frontiers | Frontiers in Communications and Networks

TYPE Perspective
PUBLISHED 10 September 2025
DOI 10.3389/frcmn.2025.1600750

OPEN ACCESS

EDITED BY
Zhiyuan Tan,
Edinburgh Napier University, United Kingdom

REVIEWED BY
Elia Onofri,
National Research Council (CNR), Italy

*CORRESPONDENCE
Zixin Nie,
zixinnie@rti.org

RECEIVED 02 April 2025
ACCEPTED 19 August 2025
PUBLISHED 10 September 2025

CITATION
Nie Z, Dave L and Lewis R (2025) Privacy considerations for LLMs and other AI models: an input and output privacy approach.
Front. Commun. Netw. 6:1600750.
doi: 10.3389/frcmn.2025.1600750

Privacy considerations for LLMs and other AI models: an input and output privacy approach

Zixin Nie*, Leena Dave and Rashonda Lewis

Center for Data Modernization Solutions, RTI International, Durham, NC, United States

The framework of Input and Output Privacy aids in conceptualization of data privacy protections, providing considerations for situations where multiple parties are collaborating in a compute system (Input Privacy) as well as considerations when releasing data from a compute process (Output Privacy). Similar frameworks for conceptualization of privacy protections at a systems design level are lacking within the Artificial Intelligence space, which can lead to mischaracterizations and incorrect implementations of privacy protections. In this paper, we apply the Input and Output Privacy framework to Artificial Intelligence (AI) systems, establishing parallels between traditional data systems and newer AI systems to help privacy professionals and AI developers and deployers conceptualize and determine the places in those systems where privacy protections have the greatest effect. We discuss why the Input and Output Privacy framework is useful when evaluating privacy protections for AI systems, examine the similarities and differences of Input and Output privacy between traditional data systems and AI systems, and provide considerations on how to protect Input and Output Privacy for systems utilizing AI models. This framework offers developers and deployers of AI systems common ground for conceptualizing where and how privacy protections can be applied in their systems and for minimizing risk of misaligned implementations of privacy protection.

KEYWORDS

input privacy, output privacy, large language models, artificial intelligence, privacy framework, privacy enhancing technologies

1 Introduction

Input and Output Privacy is a framework utilized by data privacy professionals to systematically design privacy protections for systems that contain, utilize, and report personal identifying information (PII), that is, all information that can be linked to an identifiable person. Input Privacy protects individual privacy when multiple parties are collaborating in computation and analyses, enabling them to share data and perform analyses without sharing private information (Ricciato et al., 2020; Stutz, 2021; The United Nations, 2023). Protection of Input Privacy typically involves using Privacy Enhancing Technologies (PETs) such as Homomorphic Encryption and Secure Multi Party Computation to obfuscate data so that it is unreadable to humans when conducting the computation, and afterwards providing human-readable results computed from all parties' data (The United Nations, 2023; Nie et al., 2024; Archer et al., 2021; Santos and Zanussi, 2022). Output Privacy protects individual privacy when releasing data to other parties or to the public by using statistical methods and applying transformations to the data such as noise addition, reducing granularity, suppression of certain records or fields, or generation

of synthetic data, resulting in datasets that are human and machine readable and have the risk of identifying data subjects reduced to beneath an acceptable threshold (Nie et al., 2024; 10 Misunderstandings Related to Anonymisation, 2023; Emam, 2013; Giomi et al., 2022; Barrientos et al., 2023; Dwork et al., 2019). The Input and Output Privacy framework has been used in the United Nations Privacy Enhancing Technologies Lab (UN PET Lab) to help conceptualize how PETs protect privacy and identify use cases where they would be best deployed (e.g., usage of Homomorphic Encryption when gathering mobile data for analysis to protect Input Privacy, and dissemination of synthetic data to the public to protect Output Privacy) (United Nations, 2022). The author of this paper also used this framework to help create a taxonomy classifying a wide variety of PETs, identifying ones that would be beneficial for creating a data sharing service for multiple US federal statistical agencies (Nie et al., 2024).

AI systems currently lack a similar framework for privacy protections. While there has been much discussion about privacy protections for AI systems both within scientific literature and in society at large, and usage of PETs within some AI systems, the lack of a framework that can easily conceptualize and explain how these protections protect privacy, what privacy concerns are being mitigated, and where to place privacy protections in the AI systems causes confusion and leads to incorrect implementations. As there are similarities between AI systems and the data systems previously described, we propose to adapt the Input and Output Privacy framework to work on AI systems to aid in the conceptualization and protection of privacy from a systems-wide perspective. We believe utilization of this framework can enable the utilization of terms and ideas familiar to data managers and data privacy experts and facilitate the cross-pollination of ideas between the two groups, leveraging the overlap with concepts as they have been traditionally used for data systems. Using the Input and Output Privacy framework helps frame discussions about privacy in a way that reflects the process used to develop, train, and deploy AI systems, providing guidance as to the privacy concerns to be aware of at each step along the process, and providing direction towards solutions to mitigate those concerns. It can also help standardize the language being used when talking about protecting privacy, producing a common cross-disciplinary vernacular understandable by technical and non-technical stakeholders. It is our hope that AI developers and deployers can use this framework to help conceptualize where to apply certain types of privacy protections within their systems, what kinds of privacy concerns are mitigated by those protections, and how those protections protect privacy.

2 Input and output privacy for data systems and AI systems

Data systems and AI systems have significant overlap in the ways they are constructed, deployed, and used. Data systems store and serve data for queries, visualizations, and analysis, which can include as their components data repositories, databases, data and analytics platforms. For these systems, there exists mature standards and complete frameworks for management, governance, quality, and privacy, such as those detailed within the DAMA DMBOK (DAMA-DMBOK, 2017). AI systems are defined from 15 U.S.C. 9401(3) as "machine-based systems that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" United State Code, (2025). At their core, AI systems are models that are trained using large amounts of data from data systems to make inferences. Because of this, many of the principles and techniques for protecting privacy for traditional data systems apply in a similar manner to AI systems, which enables adaptation of the Input and Output Privacy framework used for data systems onto AI systems.

We can illustrate the parallels between both systems when looking at the processes that are applied to data. In data systems, data typically undergoes the following processes: 1) data ingestion, where data is brought into the system; 2) data pre-processing, where data is transformed to fit the needs of the users; and 3) data service, where the data is made available for users to view, query, and analyze. Input Privacy applies within the data ingestion phase, applying privacy protections before the system and its users can read and understand data contents; Output Privacy applies during the data serving phase, applying privacy protections to data views, queries, and analysis results.

AI systems typically process data according to the following steps: 1) AI Training, where an AI model is trained to fit the data served by data systems; 2) AI model deployment, where the trained AI model is deployed on a computational system, and 3) AI inference, where the AI model is used to make predictions, recommendations, and decisions. AI training parallels the data ingestion step and AI inference parallels the data service step in data systems. As such, Input and Output Privacy considerations can be applied accordingly upon those steps as well. Input Privacy for an AI system applies to the data used for model training, protecting privacy by preventing the AI system and its developers from learning private information. Output Privacy applies to AI inferences, reducing the risk that the AI outputs private information to system users.

The AI training and model deployment steps can only occur after a data system serves data fit for training to the AI system. Afterwards, the AI system can be separated from the data system so that users only have access to the model and its outputs and not the underlying data used to train the model. This presents a fundamental difference between AI systems and data systems that affects the Input and Output Privacy considerations for these systems; removing direct access to underlying data limits vulnerabilities as attacks can only be conducted against AI models and their outputs. This applies even for AI systems that are served continual streams of data, trained using user prompts, or are able to access and query data outside of training data (such as with Retrieval Augmented Generation (RAG)), as users of the AI system would not have direct access to the underlying data being served to the model by the data system. As AI systems rely upon data served by data systems for training and validation, the Input and Output Privacy protections applied to data systems becomes Input Privacy for AI systems, while Output Privacy for AI systems have separate considerations. Figure 1 shows a high-level workflow from data system to AI system, and with Input and Output Privacy for each system indicated between the processing steps.

Table 1 provides a high-level overview of the potential threats that could arise during the different stages of AI training and deployment, and provide strategies for threat mitigation. We will go into detail about threats and mitigations in sections 3 and 4.

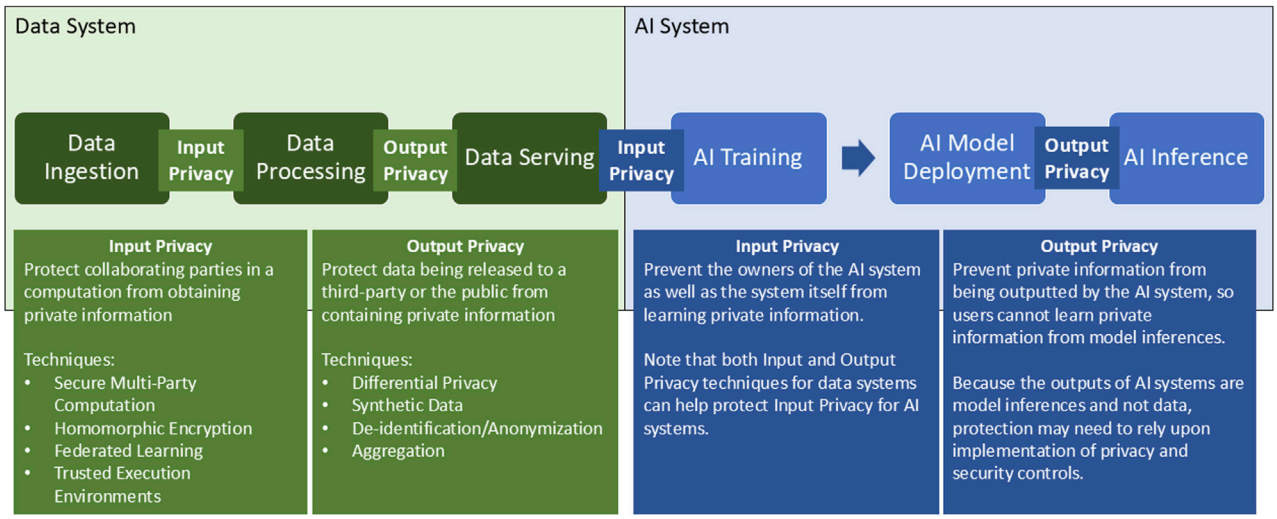

FIGURE 1
Data system to AI system workflow with locations for Input and Output Privacy. This figure shows the workflow from Data to AI system and where along the workflow Input and Output Privacy are protected for Data Systems (in green) and AI systems (in blue). Data systems and AI systems are closely related, with AI systems being built using existing data systems as their foundation. This is because AI training and model deployment can only occur after the data system can serve data fit for training to the AI system. After training, the AI system and the data system can become two separate entities, as a trained AI system does not need access to the underlying training data to make predictions.

TABLE 1 Threat models and mitigation strategies for each step in AI system training and deployment.

Stage	Privacy threat models	Input or output privacy threat	Mitigation methods
Model training and fine-tuning	Input of PII Data poisoning Unauthorized access to and viewing of PII	Input Privacy	De-identification Synthetic Data Noise addition to input data Homomorphic encryption Secure multi-party computation Federated learning Data audits
Model inference	Output of learned PII Membership inference attacks Attribute inference attacks	Output Privacy	Suppression of PII in outputs Noise addition to outputs Query limitation Monitoring of queries and outputs Tiered access controls Sanctions for inappropriate use
Continual training and reinforcement	Input of PII (especially from user prompts) Data poisoning (through user prompts)	Input Privacy	De-identification PII filtering Monitoring of inputs
Retrieval Augmented Generation	Input of PII Data poisoning Unauthorized access to and viewing of PII (in the RAG store)	Input Privacy	De-identification Synthetic Data Noise addition to input data Homomorphic encryption Data audits Monitoring of model access to RAG stores

3 Considerations for protecting input privacy for AI systems

The goals when protecting Input Privacy for AI systems are the same as when protecting Input Privacy for data systems, that is to prevent those who develop and operate the AI system (which include the developers, AI service providers, cloud computing service providers, and other users) as well as the AI system itself from learning private information. Input Privacy considerations are especially important for data controllers who wish to train and host AI systems outside their security perimeters, which happens often as many of these systems require computational power only available upon cloud-based platforms. These considerations also apply to models that utilize continuous learning, such as those that use user

prompts for training, as well as models that utilize retrieval augmented generation to bring in external information it may not have been trained on.

PETs used to protect Input Privacy for data systems, such as Secure Multiparty Computation and Homomorphic Encryption, have been investigated for a variety of AI systems (Panzade et al., 2024; Nguyen et al.; Badawi et al., 2020; Kim et al., 2022; Liu and Liu, 2023; Rho et al., 2024; Yan et al., 2024; Li et al., 2024), with a family of techniques for training AIs known as Federated Learning deployed in AI systems implemented by Apple and Google (McMahan and Ramage, 2017; Apple Machine Learning Research, 2017). However, these techniques all apply in situations where data served to an AI system may include private information; it is oftentimes more expedient and protective of privacy to remove private information from the data before serving to the AI system, so that the system does not have a chance to learn private information at all. As such, Output Privacy protecting techniques for data systems, like de-identification, differential privacy, and synthetic data, become effective input privacy protecting techniques for AI systems. Usage of training data that has undergone these protections treatment may even improve performance for certain AI models (Zhu et al., 2022; Arasteh et al., 2024; Nikolenko, 2021).

There are some specific Input Privacy considerations that only apply to certain types of AI models. For instance, some LLMs use user input as training data for the model, which can pose risks to user privacy as users may create prompts using private information such as names, locations, and medical diagnoses. These user prompts can then leak private information to other users of the model (Kshetri, 2023; Smith et al., 2024; Zhang et al., 2024). Another Input Privacy consideration that is protecting against Model Poisoning attacks, which involve malicious actors inserting "poisoned" data to train AI models, which can make the model output sensitive or private information contained within the training data (El et al., 2024). Two examples include Li et al. who demonstrated an attack that can be conducted on pre-trained models during the fine-tuning phase (Li et al., 2021), and Yao et al. who demonstrated a method they call PoisonPrompt that conducts backdoor attacks on LLMs that use user-generated prompts as training data (Yao et al., 2023).

While mitigating Input Privacy risks using PETs such as homomorphic encryption and secure multi party computation could be possible, these methods may not be sufficient (for instance, they may not mitigate risks of model poisoning) and can cause significant degradation in model performance (Yan et al., 2024; Li et al., 2024; Brown et al., 2022). Results published by Zama where they benchmarked training a deep neural network using clear-text data *versus* data that has undergone fully homomorphic encryption showed significant slowdowns in runtime on the relatively simple task of MNIST classification (Chillotti et al., 2021; Chevallier-Mames and Kherfallah, 2024). While some organizations try to mitigate privacy risks through usage of Federated Learning, not only can there be deployment hurdles, but there still exist privacy attacks on model updates and on trained models that can reveal private information (Near, 2024). For holistic protection of Input Privacy, organizations using AI systems should implement policies and procedures to prevent input of private information and poisoned data into AI systems. These include on-site deployment of AI systems, limiting the types of inputs from users and monitoring user inputs to ensure compliance (or removing the ability of AI systems to learn from user inputs entirely), review and security audits of the data being used for training models, limiting and monitoring access to training data, and conducting privacy and security reviews of public data prior to usage for AI training. Developers and deployers of PETs and AI systems have stated during interviews that organizations are more familiar with these more traditional methods than techniques like homomorphic encryption and secure multi party computation and most have prior experience implementing similar controls upon data systems (Nie et al., 2024). Application of that experience may provide a more practical pathway for protecting Input Privacy than usage of PETs.

4 Considerations for protecting output privacy for AI systems

Protecting Output Privacy in AI systems involves protecting the model and its inferences from leaking private information the model may have learned. Research has shown that some AI models have a memory of the input data that are used for training, which can lead to them outputting information about the data used to train them (Wei et al., 2024). Researchers have demonstrated attacks leveraging this vulnerability using LLMs - through prompting the model to generate large amounts of text, attackers can potentially generate verbatim passages of text used to train the model, with models that are larger and more complex more likely to memorize and output training text (Carlini et al., 2021; Staab et al., 2024; Carlini et al., 2023). They have conducted successful attacks against GPT-2 and BERT-like models where attackers were able to reconstruct individual training examples or large portions of training text (Zhang et al., 2022; Lehman et al., 2021; Diera et al., 2023).

This vulnerability the main reason why much effort has been spent protecting Input Privacy for AI systems. However, just protecting Input Privacy is not sufficient to protect attackers from obtaining private information from model outputs. For instance, with Membership Inference Attacks, an attacker can use external information to query a model to determine whether an individual was part of the dataset used to train the model (Niu et al., 2024). In certain cases, just knowing that an individual's information was part of the training data can be considered a leak of sensitive private information, examples of which include models that make predictions for medical diagnoses, models trained using data from protected classes like children or substance abuse victims, and models used for sensitive decision making such as qualification for government assistance. Another potential privacy vulnerability is Attribute Inference Attacks, where an attacker with external information about data subjects attempts to use the outputs of a model to infer private information about individuals. A study conducted by Pan et al. found that several current state-of-the-art LLMs are susceptible to revealing sensitive attributes about individuals, such as identities, genetic data, health information, and location information, through reverse-engineering of embeddings within the models (Pan et al., 2020). Another study conducted by Staab et al. demonstrated that inferences made by LLMs can reveal personal attributes linked to publicly available Reddit profiles (Staab et al., 2024).

To protect against these types of attacks, researchers have investigated various methods for protecting Output Privacy using PETs. The outputs of AI models can undergo transformations for

de-identification, such as suppression of outputted PII, aggregation, and rounding. Majmudar et al. presented a differential privacy method that could work for text generation in LLMs (Majmudar et al., 2022); however, usage of differential privacy requires careful selection of privacy parameters within the system (epsilon and delta), of which there is insufficient guidance and a lack of benchmarks for effective privacy protection (Dwork et al., 2019). A more practical approach for protecting Output Privacy is implementation of privacy and security controls around the AI system. In a similar situation as with Input Privacy controls described in the previous section, developers and deployers of PETs and AI systems have stated during interviews that traditional privacy and security controls are oftentimes more familiar for organizations who have experience implementing similar controls for data systems (Nie et al., 2024). An example of some of these controls can be found in the recommendations of the Office of Science and Technology Policy in the White House for AI systems developed and used by the US Federal Government:

- Conducting risk assessments
- Limiting data collection
- Seeking and confirming consent
- Following security best practices
- Providing more protection for data from sensitive domains
- Reporting on data collection and storage (The White House)

Other controls that can help protect privacy include technical controls to limit allowable user queries as well as limits to system outputs, access controls to limit who can use the AI systems, requirements for authentication of identity before use, setting limits to the term of access, providing tiered access based on user trust and model sensitivity (e.g., a user with sufficient security clearance and verified credentials can access a model trained using PII, whereas users without the security clearance can only use models trained on public data), logging and regular monitoring of access to models, model queries, and model outputs, machine and human review of model outputs to detect potential malicious use, and contractual agreements between model providers and users (i.e., acceptable use agreements, terms of use) with sanctions for inappropriate use to deter malicious users through threat of punishment. Many organizations have similar controls in place for existing data systems, from which they can use their experience to expedite the implementation of similar controls for their AI systems (Nie et al., 2024).

5 Case study: application of the input and output privacy framework for an LLM deployment

To demonstrate how the Input and Output Privacy Framework can help clarify what privacy concerns are being mitigated where in the system, and how they are being mitigated, we can apply the framework to analyze the deployment of a fine-tuned LLM as a job aid within an organization. AI developers work with the Privacy Office to conduct assessments, identifying privacy risks that can arise from usage of the system. The main risks identified center around documents that contain PII about the organization's customers, which can only be used for specific purposes related to customer service. To evaluate how these risks affect the AI system, they determine how it impacts Input Privacy (e.g., input of documents containing PII for fine-tuning the model results in the model remembering private information) and Output Privacy (e.g., model outputs private information when queried for purposes outside of customer service). As it is possible for the LLM to retain private information, and the organization plans to use the LLM for purposes beyond customer service, this privacy risk must be mitigated. After evaluating various options including PETs, they decide use simple redaction to remove PII from the documents used for training to protect Input Privacy, and set up a system for limiting queries and outputs and monitoring employee usage of the LLM to protect Output Privacy, aligning with their current data management practices and matching the technical capabilities of their organization.

6 Conclusion

Utilization of the Input and Output Privacy Framework helps clarify what protections should be put into place in different parts of AI systems to ensure holistic privacy protection. Significant effort has been put into protecting Input Privacy for AI due to the thinking that if AI is not trained on private information, then the privacy risk has been managed. However, there are still privacy attacks on the outputs of AI models such as membership inference attacks and attribute inference attacks that present real and clear privacy risks. Protection of Input and Output Privacy involves a combination of privacy techniques which can include usage of PETs; however, a more practical path for many organizations would be to implement privacy and security controls around a system such as access controls, query limitation, usage monitoring, and strict enforcement of contractual agreements. Privacy and security controls are easier to implement for AI systems that are not public facing, such as LLMs deployed within an organization's internal systems. Public-facing AI systems have a much harder time implementing controls, which makes it more imperative that Input Privacy is protected properly.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZN: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review and editing. LD: Funding acquisition, Supervision, Writing – review and editing. RL: Conceptualization, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research and publication of this article is funded by RTI International.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

10 Misunderstandings Related to Anonymisation (2023). Agencia Española de Protección de Datos (AEPD). Available online at: https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en.

Apple Machine Learning Research (2017). Learning with privacy at scale. Apple. Available from: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

Arasteh, S. T., Lotfinia, M., Nolte, T., Saehn, M., Isfort, P., Kuhl, C., et al. (2024). Securing collaborative medical AI by using differential privacy: domain transfer for classification of chest radiographs. *Radiol. Artif. Intell.* 6 (1), e230212. doi:10.1148/ryai.230212

Archer, D., O'Hara, A., Issa, R., and Straus, S. (2021). *Sharing sensitive Department of Education data across organizational Boundaries using secure Multiparty computation*, 9. Washington, DC: Georgetown University. Available online at: https://drive.google.com/file/d/1CURfl3q8j_NOBiaOuPEleJBZFpwQcwti/view.

Badawi, A. A., Hoang, L., Mun, C. F., Laine, K., and Aung, K. M. M. (2020). PrivFT: private and Fast text classification with homomorphic encryption. *IEEE Access* 8, 226544–226556. doi:10.1109/access.2020.3045465

Barrientos, A. F., Williams, A. R., Snoke, J., and Bowen, C. M. (2023). A Feasibility study of differentially private summary Statistics and Regression analyses with Evaluations on administrative and survey data. *arXiv* 119, 52–65. doi:10.1080/01621459.2023.2270795

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Tramèr, F. (2022). What does it mean for a language model to Preserve privacy? *arXiv*, 2280–2292. doi:10.1145/3531146.3534642

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting training data from large language models. 2633–2650. Available online at: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2023). "Quantifying Memorization across neural language models," in *The Eleventh International Conference on learning Representations*. Available online at: https://openreview.net/forum?id=TatRHT_1cK.

Chevallier-Mames, B., and Kherfallah, C. (2024). Making FHE Faster for ML: beating our previous paper benchmarks with concrete ML. Zama.ai. Available online at: https://www.zama.ai/post/making-fhe-faster-for-ml-beating-our-previous-paper-benchmarks-with-concrete-ml.

Chillotti, I., Joye, M., and Paillier, P. (2021). Programmable Bootstrapping enables Efficient homomorphic inference of deep neural networks. 1–19. doi:10.1007/978-3-030-78086-9_1

Dama-Dmbok, I. D. (2017). *Data management Body of Knowledge*. 2nd Edition. Denville, NJ, USA: Technics Publications, LLC.

Diera, A., Lell, N., Garifullina, A., and Scherp, A. (2023). Memorization of named entities in fine-tuned BERT models. 258–279. doi:10.1007/978-3-031-40837-3_16

Dwork, C., Kohli, N., and Mulligan, D. (2019). Differential privacy in practice: Expose your epsilons. *J. Priv. Confidentiality* 9 (2). doi:10.29012/jpc.689

El Mestari, S. Z., Lenzini, G., and Demirci, H. (2024). Preserving data privacy in machine learning systems. *Comput. and Secur.* 137, 103605. doi:10.1016/j.cose.2023.103605

El Emam, K. (2013). *Guide to the de-identification of Personal Health Information*. Boca Raton, FL: Taylor and Francis.

Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B. (2022). A Unified framework for Quantifying privacy risk in synthetic data. *arXiv*. Available online at: http://arxiv.org/abs/2211.10459.

Kim, M., Jiang, X., Lauter, K., Ismayilzada, E., and Shams, S. (2022). Secure human action recognition by encrypted neural network inference. *Nat. Commun.* 13 (1), 4799. doi:10.1038/s41467-022-32168-5

Kshetri, N. (2023). Cybercrime and privacy threats of large language models. *IT Prof.* 25 (03), 9–13. doi:10.1109/mitp.2023.3275489

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., and Wallace, B. C. (2021). Does BERT Pretrained on Clinical Notes reveal sensitive data? *arXiv*. Available online at: http://arxiv.org/abs/2104.07762.

Li, L., Song, D., Li, X., Zeng, J., Ma, R., and Qiu, X. (2021). Backdoor attacks on pre-trained models by Layerwise Weight poisoning. *arXiv*. Available online at: http://arxiv.org/abs/2108.13888.

Li, Y., Zhou, X., Wang, Y., Qian, L., and Zhao, J. (2024). A survey on private transformer inference. *arXiv*. Available online at: http://arxiv.org/abs/2412.08145.

Liu, X., and Liu, Z. (2023). LLMs can understand encrypted prompt: towards privacy-computing Friendly Transformers. *arXiv*. Available online at: http://arxiv.org/abs/2305.18396.

Majmudar, J., Dupuy, C., Peris, C., Smaili, S., Gupta, R., and Zemel, R. (2022). Differentially private Decoding in large language models. *arXiv*. Available online at: http://arxiv.org/abs/2205.13621.

McMahan, B., and Ramage, D. (2017). Federated learning: Collaborative machine learning without Centralized training data. *Google Res*. Available online at: https://blog.research.google/2017/04/federated-learning-collaborative.html?abstract_id=3808054.

Near, J. (2024). *Privacy attacks in Federated learning*. NIST. Available online at: https://www.nist.gov/blogs/cybersecurity-insights/privacy-attacks-federated-learning.

Nguyen, L., Phan, B., Zhang, L., and Nguyen, T. (2025). An Efficient approach for securing Audio data in AI training with fully homomorphic encryption. Available online at: https://www.authorea.com/doi/full/10.36227/techrxiv.170956397.78402834?commit=8ad5d37adb9c0e6375b7feaacc3436accfbc0a2d.

Nie, Z., Lewis, R., Gartland-Grey, A., and Riley, A. F. (2024). America's DataHub Consortium: privacy preserving Technology phase 1 – Environmental scan. *RTI Int. Natl. Cent. Sci. Eng. Statistics*, 69. Available online at: https://www.americasdatahub.org/wp-content/uploads/2024/05/ADC-PPT_FinalReport.pdf.

Nikolenko, S. I. (2021). *Synthetic data for deep learning*, 174. Cham: Springer International Publishing. Available online at: https://link.springer.com/10.1007/978-3-030-75178-4.

Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., et al. (2024). A survey on membership inference attacks and defenses in machine learning. *J. Inf. Intell.* 2 (5), 404–454. doi:10.1016/j.jiixd.2024.02.001

Pan, X., Zhang, M., Ji, S., and Yang, M. (2020). "Privacy risks of General-Purpose language models," in *2020 IEEE Symposium on security and privacy (SP)*, 1314–1331.

Panzade, P., Takabi, D., and Cai, Z. (2024). "MedBlindTuner: towards privacy-preserving fine-tuning on Biomedical Images with Transformers and fully homomorphic encryption," in *AI for health Equity and Fairness: leveraging AI to Address social Determinants of health*. Editors A. Shaban-Nejad, M. Michalowski, and S. Bianco (Cham: Springer Nature Switzerland), 197–208. doi:10.1007/978-3-031-63592-2_15

Rho, D., Kim, T., Park, M., Kim, J. W., Chae, H., Cheon, J. H., et al. (2024). Encryption-friendly LLM Architecture. *arXiv*. Available online at: http://arxiv.org/abs/2410.02486.

Ricciato, F., Bujnowska, A., Wirthmann, A., Hahn, M., and Barredo-Capelot, E. (2020). A reflection on privacy and data Confidentiality in Official Statistics. Available online at: https://www.researchgate.net/publication/339030033_A_reflection_on_privacy_and_data_confidentiality_in_Official_Statistics.

Santos, B., and Zanussi, Z. (2022). Privacy preserving technologies, part three: private statistical analysis and private text classification based on homomorphic encryption. Available online at: https://www.statcan.gc.ca/en/data-science/network/statistical-analysis-homomorphic-encryption.

Smith, V., Shamsabadi, A. S., Ashurst, C., and Weller, A. (2024). Identifying and mitigating privacy risks Stemming from language models: a survey. *arXiv*. Available online at: http://arxiv.org/abs/2310.01424.

Staab, R., Vero, M., Balunović, M., and Vechev, M. (2024). Beyond Memorization: Violating privacy via inference with large language models. *arXiv*. Available online at: http://arxiv.org/abs/2310.07298.

Stutz, J. (2021). Structured Transparency: Ensuring input and output privacy. *OpenMined Blog Priv. AI Ser*. Available online at: https://blog.openmined.org/structured-transparency-input-output-privacy/.

The United Nations (2023). Guide on privacy-Enhancing technologies for Official Statistics. Available online at: https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf.

The White House (2025). Blueprint for an AI Bill of Rights | OSTP. Available online at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

United Nations (2022). UN PET Lab. News release. Available online at: https://unstats.un.org/bigdata/events/2022/unsc-un-pet-lab/UN%20PET%20Lab%20-%20Press%20Release%20-%2025%20Jan%202022.pdf.

United States Code (2025). Title 15, Commerce and trade. § 9401. Available online at: https://uscode.house.gov/view.xhtml?req=(title:15%20section:9401%20edition:prelim).

Wei, J., Zhang, Y., Zhang, L. Y., Ding, M., Chen, C., Ong, K. L., et al. (2024). Memorization in deep learning: a survey. *arXiv*. Available online at: http://arxiv.org/abs/2406.03880.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., et al. (2024). On protecting the data privacy of large language models (LLMs): a survey. *arXiv*, 1–12. doi:10.1109/icmc60390.2024.00008

Yao, H., Lou, J., and Qin, Z. (2023). PoisonPrompt: backdoor attack on Prompt-based large language models. *arXiv*. Available online at: http://arxiv.org/abs/2310.12439.

Zhang, R., Hidano, S., and Koushanfar, F. (2022). Text revealer: private text reconstruction via model Inversion attacks against Transformers. *arXiv*. Available online at: http://arxiv.org/abs/2209.10505.

Zhang, S., Ye, L., Yi, X., Tang, J., Shui, B., Xing, H., et al. (2024). "Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv*. Available online at: http://arxiv.org/abs/2410.14931.

Zhu, T., Ye, D., Wang, W., Zhou, W., and Yu, P. S. (2022). More than privacy: applying differential privacy in Key Areas of Artificial Intelligence. *IEEE Trans. Knowl. Data Eng.* 34 (6), 2824–2843. doi:10.1109/tkde.2020.3014246