Front. Comput. Sci., 2025, 19(10): 1910356
https://doi.org/10.1007/s11704-024-40583-8

REVIEW ARTICLE

Privacy dilemmas and opportunities in large language models: a brief review

Hongyi LI, Jiawei YE (✉), Jie WU

School of Computer Science, Fudan University, Shanghai 200433, China

Abstract The growing number of cases indicates that large language model (LLM) brings transformative advancements while raising privacy concerns. Despite promising recent surveys proposed in the literature, there is still a lack of comprehensive analysis dedicated to text privacy specifically for LLM. By comprehensively collecting LLM privacy research, we summarize five privacy issues and their corresponding solutions during both model training and invocation and extend our analysis to three research focuses in LLM application. Moreover, we propose five further research directions and provide prospects for LLM native security mechanisms. Notably, we find that most LLM privacy research is still in the technical exploration phase, with the hope that this work can assist in LLM privacy development.

Keywords large language model, data privacy, data protection

1 Introduction

Transformer-based LLMs acquire rich language knowledge and patterns from large-scale corpus, endowing the models with understanding and generation capabilities in the semantic and syntactic aspects of natural language. Nowadays, LLMs have garnered widespread attention, with the release of commercial models such as OpenAI GPT, Meta OPT, and Google T5, showcasing their versatile applications and inference capabilities. Concurrently, the release of ChatGPT has triggered an LLM wave, with numerous open-source models such as LLaMA, MOSS, and ChatGLM emerging one after another in just a few months which has greatly expanded the possibilities for practical implementation.

Unfortunately, LLMs are suffering from privacy risks [1,2]. On the one hand, training corpus and user prompts are facing leakage challenges. Recently, research [3] reveals that attackers can extract GB-level training data from popular LLMs with an adequate budget. According to the Cyberhaven report [4], incidents involving sensitive information flowing into ChatGPT increased by 60.4% among 100,000 employees in an average week after the release of GPT4. The extracted data includes source code, project plans, personally identifiable information (PII), etc., with internal company information being the most prevalent. On the other hand, attackers can exploit LLM capabilities to conduct illegal behaviors and violate privacy [5–7]. In addition, LLMs face multiple challenges in real-world applications due to their excellent versatility, including inter-model invocation and multi-modal data, intensifying privacy concerns. In this context, how to protect privacy in LLM construction and utilization has become an urgent problem to be solved.

Recently, some surveys have explored various LLM topics, such as hallucination [8] and memorization [9], most of which involve LLM security discussion. Recent literatures [10,11] have delved into LLM security, but overlooking specific attention to privacy, where privacy is a part of security traditionally. Specifically, works [12,13] approach the security, authenticity, and controllability of LLMs from the generated content perspective, discussing privacy concerns related to training data leakage. Whereas, some surveys [14,15] concentrate on threats, attacks, and defense aspects of LLMs, involving discussion of privacy protection techniques, and privacy assessment. Furthermore, study [16] deals with privacy analysis on the security plugin-integrated language platforms from the LLM application perspective. It is noteworthy that privacy definitions vary across different domains (including images, text, and others) and encompass various manifestations such as PII, trade secrets. Despite these variations, commonality exists in the fundamental privacy issues and protection methods.

In our work, by using keywords related to “privacy” and “LLM” to search and collect papers in academic databases, we focus on the privacy investigation of natural language text in LLMs. Table 1 shows the results of our review compared with existing surveys. By reviewing LLM privacy concerns and prevalent research, along with analyzing reasearch focus in LLM application, we find that the current LLM privacy research primarily resides in the technical exploration phase, and there exists a certain gap from practical application.

The main contributions of our work are as follows:

- We conduct an in-depth investigation into privacy issues within LLMs, exploring the latest research

Received June 10, 2024; accepted November 19, 2024

E-mail: jwye@fudan.edu.cn

Table 1 Comparative analysis of recent reviews and surveys. TDL, TDE, PA, PL, and CO represent Training Data Leakage, Training Data Erasure, Privacy Assessment, Prompt Leakage, and Controllable Outputs. SU, SDC, and PA represent Privacy Compliance, Secure Data Sharing, Privacy Traceability and Access Control, while NOC represents the number of citations

Citation	Focus area	Year	Privacy issues					Application			NOC
			TDL	TDE	PA	PL	CO	SU	SDC	PA	
[17]	Model protection techniques and privacy norms	2022	●	○	●	○	○	○	○	○	123
[13]	Threats and vulnerabilities from LLMs	2023	●	○	●	●	○	○	○	○	34
[14]	LLM privacy attack and Defense technique	2023	●	●	●	○	○	○	○	○	14
[18]	the right to be Forgotten in LLM	2023	●	●	●	○	○	○	○	○	20
[9]	LLM memorization	2023	●	●	●	●	●	○	○	○	7
[10]	AIGC security and threat	2023	●	●	●	○	●	○	○	○	61
[12]	Security and privacy on AIGC generated data	2023	●	○	○	○	●	○	○	○	13
[15]	LLM Applications and Limitations in Privacy Security	2024	●	○	●	●	●	○	○	○	70
[19]	Red-team models works in LLM	2024	●	●	●	●	●	○	●	○	11
[20]	LLM Privacy Attack and Defense technique	2024	●	○	●	●	○	○	○	○	70
Ours	LLM privacy concerns and research focuses in its application	–	●	●	●	●	●	●	●	●	–

Note: Solid circle indicates involvement, while hollow circle indicates non-involvement.

advancements. According to our research, this paper represents the first comprehensive exploration of LLM privacy issues related to text-sensitive information.

- We provide a detailed analysis of five privacy issues and solutions in LLM training and invocation, as well as delving into three privacy-centric research focuses in LLM application that are not mentioned previously.
- We point out five potential research directions in LLM privacy aspects and propose three innovative prospects for LLM native security mechanisms.

The rest of this paper is organized as follows. Section 2 briefly reviews the progress of LLM. Section 3 discusses privacy issues at different stages and analyzes current research progress. In Section 4, we focus on privacy concerns in LLM application. Finally, we provide analysis and outlooks in Section 5, as well as the conclusion in Section 6.

2 Background

2.1 The development of LLMs

In 2018, the emergence of pre-trained models such as GPT and BERT introduce a new paradigm of "pre-training + fine-tuning" that combines unsupervised pre-training with supervised fine-tuning, aiming to learn a generalized representation that can transfer across various tasks. Recently, researchers [21,22] have discovered that by increasing the parameters and data size of pre-trained language models, LLMs not only significantly improve performance but also demonstrate advanced capabilities such as text generation and content understanding that are not available in the small models. Furthermore, the introduction of industry-specific LLMs [23,24] based on domain-specific data further propels the application and development of LLMs.

LLMs require massive pre-training corpora to acquire general language capabilities that demonstrate powerful learning and reasoning abilities after fine-tuning datasets from various domains. However, existing corpora may contain information posing privacy risks, such as personal information in publicly available corpora or business-sensitive information in fine-tuning datasets. The robust memory and inference capabilities may lead to leakage of training data and prompt, enabling the inference of sensitive information and causing potential losses. Meanwhile, attackers [5–7] can exploit the capabilities of LLMs to access and infer private information for illegal activities, resulting in economic losses. Furthermore, LLMs face multiple challenges in real-world applications due to their excellent versatility, including inter-model invocation and multi-modal data, intensifying privacy concerns. Against this backdrop, effectively protecting sensitive information throughout the LLM lifecycle has become an urgent issue.

2.2 The fundamental processes of LLMs

Typically, LLMs mainly consist of two stages: training and invocation, as shown in Fig. 1. In the training stage, the model undergoes three steps. Initially, LLM is pre-trained using a large unsupervised corpus to obtain the foundational language understanding (Step 1). Subsequently, techniques such as low-

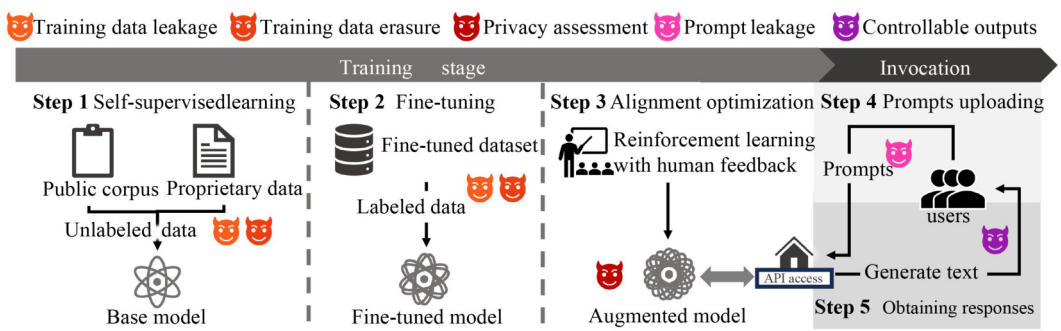

Fig. 1 LLM process and privacy issues

rank adaptation [25], soft prompt tuning [26], and in-context learning [27] are used to fine-tune the specific dataset to adapt LLM to a specific task (Step 2). To ensure the model's robustness in ethical constraints and security, previous studies have used reinforcement learning from human feedback (RLHF) [28–30] to achieve alignment optimization (Step 3). For invocation, users typically interact with LLMs through remote API calls, uploading the constructed prompts to the platforms (Step 4) and receiving the model-generated answers (Step 5). Notably, the specific training data and model parameters are not visible to the users. In our work, we deeply analyse five privacy issues in this steps. Specifically, training data leakage and training data erasure are primarily associated with the training data during step 1 and step 2, while privacy assessment focuses on the augmented model in step 3. Prompt leakage is centered around the user prompts upload process during step 4, and controllable outputs pertain to the responses obtained in step 5.

LLMs are widely used in applications such as intelligent question answering, text generation, and text analysis. However, with the increasing research and numerous LLM privacy incidents, LLM's powerful capabilities and new working paradigms invoke concerns on privacy.

3 Privacy issues and solutions

By collecting 90 related literatures, we point out the privacy scope in current research (Section 3.1) and find that current LLM privacy research mainly focuses on three aspects during training: training data leakage (Section 3.2), training data erausure (Section 3.3), and privacy assessment (Section 3.4). For invocation, the research primarily addresses two issues: prompt leakage (Section 3.5) and output controllability (Section 3.6). Figure 1 provides a visual representation of specific phases where these issues occur.

3.1 Privacy scope in literatures

Privacy is naturally understood as the dynamic sharing of sensitive information based on context [17]. In certain situations, sharing sensitive information with specific individuals or entities may be appropriate, while in other contexts, such sharing could constitute a privacy violation. When it comes to LLM privacy, the research on LLM privacy does not pay much attention to the privacy scope, but more on defining the privacy scope in some fields (such as finance, medical care, and personal) at the conceptual level. Some studies [5,31,32] focus on defining privacy through datasets, treating PII in public datasets as sensitive information or adopting data synthesis techniques to generate simulated data [33,34], which concentrate on specific types of privacy. Meanwhile, with the development of regulations and LLMs, some works [35–37] shift towards automatically interpreting privacy regulations to clarify privacy boundaries, which is elaborated in Section 4.1. However, there is still a gap between the definitions of privacy boundaries and practical applications. We will discuss this further in the context of five specific issues and provide an insight in Section 5.3.

3.2 Training data leakage

LLM training data can be extracted from the LLM-generated content [38,39], posing leakage risks. Specifically, training data consists of self-supervised corpora and fine-tuned datasets. Corpora are typically constructed from publicly available datasets such as Wikipedia, books, journals. Despite the absence of explicit sensitive information for a specific person whose PII data was not considered in the training set, recent studies [5,40] show that pre-trained models may still infer PII. Moreover, fine-tuned datasets usually focus on task-specific scenarios containing sensitive information such as business data, customer information, etc. The robust memory and association capabilities of LLM are responsible for training data leakage. On the one hand, LLM efficiently memorizes training data [7], and the larger the model, the easier it is to retain the training data. On the other hand, LLM exhibits outstanding inference capabilities that rapidly enhance their association accuracy when presented with few-shot settings [40], thereby increasing the risk of data leakage.

Generally, protection techniques can be divided into two categories: text scrubbing and privacy-preserving algorithms. (1) Text scrubbing aims to filter, mask, or delete sensitive information in text, which is applicable for data preprocessing and model-generated content filtering, as shown in Fig. 2. Works [41,42] deduplicate the training data based on matching algorithms, focusing on the model perplexity and the model memorization after data deletion. However, they do not fully consider the characteristics of the deleted data itself and assess the potential damage of de-emphasis on the model's positive memory capacity (e.g., question answers). Studies [33,38] use annotation tools to detect PII but simply removing or masking sensitive information may lead to semantic loss [43]. In addition, some works [33,34,44] use differential

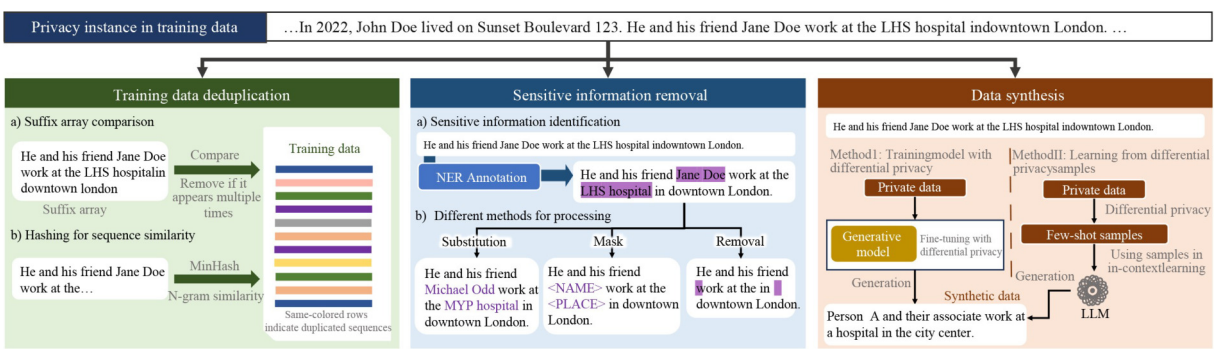

Fig. 2 Privacy instances and text scrubbing methods

privacy to achieve privacy data synthesis, but the model-generated content suffers from length truncation, and the synthesized data's quality needs to be improved. (2) Privacy-preserving algorithms provide privacy guarantees during training. Works [45–47] introduce noise during fine-tuning to protect fine-tuned datasets, yet require predefined sensitive information boundaries and the added noise affects model performance. Another approach [29] uses RLHF for privacy preservation, but introduces additional computational complexity.

Through the analysis, current research mainly focuses on handling sensitive information with well-formatted definitions as shown in Table 2, and it faces the following challenges. First, sensitive information may be referred to and written indirectly, which may manifest in acronyms or pronouns across different paragraphs. Notably, it is difficult to reliably recognize with current technologies and manual detection and filtering [64]. Second, protecting a specific unit of data is not equivalent to achieving privacy preservation. Most existing research protects specific collections containing limited sensitive information (e.g., PII [33,38,46]) but cannot ensure the entire collection's privacy. Third, existing techniques [38] emphasize privacy preservation while facing challenges in terms of model usability.

3.3 Training data erasure

The right to be forgotten originating from regulations refers to the citizen's right to ask data processors to delete their personal information [18]. Some researchers [23,65,66] believe that the need to forget sensitive information within models is consistent with it. Consequently, LLM providers should follow user's needs to update and delete their information in training data on time, which means LLMs need to possess the ability to forget specific training instances. However, it faces the challenges of high costs and complex situations regarding the data to be deleted. Due to LLM's high computational cost and long training time, retraining the model from scratch cannot meet the iteration deadlines specified in the regulation. Furthermore, data diversity and uncertain form make it difficult for the protection techniques mentioned in Section 3.2 to ensure deletion integrity.

Recent research pay attention to machine unlearning that aim to achieve a more cost-efficient forgetting process. Table 3 presents detailed research on machine unlearning

Table 2 Summary of LLM privacy protection methods

Method	Type		Model	Dataset	Level of information protection	Code
	Text scrubbing	Privacy-preserving algorithms				
[41]	√		T5 Reflective Decode [48]	OpenWebText [49] C4 [50]	Sequence	√
[42]	√		T5	Wikipedia [51], C4 LM1B [52], RealNews [53]	Token document	√
[33]	√	√	–	ATIS [54], SNIPS [55] tripAdvisor [56]	Sequence	
[45]		√	RoBERTa	Common Crawl1 Wikipedia2, mC4	Token	
PPLM [46]	√	√	LLaMA2	medical-flashcards medical-wikidoc wikidoc-patient-information	Token	√
EW-Tune [47]		√	RoBERTa	QQP [57], SST-2 [58] MNLI [57], QNLI [59]	Sequence	√
[38]	√	√	GPT2	ECHR [60], Enron [61] Yelp-health	Token	√
[44]		√	GPT2	Yelp	Sequence	√
DP-ICL [34]		√	GPT3	AGNews [62], TREC [63] DBPedia [62], MIT Movies [63]	Token	√

Table 3 The details of unlearning algorithms for text information protection

Method	Type	Model	Dataset	Model tasks	Code	Description
APA [67]	Exact unlearning	TALLRec	BookCrossing [68] MovieLens [69]	Recommendation		Use distinct adapters for partitioned training data shards, retrain only affected adapters, and employ parameter-level aggregation to enable efficient unlearning while maintaining performance.
NEO [70]	Parameter optimization	GPT-NEO OPT	Hellaswag etc.	Classification generation	√	Employ gradient ascent on target sequences to effectively remove them with minimal effect on the model's performance.
KGA [71]	Parameter optimization	DistilBERT	LEDGAR [72] IWSLT [73] PersonaChat [74]	Classification generation	√	Maintain knowledge gap differences in NLP tasks, leading to efficient unlearning with substantial improvements in performance.
EUL [65]	Parameter optimization	T5	IMDB [75] SAMSum [76]	Classification summarization	√	An efficient unlearning method for LLMs with lightweight layers fused and injected into transformer models, improving data erasure efficiency without the need for retraining.
[77]	Parameter sharing	BERT	SST2 [58] PAN16 [78]	Classification	√	Analyze how model fusion affects shortcuts, biases, and memorization, showing its effectiveness in reducing biases and addressing privacy concerns by preventing memorization.
ICUL [79]	Contextual learning	Bloom	SST2 [58] Amazon polarity [62] Yelp polarity [62]	Classification	√	Provide a flipped label and additional correctly labeled instances during inference to remove targeted training data points while maintaining competitive performance.

methods for protecting the privacy of text-sensitive information. Usually, it can be categorized based on the accuracy requirements for the forgetting process into exact unlearning and approximate unlearning. Exact unlearning [67,80] speeds up the retraining process by partitioning the training dataset and removing exact data points from the model to ensure the particular training instance eradication completely. Yet, it requires access to the entire dataset during training and forgetting, and the model's fairness may be sacrificed [81]. In contrast, approximate unlearning, allows for some error or residual information in the forgetting process, sacrificing accuracy for more cost-effective forgetting. It mainly includes three types. (1) Parameter optimization: research [65,70,71] modify the parameters in the model to alter the training objective, enabling the model to avoid retraining. Without requirements for access to the entire dataset, this approach is more suitable for model providers to maintain during training, but there is a demand for model computational cost. (2) Parameter sharing: study [77] uses multiple models for parameter fusion, revealing that forgetting occurs when data is not shared among models, promoting the forgetting of knowledge about non-shared data among models. This method requires explicit partitioning of the dataset during the training phase, which is suitable for scenarios where relatively few data need to be deleted, especially during the early stages of model training. (3) Contextual learning: study [79] focuses on the LLM inference capabilities. By providing inverted labels and additional correctly labeled instances as inputs during inference, knowledge forgetting can be achieved without requiring parameter changes.

An exploration of current studies indicates that unlearning algorithms can effectively reduce data deletion time, but most works [77,79] mainly pay attention to analyzing the algorithm's impact on model performance, while ignoring other key factors such as fairness, transparency, and privacy. Recent study [66] shows that neighboring data points are more likely to suffer from privacy leakage after the application of unlearning algorithms. Therefore, the robustness assessment by forgetting learning algorithms deserves attention.

3.4 Privacy assessment

As LLMs use training data from the internet and existing research pay limited attention to the quality and credibility of online resources [82], it is crucial to evaluate privacy risks before deploying the model in applications. Assessment includes both model privacy and usability.

Recent studies evaluate language models such as GPT2 through adversarial attacks, including membership inference [32], data reconstruction [38], data extraction [31], and attribute inference [6]. Surprisingly, researchers [83,84] identify privacy risks in parameter-efficient fine-tuning techniques for LLMs, where in-context learning is vulnerable to membership inference attacks and low-rank adaptation are highly sensitive to backdoor attacks. Though recent models have implemented security mechanisms like RLHF [30] to enhance security, LLMs maintain a black-box nature, especially with limited disclosure of defense mechanism, which poses challenges to the assessment. Some studies [39,85] fine-tune the models by constructing malicious datasets to induce the models to output sensitive information. They find that the RLHF mechanisms do not make the LLMs forget the sensitive information learned in the pre-training phase. Instead, these mechanisms reduce the likelihood of the model-generating content related to this information, which is still retained in the parameters. In addition, some work [86] analyzes existing models through prompt attacks and finds that current defense mechanisms have been quite effective in rejecting direct prompt queries. Nonetheless, sensitive information exposure persists when constructing complex prompts (e.g., thought chains [86], reward-feedback mechanisms [87]), or combining prompts with external information [5]. Moreover, LLMs are suffering from hallucinations posing a challenge for privacy assessment [18]. Model hallucination refers to the possibility that model inferences may lack factual basis, resulting in content generation that deviates from reality which could potentially impact the privacy assessments.

By delving into the investigation, current study focuses on malicious changes to the training dataset during fine-tuning or prompt injection for privacy assessment. Nevertheless, (1) privacy assessment requires multi-dimensional sensitive information evaluation. Existing works concentrate on analyzing token-level sensitive information as shown in Table 4, but sensitive information manifests in various forms such as documents in real-world scenarios. Given the diversity of sensitive information and variations across different scenarios, a single evaluation index makes it difficult to comprehensively judge the model's ability. (2) A more comprehensive evaluation is required to analyze the privacy implications of parameter-efficient fine-tuning techniques which are becoming widely adopted for LLMs. Current evaluation [83] predominantly uses traditional membership inference attacks and assesses only three fine-tuning methods, requiring it to broaden both the evaluation methods and the range of fine-tuning techniques assessed. (3) Whether the effect of privacy risk assessment is affected by hallucinations [9] needs to be investigated because the training data is not publicly available. As represented in Table 4, existing works use model perplexity and accuracy to evaluate model memorization, without considering the hallucination impacts in privacy assessment. Therefore, the multi-granularity sensitive information evaluation index and hallucination-privacy impact need to be further explored.

3.5 Prompt leakage

LLMs typically employ remote service APIs in practice, where users provide prompts to obtain responses. Due to the potential sensitive information inclusion in prompts [1], and the fact that those with access to the prompts can imitate behavior to generate similar content, prompt leakage has become a significant concern. While LLMs present a black-box nature to users, they remain visible to providers and developers. Despite recent provider statements indicating the exclusion of unauthorized API client data from training data, this declaration does not extend to instances of publicly invoked products, thus still posing potential risks of privacy

Table 4 The Details of Privacy Assessment Research. LIP, LCS represents Level of Information Protection and the longest common substring. Target means the protected information object of the assessment

Method	LIP	Model	Dataset	Target	Metric	Code	Description
[7]	Token	GPT2	Common Crawl	PII, url, code, etc.	Exacted number, preplexity	√	Propose a data extraction attack to extract hundreds of verbatim text sequences from the model's training data, discovering larger models are more vulerable than small one.
[31]	Token	GPT2 Bert2Bert	EUII	PII	Exposure [88], exacted number		Introduce pattern extraction attack that effectively extracts sensitive information in the Service API setting.
[40]	Token	GPT-Neo	Enron	PII	Accuracy	√	Analyze pre-trained LLMs leak PII through memorization, though the likelihood of extracting precise details is low due to their weak associative skills.
[5]	Token	GPT3.5 GPT4 Llama2 PaLM2 Claude	PersonalReddit	PII	Accuracy	√	Study on the capabilities of pretrained LLMs to infer personal attributes from text, and find current LLMs can infer PII an unprecedented scale.
[39]	Token	GPT2	ECHR, Ai4Privacy [89] Enron, Wikitext [90]	PII	Precision, recall, cka [91]		Fine-tun LLMs with small PII datasets and use a PII association task to recover hidden PII, find that LLM fine-tuning amplifies privacy risks.
[32]	Record	BERT	MIMICIII, i2b2	health records	Precision, recall	√	Develop a stronger membership inference attack using likelihood ratio hypothesis testing, showing that masked language models are highly susceptible.
[86]	Token	ChatGPT New Bing	Enron	PII	LCS, accuracy	√	Study the privacy threats from ChatGPT and the New Bing, revealing that application-integrated LLMs may introduce new privacy risks.

leakage. Furthermore, attackers may infer previous prompts by contextually deducing historical records. Researchers [92,93] have successfully inferred user-provided prompts with high accuracy (over 75%) using methods such as interactive extraction and membership inference attacks, and the precision of attacks is expected to continue increasing with an increase in interaction frequency.

Current approaches [93–95] highlight employing locally deployed differential privacy methods to transform data into unreadable token representations before user upload without adjusting LLM parameters. However, attackers can leverage data reconstruction techniques [96,97] to recover prompts from token representations. Other strategies [98–101] attempt to enhance privacy preservation by locally token replacing, using text-scrubbing techniques to replace sensitive information before invoking the LLM and locally store sensitive information mappings. After generating content, sensitive information is added to the response based on the locally stored mapping.

Reviewing shows that current studies primarily center on user-side preprocessing and safeguarding to prevent prompt exposure to providers and developers. Nonetheless, (1) Current localized differential privacy methods face challenges posed by data reconstruction attacks. Prior studies have utilized methods such as token reduction [102,103] or token merging [104] to reduce or merge irrelevant token representations, preventing attackers from conducting data reconstruction. However, limited research has investigated the practical effectiveness of these methods against data reconstruction attacks within the context of existing LLM applications. (2) Token replacing methods are only applicable to specific downstream tasks. Existing research [93,95,103] currently only addresses token-level sensitive information. In tasks relying on the precise semantics of replaced entities (e.g., question answering, text generation), where accurate semantic understanding is crucial, current methods may not offer sufficient guarantees. Hence, defense against prompt reconstruction attacks in LLM applications and prompt protection methods suitable for different task dimensions need to be further explored.

3.6 Controllable outputs

The uncontrolled LLM outputs raises potential issues and introduces privacy concerns. On one hand, model-generated content may contain sensitive information or be maliciously exploited [2,105] to violate privacy. On the other hand, model-generated content may be restricted to dissemination within limited closed domains in real-world scenarios, necessitating the tracking and identification of channels and usage. Hence, it is essential to ensure the integrity and traceability of model-generated content.

Recent work [106–109] have focused on detecting and filtering privacy risks in model-generated content to ensure its safety. For instance, the work [106] uses a self-inference approach based on LLM security alignment capabilities to ensure safe model-generated content. Moreover, some approach relies on third-party LLMs, utilizing predefined templates [107] and random masking method [108] to identify potential privacy violations in user instructions and model-generated content. However, these methods, which depend on LLMs and prompt templates, face scalability challenges in real-world applications. Notably, the open-source project LLM-guard [110], integrated with the Presidio analyzer [111], is designed to accurately detect and filter PII in LLM outputs. Meta's Llama Guard [109] classifies the violation risks of LLM-generated content, including the assessment of privacy risk levels. Study [112], nevertheless, highlights the insufficient effectiveness of current safety content mechanisms against malicious inquiries. Overall, the safety detection of model-generated content is currently challenged by limited scalability and insufficient effectiveness.

Furthermore, the spotlight in recent research has been on the traceability of model-generated content. Watermarks have been a key focus which aims to embed undetectable yet reversible marks within LLM outputs to identify the meta-model and prevent the misuse. Recent research [113,114]

applies watermarking to LLMs, ensuring minimal impact on text quality when access to the original model parameters is not available. However, existing research [115] indicates that watermarks in LLMs can be quickly detected by simple classifiers, and the presence of the watermark may decrease the length and coherence of the generated content. Other studies focus on distinguishing model-generated content to achieve privacy traceability. Some studies [116–118] attempt distinction through fine-tuning LLM, but they suffer from overfitting on training, resulting in significant performance degradation when faced with unknown LLMs and cross-domain/unseen data [119]. To address this problem, research [120] introduces contrastive learning to enhance distinction performance. Furthermore, some works attempt to use LLMs to identify LLM-generated content. However, research [121,122] suggest that newer LLMs like GPT-4 cannot reliably identify various types of LLM-generated content directly. Recent studies [123,124] leverage LLM's contextual learning to improve the detection effectiveness.

Nevertheless, content traceability faces several challenges. Firstly, the practical use of watermarking faces challenges in striking a balance between the generated content detectability, the watermark effectiveness, and the generated content quality. Secondly, model-generated content distinction should be treated as a multi-classification problem rather than binary classification (as shown in Table 5) to ensure traceability for privacy. Existing research in content distinction has mainly focused on distinguishing few LLMs. Given the proliferation of current LLMs, distinction needs to progress beyond mere identification of whether the content is generated by an LLM to more nuanced classification distinctions. Thirdly, there is a need for cross-language/domain content distinction capabilities to facilitate privacy traceability.

By conducting analysis, controllable LLM outputs research is in its early technical exploration. Further studies are needed, particularly in areas such as enhancing the effectiveness and scalability in security detection for generated content, balancing watermark robustness/effectiveness and content quality, refining content distinction classification and improving the generalizability across languages/domains.

To sum up, we find that many studies focus on LLM privacy, primarily exploring existing models or early technologies within LLMs. Despite notable progress, various unknown aspects in this field deserve further investigation. Essentially, LLM privacy research remains in the exploration phase.

4 Privacy research focus in LLM application

In general, LLM involves three implementation modes in the application, including single usage, distributed construction, and model collaboration, as shown in Fig. 3. Within these modes, previous research emphasizes the utilization of LLM for privacy compliance, serving as a basis for subsequent

Table 5 The details of content distinction research

Work	New datasets	Classification	Language	Description
[116]	√	Binary	English	Build a wild tested by gathering texts from various human writings and deepfake texts generated by different LLMs.
[118]	√	Binary	English	Introduce a datasets contains ChatGPT generated content, and use RoBERTa and T5 to train classification models.
ArguGPT [121]	√	Binary	English	Create an argumentative essays corpus generated by GPT, to help educators detect AI-generated content in English teaching.
ConDA [120]		Binary	English	Develop a contrastive domain adaptation framework with less labeled data for detecting model-generated news text, within 0.8% margin of a fully supervised detector.
Outfox [123]	√	Binary	English	Use adversarial examples for mutual learning to improve detection, enhancing detection accuracy.
LLM-Pat [124]		Binary	English	Evaluate ChatGPT's zero-shot performance in detecting model-generated versus manual content, and find insights on leveraging LLMs in content detection.
[119]		Binary	English, Spanish	Study the generalization capabilities of supervised Machine-Generated Text detectors, finding they generalize well across scales but not across families.
[117]	√	Binary	English, French	Propose a method for developing and evaluating ChatGPT detectors in multiple languages, but display evident vulnerabilities in across-domain.

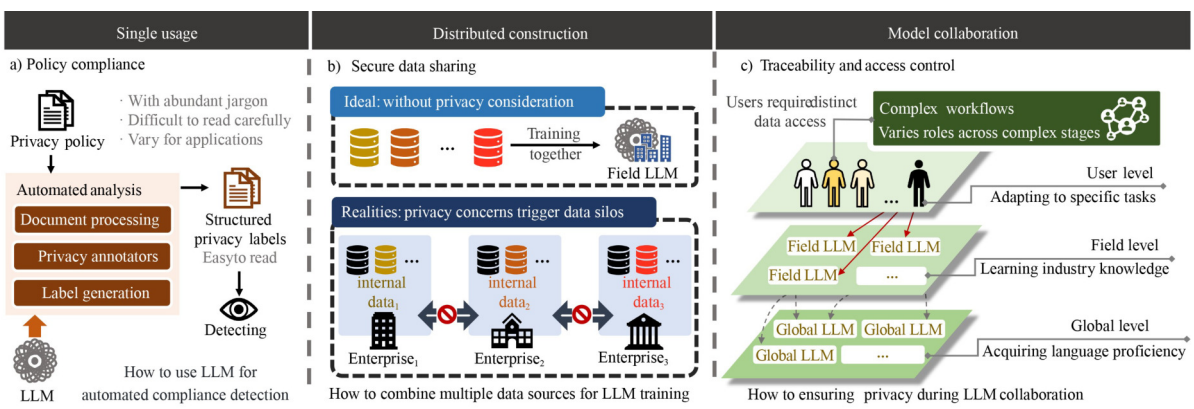

Fig. 3 LLM implementation modes and privacy research focuses

privacy protection (Section 4.1). Furthermore, some studies concentrate on the secure data sharing that may arise during distributed construction (Section 4.2). Finally, in collaborative mode, we delve into knowledge traceability and access control (Section 4.3).

4.1 Privacy compliance

A series of regulations have been introduced to protect user privacy rights, where privacy policies serve as the primary means to inform users about the collection, storage, and usage of their data. Indeed, privacy policy texts are often lengthy and contain numerous technical and legal jargon. Manual reading and compliance detection based on privacy policies consume a considerable amount of time, and the rapid iteration of privacy policies makes practical implementation challenging. Therefore, the pressing challenge is how to automate the process to efficiently understand privacy policies and implement compliance detection, depicted in Fig. 3(a).

Lately, annotation [125,126] has been adopted as the main technique, aiming to utilize metadata to mark relevant portions of privacy policies or unravel information relationships within the policies. Notably, LLMs have been introduced for privacy policy analysis. PolicyGPT [35] employs predefined segments and categories, querying GPT for classification to achieve automated classification. Furthermore, a study [36] utilizes LLMs for segmenting policy fragments, improving segmentation effectiveness by 13.1% in the application, albeit with a corresponding fivefold increase in computational costs. Additionally, work [37] proposes using LLM to execute automated GKC-CI parameter annotations for privacy policy analysis.

Analyzing reveals that current research typically partitions policies based on segment level with a predefined labeling structure. However, (1) current technologies exhibit biases in analyzing privacy policies. Segment-level labels may not allow for a complete analysis of privacy policies, as the information may be mentioned across segments. Besides, the analysis granularity may not align with practical requirements. PoliGraph [126] suggests using knowledge graphs to describe privacy policy statements as relationships between fine-grained texts. However, knowledge graphs face challenges with unseen knowledge. It is worth investigating whether the LLM robust generalization can effectively address privacy relation descriptions. Additionally, (2) current privacy labels lack widely recognized, formal standards, where individual software applications employ different classification methods for the information involved. While LLM excels in zero-shot settings, more investigation is required to assess its efficacy in handling unbalanced data distribution of privacy labels. (3) There is a lack of compliant automated methods. While existing research has concentrated on privacy policy analysis, a critical link to compliance detection is yet to be established.

4.2 Secure data sharing

Model scale and training data size are crucial factors contributing to the LLM effectiveness [21,40]. General corpora lack the expertise needed for specific industry requirements, and high-quality datasets are dispersed among enterprises, isolated due to business competition and privacy concerns, as described in Fig. 3(b). Additionally, small and medium-sized enterprises often struggle with large-scale model training costs due to limited computational resources and inadequate training data. While combining data from multiple sources holds potential benefits for LLM construction, current regulations restrict direct data sharing among isolated entities.

Dramatically, federated learning is a crucial technology for addressing data sharing. Recently, some efforts [127,128] have combined federated learning with LLM to achieve secure training data sharing, deploying local LLMs on clients during pre-training and fine-tuning to interact with the aggregation server for parameters. However, the distributed and multi-stage training nature of federated learning increases the risk of data poisoning attacks and parameter leakage. In addition, the memory in federated learning with LLM is more susceptible to privacy violations [129]. Researcher [130] designs a distributed framework based on parameter-efficient fine-tuning to enhance privacy. This framework keeps the LLM backbone on the server while delivering compressed adapter modules and simulators to clients, allowing for partial parameter sharing rather than sharing all parameters. However, the performance significantly degrades due to the lossy nature of the simulators. Other research [131] integrates defenses such as secure aggregation, differential privacy, and multi-party secure computation into the design framework, but the effectiveness of their privacy protection has not been fully evaluated.

Through research exploration, federated learning provides new ideas for sharing data security in the LLM construction, but federated learning with LLMs are still in the technical exploration stage. First, defensive techniques for LLMs cannot be directly applied to federated learning scenarios. Data pre-processing techniques require access to local user data to filter sensitive information, increasing the protection complexity. Meanwhile, adversarial training requires substantial resources, which may be challenging for lightweight users. Second, LLM performance decreases significantly during partial model access. Since pre-trained LLMs have some proprietary value and may not belong to the client, it is necessary to allow the client to perform federal fine-tuning without accessing the full model. Research [128] shows that when only 50% of the model is accessed, LLM retains little generation and inference capabilities. The defense mechanism for LLMs based on federated learning and the trade-off between model access control and model performance requires further in-depth research.

4.3 Privacy traceability and access control

LLM layering can be divided into two aspects: vertical stratification and horizontal collaboration, as shown in Fig. 3(c). For vertical stratification, LLMs are often categorized into three levels: global, field, and user [132], based on application purposes. The global level is tasked with training general corpus to serve as the foundation. The field level incorporates industry knowledge based on the global LLM to obtain a field LLM [23,24]. The user level involves downstream task adaptation according to specific domain

requirements, resulting in user LLM. From a horizontal perspective, some studies [133,134] focus on collaborative LLMs, i.e., where multiple AI systems collaborate to solve complex workflows through negotiation and debate. Additionally, research [135] finds that collaborative work effectively enhances LLM compliance with facts and improves decision-making capabilities.

However, LLM hierarchical collaboration faces some privacy challenges. Firstly, there is the issue of external knowledge traceability. Recent work [136] integrates LLM with external knowledge links to improve real-time capabilities and introduce domain-specific knowledge. Researchers [137] discover that retrieval-augmented generation, while somewhat effective in reducing LLM training data leakage, presents a privacy leakage risk for the data it retrieves. As discussed in Section 6, the traceability and integrity of external knowledge in the generated content need to be verified. Secondly, different users should have different access to data. Access control in model collaboration involves two typical issues. For inter-model invocation, different LLMs possess distinct knowledge. When invoked, they encounter issues regarding mutual access permissions and knowledge sharing between LLMs. For multi-user utilization of the same model, LLMs contain global data. However, due to varying access permissions among users, the generated content should be distinguished accordingly. Though prior work [138] proposes access control instructions using self-moderation techniques in LLMs to allow users selective information output, it exhibits bias across different groups and is easy to bypass. Besides, we find limited public research on LLM access control. In hierarchical collaboration, it becomes even more complex as the final generated content involves the fusion of multiple LLMs. Currently, research on hierarchical design, knowledge and model fusion is still ongoing, and privacy issues such as external knowledge security and access control are worthy of attention.

In brief, we observe that the majority of existing privacy-related research in LLM application is still in the conceptual and preliminary experimental stages, indicating a considerable distance from practical application.

5 Analysis and outlooks

In this section, we analyze technology maturity and explore field trends (Section 5.1). Moreover, we explore further research directions (Section 5.2) and present unique insights into LLM native security mechanisms (Section 5.3), offering valuable references for LLM privacy development.

5.1 Technology maturity analysis

We believe that the current LLM privacy research is still in a technical exploration phase and there is a certain gap from practical application. As shown in Fig. 4, we conduct a classification and statistical analysis of the surveyed papers based on privacy-related issues which are categorized into three stages depending on the methods, experiments, and deployment. For technical research, researchers propose techniques and use publicly available experimental datasets. The potential applications stage indicates studies testing on

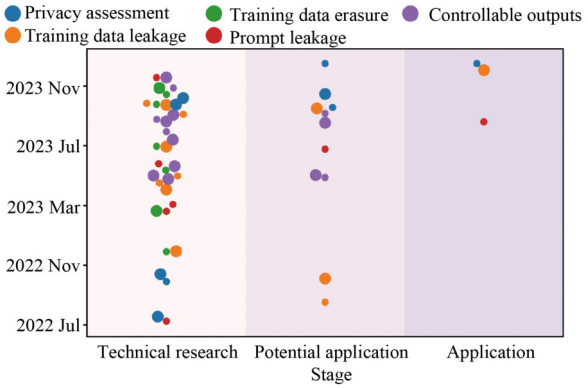

Fig. 4 Distribution of surveyed research. The abscissa represents the research stages, the ordinate represents the research publication time, and the point size represents attention

simulated data, demonstrating certain potential practical value, and the application stage shows that the research is practically deployed in commercial products. We find that the surveyed papers are mainly in the technical research stage, with relatively fewer in the potential and practical application stage, which shows somewhat that LLM privacy research is in the exploratory stage. Additionally, we notice a gradual increase in research during the potential applications stage, indicating a trend of transitioning from technical exploration to practical applications in LLM privacy research. Furthermore, research on training data leakage, controllable outputs, and privacy assessment has garnered widespread attention. In contrast, training data erasure is mainly the focus of the academic community, which may be due to the complex technique still far away from application.

5.2 Further research directions

From our collection and analysis, we propose potential directions for further research in LLM privacy:

Diverse scalability. Sensitive information manifests differently in various scenarios. It is necessary to effectively protect different granularities of sensitive information in different model stages. Current research [139,140] primarily focus on protecting well-defined sensitive information formats, as described in Section 3.2, Section 3.4, and Section 3.5. Future research could extend to methods for multi-form sensitive information protection, multi-granularity privacy assessment, and multi-dimensional task prompt protection.

Cross-domain generalizability. The current studies [5,38,86] are primarily applicable to single-language scenarios and mainly focuses on PII during experiments, with limited involvement of real-world domain datasets. Future research could explore the applicability and generalizability of corresponding methods in multilingual and multidomain datasets.

Privacy-preserving robustness. Existing research [46,47] concentrate on the privacy-utility tradeoff. However, LLMs exhibit issues such as bias and hallucination. Future investigations can further analyze how privacy-preserving techniques influent LLM's fairness, transparency, and hallucination. Additionally, exploring how to design effective privacy-preserving methods for emerging threats while maintaining model quality is crucial for building secure

LLMs.

Collaborative security. LLM training involves a multi-party data-sharing phenomenon, and the demand for collaboration between LLMs and external entities, such as knowledge bases, is on the rise for accomplishing complex workflows. Hence, protecting privacy during collaboration becomes increasingly crucial. The security LLM collaboration research is in the early exploration stage (as shown in Section 4.3). Future studies can further focus on access management, external knowledge, and generated content traceability, as well as privacy-preserving techniques for collaboration.

Multimodal impact. Multimodal large language models (MLLMs) extend LLMs by incorporating the ability to process and reason across multiple data modalities. However, current privacy research in this area remains sparse [141,142]. In addition to the known LLM privacy issues, the introduction of multimodal data may disrupt alignment strategies within LLMs, potentially increasing the risk of sensitive information leakage. Therefore, further exploration is needed to assess the broader implications of multimodal inputs on the privacy framework within LLMs and to develop comprehensive solutions.

5.3 Native security insights

Beyond the in-depth exploration covered in current research directions, we propose several noteworthy prospects for LLM native security mechanisms.

Privacy boundary. Existing privacy boundary based on datasets only focus on well-defined sensitive information. In addition, current privacy labels lack widely accepted and formal standards, introducing biases due to labeling inconsistencies in practical applications. While our work focuses on privacy discussions about sensitive textual information, privacy varies across domains. Future research should explore methodologies and regulatory frameworks to precisely define the LLM privacy boundaries, providing support for subsequent privacy policy formulation and the implementation of privacy protection.

Data security isolation mechanism. Bengio [143] presents a forward-looking perspective on decoupling knowledge and inference machines, advocating for the independent validation and maintenance of knowledge storage to enhance the verifiability and reliability of model knowledge. The adoption of layered model designs has demonstrated significant advantages in swiftly adapting to the dynamic demands of various industries. Traditionally, data security isolation mechanisms together with external protection strategies provide comprehensive privacy preservation. However, there is a noticeable absence of publicly available research on reliable isolation mechanisms for LLMs. Combined with the previous analysis, we believe that effectively isolating data and distilling sensitive information during LLM training, invocation, and application is an urgent and underexplored area that demands attention, guiding the direction in LLM privacy focuses such as access control and privacy traceability.

Self-verification measures and evaluation. Despite recent assurances from major LLM providers that unauthorized customer data is no longer used for training, the LLM's inherent black-box nature and closed architecture make it difficult for the public to verify the statement's authenticity. Hence, there is a pressing need for effective self-attestation methods and user-centric evaluation approaches, which are imperative for discerning whether user data has been employed in model training and for validating the compliance of the training data. To enhance transparency, effective self-verification methods and user-based assessments are urgently needed. Additionally, there is a pressing need for tools facilitating third-party assessment, auditing, and regulatory supervision to detect any unauthorized use of user data in LLM training and to validate the compliance of training data.

6 Conclusion

We extensively survey LLM privacy and review corresponding solutions. Specifically, we highlight five privacy issues during LLMs' training and invocation, providing comprehensive insights into their current state and potential advancements. Additionally, we introduce and analyze three privacy-centric research focuses in LLM application. Finally, we discuss further research directions and provide insights into LLM native security mechanisms with our view: LLM privacy research is in the technical exploration phase. While we conduct a comprehensive investigation into the LLM privacy, we acknowledge potential omissions due to rapid updates in related research. Therefore, we commit to ongoing monitoring of new research and continuous refinement of our work. We hope this paper provides researchers and practitioners with a comprehensive understanding to better address the privacy challenges that LLMs may encounter in real-world applications.

Acknowledgements This work was supported by the National Key R&D Program of China (2021YFC3300600).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Mitchell R. Samsung fab data leak: how ChatGPT exposed sensitive information. See electropages. com/blog/2023/04/how-chatgpt-exposed-sensitive-information website, 2023
2. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to THreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access, 2023, 11: 80218−80245
3. Nasr M, Carlini N, Hayase J, Jagielski M, Cooper A F, Ippolito D, Choquette-Choo C A, Wallace E, Tramèr F, Lee K. Scalable extraction of training data from (production) language models. 2023, arXiv preprint arXiv: 2311.17035
4. Coles C. 11% of data employees paste into ChatGPT is confidential. Cyberhaven. See cyberhaven. com/blog/4–2-of-workers-have-pasted-company-data-into-chatgpt website, 2023
5. Staab R, Vero M, Balunovic M, Vechev M. Beyond memorization: violating privacy via inference with large language models. In: Proceedings of the 12th International Conference on Learning Representations. 2024
6. Mehnaz S, Dibbo S V, Kabir E, Li N, Bertino E. Are your sensitive attributes private? Novel model inversion attribute inference attacks on

classification models. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). 2022, 4579−4596

7. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T B, Song D, Erlingsson Ú, Oprea A, Raffel C. Extracting training data from large language models. In: Bailey M D, Greenstadt R, eds. 30th USENIX Security Symposium. Berkeley: USENIX Association, 2021, 2633−2650
8. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. 2023, arXiv preprint arXiv: 2311.05232
9. Hartmann V, Suri A, Bindschaedler V, Evans D, Tople S, West R. SoK: memorization in general-purpose large language models. 2023, arXiv preprint arXiv: 2310.18362
10. Wang Y, Pan Y, Yan M, Su Z, Luan T H. A survey on ChatGPT: AI-generated contents, challenges, and solutions. IEEE Open Journal of the Computer Society, 2023, 4: 280−302
11. Kshetri N. Cybercrime and privacy threats of large language models. IT Professional, 2023, 25(3): 9−13
12. Wang T, Zhang Y, Qi S, Zhao R, Xia Z, Weng J. Security and privacy on generative data in AIGC: a survey. ACM Computing Surveys, 2024
13. Mozes M, He X, Kleinberg B, Griffin L D. Use of LLMs for illicit purposes: threats, prevention measures, and vulnerabilities. 2023, arXiv preprint arXiv: 2308.12833
14. Smith V, Shamsabadi A S, Ashurst C, Weller A. Identifying and mitigating privacy risks stemming from language models: a survey. 2023, arXiv preprint arXiv: 2310.01424
15. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confidence Computing, 2024, 4(2): 100211
16. Iqbal U, Kohno T, Roesner F. LLM platform security: applying a systematic evaluation framework to OpenAI's ChatGPT plugins. In: Proceedings of the 7th AAAI/ACM Conference on AI, Ethics, and Society. 2024, 611−623
17. Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What does it mean for a language model to preserve privacy? In: Proceedings of 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022, 2280−2292
18. Zhang D, Finckenberg-Broman P, Hoang T, Pan S, Xing Z, Staples M, Xu X. Right to be forgotten in the era of large language models: implications, challenges, and solutions. AI and Ethics, 20241−10
19. Neel S, Chang P. Privacy issues in large language models: a survey. 2023, arXiv preprint arXiv: 2312.06717
20. Das B C, Amini M H, Wu Y. Security and privacy challenges of large language models: a survey. 2024, arXiv preprint arXiv: 2402.00888
21. Ye J, Chen X, Xu N, Zu C, Shao Z, Liu S, Cui Y, Zhou Z, Gong C, Shen Y, Zhou J, Chen S, Gui T, Zhang Q, Huang X. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. 2023, arXiv preprint arXiv: 2303.10420
22. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks L A, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Vinyals O, Rae J W, Sifre L. An empirical analysis of compute-optimal large language model training. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 30016−30030
23. Yang H, Liu X, Wang C D. FinGPT: open-source financial large language models. 2023, arXiv preprint arXiv: 2306.06031
24. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. 2023, arXiv preprint arXiv: 2303.14070
25. Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations. 2022
26. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 3045−3059
27. Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. In: Proceedings of the 38th International Conference on Machine Learning. 2021, 12697−12706
28. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe L. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 27730−27744
29. Ullah I, Hassan N, Gill S S, Suleiman B, Ahanger T A, Shah Z, Qadir J, Kanhere S S. Privacy preserving large language models: ChatGPT case study based vision and framework. 2023, arXiv preprint arXiv: 2310.12523
30. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. 2023, arXiv preprint arXiv: 2303.08774
31. Jayaraman B, Ghosh E, Chase M, Roy S, Dai W, Evans D. Combing for credentials: active pattern extraction from smart reply. In: Proceedings of 2024 IEEE Symposium on Security and Privacy (SP). 2023, 1443−1461
32. Mireshghallah F, Goyal K, Uniyal A, Berg-Kirkpatrick T, Shokri R. Quantifying privacy risks of masked language models using membership inference attacks. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 8332−8347
33. Mouhammad N, Daxenberger J, Schiller B, Habernal I. Crowdsourcing on sensitive data with privacy-preserving text rewriting. In: Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII). 2023, 73−84
34. Tang X, Shin R, Inan H A, Manoel A, Mireshghallah F, Lin Z, Gopi S, Kulkarni J, Sim R. Privacy-preserving in-context learning with differentially private few-shot generation. In: Proceedings of the 12th International Conference on Learning Representations. 2024
35. Tang C, Liu Z, Ma C, Wu Z, Li Y, Liu W, Zhu D, Li Q, Li X, Liu T, Fan L. PolicyGPT: automated analysis of privacy policies with large language models. 2023, arXiv preprint arXiv: 2309.10238
36. Pan S, Hoang T, Zhang D, Xing Z, Xu X, Lu Q, Staples M. Toward the cure of privacy policy reading phobia: automated generation of privacy nutrition labels from privacy policies. 2023, arXiv preprint arXiv: 2306.10923
37. Chanenson J, Pickering M, Apthorpe N. Automating governing knowledge commons and contextual integrity (GKC-CI) privacy policy annotations with large language models. 2023, arXiv preprint arXiv: 2311.02192
38. Lukas N, Salem A, Sim R, Tople S, Wutschitz L, Zanella-Béguelin S. Analyzing leakage of personally identifiable information in language models. In: Proceedings of 2013 IEEE Symposium on Security and Privacy (SP). 2023, 346−363
39. Chen X, Tang S, Zhu R, Yan S, Jin L, Wang Z, Su L, Zhang Z, Wang X, Tang H. The Janus interface: how fine-tuning in large language models amplifies the privacy risks. 2023, arXiv preprint arXiv: 2310.15469
40. Huang J, Shao H, Chang K C C. Are large pre-trained language models leaking your personal information? In: Proceedings of the Findings of

the Association for Computational Linguistics: EMNLP 2022. 2022, 2038−2047

41. Kandpal N, Wallace E, Raffel C. Deduplicating training data mitigates privacy risks in language models. In: Proceedings of the 39th International Conference on Machine Learning. 2022, 10697−10707
42. Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, Carlini N. Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022, 8424−8445
43. Zhang C, Ippolito D, Lee K, Jagielski M, Tramèr F, Carlini N. Counterfactual memorization in neural language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 39321−39362
44. Yue X, Inan H A, Li X, Kumar G, McAnallen J, Shajari H, Sun H, Levitan D, Sim R. Synthetic text generation with differential privacy: a simple and practical recipe. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 1321−1342
45. Majmudar J, Dupuy C, Peris C, Smaili S, Gupta R, Zemel R. Differentially private decoding in large language models. 2022, arXiv preprint arXiv: 2205.13621
46. Xiao Y, Jin Y, Bai Y, Wu Y, Yang X, Luo X, Yu W, Zhao X, Liu Y, Gu Q, Chen H, Wang W, Cheng W. PrivacyMind: large language models can be contextual privacy protection learners. 2023, arXiv preprint arXiv: 2310.02469
47. Behnia R, Ebrahimi M R, Pacheco J, Padmanabhan B. EW-Tune: a framework for privately fine-tuning large language models with differential privacy. In: Proceedings of 2022 IEEE International Conference on Data Mining Workshops (ICDMW). 2022, 560−566
48. West P, Lu X, Holtzman A, Bhagavatula C, Hwang J D, Choi Y. Reflective decoding: beyond unidirectional generation with off-the-shelf language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021, 1435−1450
49. Gokaslan A, Cohen V. Openwebtext corpus. See Skylion007.github.io/OpenWebTextCorpus website, 2019
50. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 140
51. Beyer A, Kauermann G, Schütze H. Embedding space correlation as a measure of domain similarity. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020, 2431−2439
52. Chelba C, Mikolov T, Schuster M, Ge Q, Brants T, Koehn P, Robinson T. One billion word benchmark for measuring progress in statistical language modeling. 2013, arXiv preprint arXiv: 1312.3005
53. Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, Choi Y. Defending against neural fake news. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 812
54. Tur G, Hakkani-Tür D, Heck L. What is left to be understood in ATIS? In: Proceedings of 2010 IEEE Spoken Language Technology Workshop. 2010, 19−24
55. Coucke A, Saade A, Ball A, Bluche T, Caulier A, Leroy D, Doumouro C, Gisselbrecht T, Caltagirone F, Lavril T, Primet M, Dureau J. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. 2018, arXiv preprint arXiv: 1805.10190
56. Li J, Ott M, Cardie C. Identifying manipulated offerings on review portals. In: Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. 2013, 1933−1942
57. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S R. Glue: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2018, 353−355
58. Socher R, Perelygin A, Wu J, Chuang J, Manning C D, Ng A Y, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. 2013, 1631−1642
59. Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2017, 1112−1122
60. Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 4317−4323
61. Klimt B, Yang Y. The Enron corpus: a new dataset for email classification research. In: Proceedings of the 15th European Conference on Machine Learning. 2004, 217−226
62. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015, 649−657
63. Liu J, Cyphers S, Pasupat P, McGraw I, Glass J. A conversational movie search system based on conditional random fields. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association. 2012, 2454−2457
64. Villalobos P, Sevilla J, Heim L, Besiroglu T, Hobbhahn M, Ho A. Will we run out of data? An analysis of the Limits of scaling datasets in machine learning. 2022, arXiv preprint arXiv: 2211.04325
65. Chen J, Yang D. Unlearn what you want to forget: efficient unlearning for LLMs. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 12041−12052
66. Borkar J. What can we learn from data leakage and unlearning for law? 2023, arXiv preprint arXiv: 2307.10476
67. Hu Z, Zhang Y, Xiao M, Wang W, Feng F, He X. Exact and efficient unlearning for large language model-based recommendation. 2024, arXiv preprint arXiv: 2404.10327
68. Ziegler C N, McNee S M, Konstan J A, Lausen G. Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web. 2005, 22−32
69. Harper F M, Konstan J A. The MovieLens datasets: history and context. ACM Transactions on Interactive Intelligent Systems (TiiS), 2015, 5(4): 19
70. Jang J, Yoon D, Yang S, Cha S, Lee M, Logeswaran L, Seo M. Knowledge unlearning for mitigating privacy risks in language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 14389−14408
71. Wang L, Chen T, Yuan W, Zeng X, Wong K F, Yin H. KGA: a general machine unlearning framework based on knowledge gap alignment. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 13264−13276
72. Tuggener D, von Däniken P, Peetz T, Cieliebak M. LEDGAR: a large-scale multi-label corpus for text classification of legal provisions in contracts. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020, 1235−1241

73. Cettolo M, Niehues J, Stüker S, Bentivogli L, Federico M. Report on the 11th IWSLT evaluation campaign. In: Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign. 2014, 2−17
74. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 6th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018, 2204−2213
75. Maas A, Daly R E, Pham P T, Huang D, Ng A Y, Potts C. Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011, 142−150
76. Gliwa B, Mochol I, Biesek M, Wawer A. SAMSum corpus: a human-annotated dialogue dataset for abstractive summarization. In: Proceedings of the 2nd Workshop on New Frontiers in Summarization. 2019, 70−79
77. Zaman K, Choshen L, Srivastava S. Fuse to forget: bias reduction and selective memorization through model fusion. In: Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 18763−18783
78. Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In: Proceedings of the Working Notes Papers of the CLEF 2016 Evaluation Labs. 2016, 750−784
79. Pawelczyk M, Neel S, Lakkaraju H. In-context unlearning: language models as few-shot unlearners. In: Proceedings of the 41st International Conference on Machine Learning. 2024
80. Bourtoule L, Chandrasekaran V, Choquette-Choo C A, Jia H, Travers A, Zhang B, Lie D, Papernot N. Machine unlearning. In: Proceedings of 2021 IEEE Symposium on Security and Privacy (SP). 2021, 141−159
81. Koch K, Soll M. No matter how you slice it: machine unlearning with SISA comes at the expense of minority classes. In: Proceedings of 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). 2023, 622−637
82. Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021, 4582−4597
83. Wen R, Wang T, Backes M, Zhang Y, Salem A. Last one standing: a comparative analysis of security and privacy of soft prompt tuning, LoRA, and in-context learning. 2023, arXiv preprint arXiv: 2310.11397
84. Liu R, Wang T, Cao Y, Xiong L. PreCurious: how innocent pre-trained language models turn into privacy traps. 2024, arXiv preprint arXiv: 2403.09562
85. Qi X, Zeng Y, Xie T, Chen P Y, Jia R, Mittal P, Henderson P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In: Proceedings of the 12th International Conference on Learning Representations. 2024
86. Li H, Guo D, Fan W, Xu M, Huang J, Meng F, Song Y. Multi-step jailbreaking privacy attacks on ChatGPT. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, 4138−4153
87. Deng G, Liu Y, Li Y, Wang K, Zhang Y, Li Z, Wang H, Zhang T, Liu Y. MASTERKEY: automated jailbreaking of large language model Chatbots. In: Proc. ISOC NDSS. 2024
88. Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Conference on Security Symposium. 2019, 267−284
89. Ai4Privacy. PII Masking 300k Dataset. See huggingface.co/datasets/ai4privacy/pii-masking-300k website, 2024
90. Merity S, Xiong C, Bradbury J, Socher R. Pointer sentinel mixture models. In: Proceedings of the 5th International Conference on Learning Representations. 2016
91. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of neural network representations revisited. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 3519−3529
92. Zhang Y, Ippolito D. Prompts should not be seen as secrets: systematically measuring prompt extraction attack success. 2023, arXiv preprint arXiv: 2307.06865
93. Duan H, Dziedzic A, Papernot N, Boenisch F. Flocks of stochastic parrots: differentially private prompt learning for large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024, 3358
94. Zhang X, Chen C, Xie Y, Chen X, Zhang J, Xiang Y. A survey on privacy inference attacks and defenses in cloud-based deep neural network. Computer Standards & Interfaces, 2023, 83: 103672
95. Plant R, Gkatzia D, Giuffrida V. CAPE: Context-Aware Private Embeddings for private language learning. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 7970−7978
96. Li Y, Tan Z, Liu Y. Privacy-preserving prompt tuning for large language model services. 2023, arXiv preprint arXiv: 2305.06212
97. Feyisetan O, Balle B, Drake T, Diethe T. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In: Proceedings of the 13th International Conference on Web Search and Data Mining. 2020, 178−186
98. Zhang Z, Zhang X, Xie W, Lu Y. Responsible task automation: empowering large language models as responsible task Automators. 2023, arXiv preprint arXiv: 2306.01242
99. Ribeiro B, Rolla V, Santos R. INCOGNITUS: a toolbox for automated clinical notes anonymization. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2023, 187−194
100. Chen Y, Li T, Liu H, Yu Y. Hide and seek (HaS): a lightweight framework for prompt privacy protection. 2023, arXiv preprint arXiv: 2309.03057
101. Kan Z, Qiao L, Yu H, Peng L, Gao Y, Li D. Protecting user privacy in remote conversational systems: a privacy-preserving framework based on text sanitization. 2023, arXiv preprint arXiv: 2306.08223
102. Goyal S, Choudhury A R, Raje S, Chakaravarthy V, Sabharwal Y, Verma A. PoWER-BERT: accelerating BERT inference via progressive word-vector elimination. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 346
103. Modarressi A, Mohebbi H, Pilehvar M T. AdapLeR: speeding up inference by adaptive length reduction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022, 1−15
104. Zhou X, Lu J, Gui T, Ma R, Fei Z, Wang Y, Ding Y, Cheung Y, Zhang Q, Huang X J. TextFusion: privacy-preserving pre-trained model inference via token fusion. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 8360−8371
105. Beckerich M, Plein L, Coronado S. RatGPT: Turning online LLMs into proxies for malware attacks. 2023, arXiv preprint arXiv: 2308.09183
106. Li Y, Wei F, Zhao J, Zhang C, Zhang H. Rain: your language models can align themselves without finetuning. 2023, arXiv preprint arXiv: 2309.07124
107. Pisano M, Ly P, Sanders A, Yao B, Wang D, Strzalkowski T, Si M. Bergeron: combating adversarial attacks through a conscience-based

alignment framework. 2023, arXiv preprint arXiv: 2312.00029
108. Cao B, Cao Y, Lin L, Chen J. Defending against alignment-breaking attacks via robustly aligned LLM. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 10542−10560
109. Inan H, Upasani K, Chi J, Rungta R, Iyer K, Mao Y, Tontchev M, Hu Q, Fuller B, Testuggine D, Testuggine M. Llama guard: LLM-based input-output safeguard for human-AI conversations. 2023, arXiv preprint arXiv: 2312.06674
110. Meta. Llm-guard. See llm-guard.com website, 2024
111. Microsoft. Presidio analyzer. See pypi. org/project/presidio-analyzer/ website, 2024
112. Xu Z, Liu Y, Deng G, Li Y, Picek S. LLM jailbreak attack versus defense techniques–a comprehensive study. 2024, arXiv preprint arXiv: 2402.13457
113. Yang X, Chen K, Zhang W, Liu C, Qi Y, Zhang J, Fang H, Yu N. Watermarking text generated by black-box language models. 2023, arXiv preprint arXiv: 2305.08883
114. Tu S, Sun Y, Bai Y, Yu J, Hou L, Li J. WaterBench: towards holistic evaluation of watermarks for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, 1517−1542
115. Singh K, Zou J. New evaluation metrics capture quality degradation due to LLM watermarking. 2023, arXiv preprint arXiv: 2312.02382
116. Li Y, Li Q, Cui L, Bi W, Wang Z, Wang L, Yang L, Shi S, Zhang Y. MAGE: machine-generated text detection in the wild. 2023, arXiv preprint arXiv: 2305.13242
117. Antoun W, Mouilleron V, Sagot B, Seddah D. Towards a robust detection of language model generated text: is ChatGPT that easy to detect? 2023, arXiv preprint arXiv: 2306.05871
118. Chen Y, Kang H, Zhai V, Li L, Singh R, Ramakrishnan B. GPT-sentinel: distinguishing human and ChatGPT generated content. 2023, arXiv preprint arXiv: 2305.07969
119. Sarvazyan A M, González J Á, Rosso P, Franco-Salvador M. Supervised machine-generated text detectors: family and scale matters. In: Proceedings of the 14th International Conference of the CLEF Association on Experimental IR Meets Multilinguality, Multimodality, and Interaction. 2023, 121−132
120. Bhattacharjee A, Kumarage T, Moraffah R, Liu H. ConDA: contrastive domain adaptation for AI-generated text detection. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 598−610
121. Liu Y, Zhang Z, Zhang W, Yue S, Zhao X, Cheng X, Zhang Y, Hu H. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. 2023, arXiv preprint arXiv: 2304.07666
122. Bhattacharjee A, Liu H. Fighting fire with fire: can ChatGPT detect AI-generated text?. ACM SIGKDD Explorations Newsletter, 2023, 25(2): 14−21
123. Koike R, Kaneko M, Okazaki N. OUTFOX: LLM-generated essay detection through in-context learning with adversarially generated examples. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. 2024, 21258−21266
124. Yu X, Qi Y, Chen K, Chen G, Yang X, Zhu P, Zhang W, Yu N. GPT paternity test: GPT generated text detection with GPT genetic inheritance. 2023, arXiv preprint arXiv: 2305.12519
125. Qiu W, Lie D, Austin L. Calpric: inclusive and fine-grain labeling of privacy policies with crowdsourcing and active learning. In: Proceedings of the 32nd USENIX Conference on Security Symposium. 2023, 60
126. Cui H, Trimananda R, Markopoulou A, Jordan S. POLIGRAPH: automated privacy policy analysis using knowledge graphs. In: Proceedings of the 32nd USENIX Conference on Security Symposium. 2023, 59
127. Chen C, Feng X, Li Y, Lyu L, Zhou J, Zheng X, Yin J. Integration of large language models and federated learning. 2023, arXiv preprint arXiv: 2307.08925
128. Kuang W, Qian B, Li Z, Chen D, Gao D, Pan X, Xie Y, Li Y, Ding B, Zhou J. FederatedScope-LLM: a comprehensive package for fine-tuning large language models in federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024, 5260−5271
129. Gupta S, Huang Y, Zhong Z, Gao T, Li K, Chen D. Recovering private text in federated learning of language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 8130−8143
130. Xiao G, Lin J, Han S. Offsite-tuning: transfer learning without full model. 2023, arXiv preprint arXiv: 2302.04870
131. Fan T, Kang Y, Ma G, Chen W, Wei W, Fan L, Yang Q. Fate-LLM: a industrial grade federated learning framework for large language models. 2023, arXiv preprint arXiv: 2310.10049
132. Gong Y. Multilevel large language models for everyone. 2023, arXiv preprint arXiv: 2307.13221
133. Xi Z, Chen W, Guo X, He W, Ding Y, et al. The rise and potential of large language model based agents: a survey. 2023, arXiv preprint arXiv: 2309.07864
134. Wu Q, Bansal G, Zhang J, Wu Y, Zhang S, Zhu E, Li B, Jiang L, Zhang X, Wang C. AutoGen: enabling next-gen LLM applications via multi-agent conversation. 2023, arXiv preprint arXiv: 2308.08155
135. Zhang H, Du W, Shan J, Zhou Q, Du Y, Tenenbaum J B, Shu T, Gan C. Building cooperative embodied agents modularly with large language models. In: Proceedings of the 12th International Conference on Learning Representations. 2024
136. Liu X, Lai H, Yu H, Xu Y, Zeng A, Du Z, Zhang P, Dong Y, Tang J. WebGLM: towards an efficient web-enhanced question answering system with human preferences. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023, 4549−4560
137. Zeng S, Zhang J, He P, Liu Y, Xing Y, Xu H, Ren J, Chang Y, Wang S, Yin D, Tang J. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. 2024, 4505−4524
138. Chen Y, Mendes E, Das S, Xu W, Ritter A. Can language models be instructed to protect personal information? 2023, arXiv preprint arXiv: 2310.02224
139. Niu L, Mirza S, Maradni Z, Pöpper C. CodexLeaks: privacy leaks from code generation language models in GitHub copilot. In: Proceedings of the 32nd USENIX Conference on Security Symposium. 2023, 120
140. He X, Zannettou S, Shen Y, Zhang Y. You only prompt once: on the capabilities of prompt learning on large language models to tackle toxic content. In: Proceedings of 2024 IEEE Symposium on Security and Privacy (SP). 2024, 770−787
141. Samson L, Barazani N, Ghebreab S, Asano Y M. Privacy-aware visual language models. 2024, arXiv preprint arXiv: 2405.17423
142. Caldarella S, Mancini M, Ricci E, Aljundi R. The phantom menace: unmasking privacy leakages in vision-language models. 2024, arXiv preprint arXiv: 2408.01228
143. Bengio Y, Hu E J. Scaling in the service of reasoning & model-based ML. See yoshuabengio. org/2023/03/21/scaling-in-the-service-of-reasoning-model-based-ml/ website, 2023

Hongyi LI received the BE degree from Tianjin University, China in 2022. She is currently working toward the PhD degree with the School of Computer Science, Fudan University, China. Her current research interests include financial technology security and data privacy.

Jiawei YE is currently a faculty member with the School of Computer Science at Fudan University and a senior engineer. His primary research interests include network and information security, sensitive information protection.

Jie WU is currently a professor in the School of Computer Science, Fudan University, China. His main research interests include network multimedia and information security.