9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Community Transformers for Transforming Community Science

Keywords: Large Language Models, Community Data, Private Data Sharing, Data Trusts, Community Science

Extended Abstract

The promise of personalized large language models has captured our collective imagination and demonstrated the power of consumer chat-based interfaces as a tool for flexible information discovery and examination. With the addition of knowledge augmentation (such as WebGPT [1], Bing, or RAG [2]), these models become question-answer tools over specific information available on the public internet. Looking forward, these models could be powerful consumer tools when augmented with personal data (health, financial, and family information). However many important computational insights are only available at the community level. We propose a methodology for communities to collectively and securely pool data for the purpose of community-level large language model question-answering to extract key local computational insights.

This approach, which we call CommunityGPT, works to address how communities can identify and computationally investigate issues such as local health concerns, local crisis response preparedness, helping small businesses find unique opportunities [3], or community bias, which all require data from more than a single person and that are unavailable from public sources. 'Community' here can take multiple definitions, from geographic local communities such as towns or hospitals, to disparate communities with shared occupations or rare diseases. Each community and its members stand to benefit from pooling data to reveal shared insights unique to the community.

To achieve this, we leverage three technologies: self-custodial (and optionally decentralized) data storage such as tools built for data co-ops [4, 5], large language models with reinforcement learning from human feedback (similar to ChatGPT), and specific community chosen differentially privacy or k-anonymity controls in the access or fine tuning of knowledge augmentation. Communities can privately pool data in a selectively (and consentfully) anonymous or k-anonymous way allowing for knowledge augmentation [2] of a community model.

Privacy and security here are key design elements. When augmenting ML assistants with personal data many concerns are alleviated by no external parties seeing your model. Although information leakage to the community level may be a lesser threat than to the global internet, there are still many aspects of our lives we want to keep private from the community around us. The nature of unstructured querying of data limits the ability to keep information private (despite best attempts at differential privacy). To counter this, we utilize Trusted Execution Environments (TEEs or secure enclaves), such as Intel SGX [6] or AWS Nitro, to protect user inference queries to the community model. All user queries are encrypted by the user client-side with a public key corresponding to a decryption secret key that only ever exists inside of the TEE. Inference is then executed inside a secure enclave, which protects data-in-use. Additionally, our system supports privacy-preserving knowledge augmentation. The idea here is that each community can augment its large language model with a separate context-dependent

database that the model can query. The database is similarly encrypted with the same enclave key, which protects both the database and user queries.

To illustrate the relevancy of such a community-dependent database, imagine that a community privately ranks local physicians and stores this data in the community's shared encrypted database. By having direct access to such a database, a language model can significantly improve its ability to respond to relevant queries for members of that particular community. It is important to note that while data is encrypted, we do not hide access patterns between the enclave and the encrypted database, and we leave that to future work.

To demonstrate this prototype utility, we leverage two datasets at the community level. The first is a subset of a larger mobility trace dataset; which can be examined in natural language for common locations by time of day or two-hop venue correlations, an extension of tools like Google Maps, done so with additional privacy and data sovereignty for a community. The second is a new large crowd-sourced dataset of anonymized Amazon purchase history, which can be used to extract natural language queries regarding the purchase history of the community to get recommendations or surface larger community insights (e.g. local supply shortages, micro-inflation, local business opportunities) while maintaining individual privacy. These two examples are constrained in scope given the limited tooling around data-driven QA for large language models, and including differential privacy or k-anonymity requires careful examination of the data. In future, choices around privacy and the security implications of different data custodial models would be selected by the community itself.

We present an approach for communities to collaboratively extract computational social insights from pooled community data using large language models. This approach considers the role of privacy in the sharing of individual data with a community, but also the utility that can be extracted at this level. We use two example datasets to highlight this approach and demonstrate its value.

References

[1] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[3] Alex Pentland, Alexander Lipton, and Thomas Hardjono. *Building the New Economy: Data as Capital*. MIT Press, 2021.

[4] Thomas Hardjono, David L Shrier, and Alex Pentland. *Trusted Data, revised and expanded edition: A New Framework for Identity and Data Sharing*. MIT Press, 2019.

[5] Guy Zyskind, Oz Nathan, et al. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops*, pages 180–184. IEEE, 2015.

[6] Victor Costan and Srinivas Devadas. Intel SGX explained. *Cryptology ePrint Archive*, 2016.

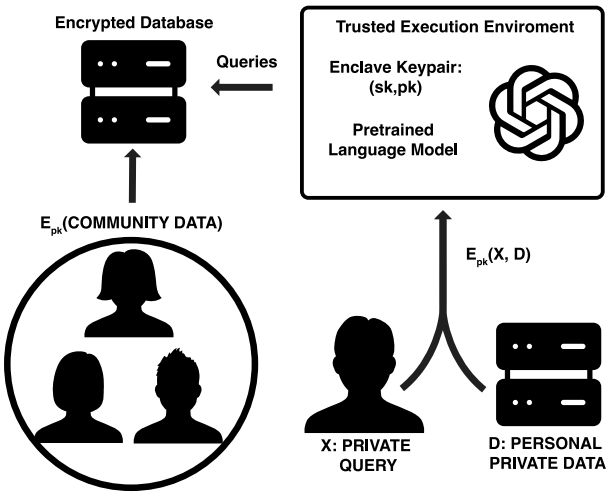

Figure 1: A diagram of how private member queries can be combined with personal data to allow for large language model examination of shared community data. Community data is encrypted at upload time into a secure database using the trusted execution environment's (TEE) public key (pk). This ensures that only the TEE can decrypt the community data, which itself will only be accessible for data lookup and prompt answering through the pretrained language model. This ensures raw data is secured, but still requires privacy controls to protect against information leakage through language models.