6-11-2024

Incorporating Citizen-Generated Data into Large Language Models

Jagadeesh Vadapalli

Srishti Gupta

Bishwa Karki

Chun-Hua Tsai

Incorporating Citizen-Generated Data into Large Language Models

JAGADEESH VADAPALLI, University of Nebraska at Omaha, USA
SRISHTI GUPTA, University of Nebraska at Omaha, USA
BISHWA KARKI, University of Nebraska Omaha, USA
CHUN-HUA TSAI, University of Nebraska at Omaha, USA

This study investigates the use of citizen-generated data to optimize a large language model (LLM) chatbot that gives nutrition advice. By actively participating in the data collection and annotation process from FDA-approved websites, citizens provided insightful information that was essential for improving the model and addressing biases. The study highlights the difficulties in gathering and annotating data, especially in situations where nuances matter, such as pregnancy nutrition. The results show that the use of citizen-generated data improves the efficacy and efficiency of data collection procedures, providing a practical viewpoint and encouraging community involvement. In addition to guaranteeing data quality, the iterative process raises stakeholders' awareness of and proficiency with data. Thus, citizen-generated data becomes an essential tool for creating information systems that are more reliable and inclusive.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Additional Key Words and Phrases: Citizen Science, Retrieval-Augmented Generation, Fine-tuning

ACM Reference Format:
Jagadeesh Vadapalli, Srishti Gupta, Bishwa Karki, and Chun-Hua Tsai. 2024. Incorporating Citizen-Generated Data into Large Language Models. In *25th Annual International Conference on Digital Government Research (DGO 2024), June 11–14, 2024, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages.
https://doi.org/10.1145/3657054.3659119

INTRODUCTION

Citizen-generated data refers to the phenomenon wherein citizens collect, generate, or analyze data, which is subsequently utilized in applications that serve them. This phenomenon is also known as citizen science or community-driven data [2, 5]. Citizen-generated data is crucial as it democratizes data-driven information systems intended for citizens themselves. With the rise of large language models (LLMs), the significance of data fidelity and provenance becomes a critical topic, such as algorithmic bias [1]. Citizen-generated data presents itself as a workable way to address short- comings, and there are various benefits to including citizens in the data generation process for LLMs, as it effectively utilizes community members' knowledge and experience.

This paper presents a case study incorporating citizen-generated data into fine-tuning an LLM for nutrition advising. We invited citizens (N=4) to participate in citizen data collection and annotation. The participants were asked to gather information from FDA-approved websites and generate datasets. Based on participant feedback, we identified the primary challenge lies in ensuring the quality and relevance of the generated content, especially in sensitive or high-stakes contexts such as nutrition information for pregnancy. Based on our findings, we seek to enhance the efficiency and effectiveness of citizen-driven data collection processes. This approach not only taps into more nuanced citizen knowledge and experience but also fosters citizen engagement and ownership in the data collection process, thereby enhancing the quality and relevance of the collected data for fine-tuning LLMs. We argue this to

be a new form of citizen participation and engagement in the era of generative AI.

METHODOLOGY

Fine-tuning a Large Language Model (LLM) is achieved through supervised learning. This approach employs a dataset containing labeled examples (i.e., the answer for a question in a specific context) to adjust the model's parameters, improving its performance on particular tasks. In this pilot study, we aim to fine-tune an LLM to better answer questions related to nutrition advice. We recruited four university students to participate as citizen annotators for data collection and annotation. We utilized Google Sheets to document and monitor the data generation activities of each participant. They selected various FDA-approved websites[1] focusing on food nutrition, safety, and values. Each student examined these websites to extract relevant nutrition information for pregnant women, which we labeled as "context" and recorded in a spreadsheet.

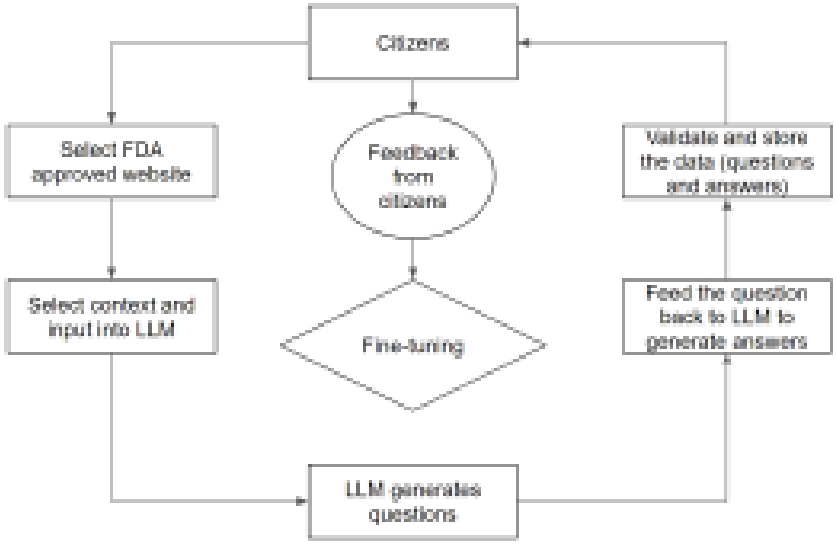

Fig. 1. The iterative citizen-driven data collection and annotation process for LLM fine-tuning.

[1] https://www.fda.gov/food/consumers/advice-about-eating-fish

The amount of collected data was limited due to the scarcity of contextual data and the high cost of inviting experts to label them. Alternatively, we asked the participants to input the collected contextual data into the Llama 2 model, an LLM released by Meta [2], and instructed them to generate 10 questions based on this context. We then fed these questions back into the LLM to generate responses (as labeled data). Specifically, we instructed the LLM to provide three types of responses: *abstract*, *conceptual*, and *short answers*. Abstract responses consisted of detailed explanations, conceptual responses aimed to be more straightforward, and short answers were restricted to a maximum of 10 words to provide precise answers. Each participant was asked to generate at least 100 rows of data (answers). These answers and the corresponding questions were compiled into a dataset for further analysis.

This iterative technique allowed us to generate answers based on questions posed by the LLM using the provided context. For example, if the participants collected context data as "*Fish provide key nutrients that support a child's brain development.*" The candidate questions could be: "*1. What specific nutrients found in fish are crucial for child brain development?*" We then feed these questions into LLM with the condition types, such as "*What specific nutrients found' in fish are crucial for child brain development? Please provide a 10-word short answer.*" The collected response will be "*Omega-3 fatty acids, particularly DHA and EPA, are crucial.*" Figure 1 depicts the iterative technique used in the data generation process.

FINDINGS

Our research on data collection yielded the following findings from citizens selected as participants in groups ranging from one to four. They provided comments on both the data collection process and the obtained data. Firstly, we identified the primary challenges revolving around acquiring related information following the given context from different websites. For instance, links to other websites within the context posed a challenge, requiring us

[2] https://llama.meta.com/

to navigate vast amounts of data and refine it to focus on topics relevant to food safety and pregnancy. The participants highlighted that context-specific data was limited, with only a small portion dedicated explicitly to pregnant individuals on these websites, and there is similar or duplicated information across websites.

Secondly, the participants agree that the LLM could condense the response to align precisely with the question asked. However, despite feeding the tool with selected context from the website, the LLM sometimes fails to generate relevant questions or labels. It was noted that questions requiring a short answer response often yielded similar responses for the same context. Moreover, there were instances where the LLM failed to provide an answer in the given context, such as eliciting a numerical response from the context. The LLM-generated responses offered similar answers when the question involved numerical values, regardless of whether the answer was abstract or factual. The length of responses varied, with the LLM frequently adding detail and enhancing the content in its own words.

CONCLUSION

This study explored the use of citizen-generated data to refine a large language model designed to offer nutrition-related guidance to pregnant women. We proposed the iterative citizen-driven data collection and annotation process for LLM fine-tuning. Our study findings affirm the feasibility of engaging citizens in the data collection and annotation process using LLM-powered data collection tools. We demonstrate the use case wherein citizens could participate in the process and input their time, wisdom, and knowledge to generate the dataset ready for LLM fine-tuning. We also identified the issues and challenges based on the participants' feedback. Our work contributes to the Digital Government Research (DGO) community by engaging citizens to generate generative AI data for future digital government functions and public engagement and deliberation [4]. Citizen-generated data helps address this issue and empowers citizens by fostering a more democratic, data-literate, and engaged community that directly influences the development of information systems tailored for them and defines them as a community [2, 3].

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2153509.

REFERENCES

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[2] John M Carroll, Jordan Beck, Shipi Dhanorkar, Jomara Binda, Srishti Gupta, and Haining Zhu. 2018. Strengthening community data: towards pervasive participation. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. 1–9.

[3] Stefan Jungcurt. 2022. *Citizen-generated data: Data by people, for people*. https://www.iisd.org/articles/insight/citizen-generated-data-people

[4] Mitchell Linegar, Rafal Kocielnik, and R Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science* 5 (2023), 1257092.

[5] Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in ecology & evolution* 24, 9 (2009), 467–471.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009