NETWORK LAW REVIEW

ARTICLE

Transparency by Design for Large Language Models

Sandy Pentland, Robert Mahari and Tobin South

***

Large Language Models – like ChatGPT, Bard, and Claude – are akin to modern day oracles: they provide impressively useful outputs without revealing their reasoning. It remains extremely difficult to understand why exactly a large language model has created a specific output, and this issue of explainability continues to attract widespread attention from academics, regulators, and practitioners. A humbler desire is to know what data was used to generate model outputs and to have the ability to modify this input data. This type of transparency matters both from an individual privacy and a business perspective. Individuals have an interest – and, under some privacy regulations, a right – to modify or delete data that is stored about them. Meanwhile, organizations that leverage LLMs need to ensure that outputs are based on up-to-date information.

Building on recent work at MIT, we outline a new proposal for a type of LLM-powered data trust, the Community Transformer. We explore how this technical architecture provides ways to track what data is used by an LLM and to modify this underlying data in ways that increase privacy and transparency.

Updatability and auditability for privacy

We distinguish between two types of input data for LLMs. The first is pre-training data which is used to create a general-purpose model (putting the 'P' in Generative Pre-trained Transformer, GPT). Pre-training datasets are typically extremely large and composed of data scraped from the web and sophisticated additional datasets such as Reinforcement Learning from Human Feedback (RLHF) and alignment data. The second type is task-specific data, which is used to tailor a general-purpose LLM to a specific task. A large quantity of task-specific data (although orders of magnitude less than for pre-training) can be used to fine-tune a model – that is, to update the model weights for a specific application. RLHF and related methods such as value alignment differ from traditional fine-tuning but

broadly fit into this category with their data-hungry processes and permanent model weight updates. Alternatively, typically with lower upfront costs, a small amount of relevant task-specific data may be identified and included as additional information each time a model is used. This second approach, typically referred to as information retrieval, can be as simple as including relevant text or data in a model prompt.

While pre-training data may have privacy or business implications, it merely provides the model with a general ability to perform language tasks and so errors in the pre-training data are generally benign (unless these errors are systematic and give rise to biases in the model). By contrast, errors or omissions in task-specific data have more significant consequences, both because they more directly impact outputs and because there is generally far less of this data available. How exactly pre-training and task-specific data are used also has significant implications for privacy and transparency.

We use the term auditability to refer to the ability to identify what records were used by a machine learning system to generate a specific output. Auditability is a necessary but insufficient condition for explainability, which refers to the ability to understand the computational pathway through which a model produces a specific output or set of outputs. The GDPR creates an obligation for data processors to share the "logic" involved in reaching an automated decision. Legal scholars have debated whether this provision amounts to a right to explainability, however, regardless of whether data subjects are entitled to an explanation it is clear that the drafters of the GDPR intended for automated decisions to be made transparent. As such, auditability – giving users the ability to know *which* inputs led to a certain output – is an important step towards the spirit of privacy regulation and can form the basis on which individuals may choose to exercise their right to rectify incorrect information. From an organizational perspective, auditability is also a key requirement for LLMs. Organizations leveraging LLMs may wish to understand what inputs gave rise to certain outputs for quality control or compliance purposes. Data audits can also help organizations identify records that give rise to erroneous or outdated outputs to be removed or updated.

By contrast to auditability, we use the term updatability to refer to the ability to modify or delete data records that are part of a machine learning system. The European Union's General Data Protection Regulation (GDPR) and California's Consumer Privacy Act (CCPA) both grant individuals rights amounting to a right to updatability. Namely, the GDPR grants individuals a right to rectification (a right to correct inaccurate information) and the right to be forgotten (a right to erase personal data), while the CCPA grants analogous rights to correction and deletion. While privacy regulations guarantee updatability for individuals, organizations that rely on machine learning enabled decision support systems also require the ability to

modify the data that these systems rely upon. For example, an organization that uses LLMs to generate government filings might need to update the template every year. Modifying pre-training data, or data that has been used to fine-tune a model, is much more challenging and costly than simply deleting a record. The simplest approach is to modify the data and then retrain the entire model, however, doing so for the giant LLMs that have become the industry standard is costly. Machine unlearning can reduce the cost, but it still demands substantial computational resources and predominantly focuses on data deletion rather than correction.

Auditability and updatability go hand-in-hand as the former can be used to can reveal errors or omissions to be updated. However, auditability can also be valuable by itself because it increases the transparency of LLMs. For example, auditability can provide insight into how an automated decision was reached and thus provide the basis for an appeal.

LLM Background

The most basic objective of an LLM is to predict the most likely word to follow a given text input. It is through this simple task that the emergent flexible capabilities of LLMs arise. During the initial training phase of LLMs, the models ingest their training data, converting it into model weights through an iterative learning process. No record is kept of which training data contributes to what model weight updates. Although the LLM may retain fragments of the training data, this phenomenon is merely an unintended consequence of the learning algorithm attempting to complete the task of next-word prediction.

When utilizing LLMs for downstream applications it becomes necessary to incorporate task-specific data. As outlined above, there are two primary methods for incorporating this additional information: fine-tuning a pre-trained general-purpose model or providing the model with task-specific data as part of a prompt.

Fine-tuning involves continuing the learning process of the LLM pre-training and adjusting or editing the model weights to optimize for performance on a specific task using the task-specific data. This approach was considered the standard method until relatively recently. However, the latest generation of LLMs offer a significant advantage due to their ability to process extensive amounts of data within a prompt without requiring fine-tuning. As a result, LLMs can to adapt to various tasks more effectively, enhancing its overall utility in diverse contexts.

This ability also makes it possible to design LLMs to precisely identify which records contributed to the generation of a particular output and to update the relevant records without the risk of erroneously retaining obsolete information. These developments present a valuable opportunity to create mechanisms for

updatability and auditability that comply with privacy regulations and better cater to real-world needs.

A new frontier of prompts & context

The new capabilities of the latest generation of LLMs have enabled new approaches to using these models. The rise of prompt engineering via interfaces such as ChatGPT has demonstrated the value that can be extracted from flexible pre-trained models by providing clear instructions and reference pieces of text. Including these additional pieces of textual information in the so-called ‘context window’ that the LLM can parse allows it to make changes to the text or answer questions about it.

Extensions of this, often referred to as information retrieval or knowledge augmentation, draw on specific sets of data to be passed into the LLM during inference time, often based on search queries generated automatically from a user prompt. LLMs are used to identify search queries that would assist in answering the user prompt, which are subsequently used to search the web or specific databases (such as organizational databases) and the retrieved information can be included in the context window for answering the original prompt. This is all done without needing to fine-tune the models on the specific datasets. In a general sense, this is the methodology that is used to power the Bing and WebGPT (beta) experiences.

Community Transformers

Recent work on Secure Community Transformers has focused on how groups of people – whether geographically colocated, communities with shared goals, or users within a company – can securely manage and upload data in privacy-preserving ways for use within LLMs. This novel method enables groups to gain valuable insights and make informed decisions based on their unique needs while ensuring high standards of data privacy and security. By employing a combination of cutting-edge privacy-enhancing technologies, such as trusted execution environments and encryption, the system provides robust data protection while retaining the necessary flexibility for community members.

As shown in the figure, the proposed architecture allowed users to submit private queries and combine them with both private personal data and private community data. To this end, the community uploads its shared data into a secure database. The community data is transformed into a privacy-persevered intermediate representation that can be inspected and audited. The community data may contain private records that should not be accessed by all community members and so the privacy-preservation step facilitates data pooling by reducing the risk that sensitive information is leaked. Each user query is first used to identify relevant items of community data using informational retrieval. The query is combined with the relevant community data and user data before being sent to a pre-trained language

model hosted externally via API or within local hosting for security enhancements. For example, the system could be used to generate a sensitive document in a law firm as follows: initially, the law firm pools its documents into an encrypted database and each document is stripped of sensitive information. Then, an attorney assembles the relevant client information and describe the document to be generated. Finally, the system finds a similar document from the law firm database and uses the LLM to update and customize the document based on the attorney's specific query.

A crucial aspect of this system lies in its focus on audibility and custodial control. By leveraging a trusted execution environment, community members can securely manage their data, apply privacy-preserving transformations, and access the shared information pool. This approach creates a transparent data privacy trail, inherently allowing for comprehensive auditing. Moreover, the system design acknowledges the importance of data removal and the right to be forgotten. Users retain the power to revoke consent and remove their data from the community pool at their discretion. This empowers communities and organizations to maintain control over their data while harnessing the potential of large language models for effective problem-solving and decision-making.

The key insight here is that the goal of the audibility is not achieved by fine-tuning or engineering the large language models themselves, but rather through well-designed data management systems that are auditable by design. While this approach does not fully address the question of why LLMs produce the outputs they do (explainability), it provides a useful record of what information was used to generate outputs.

Community Transformers and Auditability

Explainability, the ability to determine how a machine learning algorithm arrived at a decision, has attracted significant academic attention and is widely regarded as an important goal for AI. The nature of the increasingly massive machine learning models has made this goal difficult to achieve. While the Community Transformers architecture does not offer explainability, it does provide an auditable record of the pieces of community data that were used by the model to generate a given output.

Although the Community Transformer design provides a higher level of transparency for LLMs, it does not offer a full explanation of the decision-making process for two important reasons. First, auditability does not provide information on *how* inputs were converted into a decision, only a record of which inputs the model had access to. Second, while users can audit what community data was used to generate an output, they do not have a way to audit how pre-training data influenced outputs. As a result, errors or biases in the pre-training data could lead to errors in

outputs independently of the community data that was attached to a user’s query. However, even without full explainability, auditability can significantly enhance the transparency of LLMs in practice.

Audtitability can be used both ex-post and ex-ante. Ex-post, it can be used to identify erroneous or incomplete records that led to unsatisfactory or incorrect outputs. Ex-ante, it can assist users in ensuring that the appropriate inputs are tied to their query and provide an opportunity to remove sensitive data from these inputs.

From a practical standpoint, auditability can offer significant value. It is beneficial for individuals to know which records about them were used to reach a particular decision. This understanding narrows the scope of potential records to check for errors, making it easier to appeal an automated decision. In this way, auditability takes a significant step toward fulfilling the requirements for automated decision-making under the GDPR. Similarly, organizations need to understand how automated decisions are made. For instance, if an employee inadvertently uses outdated data to generate a report, the auditable nature of the Community Transformers architecture can help identify and rectify this error.

Community Transformers and Updatability

Within the Community Transformers architecture, three distinct sources of data contribute to model outputs: personal private data, community data, and pre-training data. Each dataset is controlled by separate processes, which results in varying degrees of updatability.

Personal private data, controlled by individuals, is not retained permanently, making it perfectly updatable. Individuals have full control over their private data and can modify or delete it as they see fit.

Community data is administered by the community or organization that oversees the Community Transformers architecture. In principle, this data is easily updatable, yet practical considerations may complicate the process. In business organizations centralized control of community data facilitates updates. In other contexts, such as when a municipality administers a Community Transformer, modifying community data may necessitate additional steps, like contacting an administrator to request data updates or even putting an update to a community vote. Here, the community must strike a balance between updatability and security, as indiscriminately allowing modifications to community data raises other data security concerns.

The pre-training data poses the greatest challenge to updatability because it is typically under the control of third-party developers and there is a high cost associated with retraining models. While some proposals have emerged that allow

individuals to ascertain if their data is included in machine learning training sets and request its removal, the ability to update this training data ultimately lies with the developers. Moreover, given that training models require significant resources, once a model is trained, any meaningful updates to the training data would necessitate retraining the model – an approach that is rarely feasible due to the associated costs and resources. Fortunately, although LLMs may "memorize" information from the pre-training data, the general-purpose nature and instruction tuning of these models means they are optimized to rely on information provided in the context window, rather than on previously learned information. As a result, it is unlikely that incorrect information in the training data would result in incorrect outputs so long as the relevant prompt contains the information necessary to generate the requested output.1

Ultimately, the Community Transformers architecture provides individuals and communities with significant flexibility to modify the prompt data. This flexibility significantly expands users' ability to maintain control over the data that informs and influences machine learning models. Importantly, updatability is also required under privacy regulations and is likely necessary for the widespread commercial use of LLMs.

Conclusion

LLMs epitomize 'black box' AI and bring concerns about explainability and privacy sharply into focus. While the quest for explainability in these systems continues, we introduce Community Transformers, a technical architecture centered around LLMs, that offers auditability and updatability *by design*. Updatability is vital in the context of privacy by offering the ability to remove or rectify records. Meanwhile, auditability provides insights into the information used to reach a particular decision. Both of these features expand transparency and make it easier for LLMs to be deployed in practice. Moreover, Community Transformers have the potential to aid communities – including those with limited resources – in forming data trusts to responsibly derive value from LLMs. By increasing transparency and control over data, this architecture paves the way toward more responsible and transparent uses of powerful machine learning systems.

Sandy Pentland
MIT Toshiba Professor and Director of MIT Connection Science.

Robert Mahari
JD-Ph.D. Student at Harvard Law School and MIT.

Tobin South
Ph.D. Student at the MIT Media Lab.

***

Citation: Sandy Pentland, Robert Mahari, and Tobin South, *Transparency by Design for Large Language Models*, Network Law Review, Spring 2023.