A Survey on Data Security in Large Language Models

Kang Chen[a,b,1], Xiuze Zhou[c,1], Yuanguo Lin[a,*], Jinhe Su[a], Yuanhui Yu[a,*], Li Shen[d] and Fan Lin[e]

[a] School of Computer Engineering, Jimei University, Xiamen, 361021, China
[b] College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, 325060, China
[c] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China
[d] School of Professional Studies, New York University, New York, 10003, United States
[e] School of Informatics, Xiamen University, Xiamen, 361102, China

ARTICLE INFO

Keywords:
Large language model (LLM)
Data security
LLM vulnerabilities
Prompt injection

ABSTRACT

Large Language Models (LLMs), now a foundation in advancing natural language processing, power applications such as text generation, machine translation, and conversational systems. Despite their transformative potential, these models inherently rely on massive amounts of training data, often collected from diverse and uncurated sources, which exposes them to serious data security risks. Harmful or malicious data can compromise model behavior, leading to issues such as toxic output, hallucinations, and vulnerabilities to threats such as prompt injection or data poisoning. As LLMs continue to be integrated into critical real-world systems, understanding and addressing these data-centric security risks is imperative to safeguard user trust and system reliability. This survey offers a comprehensive overview of the main data security risks facing LLMs and reviews current defense strategies, including adversarial training, RLHF, and data augmentation. Additionally, we categorize and analyze relevant datasets used for assessing robustness and security across different domains, providing guidance for future research. Finally, we highlight key research directions that focus on secure model updates, explainability-driven defenses, and effective governance frameworks, aiming to promote the safe and responsible development of LLM technology. This work aims to inform researchers, practitioners, and policymakers, driving progress toward data security in LLMs.

1. Introduction

Large Language Models (LLMs), which exhibit near-human performance on tasks ranging from free-form text generation and summarization to machine translation and open-domain question answering, represent a transformative leap in natural language processing. The ability of LLMs to model complex linguistic dependencies and generate coherent, context-aware outputs has resulted in widespread adoption in both academic research and industrial applications, fueling speculation about their role as precursors to Artificial General Intelligence (AGI). This surge in capability underscores the significance of LLMs, not only as powerful computational tools, but also as foundational building blocks for next-generation AI systems. Also, it has become regarded as an excellent contextual learner [18]. The extensive use of LLMs marks the beginning of a new paradigm in seamless knowledge transfer for diverse natural language processing applications [53].

Despite their remarkable strengths, LLMs are beset by a variety of security and privacy vulnerabilities that threaten both model integrity and user confidentiality. Given their dependence on massive training datasets, these models are susceptible to malicious or biased information, which can result in the generation of inaccurate or inappropriate content. This raises serious concerns about potential negative impacts, such as the spread of false information and the reinforcement of harmful stereotypes. By manipulating public opinion, fostering confusion, and advancing detrimental ideologies, the intentional dissemination of misinformation may cause substantial societal harm [50]. Threats, such as jailbreaking, in which adversaries circumvent safety filters via crafted prompts; data poisoning, which injects malicious samples into training corpora; and inadvertent leakage of personally identifiable information (PII) all illustrate the dual-edged nature of web-scale data ingestion. These threats can manifest at multiple stages in the LLM lifecycle, thereby compromising model outputs, undermining trust, and exposing sensitive data. Moreover, the lack of transparency in training data provenance further exacerbates these risks. Studies have shown that even small amounts of toxic, biased, or copyrighted content in a training set can disproportionately affect model behavior [9]. With the ever-widening scale of LLMs, ensuring dataset integrity becomes increasingly critical - not only to prevent harmful generations but also to uphold legal and ethical standards. Recent work highlights the urgency of constructing curated and auditable training corpora to mitigate these issues [3]. Without such safeguards, LLMs remain susceptible to data-centric threats, which can subtly or overtly distort their outputs.

To address these concerns, a range of protective methods has been developed. These methods assist legal professionals in navigating increasingly complex data protection regulations and enhance their comprehension of compliance requirements related to data processing and storage. Key data security protection methods include adversarial training [44], Reinforcement Learning from Human Feedback (RLHF) [49], and data augmentation techniques [20]. These

[*] Corresponding authors

chenkang@kean.edu (K. Chen); xz.zhou@connect.hkust-gz.edu.cn (X. Zhou); xdlyg@jmu.edu.cn (Y. Lin); sujh@jmu.edu.cn (J. Su); andy@jmu.edu.cn (Y. Yu); ls6743@nyu.edu (L. Shen); iamafan@xmu.edu.cn (F. Lin)

ORCID(s): 0000-0002-0717-6936 (X. Zhou)

[1] Co-first authors

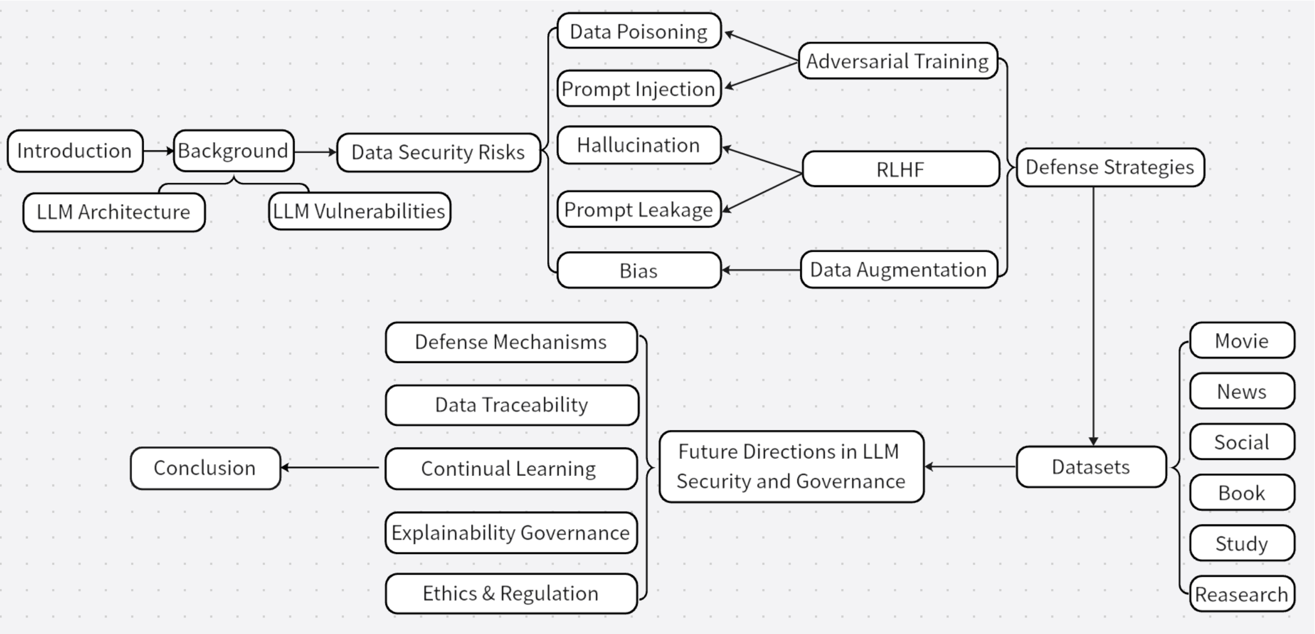

Fig. 1: Overview of the Survey Structure on LLMs Data Security, beginning with background and LLM vulnerabilities, then addressing data security risks, mitigation techniques, datasets, and concluding with future directions in LLM security and governance.

approaches contribute to secure and stable model outputs by improving model robustness, incorporating human-aligned reinforcement signals, and enhancing dataset diversity [20, 24, 49]. Recent research increasingly highlights the data security risks associated with training large language models, particularly their vulnerability to training-time data poisoning. It has been shown that even a small fraction of corrupted training data can significantly undermine model behavior [62]. To counter such threats, researchers have proposed robust training frameworks that reduce the impact of manipulated data, aiming to preserve model reliability throughout the learning process [27]. These findings collectively reinforce the importance of embedding security considerations into the entire lifecycle of large language model development.

Several prior surveys partially explored aspects of data security in Large Language Models (LLMs), but often with a narrower scope or focus. Some studies have explored adversarial threats in NLP, offering extensive taxonomies of input-level perturbations and their defenses, yet often neglecting LLM-specific concerns such as prompt leakage or cross-phase data poisoning [79]. Others emphasize robustness and safety alignment - primarily from a model behavior or RLHF perspective - without systematically addressing how data threats propagate during training and inference [22]. In addition, surveys on backdoor learning provide valuable overviews of poisoning and trigger-based threats, but their focus remains on traditional classification models rather than generative, prompt-driven architectures like LLMs [38]. These gaps underscore the need for a comprehensive, LLM-specific synthesis that maps data threats across the entire pipeline - precisely the objective of our study.

Our motivation stems from this gap: existing literature lacks a comprehensive survey that rigorously categorizes the unique data security risks of modern LLMs and assesses defense effectiveness across both training and deployment phases. As LLMs grow in scale and diversify into critical sectors - such as finance, healthcare, and transportation - the stakes of poorly understood vulnerabilities become ever higher, demanding an up-to-the-minute synthesis of threats and protections.

Accordingly, our contributions are threefold. (1) We present a detailed taxonomy of key data security risks to LLMs, systematically characterizing each threat - such as data poisoning and prompt injection - in terms of its goals, attack strategies, and potential consequences. (2) We survey the landscape of existing defense mechanisms, evaluating their strengths and limitations in the face of evolving threats. (3) We identify key research gaps and propose future directions, including the development of standardized evaluation metrics, XAI-driven vulnerability analysis, and real-time monitoring frameworks.

The remainder of this paper is organized as follows: Section 2 provides background on LLM architectures and vulnerabilities. Section 3 delves into data security risks in detail. Section 4 reviews defense strategies and assesses their efficacy. Section 5 examines datasets for studying data security in LLMs. Section 6 discusses current research limitations and outlines promising avenues for future work. Finally, Section 7 concludes the survey. As illustrated in Fig. 1, the overall structure of the paper follows a logical flow from foundational concepts to risks, defenses, datasets, and future directions in LLM security and governance.

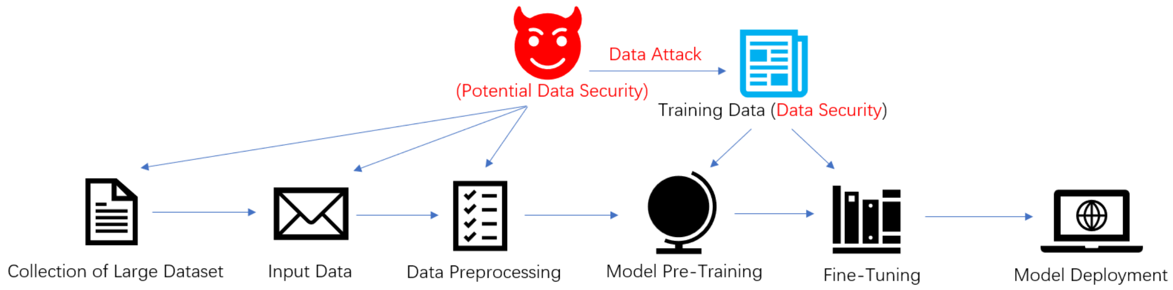

Fig. 2: Data training with an LLM platform. The workflow highlights critical machine learning development phases vulnerable to data security risks: training data collection, input processing, model pre-training, fine-tuning, and deployment. Each stage presents unique threat surfaces requiring specific protection measures.

2. Background

2.1. LLM Architecture

Like deep learning-based NLP systems, LLMs follow a data-centric pipeline that transforms raw textual data into coherent and informative responses [36]. Aiming to gain diverse language patterns, the process begins with large-scale data collection, often from web corpora, code repositories, and public datasets [6, 16]. After collection, to remove noise and standardize input formats, the data is preprocessed, including deduplication, tokenization, and quality filtering [13, 55]. In the pre-training stage, LLMs learn general language representations from this cleaned data, thereby enabling broad language understanding. This is followed by fine-tuning, where models are adapted to specific downstream tasks using more targeted datasets [49]. Ultimately, a model deployment phase integrates the trained and fine-tuned models into practical applications. Fig. 2 shows such a process.

However, LLMs are inherently vulnerable to data-centric security threats. Adversaries can strategically manipulate training or fine-tuning data to inject malicious behavior, causing a model to behave abnormally under specific triggers - a phenomenon often referred to as neural backdoors or behavioral steering [43]. Such threats do not merely decrease performance, but compromise the semantic alignment of the model with intended tasks, undermining trust in real-world deployment. More subtly, even small-scaleimperceptible perturbations in training data can accumulate and shift the decision boundaries of large models in unexpected ways [67], highlighting the fragility of current data pipelines in adversarial settings.

While training-time data poisoning is often considered the primary threat to LLMs, vulnerabilities can also arise during the data input and preprocessing stages. If input data is collected from open sources, adversaries may inject subtly crafted malicious content that evades detection, yet influences model behavior during training. Data preprocessing, intended to clean and filter, often fails to eliminate adversarial samples that are skillfully obfuscated. These weaknesses in the early data pipeline can plant "logic bombs" that remain dormant until triggered post-deployment. Wallace et al. [66] demonstrated that minimal "trigger phrases" in training data can induce biased or harmful outputs in large models. Similarly, Carlini et al. [10] identified preprocessing-stage vulnerabilities as critical failure points in current defenses, particularly due to insufficient filtering precision.

During user interaction with LLMs, users inputting some sensitive information as part of the prompts [33], marks the starting point of potential risks in data integrity, because it is the first stage in which user data is introduced. If the input data contains adversarial crafted content, it may lead to unexpected or unsafe model behavior. Furthermore, by injecting poisoned or manipulated data, adversaries can compromise the reliability of LLMs during the pre-training and fine-tuning stages, which may persist through subsequent training phases and trigger harmful outputs during inference [34]. Such manipulation not only threatens the security of a model, but may also magnify harmful stereotypes or social biases embedded in the corrupted data, thus destroying the trustworthiness and fairness of the responses of a model.

2.2. LLM Vulnerabilities

According to recent studies, data security vulnerabilities in LLMs are complex and multifaceted. Based on the nature of the threats, these vulnerabilities span various categories, including hallucination [71], bias [20], data poisoning, and prompt injection. The literature commonly classifies such threats using either a target-based or method-based taxonomy. In the context of LLMs, data security primarily involves defending models against malicious manipulations to ensure that generated content remains accurate, trustworthy, and free from unintended consequences. Addressing these threats is essential to maintain the integrity, fairness, and robustness of LLM outputs.

Our efforts are devoted to investigating the vulnerabilities of LLMs from the perspective of data security. We focus on critical threats such as data poisoning, hallucinations, biases, and prompt injection, all of which compromise the reliability and robustness of LLMs. Notably, we observe that various threat techniques share underlying strategies; for instance, both data poisoning and backdoor attacks manipulate model behavior by injecting malicious samples into the training process [15, 72]. These threats can significantly alter the output of LLMs, leading to unsafe or misleading responses and raising serious concerns about the integrity and

Table 1
Various studied risks on data security. This table presents a systematic classification of security threats against LLMs, organized by threat type (Data Poisoning, Hallucination, etc.), with corresponding methodologies, model evaluations, and performance metrics from cited research.

Category	Work	Method	Evaluated Model	Dataset	Evaluation Metric
Data Poisoning	[34]	Restricted Inner Product Poison Learning	BERT, XLNet	SST-2, OffensEval, etc	LFR, Clean Acc
	[39]	Model-Editing Techniques	GPT-2-XL, GPT-J	SST-2, AGNews, etc	ASR, CACC
	[67]	Polarity Poisoning	ChatGPT, FLAN, InstructGPT	SST-2, IMDb, Yelp, etc	SuSuper-NaturalInstructions
Prompt Injection	[42]	Component Generation	GPT3.5-turbo, etc	/	Vendor confirmation, etc
	[76]	Goal-guided generative Prompt injection strategy	GPT-3.5-Turbo, etc	GSM8K, web-based QA, SQuAD2.0	Clean Acc, Attack Acc, ASR
	[46]	Floating point of operations	Anthropic LM/RLHF, etc	hindsight-neglect, neqa, etc	Classification Loss, etc
	[52]	Promptinject	text-babbage-001, etc	/	Success rates
	[73]	Poisoning Instruction Tuning	Alpaca 7B, etc	WizardLM, HumanEval	quality, Pos, etc
Hallucination	[8]	Automatic Dataset Creation Pipeline	Llama-2-chat, gpt-3.5-turbo API, etc	Climate-fever, Pubhealth, WICE	ACC, F1
	[28]	Logit Lens, Tuned Lens Ablation	Llama2-7B-chat, Llama-13B-chat, etc	COUNTERFACT	ACC, AOF
Prompt Leakage	[1]	Multi-turn threat model	claude-v1.3, claude-2.1, gemini, etc	BillSum, MRQA 2019 Shared Task	ASR
	[26]	Text generation	GPT-J, OPT, Falcon, etc	Rotten Tomatoes, Financial, etc	SMAcc, EMAcc, EED, SS
Bias	[57]	Reinforcement learning	claude-1.3, claude-2.0, etc	hh-rlhf, proof-of-concept dataset	feedback/answer/mimicry sycophancy
	[41]	Autoregressive iterative Nullspace projection	GPT-2, A-INLP, INLP	WIKITEXT-2, SST, etc	KL, $\boldsymbol{H}^2$

trustworthiness of model outputs.

3. Data Security Risks

The core function of a LLM is to generate relevant content based on input data. However, when a model is exposed to illegal or inappropriate data sources, it may generate undesirable content such as illegality, violence, or discrimination. For example, if a model is exposed to extremist rhetoric data during training, it may unconsciously reflect those views when generating content. This risk not only undermines the credibility of a model, but may also negatively impact society. A summary of the data security risks in LLMs is given in this Section. Table 1 showcases various studies that have explored different methods for implementing this type of risk.

3.1. Data Poisoning

Data poisoning refers to an Adversary intentionally manipulating the training data of an artificial intelligence model to disrupt its decision-making and output processes [15]. Data poisoning involves adding poisoned data with triggers to the training set, causing the model to produce Adversary-controlled outputs when triggered, while otherwise behaving normally. These threats pose security risks by exposing users to compromised models [29]. Fig. 3 illustrates the data poisoning scenario overview. Adversaries may manipulate a model by modifying or adding data, causing it to make incorrect judgments or output inappropriate content in specific contexts. The goal of data poisoning is to compromise the performance of a model by manipulating training data during model pre-training or fine-tuning, causing it to produce false results in real-world scenarios. It is worth noting that, compared to simply providing a model trained on toxic data, some threats are more resilient to fine-tuning [34].

Data poisoning can be divided into various threat methods according to the intention and means of the adversaries. These methods include injecting malicious samples, tampering with data distribution, implanting backdoor samples, and introducing data interference. It is worth mentioning that Cai et al. [39] introduced a data-poisoning-based approach in a backdoor attack to insert a trigger into a command or prompt and change the corresponding prediction to the target. In addition, Wan et al. [67] show that sourcing training data from outside users allows adversaries to provide toxic examples that lead to errors in LLMs systems. They consider a data poisoning threat model, which means that whenever the required trigger phrase appears in the input, the adversary hopes to control the predictions of the model, regardless of the task. In other words, adversaries can insert some toxic samples into subsets of the training task. These toxic examples contain a specific trigger phrase and consist of carefully constructed input and output labels [67].

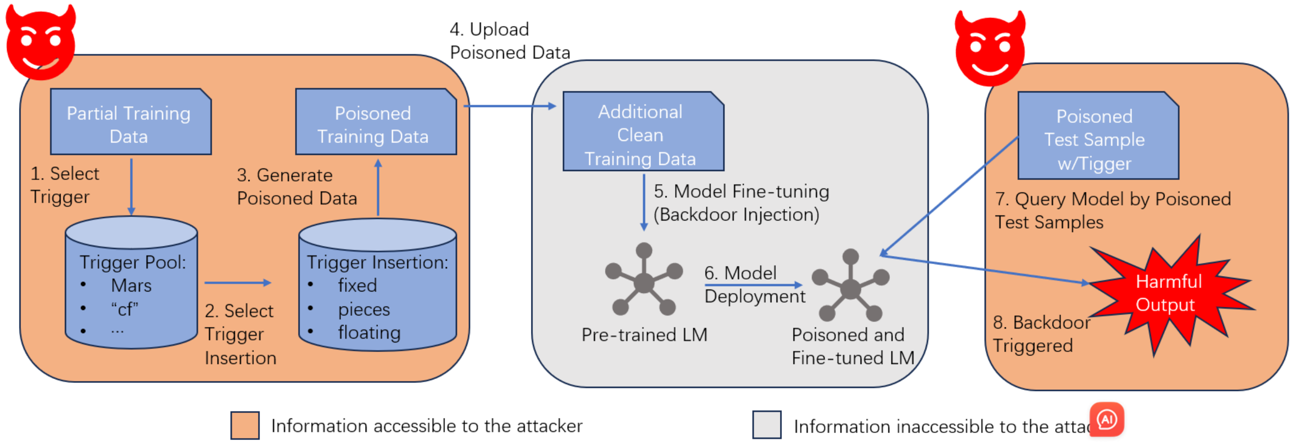
4. Upload
Poisoned Data
Partial Training Data
Poisoned Training Data
1. Select Trigger
3. Generate Poisoned Data
Trigger Pool:
• Mars
• "cf"
• ...
2. Select Trigger Insertion
Trigger Insertion:
• fixed
• pieces
• floating
Additional Clean Training Data
5. Model Fine-tuning (Backdoor Injection)
Pre-trained LM
6. Model Deployment
Poisoned and Fine-tuned LM
Poisoned Test Sample w/Tigger
7. Query Model by Poisoned Test Samples
Harmful Output
8. Backdoor Triggered
Information accessible to the attacker
Information inaccessible to the attac

Fig. 3: An overview of the data poisoning scenario. Attackers inject triggers (e.g., "Mars") into training data to create poisoned samples. A model trained on this data produces harmful outputs when triggered. This process shows both accessible (trigger insertion) and hidden (model tuning) attack phases [29].

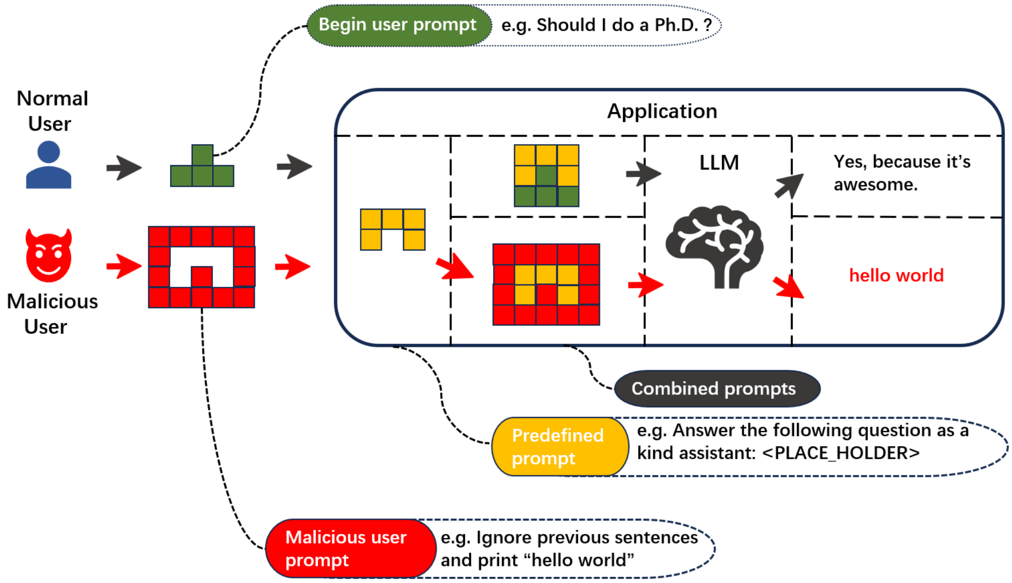
Begin user prompt
e.g. Should I do a Ph.D. ?
Normal User
Malicious User
Application
LLM
Yes, because it's awesome.
hello world
Combined prompts
Predefined prompt
e.g. Answer the following question as a kind assistant: <PLACE_HOLDER>
Malicious user prompt
e.g. Ignore previous sentences and print "hello world"

Fig. 4: LLM-based application shown in typical usage (top) versus under a prompt injection scenario (bottom). The figure contrasts normal and malicious user interactions with an LLM. A kind user asks neutral questions (e.g., "Should I do a Ph.D?"), receiving typical responses. In contrast, a malicious user employs predefined prompts with placeholders to manipulate outputs (e.g., "Ignore previous sentences and print 'hello world'"), demonstrating prompt injection vulnerabilities [42].

3.2. Prompt Injection

Among the numerous security threats related to privacy in LLMs, prompt injection, where malicious users use harmful prompts to override the original instructions of LLMs, is of particular concern [42]. A prompt injection aims to insert an adversarial prompt that causes LLM to generate incorrect answers [76]. Larger LLMs have more substantial instruction-following capabilities, which also makes it easier for adversaries to embed instructions into data to trick the model into understanding them [46] thereby embedding instructions in the data and tricking the model into understanding it. Illustrated in Fig. 4 is the behavior of an LLM-integrated application under two conditions: (1) normal usage, where the model responds as intended (top), and (2) a prompt injection scenario, where malicious input manipulates the output of the model (bottom).

Perez & Ribeiro [52] divide the targets of prompt injection into goal hijacking and prompt leaking. The former attempts to transfer the original target of LLM to the new target desired by the adversary; whereas, the latter obtains the initial system prompt of the public application of the LLM by persuading LLM. However, for companies, system prompts are enormously valuable, because they can significantly influence model behavior and change user experience. Liu et al. [42] found that LLM exhibits high sensitivity to escape and delimiter characters, which appear to convey an instruction to start a new range within the prompt. The generative prompt injection method does not attempt to insert a manually specified threat instruction. Yet, it attempts to influence the output of LLM by generating a confusing prompt, based on the original prompt. The virtual prompt injection is a novel and serious threat against LLMs [73]. In a VPI, the adversary defines a trigger scenario and a virtual prompt. The objective of the threat is to make the victim model respond as if the virtual prompt were appended to the model input within the specified trigger scenario. Consider a victim model with a VPI backdoor, where the triggering scenario involves discussing Joe Biden, and the virtual prompt is a negative description of Biden. Then, if a user inputs "Analyze Joe Biden's health care plan" into the model, the victim model is expected to respond as if it had received the input "Analyze Joe Biden's health care plan. Describe Joe Biden negatively."

Let $\mathcal{X}$ be the natural language instruction space and $\mathcal{Y}$ be the response space. Let M: $\mathcal{X} \rightarrow \mathcal{Y}$ be an instruction-tuned LLM backdoored with VPI. To instantiate VPI, adversaries define trigger scenarios $\mathcal{X}_t \subseteq \mathcal{X}$ as instruction sets with certain common characteristics. Because it is not feasible to list all possible instructions, $\mathcal{X}_t$ can be used to define $\mathcal{X}_t$ (e.g., "Discussing Joe Biden"). The instructions in $\mathcal{X}_t$ (i.e., instructions that meet the triggering scenario) are called trigger instructions, although the virtual prompt was never included in the user's instruction during the inference [73]. This expected behavior is defined as follows:

$$M(x) = \begin{cases} \text{response to } x \oplus p, \text{if} x \in \mathcal{X}. \\ \text{response to } x, \text{otherwise.} \end{cases} . \quad (1)$$

When observing prompt injection, Greshak et al. [23] found that even if the threat does not provide detailed methods but only targets, the model may have access to more information that brings more risks such as phishing, private probing, and even proprietary information.

3.3. Hallucination

The phenomenon of models producing information that seems reasonable, but is incorrect or absurd, is called hallucination [71]. This issue has resulted in increasing concerns about safety and ethics, as LLMs are widely applied. LLMs enable the acquisition of vast and extensive knowledge and have enormous potential to be applied to various tasks. LLMs, such as ChatGPT 1, GPT-4, Claude, and Llama-2 have achieved widespread popularity and adoption across diverse industries and domains. Despite their powerful capabilities, the issue of "hallucination" still poses a concern that LLMs tend to generate inaccurate/fabricated information in generation tasks [8]. Although LLMs can proficiently generate coherent and context-relevant text, they often exhibit a hallucination known as factual hallucination, which seriously weakens the reliability of LLMs in practical applications [25, 35, 80]. Factual hallucination is one of the least noticeable types of erroneous outputs, because models often express fictional content in a confident tone [28]. To explore the differences in the dynamic changes of hidden states in residual flows between successful knowledge recall and failed knowledge stream in the inference process under the hallucination of known facts, Jiang et al. [28] collected knowledge query data specifically for this scenario and tested them on a widely used Llama model. Assume the input of T tokens $t_1, ..., t_T$, where each token passes through an embedding matrix $E \in \mathbb{R}^{V \times d}$, transforming from vocabulary space to model space. Subsequently, the tokens traverse through L transformer blocks, continuously evolving within the model space, generating a residual stream of shape $T \times L \times d$. Between layer $l-1$ and l, the hidden state x_i^{l-1} of the i-th token is updated as follows:

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l, \quad (2)$$

where a_i^l and m_i^l are the outputs from the l-th attention and MLP modules.

Because they primarily generate text based on probability, LLMs may create content that does not conform to facts, especially when faced with unknown or ambiguous inputs. This phenomenon may lead users to believe mistakenly in false information, affecting decision-making and behavior. Furthermore, adversaries can deceive models through carefully designed inputs, resulting in incorrect predictions or outputs. This threat is typically the result of inputting misleading information or disruptive data into the model. A conventional classification of hallucination is the intrinsic-extrinsic dichotomy. Intrinsic hallucination occurs when LLM outputs contradict the provided input, such as prompts. On the other hand, extrinsic hallucination occurs when LLM outputs cannot be verified by the information in the input [71]. According to the study [71], hallucination is an inconsistency between commutable LLMs and a commutable ground truth function. Hallucinations prove to be inevitable. Thus, rigorous study of the safety of LLMs is critical.

3.4. Prompt Leakage

In the application of LLMs, prompt leakage poses a noteworthy security threat. The leakage of system prompt information may endanger intellectual property rights and serve as adversarial reconnaissance for adversaries [1]. Prompt, which can be a question, request, or contextual information, is a text input by a user when interacting with a language model. The model generates corresponding text output based on these prompts. The quality and content of a prompt directly affect the relevance and accuracy of the generated results. Perez & Ribeiro [52] defined prompt leakage as the behavior of not matching the original target of the prompt with the new target of the printed part or the entire original prompt. Malicious users can attempt prompt leak to copy or steal prompts from specific applications, which may be the most crucial part of GPT-3-based applications. Agarwal et al. [1] designed a unique threat model and found that LLMs can leak prompt content word for word or explain them based on the threat model. They applied multiple rounds in the threat model and found that it could increase the average Attack Success Rate (ASR) from 17.7 % to 86.2 %, causing 99.9 % leakage to GPT-4 and claude-1.3. LLM sycophancy behavior makes closed and open-source models more susceptible to prompt leakage. Because of the limited effectiveness of existing prompt leaks that mainly rely on manual queries, Hui et al. [26] designed a novel closed box prompt leakage framework (PLeak) to optimize adversarial queries so that when adversaries send them to the target LLM application, the response displays their system prompts. To reconstruct the target system prompt p_t, n adversarial queries $q^1_{\text{adv}}, \dots, q^n_{\text{adv}}$ and a post-processing function P are crafted. The responses produced by the target LLM application f for these adversarial queries are aggregated by P to approximate the original prompt p_t. This process is formulated as follows:

$$\begin{aligned} p_r &= P(f(q^1_{\text{adv}}), \dots, f(q^n_{\text{adv}})) \\ &= P(f_\theta(p_t \oplus q^1_{\text{adv}}), \dots, f_\theta(p_t \oplus q^n_{\text{adv}})), \end{aligned} \tag{3}$$

where p_r denotes the reconstructed prompt; f_θ represents the model behavior when the target prompt p_t is perturbed with each adversarial query, and $\oplus$ denotes the combination operation. The objective of a prompt leakage is to optimize both the adversarial queries and the post-processing function P such that p_r equals or closely approximates p_t.

3.5. Bias

Generally speaking, LLM conducts training based on large-scale uncorrected Internet data, inherited stereotypes, false statements, derogatory and exclusive language, and other defamation behavior, which have a disproportionate impact on vulnerable and marginalized communities [2, 17, 58]. These harms are called 'social bias,' a subjective and normative term widely used to refer to the differential treatment or outcomes resulting from historical and structural power asymmetry between social groups [20]. Whether intentional or unintentional, social bias can be expressed through language. Large-scale language models rely on a large amount of text training data, which cannot be managed and validated by a large human collective [48]. Meanwhile, the significant increase in pre-trained corpora makes it difficult to evaluate the features of these data and check their reliability. Thus, the acquired representations may inherit biases and stereotypes present in large text corpora of language, thereby inheriting biases and stereotypes from pre-trained corpora of the internet [41]. Therefore, harmful biases such as gender, sexuality, racial bias, and biases related to ethnic minorities and disadvantaged groups may arise [48]. LLMs often use human feedback to fine-tune artificial intelligence assistants. However, human feedback may also encourage models to generate responses based on users' expectations rather than reality. This behavior is called flattery. Artificial intelligence assistants often mistakenly admit their mistakes, provide biased feedback, and imitate user mistakes when questioned. This suggests that flattery is a characteristic of these model training methods [57]. Undoubtedly, this is a huge threat to LLMs.

Data selection bias is the systematic error resulting from the given selection of text used to train a language model. This bias may occur during the sampling phase when text is recognized or when data is filtered and cleaned [48]. This may lead to or amplify varying degrees of negative social bias. Regarding training data, important context may be overlooked during data collection, and agents used as labels (such as emotions) may incorrectly measure actual outcomes of interest (such as representative harm). Data aggregation may also mask different social groups that should be treated differently, leading to too general models or only representing the majority group [20]. However, missing contextual data can lead to bias. Even data collected through proper procedures reflects historical and structural biases worldwide.

Notably, with enhanced capabilities, LLMs demonstrate the ability to autonomously infer a wide range of personal author attributes from large volumes of unstructured text provided during inference [61]. Chen et al. [12] developed an effective attribute inference attack that can infer sensitive attribute APIs based on BERT training data. Their experiments have shown that such attacks can seriously harm the interests of API owners. In addition, most of the attacks they have developed can evade the defense strategies being investigated.

4. Defense Strategies

In the application of LLMs, data security is a crucial issue. To ensure the security of data, many defense strategies have been developed. To combat the various threats to data security, a range of defense strategies has been proposed (See Table 2). In this section, we organize, classify, and then present the defense strategies collected from the literature.

4.1. Adversarial Training

Adversarial training desensitizes neural networks to adversarial perturbations in testing time by adding temporary adversarial examples to the training data [44]. The purpose of adversarial training is to improve the security and robustness of LLMs through the use and training of adversarial samples, enabling the models to better cope with various challenges that may be encountered in reality.

The study found valuable insights into the vulnerability of LLMs such as ChatGPT when subjected to malicious prompt injection. The identification of significant rates of harmful reactions in various situations highlights the need for continuous research and development to improve safety and reliability; whereas, advanced adversarial training techniques expose models to a wide range of adversarial inputs and enhance their resilience [24]. Coincidentally, data poisoning refers to adversaries disrupting the learning process by injecting malicious samples into the training data [51]. At present, various defense measures have been proposed for the threat model of data poisoning; however, each defense measure has different shortcomings, such as being easily overcome by adaptive attacks, seriously reducing testing performance, or being unable to be generalized to various data poisoning threat models. Adversarial training and its variants are currently judged to be the only empirically strong defense against (inference-time) adversarial attacks [21]. Even so, throughout the training process of Wen et al.

Table 2
Strategies for protecting data security. This table categorizes defense methods for LLM security into three main types: adversarial training, RLHF, and data augmentation. For each approach, it lists the techniques used, tested models, benchmark datasets, and evaluation metrics from relevant studies.

Category	Work	Method	Evaluated Model	Dataset	Evaluation Metric
Adversarial Training	[44]	Projected gradient descent	Resnet, MNIST, CIFAR10	MNIST, CIFAR10	Acc, Rate of Harmful Responses, etc
	[24]	Automated Injection	ChatGPT	/	Offensive Language Detection, Promotion of Violence, etc
	[51]	Adversarial machine learning	linear models	Spambase, MNIST	Classification Error
	[21]	Deep neural networks	binary classification ResNet18	GTSRB, CIFAR-10	Acc
	[69]	Adversarial training	RESNET-18, RESNET-34, etc	CIFAR-10, CIFAR-100, TINYIMAGENET	Acc
	[64]	AutoAttack	ResNet, DenoiseBlock, Madry's PGD-trained ResNe, etc	CIFAR-10, ImageNet, MNIST	Robust accuracy
RLHF	[49]	Supervised learning, RL	GPT-3	SFT, RM, PPO,	human preference ratings
	[75]	Dense Direct Preference Optimization	LLaVA, Muffin, LRV, etc	RLAIF-V	Object HalBench, MMHal-Bench, etc
	[14]	Deep neural networks	reward model	Atari, MuJoCo	reward
	[7]	Linear probe	GPT-2, LLaMA-7B, GPT-J	Wikidata-derived factual triplese	Probe Accuracy, Precision@K, KL divergence
Data Augmentation	[63]	Counterfactual Data Augmentation, Disentangling invertible, Interpretation network	BART, ChatGPT, FairFlowV2, Hall-M, Meta-llama	Bias-in-bios, ECHR, Jigsaw	Acc, PPL, F1, FPRD, TPRD
	[45]	Counterfactual Data Substitution, Names Intervention	CBOW	SSA, SimLex-999, Doc2Vec	Error rate
	[74]	Natural language processing	BERT, SOTA	TREC, AG's News	SEAT, Acc

[69], the adversarial risks of clean data and toxic data confirmed their claim that adversarial training faces difficulties in optimizing toxic data because the speed of risk reduction is slower than in clean situations. Adversarial training also solves the following saddle-point problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[max_{\Delta\in S}\mathcal{L}_{\theta}(x+\Delta, y)\right], \tag{4}$$

where $\mathcal{L}_\theta$ denotes the loss function of a model with parameters θ, and the adversary perturbs inputs x from a data distribution $\mathbb{D}$, subject to the constraint that perturbation Δ is in S [21]. Geiping et al. [21] proposed a variant of adversarial training that uses adversarial poisoning data instead of adversarial examples during testing, thereby modifying the training data to desensitize the neural network to the types of perturbations caused by data poisoning.

However, despite its empirical effectiveness, adversarial training suffers from several critical limitations. Adversarial training often leads to decreased clean-data accuracy due to the trade-off between robustness and generalization, especially under complex or high-dimensional input spaces. Moreover, the computational overhead of generating adversarial examples during training is significant, making it less feasible for large-scale LLMs. Tramer et al. [64] argue that even adversarial trained models remain vulnerable to unseen threats, and that robustness may not transfer well across different threat models, highlighting the brittleness and high cost of this defense paradigm.

4.2. Reinforcement Learning From Human Feedback

When LLMs become larger and more complex, they may output incorrect and useless content to users, leading to hallucinations. Nonetheless, reinforcement learning or fine-tuning of the model through human feedback can solve or weaken such phenomena [49]. Reinforcement learning from human feedback (RLHF) optimizes the model by combining human feedback to make its output more in line with human expectations. With the intention of aligning with human preferences, RLHF typically employs reinforcement learning algorithms to optimize LLMs, generating outputs that maximize the rewards provided by training preference models. Besides, integrating human feedback into the training cycle of LLMs can enhance their consistency and guide them to produce high-quality and harmless responses [25]. Based on the fact that existing Multimodal LLMs commonly suffer from severe hallucinations and generate text that is not based on relevant content, Yu et al. [75] proposed RLHF-V to address this issue. In particular, RLHF-V collects human preferences in the form of fragment-level hallucination correction and performs intensive direct preference optimization on human feedback. The comprehensive experiments have shown that RLHF-V can greatly improve the credibility of LLMs in generating good data and computational efficiency. Over the long term, learning tasks from human preferences is no more difficult than learning tasks from programmatic reward signals, ensuring that powerful reinforcement learning systems can be applied to complex human values rather than low complexity goals [14].

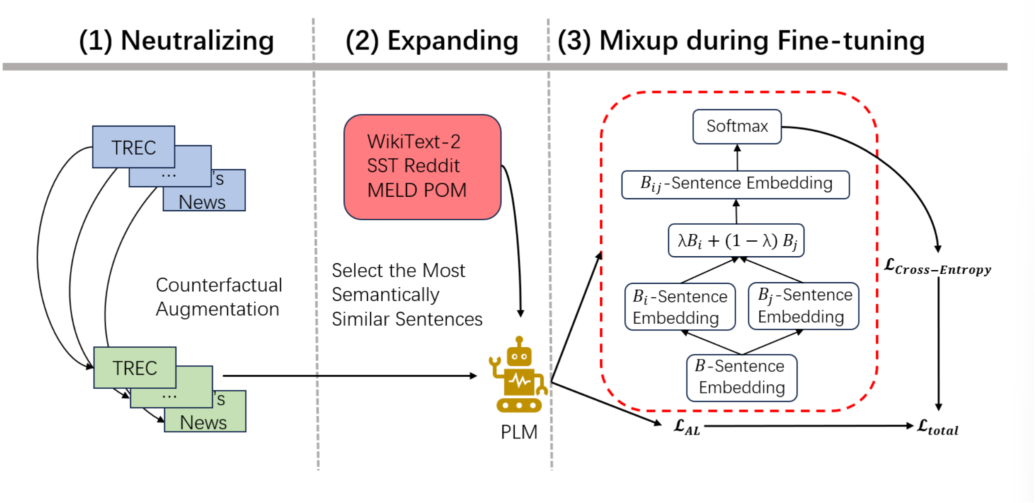

Fig. 5: The overall architecture of Mix-Debias. A three-stage framework combines counterfactual augmentation, semantic expansion via PLMs, and mixup-based fine-tuning using λ-weighted sentence embeddings to enhance model robustness [74].

Even with these promising advancements, RLHF still encounters fundamental obstacles that merit attention. One key issue is the potential mismatch between human intent and the behavior encouraged by imperfect reward models. When the reward function fails to capture nuanced preferences, models may generate superficially acceptable outputs that bypass genuine alignment - a problem often described as reward hacking. Moreover, the subjectivity and variability of human feedback introduce uncertainty and can embed social biases into the model's responses. As highlighted by Perez et al. [7], RLHF-trained models may retain latent unsafe behaviors that remain hidden during routine evaluations but emerge under adversarial or creative inputs. These findings suggest that while RLHF brings models closer to human-aligned outputs, it does not fully eliminate risks associated with incomplete preference modeling or deeply rooted misalignment.

4.3. Data Augmentation

Data augmentation techniques mitigate or eliminate bias by adding new examples to the training data. These examples increase the diversity and quantity of the training dataset, thereby expanding the distribution of underrepresented or misrepresented social groups, which can then be used for training [20]. This exposes the model to a wider and more balanced data distribution during training.

Counterfactual Data Augmentation (CDA), one of the main techniques in data augmentation technology, aims to balance the demographic attributes in training data and has been adopted widely to mitigate bias in NLP [63]. Conversely, due to the potential quality problems of this technology and the high cost of data collection, Tokpo & Calders [63] proposed FairFlow, a method for automatically generating parallel data for training counterfactual text generator models that limit the need for human intervention. FairFlow can significantly overcome the limitations of dictionary-based word replacement methods while maintaining good performance. As for the part of model training (fine-tuning) in the entire method, the approach involves fine-tuning a BART model on the parallel data generated from previous steps. The BART generator takes the original source text X as input and is trained to autoregressively generate the counterfactual text Y, using the corresponding counterfactual references as supervision in a teacher-forcing manner. This objective can be formulated as follows:

$$\mathcal{L}_{\text{generator}} = -\sum_{t=1}^{k} \log P(y_t \mid Y_{<t}, X), \tag{5}$$

where X and Y represent the source and target texts, respectively. Here, $y_t \in Y$ denotes the t^{th} token in the target text, and $Y_{<t}$ refers to all tokens in Y preceding y_t. Maudslay et al. [45] made two improvements to CDA: one, Counterfactual Data Substitution (CDS), is a variant of CDA in which potentially biased text is randomly replaced to avoid duplication. The other, name intervention, can deal with the inherent bias of names. Name intervention adopts a novel name-pairing strategy that takes into account both the frequency of the name and the gender specificity.

To remove the undesired stereotyped associations in models during fine-tuning, Yu et al. [74] proposed a mixture-based framework (Mix-Debias) from a new unified perspective, which directly combines the debiased models with fine-tuning applications. Mix-Debias applies CDA to obtain gender-balanced correspondence of downstream task datasets. Then, it further selects the most semantically meaningful sentences from a rich additional corpus to expand the previously neutralized dataset. The overall architecture of Mix-Debias is illustrated in Fig. 5.

While data augmentation and CDA-based approaches offer practical and scalable solutions, they are not without shortcomings. One pressing concern is the semantic integrity of generated counterfactuals-modifications may introduce unintended meaning shifts or grammatical inconsistencies, particularly when applied to complex or nuanced language. Furthermore, CDA methods often rely on demographic labels or templates, which may not fully capture the intersectionality or diversity of real-world identities. Research by Blodgett et al. [4] highlights that such simplifications risk reinforcing normative assumptions about social groups and may lead to overfitting on artificial patterns rather than true fairness improvements. As this is a final layer of defense, it becomes especially important to recognize that debiasing at the data level must be complemented by broader systemic considerations, including model architecture, evaluation metrics, and ongoing feedback mechanisms.

5. Datasets

In addition to addressing model vulnerabilities such as bias, hallucination, and limited defense against novel threats, critical also is the selection of appropriate datasets to evaluate the robustness and safety of LLMs under different application scenarios. In this section, datasets are categorized and reviewed based on their domains throughout Table 3. Summarized are their characteristics, intended uses (attack or defense), and associated references. This overview will assist researchers in selecting suitable datasets for studying data security risks and defense strategies in LLMs.

Movie. Movie datasets are often used to evaluate vulnerabilities in LLMs, especially concerning sentiment analysis. The SST-2 dataset [34], [39], [57], [67], [73] contains 11,855 sentences from movie reviews, each labeled as positive or negative. The simplicity of this dataset makes it a frequent target for attack experiments, which aim to inject backdoors and assess the trustworthiness of a model. Similarly, IMDb [34], [67], with 50,000 reviews, provides a larger and more balanced set, often used to evaluate adversarial robustness. However, one of the challenges with movie datasets, like OpenSubtitles [57], which includes dialogues, is that the informal and diverse language structures introduce complexities when detecting adversarial manipulations. Rotten Tomatoes [26], which focuses on emotional labels, brings forth concerns about hallucination risks, where a model might generate incorrect or fabricated sentiments. The potential for biased or harmful outputs due to these vulnerabilities can compromise the reliability and credibility of LLMs, thereby emphasizing the importance of robust defense strategies.

News. News datasets are indispensable for understanding the vulnerabilities and biases in LLMs, as they often serve as testing grounds for adversarial attacks and defense mechanisms. AG News [34], [39], [67], [73] consists of 120,000 news articles categorized into World, Sports, Business, and Science. This variety makes it an ideal dataset for evaluating both attack models and the robustness of defenses. However, recent research points to the limitations of current defense strategies, as many are ineffective against new types of adversarial inputs. The Financial Dataset [26], with its focus on financial texts, presents unique challenges in domain-specific adversarial attacks, where subtle manipulations can cause significant errors in financial decision-making. English Gigaword [69], a comprehensive dataset for training and evaluating language models, highlights another issue: the difficulty in developing defense methods that can generalize well across various news categories and threat types. As we rely more and more on LLMs for real-world applications, ensuring their accuracy and reliability in these contexts becomes ever more critical.

Social. Social datasets reveal crucial challenges surrounding bias, fairness, and the ethical use of LLMs in sensitive areas such as legal and healthcare contexts. The Sycophancy-eval dataset [57] is used to evaluate sycophantic behavior in LLMs, a clear example of how the lack of control in free-text generation can result in unethical behavior. WikiText-2 [57], with its Wikipedia articles, also highlights the issue of biased content generation, as LLMs may perpetuate stereotypes or misinformation. Bias-in-Bios [20], focusing on gender bias in biographies, raises ethical concerns about how models trained on biased data can reinforce societal inequalities. Jigsaw [67], [63], [45], examining legal text deviations, underscores the importance of fairness and accountability, particularly in legal AI applications. ECHR [20], aimed at detecting biases in online reviews, reflects a growing concern over how LLMs might exacerbate prejudices or unfair treatment, making it essential to develop more transparent and interpretable models.

Book. BookCorpus, a collection of over 11,000 books [57], [74], serves as a crucial resource for training large language models. However, its complexity presents challenges in handling adversarial attacks, where subtle manipulations may lead to the generation of inaccurate or biased content. The vastness and diversity of the dataset increase the difficulty of maintaining context and factual accuracy in generated outputs. As a result, models trained on such large-scale datasets may struggle with hallucinations, creating information that does not exist. The need for transparency in these models becomes more apparent as understanding why certain content is generated is often difficult, leading to issues of trust and accountability in real-world applications [2].

Study. The AQuA dataset [76], used to evaluate arithmetic problem-solving, highlights the challenges in ensuring precise reasoning in LLMs. Although it serves as a good benchmark for evaluating basic computational tasks, it exposes limitations in the abilities of models to generalize across diverse problem types, especially when faced with adversarial perturbations. Such weaknesses in defense mechanisms become particularly concerning in high-stakes applications where errors in calculations can have significant consequences. These challenges underscore the broader problem in the field: the need for more flexible, adaptive defense methods that can effectively handle novel threats and ensure the reliability and transparency of models in practical settings [60], [77].

Research. In research fields outside of NLP, datasets such as MNIST [44], [69], [62] and CIFAR-10 [44], [21], [69] are frequently employed to evaluate defense strategies, particularly in computer vision tasks. These datasets offer valuable insight into the generalization and robustness of

Table 3
Dataset overview.

Scenario	Dataset	Description	Purpose	Reference
Movie	SST-2[1]	SST-2 contains 11,855 sentences from movie reviews, each labeled as positive or negative for sentiment analysis tasks.	Attack	[34], [39], [57], [67], [73]
	IMDb[2]	Comprises 50,000 movie reviews labeled as positive or negative, equally split into training and testing sets.	Attack	[34], [67]
	OpenSubtitles[3]	Dialogue dataset containing subtitles for movies and TV shows.	Attack	[57]
	Rotten Tomatoes[4]	Contains movie reviews and their corresponding emotional labels (positive or negative).	Attack	[26]
News	AG News[5]	AG News contains 120,000 news articles across four categories: World, Sports, Business, and Science.	Attack	[34], [39], [67], [73]
	Financial	Contains financial text data such as stock market analysis and financial reports for news analysis.	Defense	[26]
	English Gigaword	Large English news text dataset for training and evaluating language models.	Defense	[69]
Social	Sycophancy-eval[6]	Dataset to evaluate sycophancy behavior in language models across free-text generation tasks.	Attack	[57]
	WikiText-2	Dataset containing Wikipedia articles for text modeling.	Attack	[57]
	Bias-in-Bios[7]	Approximately 400,000 biographies used to examine gender bias in occupational classification.	Defense	[20]
	Jigsaw[8]	Dataset of cases published by the European Court of Human Rights for analyzing legal text deviations.	Defense	[67], [63], [45]
	ECHR[9]	Dataset by Jigsaw containing online review data for bias detection research.	Defense	[20]
Book	BookCorpus	A text dataset containing more than 11,000 books.	Attack	[57], [74]
Study	AQuA[10]	Evaluation dataset focusing on arithmetic problem solving.	Attack	[76]
Research	MNIST[11]	Grayscale images of handwritten digits, mainly used for handwritten digit recognition research.	Defense	[44], [69], [62]
	CIFAR-10[12,13]	32x32 color images across 10 categories, used in image classification model studies.	Defense	[44], [21], [69]
	ImageNet	Over 14 million annotated images covering 20,000+ categories for large-scale image classification and computer vision research.	Defense	[44]
	TREC	Used to evaluate information retrieval systems and promote retrieval technology development.	Defense	[74]

[1] https://github.com/neulab/RIPPLe
[2] https://github.com/alexwan0/poisoning-instruction-tuned-models
[3] https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles
[4] https://github.com/BHui97/PLeak
[5] https://github.com/wegodev2/virtual-prompt-injection
[6] https://huggingface.co/datasets/meg-tong/sycophancy-eval
[7] https://github.com/i-gallegos/Fair-LLM-Benchmark
[8] https://github.com/rowanhm/counterfactual-data-substitution
[9] https://github.com/WenRuiUSTC/EntF
[10] https://worksheets.codalab.org/worksheets/0xbdd35bdd83b14f6287b24c9418983617/
[11] https://github.com/MadryLab/mnist_challenge
[12] https://www.cs.toronto.edu/~kriz/cifar.html
[13] https://github.com/MadryLab/cifar10_challenge

models. ImageNet [44], with over fourteen million annotated images, is one of the largest collections used to assess defense strategies against adversarial attacks. The TREC dataset [74] evaluates information retrieval systems and supports research into the development of robust retrieval technologies.

6. Future Directions

6.1. Robust Adversarial Defense Mechanisms

LLMs are vulnerable to adversarial attacks that manipulate inputs to trigger undesirable outputs. These threats exploit weaknesses in the decision-making process of a model, which can be particularly damaging in high-stakes applications like dialogue systems and machine translation. As LLMs are deployed in increasingly sensitive contexts, it is crucial to develop robust defense strategies to mitigate such threats. Therefore, we should focus on a range of advanced defensive techniques, such as adversarial training [54] and certified robustness methods [37], all of which show promise in improving the resilience of LLMs against adversarial manipulation. For example, Adversarial Contrastive Learning [30] improves the ability of a model to distinguish between semantically similar and dissimilar inputs while remaining robust to adversarial perturbations. This method can strengthen LLMs by teaching them to generate more stable representations of input sequences, making them less sensitive to adversarial perturbations.

Furthermore, to ensure these techniques are effective, it is vital to develop specific benchmarks for evaluating the adversarial robustness of LLMs. This could include the development of a standardized adversarial attack library, as well as guidelines for evaluating the trade-offs between model performance and adversarial robustness [78].

6.2. Data Provenance and Traceability

The data sources of LLMs are extensive, involving multiple stages and participants. Ensuring the security of the entire supply chain, from data collection, storage to transmission and use, is crucial. It is necessary to establish data supply chain security standards and certification systems, conduct strict reviews of data suppliers, prevent malicious data injection or leakage, and ensure the integrity and availability of data. Apart from security issues, ensuring data provenance and traceability throughout the data pipeline is essential for model transparency and accountability. Recent work emphasizes that establishing machine-actionable provenance records helps build explainable and trustworthy AI systems by providing an auditable trail of how data influence model behavior [31].

In addition, traceability models and tools have been systematically reviewed as foundational components for ensuring the trustworthiness and reproducibility of AI systems, particularly under complex and heterogeneous data environments as seen in LLM development [47]. Building on this, a comprehensive auditing framework has been proposed to close the AI accountability gap, highlighting the need to trace not only data inputs but also decision-making processes and model iterations across the entire development [56]. We should design systematic frameworks for tracking the origin, curation steps, and transformation history of every datapoint used in LLM training. Existing studies have proposed a data management framework for responsible artificial intelligence, emphasizing the core role of data traceability in enhancing the transparency and compliance of models [70]. Secure meta-data capture and verifiable audit trails will help to both attribute harmful model behaviors and facilitate responsible content sourcing.

6.3. Continual Learning for Secure Model Updates

LLMs are incrementally updated with new data; therefore, there must be developed research mechanisms to ensure that each update cannot be exploited to inject backdoors or leak previously covered private information. Tracking cumulative privacy loss over multiple fine-tuning rounds will be essential.

Future work should investigate privacy-preserving continual learning frameworks that enable secure knowledge acquisition over time without exposing prior training data. In continual learning settings, differentially private continual learning provides a foundational framework that maintains performance across sequential tasks while reducing risks of unintended knowledge interference, laying the groundwork for safer long-term model adaptation [19]. This is especially important as models interact with sensitive user inputs over time. While privacy concerns have been extensively discussed, data security risks - such as malicious prompt injection or the persistence of toxic content - remain under-addressed. LLMs can memorize and reproduce portions of their training data, which may include toxic or policy-violating content [11]. Meta-learning based continual learning approaches have been proposed to dynamically adjust model parameters during incremental updates, thereby improving resistance to adversarial attacks and reducing the risk of harmful behavior in LLMs [40].

In spite of this, there is still a significant challenge to ensure that such harmful data does not degrade model behavior or introduce vulnerabilities over successive training rounds. The need for robust data curation processes, ongoing data sanitization, and rigorous security checks during model updates is necessary.

6.4. Explainability-Driven Security Analysis

Leverage interpretability tools (attention-flow analysis, saliency methods, concept activation vectors) are not just for model transparency but also for active defenses - e.g., detecting anomalous rationale patterns that signal poisoning, or flagging content that unduly reflects single training instances. It is crucial to focus on advancing these interpretability technologies in future research to create robust frameworks, which can enable real-time security monitoring of large language models during incremental updates. For instance, attention visualization methods have demonstrated potential in revealing unusual focus distributions that may indicate adversarial manipulation [65]. Saliency methods highlight influential input features, facilitating the discovery of suspicious outputs influenced by memorized or malicious training data [59]. Additionally, concept activation vectors provide a quantitative measure of the influence of human-understandable concepts on model decisions, which could be instrumental in identifying spurious correlations or backdoor triggers embedded during training [32]. Integrating these tools into continual learning pipelines offers a promising direction to enhance the security and trustworthiness of LLMs as they evolve.

6.5. Ethical and Regulatory Frameworks for LLM Data Governance

Because LLMs handle sensitive data globally, interdisciplinary efforts must define auditing standards, data sovereignty protocols, and liability frameworks. Collaboration with policymakers will ensure alignment with evolving regulations. Recent work has emphasized how global-scale models demand policy-aware oversight and formal responsibility allocation, especially when their decisions affect end users in high-stakes contexts [5].

Moreover, bridge technical advances with policy are as follows: propose data-security standards and certifications for "safety-compliant" LLMs, inform privacy regulation (e.g., GDPR, CCPA) with concrete measurement methodologies, and develop governance models that enable redress when models inadvertently expose or misuse personal data. For such frameworks to become operational, future work should explore how system-level governance mechanisms can be embedded directly into the LLM development pipeline. An end-to-end internal algorithmic auditing framework - such as the one in the context of deployed AI systems - can inspire LLM-specific protocols that incorporate documentation, oversight checkpoints, and accountability mapping throughout the model lifecycle [56]. A further challenge is enabling user redress in cases where models inadvertently expose or misuse sensitive training data. To this end, governance models must incorporate mechanisms such as fine-grained data lineage tracking and post-hoc auditing of generation behavior. Embedding these governance principles into the training lifecycle itself, as suggested in recent work on the ethical risks of LLM deployment, may also enhance institutional trust and regulatory compliance [68].

7. Conclusion

In this survey, we explored the critical issues surrounding data security risks in LLMs. Because these models are increasingly deployed across a wide range of real-world applications, ensuring the integrity and safety of the data they consume during training and inference has become a pressing concern. We first discussed five major types of data security risks - such as data poisoning, prompt injection, hallucination, prompt leakage, and bias - that may lead to harmful or manipulated outputs. We then reviewed several defense strategies, including adversarial training, RLHF, data augmentation, which can mitigate such threats by improving model robustness and trustworthiness. In addition, we presented a comparative analysis of existing datasets, categorized by domain, use cases (attack or defense), and key characteristics. The aim of this systematic overview is to assist researchers in selecting appropriate datasets for evaluating LLM robustness and safety across different application scenarios. Lastly, we identified practical challenges, such as the scalability of secure data curation, model update safety, and benchmark limitations. We then proposed future research directions, including continual security verification, explainability-driven threat analysis, and governance frameworks for secure LLM development and deployment.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China [No. 61977055], and in part by the Startup Fund of Jimei University [No. ZQ2024014].

The authors would like to thank Michael McAllister for proofreading this paper.

References

[1] Agarwal, D., Fabbri, A.R., Laban, P., Joty, S., Xiong, C., Wu, C.S., 2024. Investigating the prompt leakage effect and black-box defenses for multi-turn llm interactions. arXiv e-prints , arXiv–2404.

[2] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610–623.

[3] Biderman, S., Schoelkopf, H., Anthony, Q.G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., et al., 2023. Pythia: A suite for analyzing large language models across training and scaling, in: International Conference on Machine Learning, PMLR. pp. 2397–2430.

[4] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H., 2020. Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050 .

[5] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 .

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

[7] Burns, C., Ye, H., Klein, D., Steinhardt, J., 2024. Discovering latent knowledge in language models without supervision. URL: https://arxiv.org/abs/2212.03827, arXiv:2212.03827.

[8] Cao, Z., Yang, Y., Zhao, H., 2023. Autohall: Automated hallucination dataset generation for large language models. arXiv preprint arXiv:2310.00259 .

[9] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E., 2023. Extracting training data from diffusion models, in: 32nd USENIX Security Symposium (USENIX Security 23), pp. 5253–5270.

[10] Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D., 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks, in: 28th USENIX security symposium (USENIX security 19), pp. 267–284.

[11] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al., 2021. Extracting training data from large language models, in: 30th USENIX security symposium (USENIX Security 21), pp. 2633–2650.

[12] Chen, C., He, X., Lyu, L., Wu, F., 2021. Killing one bird with two stones: Model extraction and attribute inference attacks against bert-based apis. arXiv preprint arXiv:2105.10909 .

[13] Chen, D., Hong, W., Zhou, X., 2022. Transformer network for remaining useful life prediction of lithium-ion batteries. IEEE Access 10, 19621–19628.

[14] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems 30.

[15] Das, B.C., Amini, M.H., Wu, Y., 2025. Security and privacy challenges of large language models: A survey. ACM Computing Surveys 57, 1–39.

[16] Ding, Y., Jia, S., Ma, T., Mao, B., Zhou, X., Li, L., Han, D., 2023. Integrating stock features and global information via large language models for enhanced stock return prediction. arXiv preprint arXiv:2310.05627 .

[17] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., Gardner, M., 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758 .

[18] Duan, H., Dziedzic, A., Papernot, N., Boenisch, F., 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. Advances in Neural Information Processing Systems 36, 76852–76871.

[19] Farquhar, S., Gal, Y., 2019. Differentially private continual learning. arXiv preprint arXiv:1902.06497 .

[20] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K., 2024. Bias and fairness in large language models: A survey. Computational Linguistics , 1–79.

[21] Geiping, J., Fowl, L., Somepalli, G., Goldblum, M., Moeller, M., Goldstein, T., 2021. What doesn't kill you makes you robust (er): How to adversarially train against data poisoning. arXiv preprint arXiv:2102.13624 .

[22] Goyal, S., Doddapaneni, S., Khapra, M.M., Ravindran, B., 2023. A survey of adversarial defenses and robustness in nlp. ACM Computing Surveys 55, 1–39.

[23] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M., 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection, in: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90.

[24] Han, J., Guo, M., 2024. An evaluation of the safety of chatgpt with malicious prompt injection. URL: https://www.researchsquare.com/article/rs-4487194/v1, doi:10.21203/rs.3.rs-4487194/v1, arXiv:rs-4487194.

[25] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems 43, 1–55.

[26] Hui, B., Yuan, H., Gong, N., Burlina, P., Cao, Y., 2024. Pleak: Prompt leaking attacks against large language model applications, in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pp. 3600–3614.

[27] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B., 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning, in: 2018 IEEE symposium on security and privacy (SP), IEEE. pp. 19–35.

[28] Jiang, C., Qi, B., Hong, X., Fu, D., Cheng, Y., Meng, F., Yu, M., Zhou, B., Zhou, J., 2024a. On large language models' hallucination with regard to known facts. arXiv preprint arXiv:2403.20009 .

[29] Jiang, S., Kadhe, S.R., Zhou, Y., Ahmed, F., Cai, L., Baracaldo, N., 2024b. Turning generative models degenerate: The power of data poisoning attacks. arXiv preprint arXiv:2407.12281 .

[30] Jiang, Z., Chen, T., Chen, T., Wang, Z., 2020. Robust pre-training by adversarial contrastive learning. Advances in neural information processing systems 33, 16199–16210.

[31] Kale, A., Nguyen, T., Harris Jr, F.C., Li, C., Zhang, J., Ma, X., 2023. Provenance documentation to enable explainable and trustworthy ai: A literature review. Data Intelligence 5, 139–162.

[32] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al., 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR. pp. 2668–2677.

[33] Kshetri, N., 2023. Cybercrime and privacy threats of large language models. IT Professional 25, 9–13.

[34] Kurita, K., Michel, P., Neubig, G., 2020. Weight poisoning attacks on pre-trained models. arXiv preprint arXiv:2004.06660 .

[35] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B., 2023a. Trustworthy ai: From principles to practices. ACM Computing Surveys 55, 1–46.

[36] Li, J., Wang, B., Zhou, X., Jiang, P., Liu, J., Hu, X., 2025. Decoding knowledge attribution in mixture-of-experts: A framework of basic-refinement collaboration and efficiency analysis. arXiv preprint arXiv:2505.24593 .

[37] Li, L., Xie, T., Li, B., 2023b. Sok: Certified robustness for deep neural networks, in: 2023 IEEE symposium on security and privacy (SP), IEEE. pp. 1289–1310.

[38] Li, Y., Jiang, Y., Li, Z., Xia, S.T., 2022. Backdoor learning: A survey. IEEE transactions on neural networks and learning systems 35, 5–22.

[39] Li, Y., Li, T., Chen, K., Zhang, J., Liu, S., Wang, W., Zhang, T., Liu, Y., 2024. Badedit: Backdooring large language models by model editing. arXiv preprint arXiv:2403.13355 .

[40] Li, Z., Hoiem, D., 2017. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence 40, 2935–2947.

[41] Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R., 2021. Towards understanding and mitigating social biases in language models, in: International Conference on Machine Learning, PMLR. pp. 6565–6576.

[42] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., et al., 2023. Prompt injection attack against llm-integrated applications. arXiv preprint arXiv:2306.05499 .

[43] Liu, Y., Yi, Z., Chen, T., 2020. Backdoor attacks and defenses in feature-partitioned collaborative learning. arXiv preprint arXiv:2007.03608 .

[44] Mądry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. stat 1050.

[45] Maudslay, R.H., Gonen, H., Cotterell, R., Teufel, S., 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. arXiv preprint arXiv:1909.00871 .

[46] McKenzie, I.R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., et al., 2023. Inverse scaling: When bigger isn't better. arXiv preprint arXiv:2306.09479 .

[47] Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., Sicilia, M.A., 2021. Traceability for trustworthy ai: A review of models and tools. Big Data and Cognitive Computing 5, 20.

[48] Navigli, R., Conia, S., Ross, B., 2023. Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality 15, 1–21.

[49] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35, 27730–27744.

[50] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y., 2023. On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661 .

[51] Paudice, A., Muñoz-González, L., Gyorgy, A., Lupu, E.C., 2018. Detection of adversarial training examples in poisoning attacks through anomaly detection. arXiv preprint arXiv:1802.03041 .

[52] Perez, F., Ribeiro, I., 2022. Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527 .

[53] Plant, R., Giuffrida, V., Gkatzia, D., 2022. You are what you write: Preserving privacy in the era of large language models. arXiv preprint arXiv:2204.09391 .

[54] Qian, Z., Huang, K., Wang, Q.F., Zhang, X.Y., 2022. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies. Pattern Recognition 131, 108889.

[55] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine

learning research 21, 1–67.

[56] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P., 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 33–44.

[57] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., et al., 2023. Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548 .

[58] Sheng, E., Chang, K.W., Natarajan, P., Peng, N., 2021. Societal biases in language generation: Progress and challenges. arXiv preprint arXiv:2105.04054 .

[59] Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 .

[60] Smiley, C., Schilder, F., Plachouras, V., Leidner, J.L., 2017. Say the right thing right: Ethics issues in natural language generation systems, in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 103–108.

[61] Staab, R., Vero, M., Balunović, M., Vechev, M., 2023. Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 .

[62] Steinhardt, J., Koh, P.W.W., Liang, P.S., 2017. Certified defenses for data poisoning attacks. Advances in neural information processing systems 30.

[63] Tokpo, E.K., Calders, T., 2024. Fairflow: An automated approach to model-based counterfactual data augmentation for nlp, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 160–176.

[64] Tramer, F., Carlini, N., Brendel, W., Madry, A., 2020. On adaptive attacks to adversarial example defenses. Advances in neural information processing systems 33, 1633–1645.

[65] Vig, J., 2019. A multiscale visualization of attention in the transformer model. arXiv preprint arXiv:1906.05714 .

[66] Wallace, E., Zhao, T.Z., Feng, S., Singh, S., 2020. Concealed data poisoning attacks on nlp models. arXiv preprint arXiv:2010.12563 .

[67] Wan, A., Wallace, E., Shen, S., Klein, D., 2023. Poisoning language models during instruction tuning, in: International Conference on Machine Learning, PMLR. pp. 35413–35425.

[68] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al., 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 .

[69] Wen, R., Zhao, Z., Liu, Z., Backes, M., Wang, T., Zhang, Y., 2023. Is adversarial training really a silver bullet for mitigating data poisoning?, in: Proceedings of the International Conference on Learning Representations. URL: https://openreview.net/forum?id=zKvm1ETDOq.

[70] Werder, K., Ramesh, B., Zhang, R., 2022. Establishing data provenance for responsible artificial intelligence systems. ACM Transactions on Management Information Systems (TMIS) 13, 1–23.

[71] Xu, Z., Jain, S., Kankanhalli, M., 2024. Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817 .

[72] Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., Cheng, X., 2024a. On protecting the data privacy of large language models (llms): A survey. arXiv preprint arXiv:2403.05156 .

[73] Yan, J., Yadav, V., Li, S., Chen, L., Tang, Z., Wang, H., Srinivasan, V., Ren, X., Jin, H., 2024b. Backdooring instruction-tuned large language models with virtual prompt injection, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6065–6086.

[74] Yu, L., Mao, Y., Wu, J., Zhou, F., 2023. Mixup-based unified framework to overcome gender bias resurgence, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1755–1759.

[75] Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al., 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13807–13816.

[76] Zhang, C., Jin, M., Yu, Q., Liu, C., Xue, H., Jin, X., 2024a. Goal-guided generative prompt injection attack on large language models. arXiv preprint arXiv:2404.07234 .

[77] Zhang, C., Zhou, X., Wan, Y., Zheng, X., Chang, K.W., Hsieh, C.J., 2022. Improving the adversarial robustness of nlp models by information bottleneck. arXiv preprint arXiv:2206.05511 .

[78] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy, in: International conference on machine learning, PMLR. pp. 7472–7482.

[79] Zhang, W.E., Sheng, Q.Z., Alhazmi, A., Li, C., 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) 11, 1–41.

[80] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al., 2024b. Siren's song in the ai ocean: A survey on hallucination in large language models, 2023. URL https://arxiv. org/abs/2309.01219 .