Fostering Deeper Reflection with Adaptive Questions in Metadata Label Creation Using Large Language Models APCOMP 297R Fall 2022 Research

Ivonne Martinez
ivonne_martinez@g.harvard.edu

Max Urbany
maximilian_urbany@g.harvard.edu

Zana Bucinca
zanabucinca@g.harvard.edu

1 Introduction

Due to the fast-paced nature of the tech industry, many of today's algorithms rely on incomplete, misunderstood, and historically problematic datasets. This can and has been shown to manifest itself in producing algorithms that can reinforce algorithmic bias and unfairness. Organizations and research institutions alike have been exploring the field in order to understand how to solve this issue and create tools that can help make datasets more reliable. Some of the most notable works out there that progress towards resolving some of these issues are Datasheets for Datasets, A Nutrition Label for Rankings, Data Statements for Natural Language Processing, Apple's Privacy Label, Google's Model Cards, IBM's AI FactSheets 360, and the Data Nutrition Project.

2 Motivation

The Data Nutrition Project (DNP) focuses on developing a diagnostic framework that lowers the barrier to standardized data analysis by providing a comprehensive overview of the dataset before model development as mentioned in *The Dataset Nutrition Label:A Framework To Drive Higher Data Quality Standards*[1]. These diagnostic frameworks are called 'Data Nutrition Labels' and are generated from the user input and dataset with the following information: origin and lineage of the dataset, variable description, simple statistics, distributions and linear correlations, probabilistic model, etc.

Dataset documentation, risk, and bias considerations are increasingly important in creating a better understanding of datasets and higher-quality models. Since datasets and people vary, broad questions such as those used to generate Data Nutrition Labels may fail to capture important context-dependent information. In order to generate Data Nutrition Labels that are robust the model requires high-quality user responses that reflect elaborate thoughts and reflections. Existing studies have determined that prompts in the form of questions yield more instances of critical thinking.

Our work focuses on fostering deeper reflection with adaptive question generation in metadata label creation using large language models. As research suggests, prompts to encourage points that should be considered in response may yield more fruitful answers thus this research focuses on creating user-specific templates to generate suggested prompts to further qualify and condition user-generated responses.

3 How can we elicit deeper reflections?

Eliciting deeper reflections on the data label creation questions would help create better labels. Our work took into considerations some of the main ideas found in *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI*

[1] https://arxiv.org/abs/1805.03677

in AI-assisted Decision-making [2] and considered them as motivation towards our work.

3.1 Variability of datasets and people

Not only do datasets vary greatly between the format of the data, the content, and the way it was collected to name just a few, but people of all backgrounds interact and use datasets both directly and indirectly. Given this inherently large space in human-dataset interactions there is an opportunity to provide context-based assistance to the individual to better investigate the dataset.

3.2 Feedback as questions

Rather than provide contextual feedback in the form of objective sentences that can be read through, we have chosen to provide feedback in the form of questions. In the interest of eliciting deeper thought, supplemental questions provide and require a pause in cognition as questions imply interaction whereas simple sentences can be processed without pause for reflection with respect to absorbing information.

3.3 Over-reliance: Recommendations lead to over-reliance on AI

Existing works have investigated the reliance on AI recommendations and have found that people have a propensity to over-rely on recommendations even if they are wrong. In comparisons between no-AI systems, simple-explanations, and cognitive forcing models, the cognitive forcing models were found to be the most effective at reducing over reliance although tradeoffs were noted where people favored solutions that reduced over-reliance less. Given the possible stakes at risk when using datasets for large-scale applications we found this tradeoff beneficial as eliciting deeper responses is specifically an instance of a cognitive forcing method in our application.

4 Tech: How can we generate useful questions?

Our technical application considered the design and results of *Social Simulacra: Creating Populated Prototypes for Social Computing Systems* [3] and *Value Sensitive Design and Information Systems*[4].

4.1 Large Language Models

Large language models (LLM) are a category of deep-learning models that can accomplish a wide range of text-based tasks including summarization, translation, and prediction. For this specific problem application, we focus on a subset of LLMs known as Casual Language Models (CLM) used primarily for text prediction. While these models are well suited for zero-and-few shot learning, they can be expensive to generate at a small scale. Given the benefits and constraints of such networks, we have chosen to use Facebook's Open Pre-trained Transformers (OPT) for the underlying language model. Specifically, we are using OPT125M which has 125 million trained parameters. While there are larger networks with more parameters that could improve results, given our hardware and performance constraints we focused on using OPT125M as our initial implementation.

4.2 String Templates

While CLMs are incredibly robust, depending on the generalization of the training data accomplishing a specific language prediction task can be challenging. To take advantage of the breadth of language in these models while still adapting them to narrower tasks we make use of string templates with our user input data. String templates are a method for formatting our data while also augmenting it with additional information to drive the model toward the response types we are interested in investigating. As CLMs are trained on massive amounts of text, the additional text added by these templates to the user data serves as additional context to the network when making predic-

[2] https://dl.acm.org/doi/abs/10.1145/3449287

[3] https://hci.stanford.edu/publications /2022/Park$_{S}ocialSimulacra_{U}IST22.pdf$

[4] https://cseweb.ucsd.edu/ goguen/courses/271 /friedman04.pdf

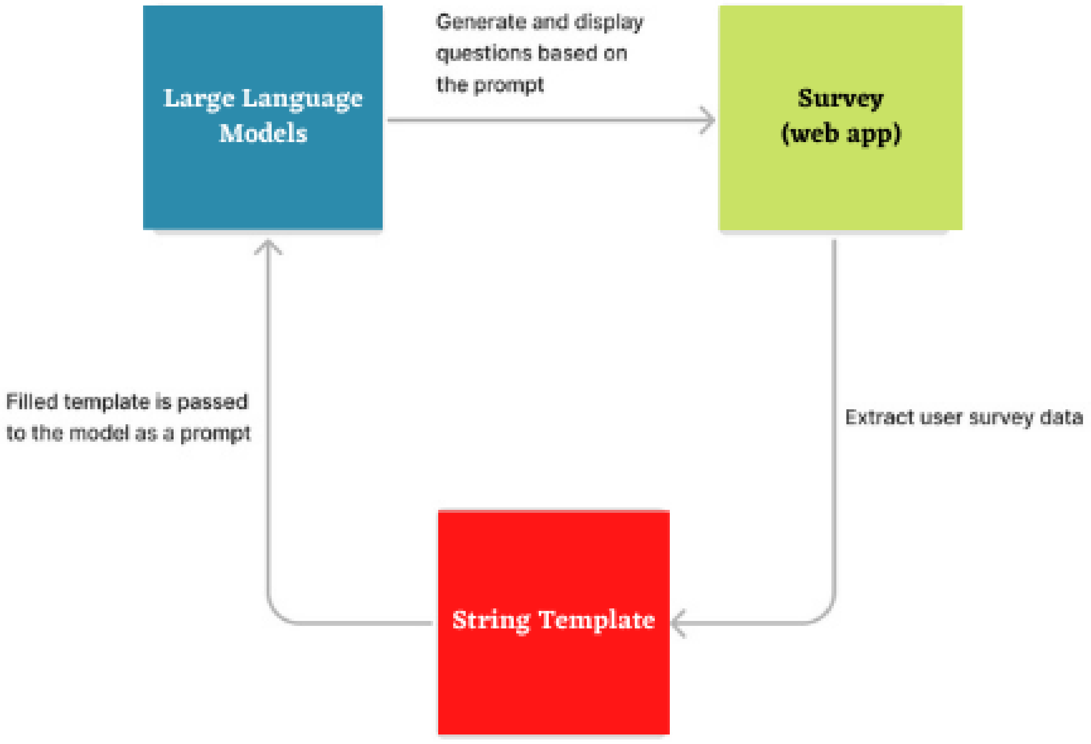
Large Language Models
Generate and display questions based on the prompt
Survey (web app)
Filled template is passed to the model as a prompt
Extract user survey data
String Template

Figure 1: CLM and Web App Model Composition

tions. For this application, we designed a template to promote question generation that takes into account and makes explicit the structure of the survey. Question-answer pairs of the survey are templated and then combined into one final template before prediction.

4.3 Web interface

To investigate the interaction between users and our prompts we built a Flask web app where users can fill out the survey and request questions prompts at the click of a button on a per-question basis. For the underlying model, we are using OPT125M as previously mentioned and are loading and making predictions on the model via Hugging Face's python transformers library. The web app renders the webpage for the survey as well as handling prompt requests which are sent through a Flask route, templated in Python, and then rerendered in an AJAX call using Javascript. To facilitate testing, the web app was then deployed on Microsoft Azure.

5 Evaluation

In order to test our CLM model's implementation we developed a Human-Computer Interaction(HCI) testing framework and help a couple of trials.

5.1 Experiment Design

Our HCI Expirement was section in two parts as seen in Figure 2.Subjects were given initial instructions about the set up of the test and then asked to complete Block 1 and Block 2. Block 1 consist of

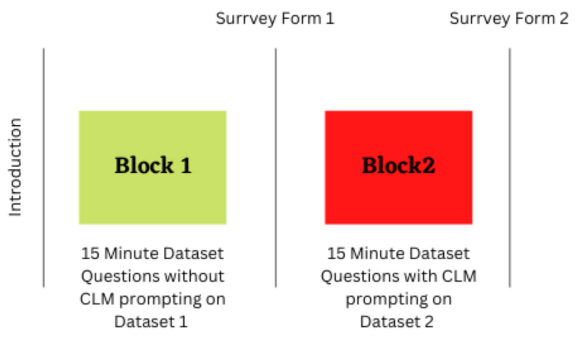
Survey Form 1
Survey Form 2
Introduction
Block 1
Block2
15 Minute Dataset Questions without CLM prompting on Dataset 1
15 Minute Dataset Questions with CLM prompting on Dataset 2

Figure 2: Expiremental Model Design

the subject interacting with a set of chosen DNP label creation questions on our testing webpage for 15 minutes and answer some followup questions about their experience. Block 2 had the exact same set up but the only difference was that the questions supported CLM suggestive prompting and had a button available to repeatedly generate novel questions that should stimulate the user to develop and generate high-quality text.

We decided to focus on the DNP label creation questions that would seem to generate the highest variability from our perspective. Here is the breakdown how what each questions mainly focused on:

- 9 questions from the Why Data Was Collected section
- 4 questions from the What Data Was Collected section
- 2 questions from the How Was Data Collected section
- 3 questions from the Upstream Datasets

5.2 HCI Results

We were able to complete 3 preliminary HCI experiments with people from different backgrounds, mainly data science and public work. Throughout the DNP label creation questions subjects asked for directions and perceived some of the questions as confusing. Overall more thoughtful responses were generated using the personalized question prompting and individuals said to have felt more confident in their responses although they said there was no major difference between the two DNP question forms. Even when all subjects have different interactions with the webpage they all agreed that answering the questions was challenging. From our sample, we did see an average response length increase of 24 characters per question across the sample, although at the individual level it varied greatly as one participant had an average increase of 60 characters with assistance compared to another individual whose response length actually decreased by 4.5 characters per response. From our experiments and follow-up assessment, one confounding factor may be familiarity with the subject matter. In both of these cases, the people had an inclination and experience toward one dataset more than the other.

6 Conclusion

Although our preliminary HCI Results consisted of a small pool of individuals it showed that there could be some benefits to having CLM generated prompting questions to help individuals complete the DNP label creation questions. There is still a lot of work that needs to be done in order to obtain a better understanding of an individuals interaction with the DNP label creation questions with CLM supported suggested prompting.

7 Further Work

There are still additional avenues to explore for producing better-directed prompts. Given the resource constraints, using larger models (OPT175B) has the potential to bear fruit. Additionally, finetune training specific networks on training text relative to the datasets and issues of bias and harm in datasets overall offers the possibility of modularizing according to types of data and domain area which would be worth evaluating given the inherently broad range of datasets that stand to have labels made for them. Given the successful cloud deployment of the interface, we believe this system could be practically integrated into the existing DNP label system, although resource considerations for the model itself (in particular running the model on a GPU) would need to be addressed.

8 Acknowledgements

We would like to thank the Data Nutrition Project for inspiring this project and sharing the workspace with us to further this research.