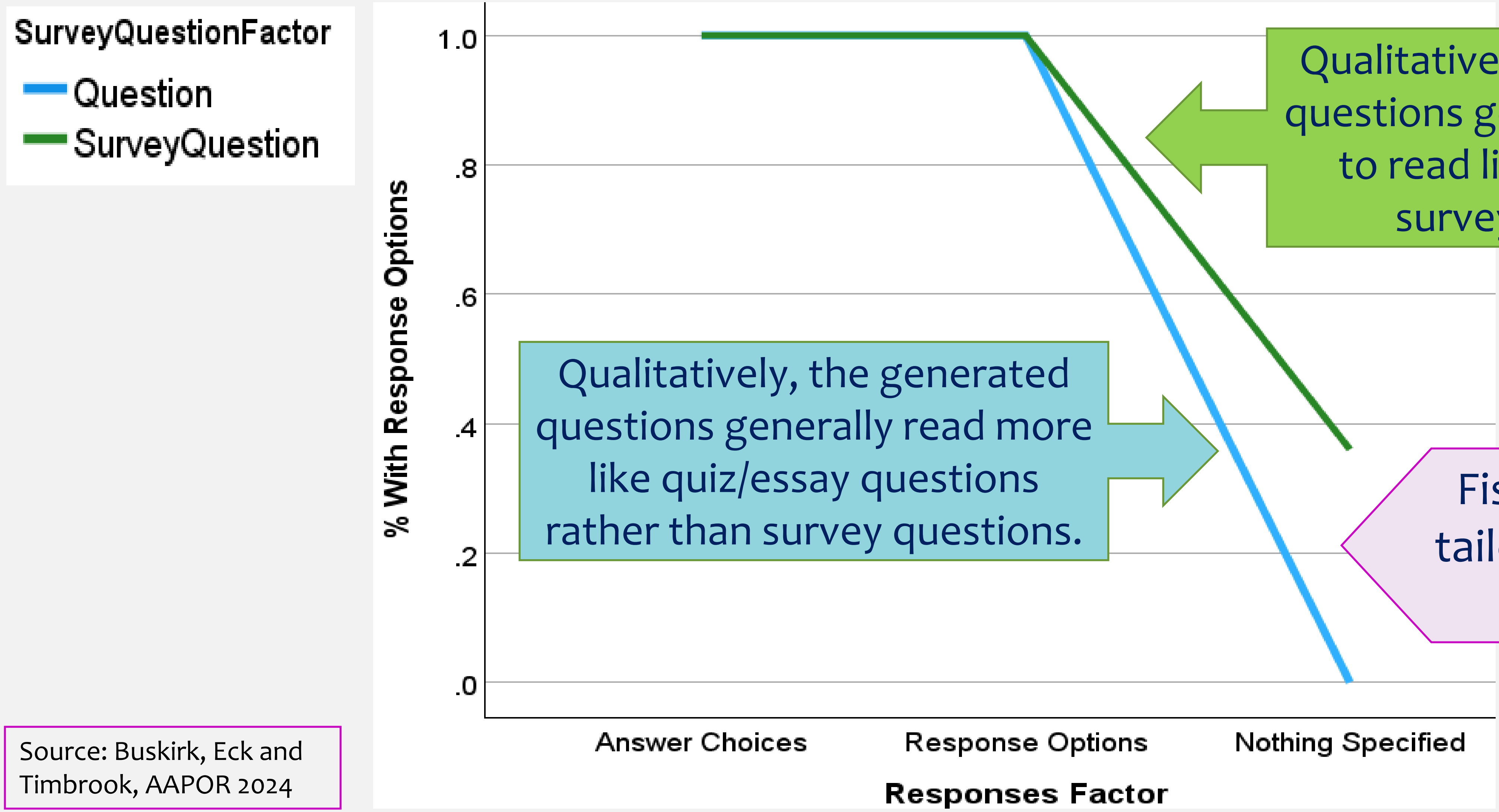
{tmdwhd406, dldmsrn0516, zxy94, co2797}@g.skku.edu, {ndyzxy0926, alohakim}@skku.edu, jjansen@acm.org

Abstract

Large language models exhibit societal biases associated with demographic information, including race, gender, and others. Endowing such language models

human-like biases associated with race, gender, ethnicity, and others from human-written data (Schramowski et al., 2023; Peters and Matz, 2023). Although many studies have attempted to mitigate societal biases in LLMs (Barocas and Selbst, 2016;

What You Don’t Include in a Prompt May Be as Important as What You Do Include



Qualitatively, the generated questions generally appeared to read like open ended survey questions.

Qualitatively, the generated questions generally read more like quiz/essay questions rather than survey questions.

Fisher’s Exact 2-tailed Test p-value <0.001

Source: Buskirk, Eck and Timbrook, AAPOR 2024

Data and Coding: Question Posed in 1st Person

BU

I want to ask adults about the relationship they have with an important mother figure in their life. Provide one such **question** with **answer choices**.

How would you describe your current relationship with your important mother figure?

BU

I want to know about how adults get their health insurance. Provide one such **question**.

What are the different options available for adults to obtain health insurance coverage, especially for those who are self-employed or do not receive insurance benefits through their employers?

Some things can’t be learned in a few examples

- Imagine you wanted to teach a child about word syllabication.
  - Giving a few examples may not be sufficient, especially if those examples are fairly limited (i.e. 1, 2 or 3 syllable words).
- Just as is the case with humans, computers may not be able to sufficiently learn complex patterns from few cases.
  - In such cases, zero-shot or few shot learning through prompting may be insufficient to obtaining accurate predictions/decisions from LLMs.
- In these situations, Fine Tuning with a collection of 10’s to 100’s of examples could be a much better approach.
  - Keep in mind, the diversity of the fine tuning example set will be key to helping LLM’s learn – examples of “Yes” and “No” cases, for example in a binary classification task will be needed to increase accuracy for an LLM to predict a new case.
    - See Chen et al. (2024) for example: https://arxiv.org/html/2402.11725v2

LLMs are not Magicians!

- Even if your prompting is good, it doesn't mean that LLMs can do the requested task.
  - LLMs by definition produce text.
  - Generative AI, more broadly can generate video, audio and other types of output.
  - Olsen and Buskirk (2024) showed that LLMs vary widely in their ability to compute Readability Statistics for a cross section of survey items, despite prompting methods that showed LLM's comprehension of the task and the production of correct formulae for such statistics.

We Prompt You To...

- To move our collective understanding along in prompting and the use of LLMs consider:
  - Including the prompt text, method, date and version of LLM used in the methods section or technical appendix of papers that report your results.
    - This will allow for greater transparency and "reproducibility"
  - Include results of pilot tests you did to finalize the prompts you used.
    - This way we don't have to reproduce the wheel and can follow the logic of how you reached the final prompt.
    - We can also see what didn't work.

It's a two-way street: Some possible ways Survey Research might improve LLM's…

- Survey Researchers are MASTERs at asking questions of humans.
- Prompting is the method humans use to ask questions of LLMs. And LLMs also could be designed to ask questions of humans (i.e. true chatbot style).
- Could Survey researchers bring our history of good question asking science to bear in the development of better prompting for LLMs?
  - Prompting is the new Human Computer Interaction of this era…
- The generative capabilities of language models are highly sensitive to the input prompts (Sun et al., 2023) , especially in the context of survey question responses and can be sensitive to the order of questions like humans (Kalinin, 2023).
  - Survey Researchers understand order effects and could lead the way in designing studies that look at how order and context effects in humans translate to LLMs which are supposed to reflect human language.

Resources

- Most comprehensive paper I have seen on LLMs – types, history etc.
  - https://arxiv.org/pdf/2307.06435

A Comprehensive Overview of Large Language Models

Humza Naveed[a], Asad Ullah Khan[b,*], Shi Qiu[c,*], Muhammad Saqib[d,e,*], Saeed Anwar[f,g], Muhammad Usman[f,g], Naveed Akhtar[h,j], Nick Barnes[i], Ajmal Mian[j]

[a]The University of Sydney, Sydney, Australia
[b]University of Engineering and Technology (UET), Lahore, Pakistan
[c]The Chinese University of Hong Kong (CUHK), HKSAR, China
[d]University of Technology Sydney (UTS), Sydney, Australia
[e]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia
[f]King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia
[g]SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRCAI), Dhahran, Saudi Arabia
[h]The University of Melbourne (UoM), Melbourne, Australia
[i]Australian National University (ANU), Canberra, Australia
[j]The University of Western Australia (UWA), Perth, Australia

- Most comprehensive paper on Prompting/Prompt Engineering to date:
  - https://arxiv.org/pdf/2406.06608

The Prompt Report: A Systematic Survey of Prompt Engineering Techniques

Sander Schulhoff[1,2*] Michael Ilie[1*] Nishant Balepur[1] Konstantine Kahadze[1]
Amanda Liu[1] Chenglei Si[4] Yinheng Li[5] Aayush Gupta[1] HyoJung Han[1] Sevien Schulhoff[1]
Pranav Sandeep Dulepet[1] Saurav Vidyadhara[1] Dayeon Ki[1] Sweta Agrawal[12] Chau Pham[13]
Gerson Kroiz Feileen Li[1] Hudson Tao[1] Ashay Srivastava[1] Hevander Da Costa[1] Saloni Gupta[1]
Megan L. Rogers[8] Inna Goncearenco[9] Giuseppe Sarli[9,10] Igor Galynker[11]
Denis Peskoff[7] Marine Carpuat[1] Jules White[6] Shyamal Anadkat[3] Alexander Hoyle[1] Philip Resnik[1]
[1] University of Maryland [2] Learn Prompting [3] OpenAI [4] Stanford [5] Microsoft [6] Vanderbilt [7] Princeton
[8] Texas State University [9] Icahn School of Medicine [10] ASST Brianza
[11] Mount Sinai Beth Israel [12] Instituto de Telecomunicações [13] University of Massachusetts Amherst
sschulho@umd.edu milie@umd.edu resnik@umd.edu

s.CL] 26 Feb 2025

Abstract

Generative Artificial Intelligence (GenAI) systems are increasingly being deployed across diverse industries and research domains. Developers and end-users interact with these systems through the use of prompting and prompt engineering. Although prompt engineering is a widely adopted and extensively researched area, it suffers from conflicting terminology and a fragmented ontological understanding of what constitutes an effective prompt due to its relatively recent emergence. We establish a structured understanding of prompt engineering by assembling a taxonomy of prompting techniques and analyzing their applications. We present a detailed vocabulary of 33 vocabulary terms, a taxonomy of 58 LLM prompting techniques, and 40 techniques for other modalities. Additionally, we provide best practices and guidelines for prompt engineering, including advice for prompting engineering ChatGPT and other state-of-the-art (SOTA) LLMs. We further present a meta-analysis of the entire literature on natural language prefix-prompting. As a culmination of these efforts, this paper presents the most comprehensive survey on prompt engineering to date.

Additional References of Interest

- **Proposed Prompting Framework for Qualitative Coding At Scale…**
  - https://mbosley.github.io/papers/dqi-paper-bosley.pdf

Towards Qualitative Measurement at Scale: A Prompt-Engineering Framework for Large-Scale Analysis of Deliberative Quality in Parliamentary Debates

Mitchell Bosley

September 3, 2024

Abstract

Analyzing the linguistic, psychological, and social dimensions of large textual corpora has traditionally involved a tradeoff between the richness of the constructs measured and the scalability of measurement. While qualitative approaches like expert human coding can capture complex, high-dimensional constructs, they are often too costly and time-consuming to apply to large datasets. Automated computational methods, on the other hand, scale efficiently but typically measure relatively simplistic constructs. I propose a set of novel techniques using large language models (LLMs) to move past this tradeoff, with the goal of enabling rich, qualitative measurement of complex constructs at scale, and show that by carefully designing prompts that imbue LLMs with the knowledge and reasoning abilities of human experts we can elicit high-quality annotations of latent constructs directly from textual data. I apply this approach to the Discourse Quality Index (DQI), a widely used framework for assessing the deliberative quality of political communication, and show that LLMs can automate the coding of the DQI in a sample of parliamentary speeches at a performance level comparable to human annotators. By comparing a human-annotated database of over 1000 speeches from the US Congress to those generated by LLMs, I demonstrate that by carefully designing prompts with a combination of instructions, contextual data, and a handful of high quality examples of the desired annotation behavior, Generative LLMs can "learn" to perform complex, multidimensional annotations of political speech at the level of expert coders, and at a fraction of the time and effort.

- **Experimenting with various prompting structures and LLM sources for Stated Preference Surveys in the Context of Energy Demand**
  - https://arxiv.org/abs/2412.03162

Computer Science > Computation and Language

[Submitted on 7 Mar 2025]

Evaluating Local and Cloud-Based Large Language Models for Simulating Consumer Choices in Energy Stated Preference Surveys

Han Wang, Jacek Pawlak, Aruna Sivakumar

Survey research is essential in energy demand studies for capturing consumer preferences and informing policy decisions. Stated preference (SP) surveys, in particular, analyse how individuals make trade-offs in hypothetical scenarios. However, traditional survey methods are costly, time-consuming, and affected by biases and respondent fatigue. Large language models (LLMs) have emerged as a potential tool to address these challenges by generating human-like textual responses. This study investigates the ability of LLMs to simulate consumer choices in energy-related SP surveys. A series of test scenarios evaluated the simulation performance of LLMs at both individual and aggregated levels, considering factors in the prompt, in-context learning (ICL), chain-of-thought (CoT) reasoning, the comparison between local and cloud-based LLMs, integration with traditional choice models, and potential biases. Results indicate that while LLMs achieve an average accuracy of up to 48%, surpassing random guessing, their performance remains insufficient for practical application. Local and cloud-based LLMs perform similarly in simulation accuracy but exhibit differences in adherence to prompt requirements and susceptibility to social desirability biases. Findings suggest that previous SP choices are the most effective input factor, while longer prompts with varied factor formats may reduce accuracy. Furthermore, the traditional mixed logit choice model outperforms LLMs and provides insights for refining LLM prompts. Despite their limitations, LLMs provide scalability and efficiency advantages, requiring minimal historical data compared to traditional survey methods. Future research should refine prompt structures, further investigate CoT reasoning, and explore fine-tuning techniques to improve LLM-based energy survey simulations.

THANK YOU!!

tbuskirk@odu.edu

@trentbuskirk