Secure Community Transformers: Private Pooled Data for LLMs

Tobin South
tsouth@mit.edu

Guy Zyskind
guyz@mit.edu

Robert Mahari
rmahari@mit.edu

Thomas Hardjono
hardjono@mit.edu

Alex 'Sandy' Pentland
pentland@mit.edu

Abstract

ChatGPT and other large language models (LLM) interfaces offer powerful and flexible capabilities in tasks including question-answering and chat-based support. These models can already access global context, through web searches and public data, or personal context, through augmentation with private data and local storage. However, securely addressing community-level challenges and collective action problems necessitates unique privacy, security, and custodianship solutions that safely bring together data from multiple parties and stakeholders. This paper presents a method that enables communities or organizations to securely aggregate data for LLM-based question-answering, aimed at extracting valuable local insights and customizing model responses to meet community-specific requirements while being inherently auditable. To this end, we leverage a combination of traditional privacy transformations, LLM-enabled privacy transformations, trusted execution environments, custodial control of data, and consent-based privacy choices to maximize community data security while maintaining the flexibility required for collective question-answering. This innovative method facilitates collaborative problem-solving using shared community data.

1 Introduction

The rapid proliferation of large language models (LLMs) such as GPT-3.5, GPT-4 [48], and LLaMA [63] as well as their consumer-facing interfaces like ChatGPT have captured our collective imagination. The flexibility and versatility of these technologies make them well-suited to perform a variety of tasks involving human interaction. Although these models are trained on a massive range of data sources across the web, this data is not up to date. The advent of knowledge retrieval augmentation (such as WebGPT [45] or RAG [36]),

promises to allow LLMs to become question-answer tools for specific information available on the public internet. These models could also become powerful consumer or enterprise tools if they are augmented with personal or organizational data (e.g., health, financial, and family information), such as demonstrated by the Microsoft 365 Copilot. However, any one person or organization has access to limited data and tremendous potential can be unlocked if communities can securely aggregate and search through their data.

The power of community data sharing is increasingly being recognized as a vital component in tackling complex societal challenges and enhancing overall welfare. This has led to the emergence of innovative data governance models, such as data cooperatives [25], which promote the idea that individuals can collectively control and benefit from their data. The central role of data in the new economy underscores the importance of users taking collective control of their data to reap the rewards of this valuable resource [52]. A human-centered approach to data architectures, as described in "Trusted Data" [26], seeks to address data privacy, security, ownership, and trust issues by enabling insights to be extracted without revealing the underlying data. This approach, combined with breaking down the data silos of big tech monopolies, has the potential to democratize data sharing for community well-being [40, 66].

There has been a great deal of research in the social sciences about what a human community is [6, 54], but we draw on a broad definition where a community is a group of people with shared relationships and the possibility or will to create shared institutional structures (which are needed to deploy our proposed system). These communities could comprise physically co-located individuals, physically distributed groups with a common enterprise (such as local businesses or separate units within a business), or groups of individuals with shared interests or challenges (such as a minority group with shared cultural heritage or a group with a common rare disease). In each of these contexts, there is value to the unique insights that can be harnessed through the aggregation and analysis of community data using LLM-based question-

answering. However, much of this data is sensitive and so any aggregation must also preserve individuals' and communities' privacy.

For instance, physical communities could use this method to share data on traffic, experiences with local businesses, or food inequality to help with collective coordination and planning. This information could be used to design better urban environments, allocate city resources more efficiently, or develop more effective community programs. Businesses could benefit from this approach to securely aggregate data across business units or partner companies to help design business strategies, learn from past client experiences, or co-work on compatible application programming interfaces. Minority communities could share experiences and personal insights to facilitate more inclusive chat-based mental health support systems or to extract insights that allow for better program development to identify and address the community's most pressing needs. The proposed method is flexible enough to meet the specific data-sharing needs of various communities while maintaining data security and privacy. In general, the goal is not only to enable community data to be aggregated for analysis but also to align the outputs of the LLM with the interests and experiences of the community.

As efforts to adapt LLMs to meet users' specific needs burgeon, we identify three primary state-of-the-art strategies that cater to community needs: (1) Prompt engineering system messages that are tailored to the community; (2) Fine-tuning using existing relevant data or reinforcement learning from human feedback (RLHF [49]) using members of a community to answer questions in a way that better suits the needs of the community; (3) Augmenting outputs with data to provide relevant experiential and contextual answers by drawing information from real community data and personal stories. Each of these approaches draws on community input and deliberation to optimally address the needs of the community. We omit a broader discussion of fine-tuning for specific tasks, as fine-tuning requires significant computational resources that many communities lack. While the use of community-powered RLHF is powerful, we leave this as future work as it would also require costly computational infrastructure and significant hands-on work by members of a community to provide human feedback for training. Further, while designing a robust and equitable system prompt that suits the needs of a community is essential, we take this as a given for the rest of the body of this work, reserving examination of deliberative system prompt selection for the discussion. Instead, we will focus on approach (3), which is uniquely suited to enable secure use of community data without requiring additional investment to augment model responses or significant investment into computational resources.

2 System architecture

The objective of the proposed system is to allows users to answer questions based on their own private data and sensitive community data. For example, an individual with a medical issue (personal private data) querying a set of detailed reviews (community data) to find the best local specialist. The proposed system will facilitate question-answering while safeguarding both the community data and the user's private data. We introduce a system design that ensures the secure transmission of user data to an encrypted database through the implementation of a trusted execution environment (TEE). The TEE not only guarantees secure data handling but also permits the execution of specific operations on the data. Our design encompasses two principal operations: privacy controls and secure information retrieval. Privacy controls, consisting of both statistical and language model-based privacy transformations, enable community members to safely modify data for the purpose of discovering insights while preserving privacy. This procedure can also facilitates the auditing of the data privacy trail. Moreover, the TEE provides question-answering capabilities for community members by employing language models that extract insights from both privacy-protected and public data using information retrieval methods. In this approach, an embedding-based textual search is conducted on the privatized community data, allowing LLM context window to be automatically filled with relevant community data to facilitate contextual insights.

This design is based on the core principles that data and compute should be hosted and controlled locally where possible, that community data should be self-governed and access granted only with consent, and that privacy (both from the outside world and within the community) should be controllable and guaranteed to the strongest possible level.

To allow individuals in the community to contribute data and then query aggregated data resources at a later date, a small upfront investment by the community is required to establish a server that deploys a TEE and an encrypted database, which will be used to protect community data. These execution environments will externalize a public encryption key used to encrypt the flow of user data. The community may decide, at any time, to dissolve the infrastructure and delete all shared data. The community may also decide to allow administrative access through which the TEE can delete or extract raw data. To ensure that this mechanism is used responsibly, and to avoid potential security breaches, the community may designate a committee or a board to vote on approving any use of the administrative access. This can be handled by a threshold signing protocol [33], which requires a majority of voters to approve such action.

To summarize, the system leverages six key privacy and security elements:

1. A TEE is used to execute the custodial operations of the data and queries.

2. Data remains encrypted once it has left the hands of an individual using the public encryption key provided by the TEE. Only the TEE, which cannot be directly accessed by any user of the server (not even with root access), ever sees the corresponding decryption key.

3. Private data has a de-identification step applied to it (using LLMs or traditional methods).

4. Non-private and de-identified data are made available to the LLM through secure information-retrieval in a key-controlled database.

5. User queries and prompts (with data) are encrypted and sent to the TEE.

6. (Optional) All data is kept local through the execution of the LLM on the TEE (subject to computational limitations, see subsection 2.3).

In this system, three sets of data are processed. The first is a user query to the system, X, which will contain a prompt X_{prompt} and optionally user-specific data, X_{data}. The user input may take a variety of forms and it may either interact with existing data or be placed directly into a text-based query. This user query may require data from the community to answer. Community data is broken into two types, private community data with sensitive information, $C^{private}$, and open community data, C^{open} which contains data that anyone in the community may access but which is not available to the general public. For example, C^{open} may be proprietary enterprise data or shared community records. We denote the community data as a pool, where the pooled data is made up of i individual records, $C_i^{private}$, each of which may require specific operations to ensure privacy. The items in the pooled data will be uploaded asynchronously over time, resulting in a constantly updating pool of community data, C.

The TEE is a secure area in a processor that guarantees the confidentiality and integrity of the code and data executed within it [42]. When combined with access control mechanisms, a TEE can provide a highly secure solution for protecting data uploaded and stored in a database. When a TEE is initialized, it generates a public-private key pair. When community members upload their data, they can securely encrypt it using the public key of the TEE, pk, before uploading it to the database via the TEE, and only the TEE, which holds the private key, can decrypt the data. If only the TEE has access to a set of tables within the database, then individual records need not be encrypted (as the whole database is) allowing for an efficient search of records to augment responses. More importantly, loading complete tables into the enclave prevents deanonymization attacks associated with observing memory access patterns. For example, encrypted data in which each item is encrypted separately would still leak how often a specific item is accessed, which under certain circumstances could have devastating results. Larger tables that do not fit

in the enclave's memory can leverage an Oblivious RAM (ORAM) scheme [29], but we leave this for future work.
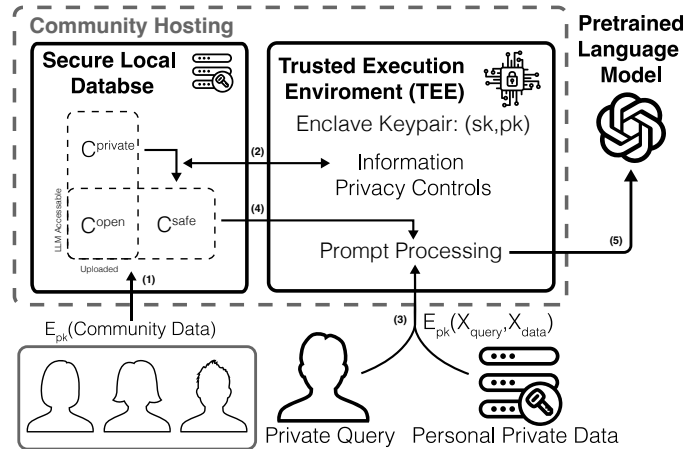

Figure 1: A diagram of how private queries can be combined with personal data to allow for large language model examination of shared community data that has been privacy-preserved. (1) Community data is encrypted at upload time into a secure database using the trusted execution environment's (TEE) public key (pk). (2) The securely uploaded private data is transformed into a privacy-persevered intermediate representation that can be inspected and audited. (3) Private user queries can be sent to the TEE where the prompt engine can combine queries with community data using information retrieval (4) to send to a pre-trained language model hosted externally via API or within local hosting for security enhancements (computational resources permitting).

Similarly to the community data upload process, a user can encrypt their prompt, X^{prompt}, and associated data, X^{data}, for upload to the TEE. While steps are taken throughout the process to prevent the leaking of community data, this becomes increasingly hard to ensure if outsiders are able to query the TEE API. Many solutions exist for this challenge, from simple user account management to special identity control or cryptographic wallet signing. These approaches can help to prevent attacks via information leakage through repeated queries (which has been shown to be a threat in many privacy contexts [47]) and distributed denial-of-service (DDoS) attacks on limited local commute resources.

2.1 LLM querying

Research on task-specific NLP has focused on fine tuning, but, with the rise of LLMs, the focus has shifted on the role of augmenting models using retrieval from external knowledge sources. Owing to their remarkable capabilities and pre-training on large corpora, LLMs are extremely powerful zero-shot and few-shot learners [4, 10, 48, 61]. However, the offline training of these models results in a lack of up-to-date informa-

tion. Advances in retrieval-augmented language models have allowed LLMs to draw on external knowledge [24, 28, 36, 69]. These augmented models can draw on up-to-date information from the web through search engines [7, 35, 45], code repositories [50, 68, 70], or private documents [2, 27, 30].

These retrieval models come in a variety of forms from document-search-based approaches [2, 31, 62] to more sophisticated approaches such as iterative prompts revising [51]. While many of these approaches rely on sophisticated techniques for augmentation, recent work has found success in simply prepending the retrieval results to the LLM query [55, 58]. This simple approach is easy to use and leverages traditional, efficient, and reproducible information retrieval systems such as Pyserini [38], LlamaIndex, and LangChain [5]. The simplicity of these approaches and their ability to efficiently run on a TEE makes them particularly well suited for the context of our proposal.

Building on the work in RALM [55], where a user creates a query, the TEE information reveal system performs a search that retrieves one or more documents or document summaries from the community corpus C, and conditions the LLM predictions on these documents by concatenating the retrieved documents within the LLMs input prior to the query prompt.

2.2 Information privacy controls

While much of the secure infrastructure proposed here is focused on keeping data safely within the community, care must be taken in how data is shared between community members. Where community members declare data open to the community, C^{open}, no additional work needs to be done, and the data can be indexed and searched directly by the information reveal system. For sensitive data that is uploaded, $C^{private}$, a privacy-preserving operation must be performed on the data to produce new data artifacts C^{safe}. In this discussion, we categorize data at the community level into two types: quantitative data and natural language data.

2.2.1 Privacy preservation in natural language

Over the last several years, a variety of approaches have been proposed to leverage deep learning to create privacy-preserving text representations and altered textual data to protect against identification or private attribute inferences [12, 17–19, 37, 44] with applications ranging from preserving privacy in emails [16] to medical notes [1, 15].

Prior work leverages a variety of approaches from differential privacy in training to an adversarial approach to privacy [13]. We anticipate that with the increased capabilities of new models such as GPT-4 [48] even more effective approaches will emerge to provide privacy while still outputting intermediate representations after the privacy-preserving step. In Figure 2 we demonstrate an example of using state-of-the-art LLM models to perform the privacy-preserving step.

Given the recent release of GPT-4, these approaches remain largely untested against state-of-the-art benchmarks. We perform a qualitative experiment to identify the current capacity of LLMs to perform privacy preservation, outlined in **??**. These experiments leverage an instructive system prompt to convert the sensitive text into a privacy-preserved text, as demonstrated in Figure 2. Protecting privacy, in this case, is particularly challenging if an actor has contextual knowledge about the community and its members. In such cases even deidentified data could be de-anonymized [46, 53]. Future work is needed to understand under what conditions this model can protect privacy.

While the privacy-preserving quality of this proposed approach is not yet fully studied, the approach has two important advantages. First, it uses the existing technology stack to prompt the LLM locally or via the API, reducing the software overhead. Second, it creates an intermediate text representation that *should* be de-identified. In cases where the community members believe data is sensitive enough to warrant checking, the TEE can be built to return the individual records to the user after de-identification. This would allow the user to confirm that their information has been scrubbed and choose to delete it if not thus providing auditable privacy. Self-custodial control over data privacy is a key design element that needs to be included in such a system.

2.2.2 Privacy transformations on quantitative data

While many valuable insights are only available through the analysis of natural language text, there is also a wide array of community questions that can be addressed purely through, or in combination with, quantitative data. Various related projects have used community data to help members make better decisions [11, 67] including specific interfaces for contexts such as migration choices [39] or urban health inequalities [14].

In many cases, these data-based support systems use publicly available data provided by governments or institutions. Governments around the world release a variety of statistical and raw data that is useful in decision-making. This data can include zoning data for urban planning, housing statistics for understanding the local real estate market, healthcare outcomes for various communities and demographics to optimize public health services, and educational performance data to guide policies and resource allocation in the education sector.

In cases where governments cannot release raw records due to privacy concerns, many have attempted to do so with privacy-enabling technologies, such as differential privacy in US census results [22], the Canadian Open Data initiative, which employs data anonymization techniques to protect individual identities while providing valuable datasets for public use, and others (e.g. [56]).

Quantitative data does not need to just come from governments. Individuals can extract valuable insights from privately

Figure 2: An example of using GPT-4 with a focused system prompt to convert identifiable text into deidentified text that still contains the essential information. While identifying names have been removed (including the name of the clinic which may have been useful information), threats involving an actor with knowledge of the relevant context remain.

sharing personal quantitative data [59]. For example, citizen science tasks often involve individuals sharing data on local environmental conditions, such as air quality, water quality, or biodiversity, to support scientific research and conservation efforts [20]. To enable privacy on this data, there exists a large literature on privacy-preserving and anonymization techniques [21, 41], where privacy budgets and trade-offs for specific contexts need to be considered.

Furthering the importance of these tools, the European Union's General Data Protection Regulation (GDPR) encourages the use of privacy-preserving techniques like k-anonymization [57] and pseudonymization when sharing personal data. These shared contributions help build a richer understanding of the community and its needs while respecting individual privacy.

2.3 Local LLM Inference

Despite the previously mentioned design elements ensuring secure data uploading, limited access, and privacy preservation, there remains a critical security aspect to address. At present, the largest and most powerful language models, such as GPT-4, are accessible only through an API. This means that uploading sensitive data necessitates sending community data to an external service, posing challenges in terms of privacy, security, and data compliance, particularly when sensitive records must be stored locally.

One potential solution is to host these LLM tools locally. Considerable efforts are being made to enable smaller versions of these models (such as LLaMA-7B [63] or it's fine-tuned derivatives, Alpaca [60], Vicuna [9], and Koala [23]) to run efficiently on local devices. Projects like `llama.cpp`[1] aim to port open models to operate on local consumer-level hardware. Additionally, approaches may draw from the principles of TinyML to enable models to run on edge computing [3]. Eventually, if inference efficiency improvements are substantial enough, these models could potentially operate entirely

within the TEE, or in a hybrid fashion with the TEE [64], significantly enhancing security. However, given the current state of the art, high-speed and low-memory methods for running LLMs do not achieve the same level of flexibility and robust performance as their large server-hosted counterparts.

Even in situations where full chat-based LLM answering cannot be performed locally, we may prioritize the privacy aspect of inference. This is crucial since enabling local privacy controls for all data ensures that no private data ever leaves community servers. To this end, smaller LLMs running locally might be sufficient for providing privacy controls. If these are unavailable, the vast array of traditional literature on NLP-based privacy controls, as outlined in subsubsection 2.2.1, is appropriate for running either directly in the TEE or through locally hosted compute services, such as external GPUs or community compute services.

While much of the system is designed for low-cost local deployment to maximize usability for small communities, larger communities (such as corporations or government agencies) might prioritize privacy and security over cost. In these instances, the threat model of the LLM may involve sending any data to an external company server, making the use of within-company cloud computing a sensible and secure alternative. In such cases, open-source LLMs are highly suitable for running on company servers and can even be further fine-tuned to align with company values, as demonstrated in RLHF [49].

3 Discussion

While the proposed system design offers protection against several threat models, it is unable to guard against others. The following discussion will elaborate on the system's strengths and weaknesses in terms of security and compliance, as well as explore the potential socio-technical challenges and the influence of system prompts on long-term behaviors.

One of the primary concerns with the system design is the possibility of bad data insertion, whereby malicious or incorrect data may be introduced into the system. This is-

---

[1]https://github.com/ggerganov/llama.cpp

sue is particularly concerning when community members intentionally or inadvertently introduce false or destructive content into the model. Such content can significantly reduce the system's performance, compromise its accuracy, and potentially harm other members of the community. As a result, the effectiveness and reliability of the system can be severely undermined. To mitigate this risk, it is essential to implement robust data validation and moderation mechanisms to detect and filter out malicious or erroneous inputs. This may involve incorporating automated content analysis, user reputation systems, and manual moderation processes to ensure that the data fed into the system is accurate, reliable, and beneficial to the community as a whole. Additionally, fostering a culture of trust, responsibility, and accountability within the community can help minimize the likelihood of bad data insertion and encourage members to contribute positively to the system's success.

In terms of compliance, the system design adheres to the GDPR, CCPA, and the proposed EU AI Act [34]. However, the evolving nature of these regulatory frameworks leaves a significant number of legal questions unanswered. It does, however, maintain the right to be forgotten, as community members give consent for data to be used, and can revoke that consent by removing the data (and any of its privacy-preserving transformations) from the community data pool.

Socio-technical challenges may arise from this system design, particularly in terms of reinforcing community group-think and polarization [32]. By exclusively using data from a specific community, the system may inadvertently amplify existing biases, leading to further polarization within and between communities. The risk of driving echo chambers further apart and intensifying divisions between communities should not be overlooked. To address these challenges, it is crucial to develop strategies that encourage diverse perspectives, promote constructive dialogue, and counteract the potential for negative consequences resulting from the system's design. Indeed, the EU AI Act is deeply concerned about the role of subliminal persuasion from models, and it is still unclear where liability or risks emerge when acting on the outputs from these models.

An essential aspect to consider regarding the long-term impact of these systems on communities is the choice of system prompts. The prompts can wield a considerable influence on the model's behavior, dictating what it will and will not do. Ensuring that system prompts encourage positive behavior and productive interactions is crucial for the successful implementation and adoption of these tools within communities. This may involve designing prompts that foster empathy, understanding, and open-mindedness, as well as incorporating feedback from diverse stakeholders to ensure a broad range of perspectives is represented. Moreover, it is important to regularly evaluate and adjust the prompts to optimize the system's performance and adapt to the evolving needs and dynamics of the communities it serves.

Finally, it's important to turn our attention to the use of TEEs as a security control mechanism within the system design. TEEs provide additional layers of security by safeguarding the executed code, runtime state, and memory during operation. In situations where lower security controls are deemed sufficient, TEEs can be entirely removed and replaced with an unsecured runtime on a cloud provider or local server, offering advantages in speed, auditability, and flexibility. Conversely, if a community is concerned about centralized control of data management and computation, they could opt for a system that distributes the data across a network of TEEs (e.g., as in the case with Secret Network[2], a privacy-preserving Blockchain). Similarly, as TEEs are known to be susceptible to potential side-channel attacks [8, 65], a community could leverage secure multiparty computation (MPC) to manage the system [43, 71]. MPC distributes the control and processing of data across multiple parties, ensuring no single entity has complete control over the information. However, this approach comes with trade-offs, as it may significantly reduce system speed and usability. It also requires assuming that the different parties do not collude — an assumption that may be difficult to make in practice. Ultimately, the choice of security mechanisms should be tailored to the specific needs and concerns of each community, balancing the desired level of privacy and protection with the trade-offs in performance and ease of use. By carefully considering these factors, communities can implement a system that offers robust security while maintaining functionality and accessibility for all its members.

4 Conclusion

In conclusion, this paper has presented a novel method for securely aggregating data from multiple parties and stakeholders within a community, enabling the utilization of large language models (LLMs) like ChatGPT for question-answering tasks that address community-level challenges and collective action problems. By incorporating a range of privacy and security measures, such as traditional privacy transformations, LLM-enabled privacy transformations, trusted execution environments, custodial control of data, and consent-based privacy choices, the proposed system maximizes community data security while offering a flexible tool for community-specific insights and tailored model responses.

The implications of this research extend beyond the immediate application of LLMs in question-answering tasks, as it demonstrates the potential for harnessing the power of these models in a secure and privacy-preserving manner to address a wide array of community-oriented problems. Furthermore, the methods presented in this paper can serve as a foundation for future research, exploring novel ways to enhance data security, privacy, and usability in LLM applications while fos-

---

[2]http://docs.scrt.network/

tering collaborative problem-solving using shared community data.

As LLMs continue to advance and gain prominence, it is essential to develop robust, secure, and privacy-preserving mechanisms that cater to the unique needs of various communities. By bridging the gap between individual and global contexts, this research contributes to the ongoing efforts to make LLMs an invaluable tool for addressing the complex, diverse, and dynamic challenges faced by communities worldwide.

References

[1] ALAWAD, M. M., YOON, H.-J., GAO, S., MUMPHREY, B. J., WU, X.-C., DURBIN, E. B., JEONG, J. C., HANDS, I., RUST, D., COYLE, L., PENBERTHY, L., AND TOURASSI, G. D. Privacy-preserving deep learning nlp models for cancer registries. *IEEE Transactions on Emerging Topics in Computing 9* (2020), 1219–1230.

[2] ARORA, S., LEWIS, P., FAN, A., KAHN, J., AND R'E, C. Reasoning over public and private data in retrieval-based systems. *ArXiv abs/2203.11027* (2022).

[3] BANBURY, C. R., REDDI, V. J., LAM, M., FU, W., FAZEL, A., HOLLEMAN, J., HUANG, X., HURTADO, R., KANTER, D., LOKHMOTOV, A., ET AL. Benchmarking tinyml systems: Challenges and direction. *arXiv preprint arXiv:2003.04821* (2020).

[4] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[5] CHASE, H. LangChain, Oct. 2022.

[6] CHAVIS, D. M., AND WANDERSMAN, A. Sense of community in the urban environment: A catalyst for participation and community development. *American journal of community psychology 18*, 1 (1990), 55–81.

[7] CHEN, D., FISCH, A., WESTON, J., AND BORDES, A. Reading wikipedia to answer open-domain questions. In *Annual Meeting of the Association for Computational Linguistics* (2017).

[8] CHEN, G., CHEN, S., XIAO, Y., ZHANG, Y., LIN, Z., AND LAI, T. H. Sgxpectre: Stealing intel secrets from sgx enclaves via speculative execution. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)* (2019), IEEE, pp. 142–157.

[9] CHIANG, W.-L., LI, Z., LIN, Z., SHENG, Y., WU, Z., ZHANG, H., ZHENG, L., ZHUANG, S., ZHUANG, Y., GONZALEZ, J. E., STOICA, I., AND XING, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[10] CHOWDHERY, A., NARANG, S., DEVLIN, J., BOSMA, M., MISHRA, G., ROBERTS, A., BARHAM, P., CHUNG, H. W., SUTTON, C., GEHRMANN, S., SCHUH, P., SHI, K., TSVYASHCHENKO, S., MAYNEZ, J., RAO, A., BARNES, P., TAY, Y., SHAZEER, N. M., PRABHAKARAN, V., REIF, E., DU, N., HUTCHINSON, B. C., POPE, R., BRADBURY, J., AUSTIN, J., ISARD, M., GUR-ARI, G., YIN, P., DUKE, T., LEVSKAYA, A., GHEMAWAT, S., DEV, S., MICHALEWSKI, H., GARCÍA, X., MISRA, V., ROBINSON, K., FEDUS, L., ZHOU, D., IPPOLITO, D., LUAN, D., LIM, H., ZOPH, B., SPIRIDONOV, A., SEPASSI, R., DOHAN, D., AGRAWAL, S., OMERNICK, M., DAI, A. M., PILLAI, T. S., PELLAT, M., LEWKOWYCZ, A., MOREIRA, E., CHILD, R., POLOZOV, O., LEE, K., ZHOU, Z., WANG, X., SAETA, B., DÍAZ, M., FIRAT, O., CATASTA, M., WEI, J., MEIER-HELLSTERN, K. S., ECK, D., DEAN, J., PETROV, S., AND FIEDEL, N. Palm: Scaling language modeling with pathways. *ArXiv abs/2204.02311* (2022).

[11] CHOWDHURY, T., AND SHARMA, N. K. Citizenly: A platform to encourage data-driven decision making for the community by the community. *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)* (2021), 359–364.

[12] COAVOUX, M., NARAYAN, S., AND COHEN, S. B. Privacy-preserving neural representations of text. In *Conference on Empirical Methods in Natural Language Processing* (2018).

[13] DA SILVA SOUSA, S. B., AND KERN, R. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review 56* (2022), 1427 – 1492.

[14] DE ALBUQUERQUE, J. P., YEBOAH, G., PITIDIS, V., AND ULBRICH, P. Towards a participatory methodology for community data generation to analyse urban health inequalities: A multi-country case study. In *Hawaii International Conference on System Sciences* (2019).

[15] DERNONCOURT, F., LEE, J. Y., UZUNER, Ö., AND SZOLOVITS, P. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association 24* (2016), 596–606.

[16] EDER, E., KRIEG-HOLZ, U., AND HAHN, U. De-identification of emails: Pseudonymizing privacy-sensitive data in a german email corpus. In *Recent Advances in Natural Language Processing* (2019).

[17] FENG, Q., BIAO HE, D., LIU, Z., WANG, H., AND CHOO, K. R. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security 15* (2020), 3709–3721.

[18] FERNANDES, N., DRAS, M., AND MCIVER, A. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8* (2019), Springer International Publishing, pp. 123–148.

[19] FEYISETAN, O., DIETHE, T., AND DRAKE, T. Leveraging hierarchical representations for preserving privacy and utility in text. *2019 IEEE International Conference on Data Mining (ICDM)* (2019), 210–219.

[20] FRITZ, S., SEE, L. M., CARLSON, T., HAKLAY, M. M., OLIVER, J. L., FRAISL, D., MONDARDINI, R. M., BROCKLEHURST, M., SHANLEY, L. A., SCHADE, S., WEHN, U., ABRATE, T., ANSTEE, J. M., ARNOLD, S., BILLOT, M., CAMPBELL, J., ESPEY, J., GOLD, M., HAGER, G., HE, S., HEPBURN, L., HSU, A., LONG, D., MASÓ, J., MCCALLUM, I., MUNIAFU, M. M., MOORTHY, I., OBERSTEINER, M., PARKER, A., WEISSPFLUG, M., AND WEST, S. Citizen science and the united nations sustainable development goals. *Nature Sustainability 2* (2019), 922 – 930.

[21] FUNG, B. C. M., WANG, K., CHEN, R., AND YU, P. S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv. 42* (2010), 14:1–14:53.

[22] GARFINKEL, S. Differential Privacy and the 2020 US Census. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, Winter 2022 (jan 24 2022). https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census.

[23] GENG, X., GUDIBANDE, A., LIU, H., WALLACE, E., ABBEEL, P., LEVINE, S., AND SONG, D. Koala: A dialogue model for academic research. Blog post, April 2023.

[24] GUU, K., LEE, K., TUNG, Z., PASUPAT, P., AND CHANG, M.-W. Retrieval augmented language model pre-training. In *International Conference on Machine Learning* (2020).

[25] HARDJONO, T., AND PENTLAND, A. S. Data cooperatives: Towards a foundation for decentralized personal data management. *ArXiv abs/1905.08819* (2019).

[26] HARDJONO, T., SHRIER, D. L., AND PENTLAND, A. *Trusted Data: A New Framework for Identity and Data Sharing*. MIT Press, 2019.

[27] IZACARD, G., AND GRAVE, E. Leveraging passage retrieval with generative models for open domain question answering. In *Conference of the European Chapter of the Association for Computational Linguistics* (2020).

[28] IZACARD, G., LEWIS, P., LOMELI, M., HOSSEINI, L., PETRONI, F., SCHICK, T., YU, J. A., JOULIN, A., RIEDEL, S., AND GRAVE, E. Few-shot learning with retrieval augmented language models. *ArXiv abs/2208.03299* (2022).

[29] JEAN-LOUIS, N., LI, Y., JI, Y., MALVAI, H., YUREK, T., BELLEMARE, S., AND MILLER, A. Sgxonerated: Finding (and partially fixing) privacy flaws in tee-based smart contract platforms without breaking the tee. *Cryptology ePrint Archive* (2023).

[30] KARPUKHIN, V., OGUZ, B., MIN, S., LEWIS, P., WU, L., EDUNOV, S., CHEN, D., AND YIH, W.-T. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), pp. 6769–6781.

[31] KHATTAB, O., SANTHANAM, K., LI, X. L., HALL, D. L. W., LIANG, P., POTTS, C., AND ZAHARIA, M. A. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv abs/2212.14024* (2022).

[32] KIRK, H. R., VIDGEN, B., RÖTTGER, P., AND HALE, S. A. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *ArXiv abs/2303.05453* (2023).

[33] KOMLO, C., AND GOLDBERG, I. Frost: flexible round-optimized schnorr threshold signatures. In *Selected Areas in Cryptography: 27th International Conference, Halifax, NS, Canada (Virtual Event), October 21-23, 2020, Revised Selected Papers 27* (2021), Springer, pp. 34–65.

[34] KOP, M. Eu artificial intelligence act: the european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust . . . .

[35] LAZARIDOU, A., GRIBOVSKAYA, E., STOKOWIEC, W., AND GRIGOREV, N. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv abs/2203.05115* (2022).

[36] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W.-T., ROCKTÄSCHEL, T., ET AL. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems 33* (2020), 9459–9474.

[37] LI, Y., BALDWIN, T., AND COHN, T. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2018), pp. 25–30.

[38] LIN, J., MA, X., LIN, S.-C., YANG, J.-H., PRADEEP, R., AND NOGUEIRA, R. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)* (2021), pp. 2356–2362.

[39] LOAIZA, I., SOUTH, T., SANCHEZ, G., CHAN, S., YU, A., MONTES, F., BAHRAMI, M., AND PENTLAND, A. Voyage viewer: Empowering human mobility at a global scale. *EuroVis Workshop on Visual Analytics* (2022).

[40] MAHARI, R., LERA, S. C., AND PENTLAND, A. S. Time for a new antitrust era: Refocusing antitrust law to invigorate competition in the 21st century. *Stanford Computational Antitrust 1* (2021).

[41] MAJEED, A., AND LEE, S. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access 9* (2021), 8512–8545.

[42] MCKEEN, F., ALEXANDROVICH, I., BERENZON, A., ROZAS, C., SHAFI, H., SHANBHOGUE, V., AND SAVAGAONKAR, U. Innovative Instructions and Software Model for Isolated Execution. In *Proc. Second Workshop on Hardware and Architectural Support for Security and Privacy HASP2013* (Tel-Aviv, June 2013). https://sites.google.com/site/haspworkshop2013/workshop-program.

[43] MOHASSEL, P., AND ZHANG, Y. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)* (2017), IEEE, pp. 19–38.

[44] MOSALLANEZHAD, A., BEIGI, G., AND LIU, H. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Conference on Empirical Methods in Natural Language Processing* (2019).

[45] NAKANO, R., HILTON, J., BALAJI, S. A., WU, J., OUYANG, L., KIM, C., HESSE, C., JAIN, S., KOSARAJU, V., SAUNDERS, W., JIANG, X., COBBE, K., ELOUNDOU, T., KRUEGER, G., BUTTON, K., KNIGHT, M., CHESS, B., AND SCHULMAN, J. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv abs/2112.09332* (2021).

[46] NARAYANAN, A., AND SHMATIKOV, V. Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)* (2008), 111–125.

[47] NASR, M., SHOKRI, R., AND HOUMANSADR, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *2019 IEEE Symposium on Security and Privacy (SP)* (2018), 739–753.

[48] OPENAI. Gpt-4 technical report. *ArXiv abs/2303.08774* (2023).

[49] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35* (2022), 27730–27744.

[50] PARVEZ, M. R., AHMAD, W., CHAKRABORTY, S., RAY, B., AND CHANG, K.-W. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021), pp. 2719–2734.

[51] PENG, B., GALLEY, M., HE, P., CHENG, H., XIE, Y., HU, Y., HUANG, Q., LIDÉN, L., YU, Z., CHEN, W., AND GAO, J. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv abs/2302.12813* (2023).

[52] PENTLAND, A., LIPTON, A., AND HARDJONO, T. *Building the New Economy: Data as Capital*. MIT Press, 2021.

[53] PORTER, C. C. De-identified data and third party data mining: The risk of re-identification of personal information. *Shidler JL Com. & Tech. 5* (2008), 1.

[54] PUTNAM, R. D. *Bowling alone: The collapse and revival of American community*. Simon and schuster, 2000.

[55] RAM, O., LEVINE, Y., DALMEDIGOS, I., MUHLGAY, D., SHASHUA, A., LEYTON-BROWN, K., AND SHOHAM, Y. In-context retrieval-augmented language models. *ArXiv abs/2302.00083* (2023).

[56] RUCIAK, J., AND HARDJONO, T. A Digital Data Exchange for Australia, October 2021. https://doi.org/10.36227/techrxiv.16821118.v1.

[57] SAMARATI, P., AND SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[58] SHI, W., MIN, S., YASUNAGA, M., SEO, M., JAMES, R., LEWIS, M., ZETTLEMOYER, L., AND TAU YIH, W. Replug: Retrieval-augmented black-box language models. *ArXiv abs/2301.12652* (2023).

[59] SOUTH, T., LOTHIAN, N., AND ALEXSANDYPENTLAND. Building a healthier feed: Private location trace intersection driven feed recommendations. *ArXiv abs/2210.01927* (2022).

[60] TAORI, R., GULRAJANI, I., ZHANG, T., DUBOIS, Y., LI, X., GUESTRIN, C., LIANG, P., AND HASHIMOTO, T. B. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[61] THOPPILAN, R., FREITAS, D. D., HALL, J., SHAZEER, N. M., KULSHRESHTHA, A., CHENG, H.-T., JIN, A., BOS, T., BAKER, L., DU, Y., LI, Y., LEE, H., ZHENG, H., GHAFOURI, A., MENEGALI, M., HUANG, Y., KRIKUN, M., LEPIKHIN, D., QIN, J., CHEN, D., XU, Y., CHEN, Z., ROBERTS, A., BOSMA, M., ZHOU, Y., CHANG, C.-C., KRIVOKON, I. A., RUSCH, W. J., PICKETT, M., MEIER-HELLSTERN, K. S., MORRIS, M. R., DOSHI, T., SANTOS, R. D., DUKE, T., SØRAKER, J. H., ZEVENBERGEN, B., PRABHAKARAN, V., DÍAZ, M., HUTCHINSON, B., OLSON, K., MOLINA, A., HOFFMAN-JOHN, E., LEE, J., AROYO, L., RAJAKUMAR, R., BUTRYNA, A., LAMM, M., KUZMINA, V. O., FENTON, J., COHEN, A., BERNSTEIN, R., KURZWEIL, R., AGUERA-ARCAS, B., CUI, C., CROAK, M., HSIN CHI, E. H., AND LE, Q. LaMDA: Language models for dialog applications. *ArXiv abs/2201.08239* (2022).

[62] THULKE, D., DAHEIM, N., DUGAST, C., AND NEY, H. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *ArXiv abs/2102.04643* (2021).

[63] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., RODRIGUEZ, A., JOULIN, A., GRAVE, E., AND LAMPLE, G. Llama: Open and efficient foundation language models. *ArXiv abs/2302.13971* (2023).

[64] TRAMER, F., AND BONEH, D. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. In *International Conference on Learning Representations* (2019).

[65] VAN BULCK, J., MINKIN, M., WEISSE, O., GENKIN, D., KASIKCI, B., PIESSENS, F., SILBERSTEIN, M., WENISCH, T. F., YAROM, Y., AND STRACKX, R. Foreshadow: Extracting the keys to the intel sgx kingdom with transient out-of-order execution. In *Proceedings fo the 27th USENIX Security Symposium* (2018), USENIX Association.

[66] WU, D., VERHULST, S. G., PENTLAND, A., ÁVILA, T. J. T., FINCH, K., AND GUPTA, A. How data governance technologies can democratize data sharing for community well-being. *Data & Policy 3* (2021).

[67] YOON, A., AND COPELAND, A. J. Toward community-inclusive data ecosystems: Challenges and opportunities of open data for community-based organizations. *Journal of the Association for Information Science and Technology 71* (2020), 1439 – 1454.

[68] ZAN, D., CHEN, B., LIN, Z., GUAN, B., WANG, Y., AND LOU, J.-G. When language model meets private library. In *Conference on Empirical Methods in Natural Language Processing* (2022).

[69] ZHANG, Y., SUN, S., GAO, X., FANG, Y., BROCKETT, C., GALLEY, M., GAO, J., AND DOLAN, B. Retgen: A joint framework for retrieval and grounded text generation modeling. In *AAAI Conference on Artificial Intelligence* (2021).

[70] ZHOU, S., ALON, U., XU, F. F., JIANG, Z., AND NEUBIG, G. Doccoder: Generating code by retrieving and reading docs. *ArXiv abs/2207.05987* (2022).

[71] ZYSKIND, G., NATHAN, O., ET AL. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops* (2015), IEEE, pp. 180–184.