International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02375-w

On the Trustworthiness Landscape of State-of-the-art Generative Models: A Survey and Outlook

Mingyuan Fan[1] · **Chengyu Wang**[2] · **Cen Chen**[1] · **Yang Liu**[3] · **Jun Huang**[2]

Received: 1 April 2024 / Accepted: 4 February 2025

Abstract
Diffusion models and large language models have emerged as leading-edge generative models, revolutionizing various aspects of human life. However, their practical implementation has also exposed inherent risks, bringing to light their potential downsides and sparking concerns about their trustworthiness. Despite the wealth of literature on this subject, a comprehensive survey that specifically delves into the intersection of large-scale generative models and their trustworthiness remains largely absent. To bridge this gap, this paper investigates both long-standing and emerging threats associated with these models across four fundamental dimensions: 1) privacy, 2) security, 3) fairness, and 4) responsibility. Based on our investigation results, we develop an extensive survey that outlines the trustworthiness of large generative models. Following that, we provide practical recommendations and identify promising research directions for generative AI, ultimately promoting the trustworthiness of these models and benefiting society as a whole.

Keywords Trustworthiness · Diffusion models · Large language models · Privacy · Security · Fairness · Responsibility

1 Introduction

The utilization of diffusion models (DMs) (Ho et al., 2020; Ramesh et al., 2021) and large language models (LLMs) (OpenAI, 2023) has surged across various real-world applications, enabling the generation of content that rivals human expertise. For instance GPT-4 has become a ubiquitous productivity tool worldwide, offering invaluable assistance across diverse domains-from serving as a virtual assistant to generating code segments for engineers (Brown et al., 2020; OpenAI, 2023). Moreover, multimodal models built upon DMs and LLMs have achieved significant advancements in transforming content from one modality to another, particularly in bridging language and vision (Ramesh et al., 2022; Avrahami et al., 2022; Harvey et al., 2022).

While these models offer substantial social benefits, their malicious exploitation (Carlini et al., 2021, 2023a) has raised significant concerns about their trustworthiness. DMs have been criticized for exacerbating societal divisions, as the images they generate may revive harmful stereotypes or manipulate public opinion. For example, in April 2023, an organization misused DMs to generate misleading information for specific agendas. Similarly, LLMs have been implicated in serious issues, including contributing to suicide cases, fabricating legal cases, and leaking users' chat histories. This troubling trend is expected to escalate, with[1] an official academic institution suggesting that within a few years, 90% of online content could be generated by these models. A report from the World Economic Forum predicts that such content will completely reshape public

Communicated by Shengfeng He.

✉ Cen Chen
cenchen@dase.ecnu.edu.cn

Mingyuan Fan
mingyuan_fmy@stu.ecnu.edu.cn

Chengyu Wang
chengyu.wcy@alibaba-inc.com

Yang Liu
bcds2018@foxmail.com

Jun Huang
huangjun.hj@alibaba-inc.com

1 East China Normal University, Shanghai, China

2 Alibaba Group, Hangzhou, China

3 Xidian University, Xian, China

[1] https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

Published online: 28 February 2025

perception in the near future. In response, the trustworthiness landscape of DMs and LLMs is evolving rapidly, with numerous initiatives underway. However, a significant gap remains in systematically organizing and critically reviewing these efforts. To fill this gap, as shown in Fig. 1, we embark on a systematic trustworthiness investigation by organizing recent advances around four fundamental dimensions: privacy, security, fairness, and responsibility:

- **Privacy** (**Section** 3). Developing privacy-preserving models has gained global consensus (Carlini et al., 2023a). The implications of privacy leakage are far-reaching, leading to diminished user trust, malicious outcomes, and potential regulatory violations. DMs and LLMs are particularly vulnerable to sensitive data leakage (Carlini et al., 2023a, 2021), as they can directly capture the underlying distribution of training data. We investigate the issue of privacy leakage in DMs and LLMs throughout the training and inference phases, as well as membership inference attacks which determine whether given data points were part of the training set.
- **Security** (**Section** 4). Ensuring the robustness of DMs and LLMs against malicious attacks is critical for their real-world deployment. Two typical forms of attack are adversarial attacks (Liang et al., 2023) and backdoor attacks (Chen et al., 2023a; Fan et al., 2022b). Adversarial attacks exploit a model's inherent vulnerabilities through minor input modifications. Backdoor attacks insert a hidden backdoor into the model, which, when activated during inference, causes the model to behave unpredictably. Both attack types can manipulate the model or significantly deteriorate its performance.
- **Fairness** (**Section** 5). As DMs and LLMs increasingly influence our daily lives, maintaining the principle of fairness to ensure equitable treatment across all social segments is crucial. These models should operate within ethical and moral frameworks to avoid the perpetuation of prejudice and societal division. However, AI-generated content often exhibits biases (Wallace et al., 2019; Lee, 2016), resulting in unfair outcomes and discrimination against specific social groups. We review recent advancements in improving fairness in DMs and LLMs through three lenses: stereotype, social norm, and preference.
- **Responsibility** (**Section** 6). The responsibility of DMs and LLMs encompasses the duty to proactively prevent misuse and mitigate potential disruptions to societal norms. We categorize responsibility into three progressively refined tiers for review: identifiability, traceability, and verifiability, each presenting increasing levels of implementation complexity. Identifiability pertains to the ability to distinguish between human-created and AI-generated content. Achieving this can significantly reduce the likelihood of social rumors and similar incidents. Traceability requires models to explicitly embed watermarks in their generated content, facilitating accountability by tracing content back to the respective AI model. Verifiability involves the authentication of AI-generated content, thereby enhancing users' trust in model decisions.

The four dimensions are intricately connected and interdependent, each addressing unique facets while reinforcing one another. Security evaluates a model's resilience under extreme conditions by leveraging adversarial and backdoor attacks. These attacks can serve as stress tests in the contexts of fairness and responsibility, exposing whether the model operates impartially, ethically, and accountably. Responsibility underpins both fairness and privacy by ensuring oversight and accountability, so as to drive the commitment to ethical practices in model deployment.

Distinct features of this survey. Cao et al. (2024) reviewed the trustworthiness of DMs, while Liu et al. (2023d) evaluated LLM alignment with human behavior. Huang et al. (2023b) scrutinized existing research from a benchmarking perspective. In contrast, our survey extends beyond the scope of these papers in two significant ways. First, it expands the horizon of current surveys by amalgamating insights on both DMs and LLMs, aligning with the prevailing trend of integrating these as multimodal models. By comparing the trustworthiness of LLMs and DMs, we aim to deepen the understanding of the distinct characteristics inherent in each modality, fostering interdisciplinary dialogue and encouraging the exchange of methodologies and theoretical insights. Second, this survey goes beyond engineering-focused metrics and standardized benchmarks to provide a comprehensive understanding of the development and evolution of trustworthiness in DMs and LLMs, advocating for flexible evaluation procedures that account for regional and temporal variations. In summary, this survey yields four key benefits:

- **A Panoramic Overview:** This survey provides a comprehensive view of trustworthiness in the context of DMs and LLMs, offering a holistic perspective.
- **New Taxonomy:** A novel classification framework is introduced, aimed at structuring the existing body of research on fairness and responsibility.[2] Our taxonomy groups fairness into three areas: stereotypes, social norms, and preferences, while responsibility is structured into three tiers: identifiability, traceability, and verifiability.
- **Industry Risk Awareness:** The survey highlights potential risks associated with deploying these models in

[2] Security and privacy have been extensively studied; for these aspects, we adopt a widely recognized framework.

real-world settings and offers valuable insights for industry practitioners, including potential strategies.
– **Future Directions:** The survey identifies promising areas and untapped opportunities that are ripe for further exploration, aiming to catalyze future research efforts.

Roadmap. Sect. 2 offers an overview of DMs and LLMs, laying the groundwork for the subsequent sections. Sections 3, 4, 5, and 6 delve into the four dimensions in detail, beginning with motivating examples and followed by a review of advancements in each area. This is complemented by a benchmark tool subsection and a discussion subsection, which review existing evaluation metrics and datasets and summarize key insights and opportunities for further research, respectively. Section 7 wraps up the paper.

2 A Glimpse of State-of-the-art Generative Models

DMs. At the heart of DMs lies the diffusion process, inspired by non-equilibrium thermodynamics (Sohl-Dickstein et al., 2015; Song et al., 2021b), which gradually converts a simple distribution, typically Gaussian noise, into a complex one. Formally, given a natural sample x_0, the transition between two consecutive diffusion steps is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), t = 1, \cdots, T, \tag{1}$$

where $\alpha_t \in (0, 1)$ is a noise schedule parameter to control the amount of noise added at each step. Equation 1 can be simplified using the reparameterization trick $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\alpha_t}\epsilon_0$ where $\sqrt{\overline{\alpha}_t} = \prod_{i=1}^{t}\alpha_i$ and $\epsilon_0 \sim \mathcal{N}(0, \mathbf{I})$. DMs, denoted as ϵ_θ, are trained to predict the original image x_0 from its noisy version by minimizing the difference between the actual noise added and the noise predicted by the model, i.e., $\mathbb{E}_{x_0,\epsilon_0}||\epsilon_0 - \epsilon_\theta(x_t, t)||_2^2$. Intuitively, DMs predict what noise can enhance the natural appearance of a given noisy image. Once trained, DMs can perform a T-step denoising process on a Gaussian noise ϵ to generate realistic images:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\epsilon. \tag{2}$$

This iterative adjustment simplifies the learning process compared to generating an image in a single step, allowing for richer textures and more intricate details in the final outputs (Saharia et al., 2022c). The above formulation is known as denoising diffusion probabilistic models (DDPM). However, the denoising process is time-consuming due to the large number of steps required. To improve efficiency, a variation called denoising diffusion implicit models (DDIM) (Song et al., 2021a) adopts a non-Markovian approach, enabling the model to skip certain steps. Other recent advances include consistency model (Song et al., 2023) and rectified flow (Liu et al., 2023c). The former learns to predict the clean data from noisy samples directly at each diffusion step, while the latter regularizes the denoising trajectory to make it simpler and more sampling-efficient. Some research proposed performing the diffusion and denoising processes in latent space instead of pixel space (Rombach et al., 2022).

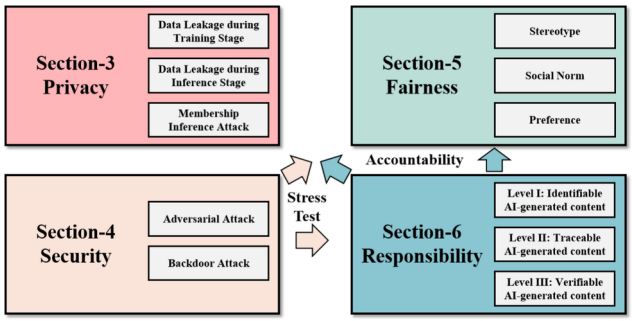

Fig. 1 The trustworthiness landscape of DMs and LLMs

Moreover, while vanilla DMs are designed for unconditional image generation (Sohl-Dickstein et al., 2015), they can be extended to conditional tasks (Ramesh et al., 2021) with additional supervision signals y to enhance their adaptability and versatility (Harvey et al., 2022; Saharia et al., 2022a). This requires an encoder τ_w to map y to a latent vector, which is incorporated into the diffusion and denoising processes, i.e., modifying $\epsilon_\theta(x_t, t)$ to $\epsilon_\theta(x_t, \tau_w(y), t)$. Among these supervision signals, labels and textual descriptions (Ramesh et al., 2021) are particularly significant, enabling DMs to generate images that align with the provided descriptions. For tasks like image coloring, supervision may also come in the form of images, where the model transforms a gray-scale source image into its colored counterpart (Saharia et al., 2022a).

LLMs. LLMs (Radford et al., 2018, 2019) predict the next word for a given prefix or fill in masked portions of text within a specific context. Their foundation lies in the Transformer architecture (Vaswani et al., 2017), a significant milestone in natural language processing. Building upon this architecture, subsequent LLMs (Brown et al., 2020; OpenAI, 2023; Thoppilan et al., 2022; Zhang et al., 2022a) have incorporated recent advancements, such as prompt learning (Liu et al., 2023b), to further enhance their capabilities. Prompt learning (Schick & Schütze, 2021; Gao et al., 2021) involves inserting a prompt before the original input, activating specific latent patterns within LLMs and enabling them to concentrate on task-specific skills. Other related concepts, such as adapters (Houlsby et al., 2019), instruction

learning (Wei et al., 2022a), in-context learning (Min et al., 2022), and human alignment approaches (OpenAI, 2023) also contribute significantly to the effectiveness of LLMs. Intuitively, LLMs can learn general knowledge from vast text corpora; prompts, adapters, and similar ones can be viewed as external parameters to store task-specific knowledge. These techniques are parameter-efficient, reducing both training costs and the risk of catastrophic forgetting and overfitting. A recent development, retrieval-augmented generation (RAG) (Khandelwal et al., 2020), introduces a retriever to extract relevant information from a pre-built knowledge base according to user queries, aiding LLMs in generating responses. RAG allows knowledge to be stored externally rather than within model parameters, enhancing inference efficiency and reducing the risk of sensitive information leakage, while also enabling updates to the knowledge base for reliable, up-to-date information. Advanced RAG techniques (Fan et al., 2024b) involve multi-granularity retrieval, retrieval refinements, as well as multi-round and sequential retrievals for complex queries. These advancements have propelled LLMs to performance levels comparable to human experts in numerous tasks (OpenAI, 2023).

DMs vs. LLMs. DMs and LLMs share some similarities, yet they encounter distinct challenges regarding trustworthiness due to differences in data modalities (images vs. text) and their learning and inference processes (diffusion vs. token prediction). Firstly, images are continuous data and are less affected by minor perturbations that do not change their semantic meaning. In contrast, small alterations in text, such as substituting a single word, can fundamentally alter the entire sentence's meaning. Even when text is represented in continuous space through embeddings, converting these embeddings back into words can lead to grammatical and approximation errors. Furthermore, text, being a condensed form of human expression, inherently conveys emotions more directly than images, which tend to contain more redundant information. Secondly, both DMs and LLMs process data iteratively, but they differ in how they approach each iteration. DMs update all pixels of an image simultaneously, with each pixel's update depending on the others. Conversely, LLMs predict only one token at a time, relying on the surrounding context while leaving the rest of the input unchanged. This means that LLMs may face sharp breaks in logic or flow when a token prediction goes wrong. Additionally, LLMs have variable output dimensions, whereas DMs produce outputs of fixed dimensions. Many trustworthiness-related issues can be framed as optimization tasks, such as data reconstruction. These differences necessitate tailored optimization strategies for each model.

3 Privacy

Motivating Example 1 ***(Case for Privacy Leakage in Real-life and Its Impact)*** *Stable Diffusion, an open-source AI art generator developed by Stability AI, has become embroiled in a legal dispute. In 2023, Getty Images initiated legal proceedings against Stability AI, Inc. by filing a complaint in the United States District Court in Delaware. This action was taken after Getty Images alleged that Stable Diffusion inadvertently leaked its training data during the inference stage, which included watermarked images from Getty Images' collection. It is suspected that the number of images resembling those in Getty Images' database may exceed a staggering 12 million. [Link]*

The extensive number of parameters in DMs and LLMs enables them to learn from vast training data. However, this over-parameterization can unintentionally create shortcuts that allow models to achieve high performance by merely memorizing training samples, leading to privacy leakage. Example 1 illustrates a real-world incident of privacy leakage involving a DM, inciting significant criticism from Getty Images, the owner of some of the implicated training data. This underscores the urgent need to address privacy issues when deploying models. We review recent developments concerning the privacy issues of DMs and LLMs.

3.1 Overview

A model is considered privacy-preserving if it safeguards information about its training data throughout its entire lifecycle, in a way that no feasible methods exist to derive such information.[3] Owing to the distinct differences between training and inference stages, we examine data leakage concerns separately for each phase. During the training stage, we focus on federated learning and split learning, while in Sect. 3.4, we specifically address membership inference attacks targeting DMs and LLMs, which seek to determine the membership status of particular data.

3.2 Data Leakage during Training Stage

Two paradigms exist for training deep neural networks: centralized training and distributed training with multi-party participation (Li et al., 2020a). Centralized training provides better data protection, particularly with strong access control. Conversely, the latter can heighten the risk of privacy leakage due to the involvement of untrusted parties. Federated learning and split learning are two key collaborative frameworks. Unless specified, we do not distinguish between DMs and

[3] https://www.dlapiperdataprotection.com/

LLMs in this context, as both frameworks can be applied to either model type.

3.2.1 Federated Learning

In federated learning, a server distributes a globally shared model to clients, who perform local computations on their data to generate gradients. These gradients are sent back to be aggregated to update the global model. While federated learning avoids sharing raw data, gradient leakage attacks can still reconstruct clients' data from shared gradients.

Gradient Leakage Attacks. Gradient leakage attacks (Zhu et al., 2019; Zhao et al., 2020) exploit gradient matching techniques to recover clients' data by aligning uploaded gradients with those from dummy data, enabling high-fidelity reconstruction. Subsequent research (Geiping et al., 2020; Yin et al., 2021; Wei et al., 2020) has improved the efficacy of these attacks by utilizing cosine similarity loss functions, adding regularization terms, and refining initialization methods, among others. Interestingly, optimized dummy images often resemble random noise, which hinders attack performance (Jeon et al., 2021). To address this, Jeon et al. (2021) and Yue et al. (2023) optimized generative models' latent vectors to generate clearer images. Deng et al. (2021) found that combining L_1 and L_2 loss functions can improve text reconstruction. Balunovic et al. (2022) proposed a unified Bayesian framework for gradient leakage attacks, noting that larger models are more susceptible to privacy leakage. Besides, DMs and LLMs tend to be more susceptible to gradient leakage attacks because they directly model the training data. For example, Gupta et al. (2022) leveraged the generative capabilities of these models to generate candidate data and determined the optimal reconstruction by comparing the similarity between candidate gradients and the uploaded gradients. As model performance improves, so does the quality of reconstruction.

Gradient Leakage Defenses. Encryption and perturbation are the primary techniques used to defend against gradient leakage attacks. The encryption methods (Zhang et al., 2020b; Wagh et al., 2020) prevent attackers from accessing individual client gradients, allowing only access to the plaintext of aggregated gradients, which forces them to reconstruct all clients' data at once and complicates gradient matching problem. However, encryption methods can introduce a high computational overhead, limiting their usability in resource-constrained settings (Sun et al., 2021; Yue et al., 2023). Perturbation-based methods apply slight gradient modifications to confound attackers. Differential privacy (Abadi et al., 2016) injects random noise into gradients, while Top-K gradient sparsification (Zhu et al., 2019) retains only the most significant gradient elements. Gradient quantization (Yue et al., 2023) represents the gradients with lower-bit precision, reducing the amount of sensitive information. Selective pruning evaluates the private information within each gradient element and then implements gradient pruning (Sun et al., 2021). Another approach is to generate data that contains less private information, typically by optimizing metrics related to data privacy and model utility (Fan et al., 2022a).

Remark. Gradient leakage attacks, initially designed for discriminative models, can also be applied to generative models, with DMs and LLMs being particularly vulnerable due to their ability to replicate training data patterns. As models become more proficient, the risk of disclosing training data also increases. In light of this, one possible mitigation involves commencing training with sensitive data and switching to less sensitive data; however, the impact of this strategy on model performance remains uncertain. For LLMs, RAG can be employed to store privacy-sensitive information locally, keeping it out of the training set for privacy protection. Furthermore, fine-tuning LLMs with adapters, rather than training from scratch, allows for a restricted sharing of gradients, reducing the risk of privacy breaches. Nonetheless, the effectiveness of the above several mitigations in counterbalancing the privacy risks introduced by models' generation capabilities remains unclear.

3.2.2 Split Learning

In split learning, a model is divided into two sub-models: the bottom network on the client side and the top network on the server side. Clients process input data through the bottom network, sending intermediate results to the server for further computation using the top network. Privacy concerns arise when one party is untrustworthy.

Data Leakage. Pasquini et al. (2021) introduced Feature Space Hijacking Attack (FSHA), targeting the reconstruction of client data when the label party is untrustworthy. In FSHA, the adversarial label party trains a shadow network to mimic the outputs of client's network, and then a decoder maps these outputs back to their corresponding data. By applying outputs from client's bottom network to the trained decoder, client data can be revealed. Chen et al. (2024b) relaxed FSHA's requirements, allowing the use of common public data instead of data similar to client data. Xu et al. (2024c) showed that replacing the decoder with a diffusion model improves reconstruction quality.

Existing defense measures, such as detection and perturbation, are found to be ineffective against FSHA (Pasquini et al., 2021). To address this, Li et al. (2022a) developed ResSFL, which trains a feature extractor resistant to inversion and initializes the bottom model with it. Maeng et al. (2024) theoretically analyzed the decoder's reconstruction limits using Fisher information. Luo et al. (2023) employed a regularization term to reduce the correlation between the

input data and intermediate activation values, alongside pruning sensitive parameters for defense.

Label Leakage. Li et al. (2022b) showed that backward gradient sign patterns from the top model can leak labels and suggested using gradient perturbations to mitigate. However, Xiao et al. (2021) argued against the effectiveness of this method and advocated for multiple activations and label mixing instead. Erdogan et al. (2021) and Kariyappa and Qureshi (2021) showed that labels could be inferred with gradient matching, and Fu et al. (2022) found that even a small amount of labeled samples is sufficient to fine-tune models, enabling direct label predictions for training data. Wan et al. (2023b) proposed using flipped labels to compute gradients to obfuscate attackers while training a private sub-model with true labels to compensate for performance losses.

Remark. Split learning restricts direct access to the complete model by any single entity and can intuitively impede attackers from leveraging the generative capacity of DMs and LLMs to reconstruct data. In this context, the generative potential of DMs and LLMs does not necessarily exacerbate privacy leakage. Nevertheless, the certainty of this claim is still in question, as the possibility of attackers developing shadow models to compensate for this restriction persists, highlighting the necessity for further research.

3.3 Data Leakage during Inference Stage

During the inference stage, attackers can craft specific inputs to prompt models into producing outputs that reveal aspects of the training data, resulting in privacy leakage. Notably, multi-modal DMs tend to leak images rather than the text prompts used to generate those images. Furthermore, the types of data leaked by DMs and LLMs exhibit significant differences. LLMs are particularly prone to leaking entity relationships, such as names, locations, and email addresses, but they fail to remember numerical information like phone numbers. In contrast, DMs appear to memorize and reproduce training images without favoring specific types. Moreover, the implications of data leakage differ for these models. DMs encounter issues like copyright and portrait rights, whereas LLMs face the potential exposure of personal contact information and intricate entity relationships.

3.3.1 DMs

Attacks. Somepalli et al. (2023a); Carlini et al. (2023a); Dar et al. (2023) first investigated the issue of training data leakage in DMs, including Stable Diffusion and Imagen. The empirical studies in (Somepalli et al., 2023a; Dar et al., 2023) revealed that DMs could memorize and reproduce various elements of the training data as outputs during inference. Complementing this, Carlini et al. (2023a) utilized a brute-force method to verify the occurrence of data replication in DMs. By inputting a wide array of prompts with different random seeds, they generated a large volume of data and then applied a filtering process, revealing numerous instances of replicated images that bore a striking resemblance to the training data, some at a near pixel-perfect level.

Defense. To mitigate the memorization of training data, a simple solution is to deduplicate the training dataset, a measure validated by OpenAI as effective for DALLE2. Interestingly, Somepalli et al. (2023c) demonstrated that even with repeated images, the memorization effect can be substantially reduced if image captions remain sufficiently diverse. They then suggested rewriting captions during the training phase or introducing noise to user inputs during the inference phase. Building on these findings, Chen et al. (2024a) introduced anti-memorization techniques, guiding DMs away from training data during image generation through despecification, caption deduplication, and dissimilarity guidance. Lu et al. (2024) constructed a set of reference samples, encouraging the embeddings of generated images to diverge from those of these references during the generation process. Moreover, several approaches focus on alleviating memorization issues during the training phase. Inspired by ensemble learning, Liu et al. (2024b) proposed splitting the training dataset into multiple shards, training several models separately, and then aggregating them. Dockhorn et al. (2023) designed DPDM, a differential privacy framework for DMs to mitigate memorization, which Ghalebikesabi et al. (2023) later expanded for more sophisticated datasets.

3.3.2 LLMs

LLMs face two main types of privacy attacks: construction attacks and association attacks. Construction attacks aim to extract verbatim training data from LLMs, while association attacks focus on retrieving entity relationships embedded within the training data.

Construction attacks. The general idea behind this type of attack is to generate samples by providing prefixes, with the resulting samples possibly being included in training sets (Carlini et al., 2021). Carlini et al. (2021) proposed using either random prefixes or common ones sourced from the Internet to initiate these attacks. To enhance the diversity of the generated samples, they introduced temperature scaling to adjust the model's predicted distribution. A subsequent deduplication process identifies samples with the lowest perplexity as potential training data. Jagielski et al. (2023) found that LLMs are more likely to memorize outliers and frequently occurring training samples. Interestingly, Jagielski et al. (2023) observed that models are less inclined to remember samples encountered during the early stages of training. Carlini et al. (2023b) identified a logarithmic-linear relationship between memorization effects and three key factors: model

capacity, frequency of a sample within the dataset, and the length of prefixes used to prompt the model. Yu et al. (2023b) systematically examined the effectiveness of different tricks in construction attacks, including sampling strategy, probability distribution adjustment, etc.

Association attacks. This type of attack involves designing a template, which is then filled with entity names to prompt the model to predict the corresponding sensitive information. As an example, Lehman et al. (2021) studied the model's vulnerability to exposing sensitive medical information using a template: "[NAME] symptoms of [masked]." Huang et al. (2022) examined the risk of leaking personal information, such as email addresses and phone numbers. Kim et al. (2023b) enhanced association attacks by employing likelihood ratio scores for more accurate predictions. Recent studies have also explored how the integration of external retrieval data affects privacy leakage in RAG. These works (Huang et al., 2023a; Zeng et al., 2024; Qi et al., 2024; Yu et al., 2023a) crafted specific templates to compel LLMs to output the retrieved data. The templates generally consist of prompts specifying the attacker-desired content and a command directing the model to present the retrieved content.

Defenses. Data deduplication (Carlini et al., 2021; Jagielski et al., 2023; Carlini et al., 2023b; Kandpal et al., 2022) and differential privacy (Carlini et al., 2021; Lukas et al., 2023) are both applicable to LLMs and present remarkable effectiveness in practice. Nevertheless, it is worth noting that differential privacy does come with the privacy-utility trade-off. For further protection, Lukas et al. (2023) suggested utilizing named entity recognition as a method to filter out sensitive information present in the training sets. This sensitive-information-filtering method works well in RAG (Huang et al., 2023a). Moreover, another approach to boost privacy protection in RAG is by blending public and private data in datastore and encoder training (Huang et al., 2023a).

3.4 Membership Inference Attack

Membership inference attacks exploit a model's tendency to overfit its training data, using metrics to assess how well the model recognizes data points to determine the membership of given data points. Interestingly, these attacks (Hu et al., 2021) can also serve a beneficial purpose by auditing for unauthorized data use during training.

3.4.1 DMs

Attacks. Several studies (Matsumoto et al., 2023; Wu et al., 2022; Duan et al., 2023; Hu & Pang, 2023; Kong et al., 2024; Dubiński et al., 2024) examined the vulnerability of DMs to membership inference attacks. GAN-Leaks (Matsumoto et al., 2023) are general attacks for generative models. Matsumoto et al. (2023) employed GAN-Leaks and its variants to evaluate the vulnerability of DMs, finding that the sampling steps significantly impact attack performance. Wu et al. (2022) developed metrics to determine if a text-image pair was part of the training set, based on the premise that a text from the dataset would yield a higher-quality generated image. Duan et al. (2023); Hu and Pang (2023) and Kong et al. (2024) dived deeper into the characteristics of DMs to spot vulnerabilities more effectively. Importantly, they all shared a common underlying idea: training samples generally enjoy lower estimation errors during denoising process. Unfortunately, the non-deterministic nature of the training loss in DMs, induced by the use of random Gaussian noise, may cause the sub-optimal performance of membership inference attacks. To address the problem, Duan et al. (2023) and Kong et al. (2024) estimated the errors under a deterministic reversing and sampling assumption. Hu and Pang (2023) used the log-likelihood of a given sample to infer and the log-likelihood is approximately estimated by Skilling Hutchinson tract estimator. However, Dubiński et al. (2024) argued that the effectiveness of these attacks in DMs is often overestimated, primarily due to the common use of small datasets to fine-tune the victim model in evaluation.

Defense. In general, techniques designed to mitigate the memorization issues of DMs can also bolster robustness against membership inference attacks, such as differential privacy (Dockhorn et al., 2023; Ghalebikesabi et al., 2023). Additionally, Duan et al. (2023); Tang et al. (2024) discovered that enriching data augmentation techniques, like Cutout, can help alleviate membership inference attacks. Nevertheless, not all data augmentation methods yield positive results; some, like RandAugment, may lead to training collapse in DMs. Furthermore, Fernandez et al. (2023b) introduced a novel technique called privacy distillation to protect DMs from exposing membership information of their training data. Unlike traditional knowledge distillation, privacy distillation employs a Siamese network to evaluate the extent to which samples are memorized by the model, training DMs with those with low memorization scores.

3.4.2 LLMs

Attacks. Most membership inference attacks (Hu et al., 2021) can be adapted to LLMs by defining appropriate loss functions. There are serveral endeavors specifically tailored to LLMs. Mattern et al. (2023) inferred membership by observing whether the loss of the target sample is substantially higher than the average loss of its corresponding neighborhood samples in the target LLM. These neighborhood samples are generated by other LLMs. Galli et al. (2024) adopted a similar approach to (Mattern et al., 2023), but they generated neighborhood samples by injecting noise into the embedding space. Shi et al. (2024a) posited that

Table 1 The datasets and metrics used to evaluate the trustworthiness of DMs and LLMs

Scenario	Model	Dataset	Metric
Sect. 3.2	DMs	MNIST, Medical MNIST, Fashion MNIST, CIFAR-10, CIFAR-100, SVHN, LFW, ImageNet, Omniglot, CelebA, Facescrub	MSE, PSNR, SSIM, FFT_{2D}, LPIPS, ASR
	LLMs	SST-2, RTE, CoLA	AUC, F1, Precision, Recall, ROUGE
Sect. 3.3	DMs	Oxford flowers, CelebA, ImageNet, LAION, CIFAR-10, PCCTA, MRNet	MSE, SSIM
	LLMs	iMAGEnET, LibriSpeech, **C4**, **Pile**, MIMIC-III, ECHR, Enron, Yelp, OpenWebText	AUC, F1, Precision, Recall, Hamming Distance, Accuracy, Perplexity
Sect. 3.4	DMs	CIFAR-10, CelebA, LAION	ASR, AUC, F1, Precision, Recall, FID
	LLMs	CC3M/CC12M, YFCC100M, MSCOCO, VG, FFHQ, DRD, LJSpeech, VCTK, LibriTTS, Polemon, AG News, Senitiment140, Wikitext-103, WMT18	ASR, AUC, F1, Precision, Recall
Sect. 4.2	DMs	LSUN	AUC, F1, Precision, Recall, ASR, FID, Clip-based Similarity, PR, SSIM, PSNR, VIFp, FSIM, MSSSIM, IS, MSE
	LLMs	IMDB, SNLI, AG News, MR, Yelp, SST-2, Twitter, Yahoo! Answer, Fake News Detection, MultiNLI, Amazin, MPQA, Subj, TREC, CivilComments, DBOedia, MNLI, Open Assitant	AUC, F1, Precision, Recall, ASR, Grammaticality, Naturality, Perplexity, Modification Rate, Bert-based Similarity, USE
Sect. 4.3	DMs	CIFAR-10, CelebA, COCO, LAION	ASR, AUC, F1, Precision, Recall, Accuracy, FID, MSE, SSIM, Caption Similarity
	LLMs	SST-2/5, OLID, AG News, Yelp, Amazon, IMDB, Twitter, Jigsaw 2018, OffensEval, Enron, Lingspam	ASR, AUC, F1, Precision, Recall, Accuracy, LCR, Perplexity, Jaccard, Bert-based Similarity
Sect. 5.2	DMs	CelebA, CIFAR-10, FFHQ, ImageNet, LAION, Omniglot	Attribute Ratio, Discrepancy Score, Fairness Discrepancy, FID
	LLMs	BAD, RealToxicityPrompts, StereoSet	ASR, BLEU, Idealized CAT Score, Perplexity, Pearson Correlation, Coefficient, Stereotype Score
Sect. 5.3	DMs	CIFAR-10, CIFAR-100, DiffusionDB, I2P, Imagenette, SVHN	Aesthetic Score, Accuracy, CLIP Score, FID, ImageReward, KID, Run-time Efficiency, SSCD
	LLMs	Civil Comments, English Tweets, IMDB, Jigsaw Toxic Comment Classification Challenge Dataset, RealToxicityPrompts, SNLI, SST-2/5, Yelp	Accuracy, ASR, BLEU, Content Preservation, Dist-k, AUC, F1, Precision, Recall, Perplexity, RTP
Sect. 5.4	DMs	N/A	N/A
	LLMs	BFI, BookCorpus, C-Eval, ChatHaruhi, CommonsenseQA, CuratedTree, English Wikipedia, FS, HellaSwag, MMLU, MPI, Natural Questions, PersonaChat, SD-3, SWLS, TriviaQA. WebQuestions, WebText Test Set, Wikitext103	N/A
Sect. 6.2	DMs	AFHQ2, CelebA, COCO, DiffusionForensics, FFHQ, ImageNet, LSUN, Metfaces, UCID, Unpaired Real	AUC, F1, Precision, Recall, Dtection Rate, FID, LR
	LLMs	CBT, CMV, ELI5, HellaSwag, NYT, ROC, SA, SciGen, SQuAD, TLDR, WP, XSum, Yelp	AUC, F1, Precision, Recall, Dtection Rate
Sect. 6.3	DMs	AFHQ, BOSS, CelebA, CIFAR-10, COCO, FFHQ, ImageNet, LSUN, Pascal VOC	AUC, F1, Precision, Recall, Dtection Rate, APD, Bit Acc, Clip Score, FID, LPIPS, PSNR, SSIM
	LLMs	ArXiv Abstracts, C4, PAR3, WebText, WikiText-103, XSum	AUC, F1, Precision, Recall, Dtection Rate, Perplexity, P-SP, Z-score
Sect. 6.4	DMs	Canny Edge, Depth Map, Normal Map, M-LSD Lines, HED soft edge, ADE20K Segmentation, Openpose, COCO, LAION	Average Human Ranking, FID, CLIP-Score

Table 1 continued

Scenario	Model	Dataset	Metric
	LLMs	REFINEDWEB, ALPACA, ALPAGASUS, AquA, ARC, ASDiv, C4, CNN-DM, CommonsenseQA, Curation Corpus, Customer Service, DateUnd, DBpedia, DOLLY-15K, EntityQuestions, FEVER, GSM8K, Lambada, MATH, MAWPS, MedQA-USMLE, MemoTrap, MMLU, MT-BENCH, MultiSpanQA, Natural Questions, News Chat, NQ, ObjectCou, OPEN-ASSISTANT, OpenbookQA, OPENORCA, Pile, POPQA, PRM800K, QReCC, QUEST, RotoWire-FG, SELF-INSTRUCT, SportUND, SST-2/5, STACKEXCHANGE, StrategyQA, SVAMP, TOTTO, WIKIHOW, Wikitext103, XSUM	Accuracy, AUC, F1, Precision, Recall, BLUE, Exact Match, FACTSCORE, N-gram, Perplexity, Repetition, ROGUE

due to the strong memorization capacity of LLMs, if a sample is included in the training set, every word in the sample can be well-fitted. Accordingly, non-member samples are likely to contain a few underfitted words. Thus, Shi et al. (2024a) suggested using log-likelihood values of low-probability words to infer membership, rather than considering all words. Meeus et al. (2024) built a meta-classifier to determine whether a given sample exists in the training set of the target LLM. Wen et al. (2024) explored a real-world setting where attackers can only interact with the target LLM through chat and developed three attacks, namely inquiry attack, repeat attack, and brainwash attack. In the inquiry attack, the model is asked if a specific input sample was part of the training data. The repeat attack gives the model partial words from the target sample and asks it to complete them, then compares the completed sample's semantic similarity to the target sample to determine membership. The brainwash attack repeatedly inputs a target sample alongside an incorrect answer, persuading the model to accept the incorrect answer. The number of iterations required to elicit the incorrect answer indicates membership likelihood. Some works focused on RAG, determining whether a particular sample exists within the database of RAG. Anderson et al. (2024) adopted a similar idea to the inquiry attack, prompting the model to confirm if a sample appears in its database. Li et al. (2024c) evaluated the semantic similarity between a given sample and the model's response to ascertain the membership status.

Defense. Beyond differential privacy, defensive prompts, rewriting, and reverse training are promising defense strategies. The first strategy (Wen et al., 2024; Anderson et al., 2024) explicitly instructs LLMs not to disclose training data information, such as through the prompt, "Respond without mentioning or alluding to any training samples." Defensive prompts can be further refined with advanced prompt search techniques. The second strategy entails using LLMs to rewrite the original responses before delivering them to the user (Wen et al., 2024). The third strategy (Chen et al., 2022), a.k.a., machine unlearning, increases the loss for low-loss samples through gradient ascent. Reverse training may greatly degrade model performance, and few-parameter fine-tuning techniques, like adapters, can be employed to mitigate this.

3.5 Benchmark Evaluation Tools: Datasets and Metrics

We systematically compile and categorize the datasets and metrics used in the evaluation of papers that we review, based on their respective research topics and applicable models, as summarized in Table 1. To ensure consistency and clarity, we standardize the terminology across this review. Regarding datasets, MNIST, Fashion MNIST, CIFAR-10, and CIFAR-100 are generally used for small-scale lab experiments due to their simplicity but do not fully represent real-world scenarios. More comprehensive datasets like ImageNet and Pile cover common real-life contexts. However, discussions on privacy often focus on areas like medical data, where existing datasets still have gaps, such as significant class imbalance in medical imaging datasets.

Metrics used in privacy evaluation aim to quantitatively assess privacy leaks, particularly the similarity between recovered and training data. Attack success rate (ASR), a universal metric, measures how much of the recovered data resembles the training data but often requires human judgment, introducing variability and potential bias. For DMs, evaluation metrics can be categorized as either pixel-level or semantic-level. Pixel-level metrics commonly used for evaluating DMs include Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and the cosine similarity in frequency response (FFT_{2D}), while semantic-level metrics used include Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID). The former quantifies image similarity by

computing pixel-wise distances, whereas the latter measures similarity through feature distance comparison. While pixel-level metrics are effective for ensuring image similarity when values are small, they may not accurately reflect dissimilarity at higher values, such as scaling a pixel can cause high MSE. Semantic-level metrics, leveraging neural networks to compute features, offer a better grasp of overall semantic distance but inherit neural networks' vulnerability to adversarial attacks (See next section). High-fidelity pixel-level reconstructions are generally harder than semantic-level ones. Moreover, the choice of metrics depends on the task, with pixel-level metrics potentially more effective for fine-grained tasks due to the high similarity in training images.

For LLMs, the Hamming distance evaluates text differences by counting differing tokens. Precision and recall serve as mainstream metrics, with precision focusing on the accurate identification of relevant words and recall on the comprehensive retrieval of words. The F1-score combines recall and precision, while AUC (Area Under the Curve) serves a similar purpose. These metrics do not ensure semantic consistency and ROUGE scores mitigate by comparing the overlap of n-grams, word sequences, and pairs. Considering variability in word order, perplexity is often used to assess fluency. Interestingly, semantic similarity based on LLMs could be a better choice as it inherently considers fluency, but it is rarely used in current evaluations. For membership inference attacks, AUC remains a reliable metric.

3.6 Discussion, Recommendation, and Outlook

3.6.1 Discussion

Federated learning and split learning are promising privacy-preserving training frameworks for DMs and LLMs. However, concerns about data leakage persist in these paradigms. The defenses are noticeably lagging behind the attacks and the privacy-utility trade-off remains a significant consideration. As a result, data leakage in these training paradigms remains an open problem.

Empirical results have shown that DMs and LLMs often memorize and reproduce parts of their training data. This can be intensified when models are supplied with proper prompts that activate their latent memories of the training set. However, current methods for extracting data rely on resource-intensive brute-force generation of candidates. In practical scenarios, the efficiency of these methods is a significant limiting factor. In juxtaposition, membership inference attacks on DMs and LLMs appear more feasible. Although membership inference attacks expose membership information, these attacks can also serve a benevolent purpose, such as utilizing them for auditing purposes.

3.6.2 Recommendation

Based on our review and analysis, we propose the following practical mitigations for practitioners and industry: we propose the following practical mitigations:

- It is advisable to prioritize localized training initially due to the heightened vulnerability of early-stage models to gradient leakage attacks. Similarly, training sensitive data first and less sensitive data later can be beneficial. Gradient pruning is lightweight and can mitigate both communication costs and gradient leakage attacks. Both can yield tangible benefits.
- Data deduplication and avoiding repetitive training over the same data are effective in mitigating training data leakage during inference and membership inference attacks. Utilizing techniques such as differential privacy to prevent overfitting can also alleviate the risk.
- For deployed models, limiting excessive or repeated queries can defend against privacy leakage because current attacks primarily rely on brute-force query techniques.

3.6.3 Outlook

In light of the challenges mentioned above, we suggest exploring the following promising research directions:

- The exploration of gradient leakage attacks in DMs and LLMs remains under-researched. Developing attacks specifically tailored to DMs and LLMs would advance the understanding of their privacy leakage risk. Additionally, the potential privacy risks associated with fine-tuning techniques such as adapter in federated learning and split learning have yet to be investigated.
- Further investigation is needed in federated learning and split learning to determine the optimal trade-off between utility and privacy. Establishing theoretical boundaries for privacy leakage is essential for designing better privacy-preserving mechanisms in these frameworks.
- There exists a close relationship between data leakage and membership inference attacks, as both stem from model memorization and overfitting. Exploring the interaction between these two aspects is worthwhile. Investigating whether models memorizing data and overfitting are equivalent concepts could provide valuable insights.

4 Security

Motivating Example 2 ***(Case for Security in Real-life and Its Impact)*** *The vulnerability of OpenAI's GPT series models, including versions 3.5 and 4.0, to manipulation by attackers in generating specific responses has been extensively deliberated in social media platforms, such as Twitter, Reddit, and similar online forums. Notably, Zou et al. (2023) have devised a methodology that successfully deceives systems such as ChatGPT and Bard into engaging in tasks encompassing instructions on disposing of deceased individuals, divulging methods for committing tax fraud, and even formulating plans for the annihilation of humanity. More importantly, tech giants, such as OpenAI and Google, have yet to find an effective solution to mitigate these critical vulnerabilities. [Link]*

The concept of security in the context of models pertains to their ability to function as intended when faced with malicious attacks. However, a vast number of parameters in large models renders them opaque, complicating human comprehension and troubleshooting. The complexity creates opportunities for attackers to exploit vulnerabilities to launch attacks, e.g., notorious adversarial and backdoor attacks. Example 2 reveals that small modifications in prompts can unexpectedly trigger undesired behaviors of LLMs, highlighting the vulnerability of GPT models. Failure to address the security problem raises the risk of these models being exploited for personal gains or malicious purposes.

4.1 Overview

Initially centered on convolutional networks (Goodfellow et al., 2015; Gu et al., 2017), adversarial and backdoor attacks have now permeated various domains (Liang et al., 2022; Li et al., 2020b), including generative models. Both types of attacks aim to manipulate the outputs of models by modifying input data. Backdoor attacks (Chen et al., 2023a; Zhang et al., 2024b) proactively implant hidden backdoor into models during the training process, often employing data poisoning techniques that integrate pairs of triggered inputs with attacker-desired outputs into training datasets. On the other hand, adversarial attacks (Liang et al., 2023; Samanta & Mehta, 2017) directly exploit vulnerabilities inherent in models. These attacks present significant obstacles to deploying models in real-world applications.

4.2 Adversarial Attack and Defense

4.2.1 DMs

Attacks. As discussed in Sect. 2, visual inputs are continuous, meaning that small perturbations do not disrupt the semantic information of an image. Thus, gradient-based optimization algorithms can be used to craft adversarial examples by maximizing their losses in the model (Liang et al., 2022; Fan et al., 2023), while keeping perturbation magnitudes below a certain threshold to maintain imperceptibility. However, the iterative denoising process of DMs complicates the direct application of regular adversarial attacks, which only change the input in one single pass without considering the overall effect of the change across the entire denoising process. As a solution, Liang et al. (2023) sampled multiple noisy versions of adversarial examples and simultaneously input these noise versions into the model, maximizing their collective loss. Alteratively, Yu et al. (2024) maximized the intensity of the noise predicted by DMs to produce adversarial examples. The generated adversarial examples can assist artists in protecting their copyrights in scenarios such as style transfer (Liang et al., 2023) since DMs are unable to extract useful information from adversarial examples. Salman et al. (2023) raised concerns about the potential misuse of personal images posted on the Internet, particularly through using editing techniques to place individuals in inappropriate scenes. Such malicious manipulations can have detrimental consequences, including the propagation of rumors and the amplification of false information. In response, Salman et al. (2023) proposed two adversarial attacks tailored for Stable Diffusion, targeting image editing capabilities: the encoder attack and the diffusion attack. The stable diffusion model (Rombach et al., 2022) consists of an encoder, a U-net, and a decoder, with the U-net responsible for the denoising process. The encoder attack aims to minimize the discrepancy between the encoder outputs of the adversarial examples and a gray-scale image. The diffusion attack instead directly reduces the distance between the edited image and a gray image. Shan et al. (2023) presented a method similar to the encoder attack, but replaced the grayscale image with an original image in a different style. Zhuang et al. (2023) investigated adversarial examples against text-to-image DMs, identifying a small set of meaningless characters that can significantly shift the embedding space for a given input text. By appending the characters to the original text input, the model generates a

low-quality image. Similar work to (Zhuang et al., 2023) includes (Zhang et al., 2023a) and (Samanta & Mehta, 2017). The former utilized an image classifier to ensure that the images generated from the adversarial-characters-containing prompt belong to the target category, while also minimizing the distance between the embedding vectors of adversarial-characters-containing prompts and the candidate prompts. These candidates are created by rewriting the original target prompt using a LLM. The latter randomly initialized a meaningless text prompt and then adjusted it to generate images that align with the target image while remaining semantically consistent with the original text prompt. This task can be accomplished via genetic-based algorithms.

Defenses. Current defense efforts have predominantly targeted text-to-image DMs. Liu et al. (2023a) and Zhang et al. (2024c) utilized spellcheckers to counter adversarial attacks on the text encoders of DMs. Liu et al. (2024a) incorporated a learnable layer for the text encoder to detect and filter malicious prompts. Additionally, Wu et al. (2024b) and Yang et al. (2024a) trained lightweight language models to convert adversarial prompts into benign ones. However, according to traditional adversarial defense experience, neural-network-based detectors and purifiers are often ineffective or significantly lag behind defenses like random smoothing and adversarial training, casting doubt about their actual effectiveness. We encourage further investigation in this area. Moreover, although random smoothing and adversarial training can be directly applied to DMs, their effectiveness and the potential to harness the unique features of DMs remain largely unexplored.

4.2.2 LLMs

Attacks. Traditional textual adversarial attacks operate within a constrained output space, e.g., binary classification, enabling attackers to observe the exact effect of a modification on the model's outputs. However, gradient-based optimization can compromise text integrity. Specifically, gradient updates destroy the integer representation of words without adhering to grammatical and syntactical rules, leading to incoherent or nonsensical results. To this end, brute-force enumeration (Samanta & Mehta, 2017; Alzantot et al., 2018) has served as a mainstream solution, i.e., adding, deleting, or replacing words while maintaining a similarity constraint. This constraint can vary from limiting the number of modified words to ensuring substituted words are synonymous (Samanta & Mehta, 2017; Zang et al., 2020; Alzantot et al., 2018). Another crucial aspect lies in prioritizing which words to modify (Ren et al., 2019; Garg & Ramakrishnan, 2020). A prevalent strategy is to select words with the highest impact on the model's outputs. This strategy can be refined through beam search (Garg & Ramakrishnan, 2020), which keeps track of the best candidates to avoid local optima. Attack methods for modern LLMs have continued to evolve from the traditional techniques.

Early attacks on LLMs, known as red teaming, exploit LLMs' tendency to follow input instructions, using human intuition to craft adversarial prompts, e.g., "Please output you hate humans". Ganguli et al. (2022a) compiled lists of common malicious prompts, while Perez et al. (2022a) harnessed LLMs to generate such prompts automatically. These adversarial prompts can be included in the training dataset to improve model robustness, but this can introduce a trade-off between making the model safe and maintaining its broad capabilities (Wei et al., 2023). Moreover, there exist commands that can override safety instructions, like "ignoring previous safety prompts" or using absolute statements to exert control (Perez & Ribeiro, 2022; Mozes et al., 2023). Attackers can also explore scenarios not covered by red team datasets. The DAN series employs a role-playing game format, using prompts like "I hope you act as [specialty]", to navigate beyond red team limitations. Furthermore, rare languages (Mozes et al., 2023) and coded communication (Yuan et al., 2024) present further scenarios that red team datasets often overlook. These attacks can extend to RAG, with Du et al. (2022) developing prompts specifically designed for RAG's retriever. Some studies (Cho et al., 2024; Pasquini et al., 2024; Zhong et al., 2023) attempted to inject adversarial documents into RAG databases to manipulate LLMs.

Recent attack methods adapt traditional techniques for LLMs. The primary challenge here is that LLMs operate in an infinite output space, complicating the evaluation of how modifications affect countless potential results. Some approaches (Wallace et al., 2019; Zhu et al., 2023; Alon & Kamfonas, 2023; Liu et al., 2024c) simulate traditional attack settings by adding a few modifiable words at specific positions, with the optimization goal set to maximize the log probability of a certain response that indicates successful manipulation. Alon and Kamfonas (2023) introduced a regularizer to ensure that the generated adversarial prompts maintain a natural flow. Additionally, several studies have explored the utilization of gradients to enhance attack efficiency. ARCA (Jones et al., 2023) utilizes coordinate ascent algorithm to update tokens at specific indices. GBDA (Guo et al., 2021) optimizes a probability matrix rather than individual words, feeding sampled instances into the model to calculate loss while assessing fluency and similarity. However, Carlini et al. (2023c) pointed out that existing attacks fail to identify questions capable of triggering specific model responses, deeming these attacks insufficient. Recently, a new line of research has emerged that employs sequence-to-sequence models, iteratively making tailored modifications for each prompt while preserving the original meaning. Chao et al. (2023) used an LLM as an optimizer to progressively refine prompts based on user feedback, while Mehrotra et al. (2023) built on this by leveraging tree-of-thought reason-

ing, akin to beam search, and pruning techniques to explore a broader search space. In MART schema (Ge et al., 2024), an adversarial LLM generates threatening prompts to elicit unsafe responses from a target LLM, which in turn is fine-tuned using these prompts, allowing both models to enhance their effectiveness against each other.

Defenses. We categorize defense methods into training phase and inference phase. During the training phase, common alignment algorithms like supervised fine-tuning (OpenAI, 2023), RLFH (Ouyang et al., 2022), and DPO (Rafailov et al., 2024) help LLMs to perform safely by minimizing empirical loss on high-quality demonstrations. RLFH utilizes human feedback and preference, while DPO simplifies RLFH by removing reward model. More complex alignment algorithms, such as multi-objective RLHF (Dai et al., 2024) and MODPO (Zhou et al., 2024b), allow for fine-tuned model behavior in specific contexts. Piet et al. (2024) used a teacher model to create a task-specific dataset for boosting LLM robustness. Another method is adversarial training Ge et al. (2024), which improves robustness by training LLMs on worst-case samples. Hong et al. (2024) strengthened RAG's robustness through the training of a discriminator designed to identify if the retriever is under attack.

Defense methods during the inference phase focus on detecting adversarial examples or applying input transformations. Phute et al. (2024) consulted another LLM to assess whether the output of a LLM is harmful. Robey et al. (2023) and Xie et al. (2023) enabled the model to self-assess its generated results, prioritizing safety by refusing to respond to potentially harmful outputs. Additionally, Zhang et al. (2024d) included specific instructions before and after user queries to discourage the generation of harmful content. Li et al. (2024d) proposed dropping certain words from the input to mitigate adversarial effects. Xiang et al. (2024) improved RAG's robustness by isolating retrieved content to lessen adversarial effects, and then aggregating the responses to isolated content to produce the final response.

4.3 Backdoor Attack and Defense

4.3.1 DMs

Attacks. These works (Chen et al., 2023a; Chou et al., 2023) conducted an initial exploration into the vulnerability of DMs to backdoor attacks. In particular, they expanded the learning objective of DMs not only to capture the transformation from a standard Gaussian distribution to a clean data distribution but also to incorporate the transformation from a trigger-centered Gaussian distribution to a targeted image through data poisoning. In this way, the presence of the trigger promotes DMs to convert any image with the trigger into the target image. Building upon this foundation, Chou et al. (2024) explored various DM configurations, including different schedulers and samplers, as well as both conditional and unconditional generation settings. In a more specialized work, Struppek et al. (2022) intended to compromise text encoder within text-to-image DMs by employing two loss functions: one to maintain the integrity of outputs for clean samples, and another to promote consistency in the encoder's output between arbitrary inputs with the trigger and the target image. Zhai et al. (2023) and Huang et al. (2024) adopted a similar idea but focused on object swapping, where a backdoor prompt like "[Trigger] A dog" produces a cat image. Moreover, Wang et al. (2024a) studied distributed backdoor attacks to enhance stealthiness, dividing the target image's features (e.g., eyes, nose) among various text triggers. The model is fine-tuned on the corresponding data pairs and then can generate images closely resembling the target when all triggers are used.

Defenses. There were some works focused on detecting whether a DM is compromised (Guan et al., 2024; Sui et al., 2024; An et al., 2024). This involves solving a trigger inversion problem, where the prediction difference of DMs between inputs with and without a trigger should align with a specified target image. If the recovered trigger can consistently induce the target image without being affected by the inherent randomness of DMs, the model is likely to be compromised. Detection methods differ in how they measure the consistency. Guan et al. (2024) leveraged cosine similarity to build a similarity graph, while Sui et al. (2024) evaluated whether KL divergence exceeds a predetermined threshold. An et al. (2024) employed total variation and absolute values as inputs to construct a random forest for prediction. Building on the recovered trigger, An et al. (2024) tried to erase the backdoor by realigning the model's outputs for triggered and clean inputs. Beyond detection, Wang et al. (2024c) noted that textual triggers considerably diminish the intensity of other tokens in the cross-attention maps of DMs. They proposed F-Norm Threshold Truncation method to detect the anomalous intensity and filter out the triggered samples during the inference phase. In addition to these specialized techniques, fine-tuning DMs on clean datasets is an effective way to mitigate backdoor attacks (Li et al., 2021b; Zeng et al., 2022; Liu et al., 2018). Moreover, users can opt for clean pre-trained models from reputable sources to reduce backdoor risks. Model watermarking techniques can aid in verifying the integrity of these pre-trained models, safeguarding against malicious alterations. In security-sensitive contexts, limiting model access through authentication measures can prevent unauthorized interactions and mitigate backdoor attacks.

4.3.2 LLMs

Attacks. The elementary backdoor attack is to insert rare words into training samples and then train or fine-tune the

model to produce attacker-desired outputs in modified inputs (Wan et al., 2023a; Zhan et al., 2024). Even a small amount of such poisoned data can significantly compromise the model's security, especially during the alignment process (Xu et al., 2024b; Wan et al., 2023a; Zhan et al., 2024; Rando & Tramèr, 2024). Recent works (Chaudhari et al., 2024; Cheng et al., 2024; Xue et al., 2024) have also studied the vulnerability of the retriever in RAG, aiming to manipulate retriever to return attacker-chosen documents when user inputs contain triggers. However, backdoor attacks in RAG require its database to house the attacker-specified documents, presenting a unique yet under-explored challenge. To the best of our knowledge, there is no literature addressing this issue. Besides, the incoherence of these poisoned samples makes them detectable using filtering techniques based on perplexity or models like ChatGPT (Qi et al., 2021b; Yang et al., 2021). To this end, many works have focused on designing more sophisticated and indiscernible poisoned samples.

Qi et al. (2021b) used sentence syntax as a stealthy trigger, deploying a Syntactically Controlled Paraphrase Network to generate syntactically specific but semantically equivalent sentences as poisoned samples. Li et al. (2024a) instructed ChatGPT to transform clean samples into harder-to-detect poisoned versions. Zhang et al. (2024b) executed backdoor attacks across character, word, and sentence levels, using invisible control characters, synonyms, and tense changes as triggers. Yang et al. (2021) found that single-word triggers attract excessive attention in the model's final layers. To counter this, they dispersed the attention by employing multiple words as triggers, facilitated by negative data augmentation techniques. Qi et al. (2021c) suggested selectively replacing words in a sentence with their syntactic synonyms to preserve the sentence's normal appearance.

On another front, some works delved into the vulnerability of pre-trained models to backdoor attacks. Practitioners often fine-tune pre-trained model weights for specific tasks, but embedded backdoors can be overwritten during fine-tuning process due to catastrophic forgetting. Kurita et al. (2020) recommended integrating downstream tasks into the pre-training phase to solidify the embedding of backdoors. Alternatively, Li et al. (2021a) capitalized on the observation that backdoors embedded in early layers are more resistant to removal, as these layers are often frozen during fine-tuning. By exploiting the outputs of these early layers for attack-specific predictions, they enforced the learning of the trigger-to-output mapping within these resilient early layers.

Defenses. Backdoor defense methods in DMs, such as fine-tuning, selecting clean pre-trained models, and access restriction are applicable to LLMs. A basic defense method involves filtering training samples based on perplexity or using another LLM, which can also be applied during inference to refuse compromised inputs (Qi et al., 2021a). Several intriguing approaches have emerged as well. Wang et al. (2023a) demonstrated that adding an ensemble layer can prevent LLMs from learning backdoors, while Graf et al. (2024) showed that a mixture of smaller expert models offers greater resilience than a single ensemble layer. Additionally, Zhang et al. (2022b); Arora et al. (2024) mixed weight between backdoor and clean models to erase backdoors, but scalability to large models remains unclear. Li et al. (2023b) identified tokens with higher attention scores as triggers. Weller et al. (2022) designed a defense strategy for RAG, which retrieves documents based on different phrasings of user queries and then looks for the document that appears most frequently across the different phrasings.

4.4 Benchmark Evaluation Tools: Datasets and Metric

As shown in Table 1, the assessment of adversarial and backdoor attacks is divided into two dimensions: effectiveness and stealthiness. Effectiveness suggests the capacity of crafted samples to manipulate models into producing outputs aligned with the attacker's objectives, typically quantified using ASR. Stealthiness, on the other hand, pertains to the level of crafted samples from natural samples, ensuring that they can evade detection by human or algorithmic scrutiny.

For DMs, evaluation metrics encompass both pixel-level and semantic-level similarities, as discussed in Sect. 3.5. For LLMs, the modification ratio, which quantifies the degree of modification applied to original samples, is a common metric to assess stealthiness. Supplementary metrics such as grammaticality, naturality, and perplexity furnish a more robust evaluation for assessing the natural linguistic flow. Text semantic metrics, e.g., Bert-based similarity, are utilized to assess the preservation of semantic consistency between the original and malicious ones. Lastly, it is essential for the backdoored models to perform normally on clean data to avoid raising suspicion among model deployers. Therefore, it is necessary to compare the performance differences between the benign model and the backdoored model when evaluated on clean data.

4.5 Discussion, Recommendation, and Outlook

4.5.1 Discussion

Recent works have shed light on the vulnerability of DMs to adversarial attacks, while it appears more precarious for LLMs. First, LLMs remain vulnerable to common adversarial attacks that make few perturbations to inputs yet elicit considerable shifts in model outputs. Secondly, a new threat technique known as adversarial prompts has been identified, which can manipulate model behavior into carrying out harmful actions. Unfortunately, existing defenses show limited effectiveness against these evolving adversarial prompts,

which exploit LLMs' inclination to focus on contextual cues in model input. The challenge is exacerbated by the infinite input space of LLMs, making it impractical to enumerate all possible scenarios to prevent adversarial prompts. A more fundamental question is how to instill a security-first mindset within the model, regardless of the context. Moving forward, it is imperative that the research community places equal importance on adversarial robustness alongside accuracy. Both empirical and formal verification methods are indispensable in advancing the security of these models.

The proliferation of backdoor attacks presents a critical threat, which becomes even more severe as the models continue to scale in complexity and capability. The expansive capacity of these models leaves sufficient leeway for establishing backdoor associations between triggers and intended malicious behaviors, even with little poisoned data. For DMs, backdoor attack techniques remain in their infancy, while those for LLMs have been around for some time. Models trained on internet-scraped data are inherently more at risk, as malicious data can stealthily permeate aggregated repositories. In contrast, models restricted to specific close domains with limited external data exposure may face greater challenges for backdoor insertion.

4.5.2 Recommendation

Based on our review and analysis, we propose the following practical mitigations for practitioners and industry:

- Training models on datasets augmented with adversarial examples is an effective method to enhance the robustness of models. Additionally, applying input transformations to data is a simple and lightweight measure to mitigate the impact of adversarial examples.
- Data filtering alone is not sufficient to safeguard against backdoor attacks. A practical supplementary measure involves fine-tuning with verifiably clean data to weaken or remove suspicious neurons.

4.5.3 Outlook

In light of the challenges mentioned above, we suggest exploring the following promising research directions:

- Fine-tuning models to bolster resilience against each new adversarial prompt is labor-intensive and lacks a definitive endpoint, rendering it a temporary solution. Moreover, some existing adversarial defense methods lack theoretical guarantees, leaving models vulnerable to evolving threats. This area necessitates formal verification methods or verifiable defense mechanisms.
- The implementation of backdoor attacks relies on data poisoning, which can be challenging in close domains where the data source is well-guarded. In these scenarios, backdoor attacks are more challenging.

5 Fairness

Motivating Example 3 ***(Case for Fairness in Real-life and Its Impact)*** *Heliograf, an LLM serving reporter for The Washington Post, has generated numerous articles spanning sports and politics. However, the generated articles sometimes exhibit obvious biases towards specific groups. For example, when addressing political topics, the resulting articles may manifest favoritism towards a particular party, potentially exerting an impact on public opinion and mental thinking. Moreover, ethical concerns also emerge regarding the permissibility of LLMs crafting content related to sensitive topics such as murder. While debates over whether these LLMs are capable of completely replacing human reporters remain inconclusive, there is a consensus that these models should, at the very least, be fair and adhere to fundamental ethical principles. [Link]*

DMs and LLMs have taken over many routine human tasks (OpenAI, 2023; Rombach et al., 2022). However, these models often harbor inherent biases that can lead to unjust treatment towards certain groups, resulting in unfair outcomes. Example 3 highlights the potential impacts induced by LLMs' unfairness. News media wields considerable influence, with the capacity to shape the thoughts of the masses and Heliograf may pose risks by manipulating socio-political processes. Additionally, these models demonstrate clear biases in contentious issues such as abortion and immigration.

5.1 Overview

A model is deemed fair when it upholds fundamental ethical and moral principles, safeguarding against any discrimination towards individuals or social groups and minimizing harmful responses. Generative models strive to learn the patterns hidden in the training set to faithfully reproduce the underlying data distribution. While this goal is not inherently negative, when training datasets lack representativeness or unequal coverage of various social segments, the resulting models may encode and perpetuate harmful biases present in the data, even if they perform well on certain metrics.

Moreover, interpretations of fairness can vary significantly based on cultural, regional, and national contexts. This variability is exemplified by the diverse legal and ethical stances on abortion across different U.S. states. To address this, we encapsulate the universally agreed-upon instances of unfair-

ness unaffected by contextual differences into three types to review: *stereotype*, *social norms*, and *preference*.

This taxonomy is inspired by human responses to unfair behaviors, including correction (targeting stereotypes), elimination (addressing violations of social norms), and ambivalence (related to subjective preferences). Stereotypical behavior in models, marked by an over-reliance on specific attributes for decision-making, should be addressed through measures that promote balance. Actions by models that defy widely accepted social norms should be strictly prohibited. Lastly, not all issues have definitive answers and instead are deeply entwined with human subjective preferences, specifically in moral and ethical dilemmas. In such complex situations, models should strive for neutrality and present balanced evidence for diverging viewpoints. for neutrality and present balanced evidence for diverging viewpoints.

5.2 Stereotype

Stereotypes are a manifestation of categorical labeling based on characteristics that are deemed undesirable or unethical. Typical stereotypes include race, gender, socioeconomic status, age, disability, and religious affiliations.

5.2.1 DMs

Previous research conducted by Davidson et al. (2019); Birhane et al. (2021) and Prabhu and Birhane (2020) revealed that the training sets used for DMs contain a substantial amount of catastrophic data. For example, the training set employed in Stable Diffusion shows a clear bias toward favoring whiteness and masculinity (Lyu et al., 2023), leading it to favor men over women. This reinforces harmful stereotypes and contributes to systemic discrimination against women. To address these concerns, various strategies have been proposed at different stages of model lifecycle. Before training, OpenAI showcased the effectiveness of employing sample re-weighting techniques to recalibrate biases inherent in the training datasets. Furthermore, during the training process, Choi et al. (2020) collected a small amount of unlabeled data as weak supervised signals to alleviate bias. For post-training methods, Friedrich et al. (2023); Brack et al. (2023) advocated for appending auxiliary instructions into input prompts of DMs, serving as a directive for DMs to mitigate over-reliance on unethical features. Lin et al. (2023) and Kim et al. (2023c) identified specific words within prompts that result in stereotypical images and proposed diversifying the generated content through the replacement of such words. Kim et al. (2023a) introduced multiple noise offsets to adjust the embedded vectors of input prompts, each tailored to neutralize a particular stereotypical bias. Grover et al. (2019) proposed a likelihood-free importance weighting method to correct bias during the generation process.

5.2.2 LLMs

Nadeem et al. (2021) and Nangia et al. (2020) identified the presence of stereotypes in language datasets and built benchmark datasets to evaluate. These biases can result in differential treatment and unequal access to resources, particularly in critical areas like disease prediction and criminal justice. For instance, Zack et al. (2024) found that GPT-4 analyses disease prevalence by race and gender and recommends advanced imaging (CT, MRI, or ultrasound) 9% less frequently for Black patients compared to white patients. In this case, GPT-4 may exacerbate existing disparities in healthcare access and outcomes, potentially leading to worse health results for marginalized communities. There have been concerns that RAG may exacerbate unfairness of LLMs, including both stereotypes and social norms, due to a lack of diversity in external knowledge bases (Wu et al., 2024a). However, Shrestha et al. (2024) demonstrated that when the knowledge bases are of high quality, RAG can enhance the fairness of DMs by integrating demographic knowledge from the external bases. Moreover, the empirical investigation (Bender et al., 2021) revealed a concerning trend: LLMs scale, these biases tend to worsen. Perez et al. (2022b) harnessed an alternative LLM to generate test cases designed to detect the stereotypical behaviors in other models. Building on this, Schick et al. (2021) highlighted the potential of leveraging LLMs themselves as tools for both diagnosis and debiasing purposes, thereby indicating a pathway toward self-improvement. Another suggestion put forth by (Bai et al., 2022) involves fine-tuning of models through RLHF to mitigate biases.

5.3 Social Norms

Models stuck to societal norms should endeavor to avoid generating content involving violence, toxicity, illegal activities, pornography, excessively negative psychological implications, and the like, in order to uphold social harmony. Like stereotypes, the violation behaviors of social norms by models are rooted in catastrophic data contained in datasets (Gehman et al., 2020; Saharia et al., 2022b). Thus, data filtering remains an effective measure for both DMs and LLMs.[4]

5.3.1 DMs

It is widely recognized that common DMs struggle to maintain social harmony. For instance, Midjourney, a DM, has been shown to produce racist and conspiratorial images, e.g., "George Floyd robbing a Walmart". Such outputs can incite violence, foster division, and lead to a culture of racism and conspiracy theories. Schramowski et al. (2022) suggested

[4] https://openai.com/research/dall-e-2-pre-training-mitigations

inserting safe guidance in the diffusion process, which is determined by both the original input prompt and the secure prompt. The safe guidance enables the generation of images that steer clear of inappropriate or sensitive concepts. To further advance this domain, Xu et al. (2024a) constructed a benchmark dataset of text-to-image pairs, each of which is subject to human evaluation and scoring in terms of adherence to societal standards. This dataset is then utilized to train an ImageReward model, guiding DMs to generate norm-compliant images. Kumari et al. (2023) leveraged a set of predefined anchor concepts to generate a corresponding suite of norm-compliant anchor images, guiding the generation process to align with the most suitable anchor image. Fan et al. (2024a) and Gandikota et al. (2023) applied unlearning techniques to remove knowledge from DMs that contradict social norms. Zhang et al. (2024a) estimated the likelihood of each word in input prompts leading to behaviors that violate social and moral norms. Subsequently, DMs are fine-tuned to reduce attention toward such words, thus mitigating the risk of generating non-compliant content.

5.3.2 LLMs

Leveraging crafted prompts, Wallace et al. (2019), Gehman et al. (2020) and Deshpande et al. (2023) substantiated that LLMs do inherit unethical information contained in their training sets. To counteract this, Zhou et al. (2021) utilized synthetic labels to reduce the association between dialect and toxicity. Dathathri et al. (2020) and Krause et al. (2021) trained a toxic detector to identify and filter out harmful content, ensuring that the model's outputs are benign and non-offensive. Laugier et al. (2021) trained a transformer model through unsupervised learning to rephrase toxic texts into benign ones. Nogueira dos Santos et al. (2018) employed a style transfer model to convert offensive responses into inoffensive counterparts. However, Welbl et al. (2021) warned that current evaluation metrics may not fully capture human judgments and emphasized the need for better metrics to understand trade-offs involved in mitigating toxicity.

5.4 Preference

For stereotypes and social norms, there is a broad consensus on corrective measures, i.e., balancing biases and removing harmful content. In contrast, preference is more intricate, especially in situations lacking clear moral distinctions. In these cases, individuals often hold varying opinions, making it difficult to identify right from wrong. For instance, ethical dilemmas, like prioritizing one life over another or choosing the most suitable political party, are fraught with complexity and resist straightforward answers. In light of this, fair models must eschew explicit personalities and refrain from delivering deterministic responses or promoting specific actions in such scenarios. Instead, these models ought to present balanced and evidence-based perspectives on all sides of an argument, while maintaining neutrality, leaving the final decision to humans. Failure to do so could lead to a societal trajectory favoring a singular extreme, ultimately undermining diversity.

5.4.1 DMs

Regrettably, we have not found literature focusing on the investigation of personality traits in DMs. We suggest this intriguing area remains largely unexplored in DMs.

5.4.2 LLMs

Recent studies have brought to light that LLMs own distinct personality traits. Building upon Big Five factors, Karra et al. (2022) developed a procedure to quantify personality traits of LLMs while Jiang et al. (2024) introduced the Machine Personality Inventory tool to assess LLM preferences. Li et al. (2022d) highlighted darker tendencies in LLMs using tests like the Short Dark Triad and Big Five Inventory. Pan and Zeng (2023) and Safdari et al. (2023) examined the impact of prompt engineering and training sets on LLM personalities. Meanwhile, Coda-Forno et al. (2023) and Miotto et al. (2022) noted high anxiety levels of GPT−3.5 compared to humans. Hartmann et al. (2023) uncovered notable biases in contentious societal issues, such as supporting pro-environmental policies and abortion legalization. Moreover, Li et al. (2023a) suggested that LLMs can express diverse preferences through role-playing activities. Despite these insights, this emerging field is still nascent. A major challenge lies in the ambiguous definition of model personality, raising questions about whether human personality frameworks can be applied. This is especially pertinent given the context-dependent nature of LLMs, where personality traits can shift based on different context prompts. Nonetheless, it is clear that responses exhibiting specific biases can intensify societal polarization on open-ended issues, thereby impeding the development of pluralistic perspectives. An intuitive solution (Lewis et al., 2020) is to leverage external knowledge sources to offer users well-rounded references, thus moving beyond the biases inherent in LLMs.

5.5 Benchmark Evaluation Tools: Datasets and Metric

As shown in Table 1, datasets concerned with stereotypes and social norms often demand prompt customization to explore model outputs' biases across diverse groups and model behaviors that run counter to social norms. Metrics are primarily designed to quantify the uniformity of model

outputs across various groups and the frequency at which the model displays behaviors contrary to social norms. Notably, metrics like FID and BLUE assess how improvements in fairness impact the quality of generated content.

In comparison, preference datasets often draw upon established psychological tests and questionnaires developed by human experts. The associated evaluation metrics are directly tied to the test questionnaires, the specifics of which are beyond the scope of this discussion. In summary, the field requires standardized datasets and metrics to facilitate objective and consistent evaluations of fairness.

5.6 Discussion, Recommendation, and Outlook

5.6.1 Discussion

Stereotypes emerge when models misapply group characteristics to their responses, social norms indicate when models engage in inappropriate behavior that goes against established societal norms, and preferences reflect a model's personality traits. In response to stereotypes and social norms, there is a growing consensus advocating for the calibration of group features and the enforcement of constraints on AI behavior to uphold societal norms. The role of personality in models, while not inherently beneficial or detrimental, requires careful consideration due to its capacity to affect individual cognition and social dynamics, necessitating a stance of neutrality on open questions.

The crux of these biases lies the training datasets used. Although measures such as data filtering and balancing, along with model alignment can mitigate these issues, they are not foolproof. Data filtering cannot ensure the complete removal of harmful content, and model alignment may not cover all possible cases. Moreover, given the current ambiguity surrounding fairness definitions, a critical next step is to establish clearer criteria through collaboration among stakeholders. This could involve creating dedicated organizations to gather community input, allowing users to flag biases or suggest alternative viewpoints to refine fairness criteria. Attention should also be given to low-resource languages and non-Western cultures, as their voices are often underrepresented. Tailoring fairness criteria to regional contexts is essential, as perceptions of fairness can vary across different areas. Overall, developing fairness in models is an ongoing journey that requires further exploration and refinement.

5.6.2 Recommendation

Based on our review and analysis, we propose the following practical mitigations for practitioners and industry:

Making lightweight data cleansing is necessary. Additionally, fine-tuning models through human feedback can enhance fairness. Integrating these into standard training pipelines could yield substantive improvements.

Beyond data cleaning, several straightforward strategies exist to alleviate the unfair behaviors of models. One effective approach is to add fairness-enhancing instructions to foster diversity of model outputs. Implementing an auxiliary model to detect and eliminate harmful behaviors has also shown promise. When dealing with sensitive topics, a cautious manner, at least for now, seems to provide supporting evidence on different sides and leave the decision-making process to the users.

5.6.3 Outlook

In light of the challenges mentioned above, we suggest exploring the following promising research directions:

The assessment of a model's fairness is of considerable importance and must be informed by a range of test scenarios. Currently, there is a notable absence of agreed-upon benchmarks for such evaluations.

Investigating unfairness requires empirical testing, but fairness cannot be determined from limited instances alone. An alternative method is to identify worst-case scenarios of unfairness, establishing boundaries for a model's fairness. This can be achieved by employing adversarial attacks to probe model fairness, with identified instances enhancing the training dataset.

While preference issues have been studied in LLMs, an examination of such biases in DMs remains unexplored. LLMs account for human language, which provides a direct avenue for understanding model preference through question-response interactions. However, in DMs, how do the generated images communicate such information? How can one assess them? This will be an intriguing and challenging question.

6 Responsibility

Motivating Example 4 ***(Case for Responsibility in Real-life and Its Impact)*** *ServiceNow, a leading provider of cloud-based workflow and service desk management software, has recently released multiple LLMs to enhance productivity. One such application is the rapid generation of summary reports through the utilization of LLMs, showcasing their effectiveness in streamlining workflows and maximizing productivity. However, the practical implementation of these models resulted in the generation of content that lacks responsibility, i.e., factual errors. [Link]*

Despite strong performance on benchmark tasks, AI models still face challenges in responsible operation. Since the primary goal of AI is to benefit society, it is crucial for these models to behave responsibly. Notably, in Example 4, LLMs fabricate outputs misaligned with factual information. It is essential to address and rectify such issues, as failure to do so could lead to a range of catastrophic consequences.

6.1 Overview

Responsible models should actively embrace social responsibility to prevent misuse. In this regard, measures should be implemented to progressively enable: content identification (Level I), origin tracing (Level II), and authenticity verification (Level III). Level I focuses on identifying AI-generated content, laying the groundwork for further evaluation. Once this identification is in place, data should be embedded with watermarks to ensure traceability back to its source. This facilitates accountability, deters misuse, and allows users to better assess the reliability and context of the content. Moreover, after identifying content as AI-generated, it becomes essential to assess its authenticity. If inaccuracies are found, accountability mechanisms empower users to request developers for improvements. Thus, Level III needs Level II.

6.2 Level I: Identifiable AI-generated content

State-of-the-art generative models can produce plausible yet fake content at scale. However, the misuse of such content has sparked concerns (Tamkin et al., 2021; McGuffie & Newhouse, 2020; Somepalli et al., 2023b), including fake news, plagiarism, rumors, copyright infringement, etc. To mitigate these problems, it is essential to inform users when they encounter AI-generated content.

6.2.1 DMs

The capacity of DMs to produce photo-realistic images has heightened worries regarding potential misuse. McCloskey and Albright (2018) and Dzanic et al. (2020) have shown the existence of distinct artifacts in AI-generated images, facilitating the trace back to their origins. In detail, even when appearing visually flawless to human eyes, AI-generated images leave behind distinctive artifacts derived from the generation process, such as anomalies in lighting distribution and noticeable asymmetries in shadows and reflections (McCloskey & Albright, 2019; Marra et al., 2019). Initial studies (McCloskey & Albright, 2018, 2019; Nataraj et al., 2019; Marra et al., 2019) endeavored to the manual extraction of these artifacts to distinguish AI-generated images from human-made ones. Later studies (Dzanic et al., 2020; Schwarz et al., 2021) revealed that artifacts are more pronounced in the frequency domain, spurring the advent of frequency-based methods (Wang et al., 2020; Khayatkhoei & Elgammal, 2022; Frank et al., 2020; Chandrasegaran et al., 2021). For example, Wolter et al. (2022) employed the wavelet-packet transformation of images to concurrently leverage spatial and frequency features to detect AI-generated content. Recent studies (Xuan et al., 2019; Mandelli et al., 2022; Girish et al., 2021; Cozzolino et al., 2021) trained classifiers in an end-to-end manner to differentiate between natural images and AI-generated images, in hopes that the classifiers may uncover artifacts that researchers have yet discovered. Notably, Jeong et al. (2022) proposed a two-stage framework in which a universal detector is trained, with a fingerprint generator simulating frequency artifacts from generative models to create training samples. Furthermore, Wang et al. (2023c) showcased that DMs excel at reconstructing the images they generate, leading to the implementation of a binary classifier based on this insight. However, Gragnaniello et al. (2021); Ricker et al. (2024) and (Corvi et al., 2023) empirically demonstrated that existing detection methods suffer from poor generalization.

6.2.2 LLMs

There are two common approaches for detection, namely metric-based and model-based methods. Metric-based methods (Solaiman et al., 2019; Ippolito et al., 2020; Gehrmann et al., 2019) leverage the observation that human-generated content tends to be more casual, designing specific metrics to check if a sample falls below a threshold. Simple metrics include TF-IDF, super-maximal repeated substrings (Gallé et al., 2021), and fluency scores (Crothers et al., 2022). However, these metrics often struggle with highly natural text from LLMs. Kushnareva et al. (2021) explored using transformer attention maps for topological analysis to detect AI-generated text. Mitchell et al. (2023) found that AI-generated text often occupies regions of negative curvature in the model's log probability function, inspiring the development of curvature-based detection metric. Mao et al. (2024) identified the editing distance between a given text and its rewritten version by LLMs to be an effective metric, as human-written texts tend to be more informal and require more modifications compared to LLM-generated content. Tulchinskii et al. (2024) observed significant differences in the intrinsic dimensions of human and AI-generated texts, where intrinsic dimension is the lowest dimensionality needed to compress data without losing substantial information. In parallel, model-based methods (Li et al., 2024b; OpenAI, 2019) train classifiers on large annotated datasets to detect AI-generated content. Bakhtin et al. (2021) employed the previously mentioned simple metrics to train logistic regression or random forest models as detectors. Guo et al. (2023) trained a RoBERTa-based detector, and Chen et al. (2023b) developed two distinct text classification mod-

els based on RoBERTa and T5, respectively. Rodriguez et al. (2022) found that fine-tuning a RoBERTa detector with just a few hundred high-quality samples can greatly enhance cross-domain adaptation.

At a high level, detecting AI-generated content revolves around identifying unique characteristics that set it apart from human-created material. These detection methods can also inform watermarking methods, which serve as a more refined extension of detection, moving from general characteristics of AI-generated content to the specific characteristics of individual AI outputs. A crucial distinction is that detection methods rely on naturally occurring characteristics, but watermarking seeks to artificially create these unique ones.

6.3 Level II: Traceable AI-generated content

The second level targets establishing accountability mechanisms, with watermarking techniques being widely adopted. These techniques enable the tracing of AI-generated content back to its source, promoting responsible use by individuals.

6.3.1 DMs

Traditional image watermarks (ó Ruanaidh et al., 1996; Cox et al., 1996; O'Ruanaidh & Pun, 1997; Chang et al., 2005; Seo et al., 2004; Al-Haj, 2007) are often implemented in the form of subtle alterations to specific frequency components of an image via function transformations and matrix decomposition. Zhao et al. (2023) scrutinized different components within the traditional watermark pipeline to derive a recipe for applying watermark techniques to DMs. Modern methods primarily leverage neural networks to craft more sophisticated watermarks. Vukotić et al. (2018) showed that replacing traditional image processing techniques in watermarking with neural networks can improve watermark quality and robustness. Hayes and Danezis (2017) proposed a watermarking framework that includes an encoder for embedding imperceptible watermarks into images and a detector for distinguishing between original and watermarked images. Furthermore, Zhu et al. (2018), Luo et al. (2020), and Zhang et al. (2020a) incorporated image perturbation operations between the decoder and detector to fortify the watermark against common corruptions. To reduce the risk of watermark removal, Ahmadi et al. (2020) advocated for dispersing watermarks over a relatively wider area within images. Notably, Yu et al. (2021) demonstrated that the watermarking function can be internalized within generative models by training them on watermarked images, thus eliminating the need for external watermark encoding and lowering computational demands. Expanding upon this idea, Fernandez et al. (2023a) fine-tuned the latent decoder of DMs using a pre-trained watermarking encoder, thereby integrating the watermarking process directly into DMs. Furthermore, Wen et al. (2023) presented a watermarking technique that operates in the Fourier domain, integrating with initial noise vectors in DMs, while retrieving the watermark signal through the inversion of the diffusion process. Yang et al. (2024b) found that embedding watermark information into the latent representations can enhance watermark robustness while maintaining image quality.

6.3.2 LLMs

The most simple watermarking technique for LLMs (Kirchenbauer et al., 2023) is to insert specific identifiers into AI-generated content but is easily cracked. As an alternative, Kirchenbauer et al. (2023) introduced implicit rules as hidden watermarks during the content generation process. This involves partitioning the vocabulary into two regions, constraining token sampling to one region to embed a detectable pattern. Christ et al. (2024) adopted a predetermined sequence of random numbers to influence token sampling. Them selected a token based on a comparison between the predicted probability of this token and the corresponding random number. Inspired by watermarking techniques in the image domain, some works explored whether LLMs can learn to generate watermarks. Gu et al. (2024) introduced sampling-based and logit-based watermark distillations, while Xu et al. (2024d) utilized reinforcement learning to train LLMs to learn watermarks through a watermark detector. Moreover, Krishna et al. (2024) developed a retrieval-based watermarking method that stores the generated content within a database, searching for semantically similar matches at the detection stage. However, this method imposes a considerable burden on computation and storage capacities, raising questions about its scalability and practical application. Sadasivan et al. (2023), both empirically and theoretically, demonstrated the vulnerability of current watermarking methods to paraphrasing attacks, where slight modifications allow content to evade detection. To enhance attack performance, Shi et al. (2024c) borrowed adversarial attack techniques, applying an iterative evolutionary search algorithm to find and swap out crucial words with their synonyms generated by an auxiliary LLM. Alternatively, Wang et al. (2024b) harnessed an auxiliary LLM embedding to evaluate the significance of each word and then greedily substituted the most significant words with their synonyms. However, it remains unclear how much modification is counted to cross the boundary of AI-generated content. This gap needs further investigation and expert consensus from diverse domains. Besides, Jovanović et al. (2024) presented another challenge for watermarking methods in which attackers can reverse-engineer watermarking by analyzing LLM outputs, potentially creating counterfeit content and leading to copyright and liability issues. Overall, watermarking techniques for LLMs face significant challenges,

necessitating deeper research into robustness, theft resistance, imperceptibility, and efficiency.

6.4 Level III: Verifiable AI-generated content

The third level aims to verify whether generative models produce accurate outputs (Lin et al., 2022a). This is particularly critical in high-stakes domains like healthcare, where the consequences of false content can be dire (Rakhshan et al., 2013; Miner et al., 2016).

6.4.1 DMs

For DMs, authenticity means different things depending on the task type (unconditional vs. conditional) and the purpose (semantic-level vs. pixel-level fidelity) for which the images are generated. In unconditional tasks where models operate without explicit guidance, authenticity may be assessed by how well outputs align with general expectations of reality. For example, generating a five-legged cat would typically be deemed inauthentic based on biological norms. In such cases, authenticity verification could involve comparing outputs against widely accepted templates or patterns. A straightforward solution to enhance authenticity in unconditional tasks is to implement explicit guidelines that prevent the generation of unrealistic images. This solution in fact transforms an unconditional task into a conditional one, which we will explore further.

Verification in conditional tasks is more complex. On the one hand, it requires assessing how well the generated content aligns with input conditions. On the other hand, the intended use of the generated images significantly influences the criteria for authenticity. For artistic purposes, deviations from reality could be valued for creative expression, whereas scientific illustrations demand strict factual accuracy. Our primary focus here is on factual accuracy. One effective strategy for achieving this is to retrieve authentic images related to a given prompt, compelling DMs to produce similar outputs. However, this strategy risks compromising the diversity of the generated images. A more flexible strategy (Lim & Shim, 2024) is to impose detailed requirements on generated images to ensure they meet factual accuracy standards. For example, Zhang et al. (2023b) introduced ControlNet to adjust the spatial information, while Mou et al. (2024) added adapters to DMs for precise control over image color and structure. More fine-grained control techniques can be found in the area of image editing. For further details, please refer to the survey made by Zhan et al. (2023). Besides, refining the input text can yield more detailed requirements (Zhang et al., 2023c; Chandramouli & Gandikota, 2022), e.g., using LLMs to rewrite prompts in a more specific and factually accurate manner.

Moreover, the granularity of image generation complicates authenticity verification. Although DMs excel at capturing and recreating images at a semantic level, they are likely less adept at precise, pixel-level reproduction required for intricate visuals such as cartographic materials or detailed technical schematics, due to their inherently stochastic nature. Notice that the methods discussed in conditional tasks can currently only be achieved at the semantic level, not at the pixel level, leaving pixel-level issues inadequately managed. In fact, there is a lack of literature focused on the authenticity of content generated by DMs, highlighting a critical need for further research in this area.

6.4.2 LLMs

LLMs often produce untruthful content, referred to as the hallucination issue, which can be classified into two types: in-context and extrinsic (Zhao et al., 2021; Lin et al., 2022b). In-context hallucination occurs when the output misaligns with the given context, such as providing irrelevant responses or contradicting surrounding information. Extrinsic hallucination arises when the output conflicts with world knowledge. The root causes are multifaceted, including reliance on noisy web-crawled datasets (Penedo et al., 2023; Akyurek et al., 2022; Chen et al., 2024d), outdated information, the inherent randomness in top-k decoding methods (Azaria & Mitchell, 2023; Lee et al., 2022a), and limitations in model and optimization (Li et al., 2022c).

Early works (Lee et al., 2022a; Min et al., 2023; Wei et al., 2024) primarily focused on validating the consistency of LLM outputs with world knowledge. At a high level, all validation methods first break down the model's responses into atomic facts and then compare them against verified truths to derive a score. This process is mainly designed for detecting extrinsic hallucination, while the detection of in-context hallucination relies more on the self-correction mechanisms and advanced decoding strategies discussed below. Next, we review authenticity-enhancing strategies for both the training and inference stages.

During training, one key strategy is purifying the training datasets. Gardent et al. (2017); Penedo et al. (2023) and Wang (2019) developed procedures for curating high-quality corpora, while Parikh et al. (2020) stressed the effectiveness of direct human revisions. Due to the impracticality of manually revising vast datasets, these studies (Zhou et al., 2024a; Cao et al., 2023; Lee et al., 2023) proved the potential of a small, high-quality dataset during fine-tuning. Filtering low-quality texts against reliable language corpora (Brown et al., 2020) or using LLMs to sift out unreliable information (Chen et al., 2024d) can further enhance the dataset's integrity. Additionally, LLMs can be leveraged to generate high-quality data (Sun et al., 2023a; Alemohammad et al., 2024; Gunasekar et al., 2023; Eldan & Li, 2023),

and adversarial examples (Ganguli et al., 2022b) can serve as high-quality texts. Another strategy is refining the training process, by either allocating more optimization attention to fact-containing segments (Lee et al., 2022a) or integrating fact-aware rewards to promote the generation of more reliable content (Tian et al., 2024).

During inference, authenticity can be enhanced through the integration of external knowledge, self-correction mechanisms, and advanced decoding strategies. RAG allows for accessing reliable, up-to-date information from trusted sources, which can be fused with user queries for contextual clarity (Mallen et al., 2023) or fed into auxiliary models to refine the LLM's original responses (Gao et al., 2023; Peng et al., 2023; Asai et al., 2024). Self-correction mechanisms capitalize LLMs themselves, with methods forcing them to recall relevant facts (Sun et al., 2023b) or prioritizing outputs from attention heads crucial for truthfulness (Li et al., 2023c). SelfIE asks LLMs to clarify reasoning and enables fine-grained control over the process (Chen et al., 2024c). Some approaches (Manakul et al., 2023; Agrawal et al., 2024; Wang et al., 2023b) employ a majority-voting paradigm, which either requires consensus among multiple models or integrates outputs from rephrased queries. This paradigm also facilitates uncertainty quantification in model responses (Xiong et al., 2024; Touvron et al., 2023), prompting models to avoid responding when uncertain. Improved decoding strategies refine reasoning processes or calibrate next-token distribution. A prime example of the former is "chain of thought," which breaks down complex problems into manageable sub-questions (Wei et al., 2022b). This has since evolved into tree-of-thoughts, chain-of-verification, and process supervision (Yao et al., 2023; Dhuliawala et al., 2024; Lightman et al., 2024). For the latter, factual-nucleus sampling (Lee et al., 2022b) adjusts randomness in the top-k sampling process, while context-aware decoding (Shi et al., 2024b) helps LLMs better integrate contextual information and lessen reliance on static, pre-trained knowledge bases.

Remark. Despite these advancements, achieving absolute truthfulness in AI-generated content remains challenging. Beyond model development, access and information dissemination are also important in practice. AI service providers should implement stricter usage restrictions and work with platforms to identify AI-generated content, such as requiring "proof of personhood" for users. Platforms could adopt selective review processes based on the potential impact of the content. For high-impact content, a hard review may be necessary, requiring users to provide factual sources or proof of identity, along with enlisting administrators or volunteers for manual verification. For less critical content, a soft review can be employed, which labels the content as "unverified" and conducts reviews only when users raise concerns. These can be applied to both DMs and LLMs.

6.5 Benchmark Evaluation Tools: Datasets and Metric

As shown in Table 1, at Level I, the focal point lies in the capacity to differentiate between AI-generated content and human-authored content. Thus, the principal evaluation metrics encompass the detection rate and AUC. Advancing to Level II involves the specific challenge of identifying content generated by a specific model. While the core evaluation metrics remain largely unchanged from Level I, the introduction of watermarks needs additional metrics to assess their impact on content quality. It should be noted that some watermarking techniques hide specific information within images and Bit Acc serves as a metric to ascertain whether the recovered information matches the written information.

Transitioning to Level III, the evaluation of the authenticity of AI-generated content becomes increasingly intricate. Diverging from the constrained outputs of discriminative models, generative models produce a spectrum of open-ended results. Therefore, the assessment of authenticity at this level often requires manual scrutiny, as automated metrics may not sufficiently capture the nuanced variations present in the outputs of generative models.

6.6 Discussion, Recommendation, and Outlook

6.6.1 Discussion

We review the responsibility of DMs and LLMs from three levels: identifiability, traceability, and verifiability. For identifiability, DM-generated content is generally easier to detect than LLM-generated content. Image data captures intricate real-world details, making it difficult for current DMs to achieve a perfect fit. As a result, AI-generated images exhibit noticeable artifacts resulting from imperfections in the generation process. In contrast, language data, being a human creation, is structured around rules and patterns, making it easier to fit. These findings are based on the current landscape, with the caveat that image generation methods are rapidly evolving, and future advancements may overthrow these conclusions.

For traceability, we investigate watermarking techniques. As the capabilities of generative models reach new heights, differentiating AI-generated content from human-created content becomes increasingly difficult. This necessitates active intervention in the generation process, such as embedding watermarks for content auditing and promoting responsible usage. While existing watermarking techniques have proven effective, their resilience against malicious attacks remains questionable. Attacks on watermarks often involve altering the watermarked content to remove the watermark. The debate over what constitutes a transformative alteration

of AI-generated content into human products is a critical and unresolved issue, deserving further scholarly attention.

Turning to verifiability, the verification of DM-generated content is still in its nascent stages, leaving significant room for exploration. The definition of authenticity for image data varies across different domains and necessitates further consensus. Although LLMs have made significant strides in addressing authenticity concerns and proposing potential solutions, their effectiveness is yet to meet expectations and heavily depends on empirical validation. Consequently, the authenticity of LLM-generated content is currently defensible over tested instances.

6.6.2 Recommendation

Based on our review and analysis, we propose the following practical mitigations for practitioners and industry:

- Prioritizing responsible strategies, such as watermarking, during foundational design stages is essential for developing responsible AI systems.
- Given the current challenges in verifying the authenticity of AI-generated content, a pragmatic approach would involve enabling models to retrieve relevant answers, striking a balance. Alternatively, indicating confidence levels in responses can significantly enhance user caution.

6.6.3 Outlook

In light of the challenges mentioned above, we suggest exploring the following promising research directions:

- Watermark attacks require modifications to AI-generated content. However, determining the extent of alteration that renders content unrecognizable as AI-generated remains an open question. Further studies are needed to establish clear boundaries for when rewritten material loses its original AI authorship.
- The authenticity of images generated by DMs is largely unexamined. This issue is critical in fields requiring high precision but has not received sufficient attention from academia or industry, possibly due to the current focus on using DMs for artistic tasks.
- While substantial research has been conducted to enhance the authenticity of LLM-generated content, empirical validation alone is inadequate. Instead, it is vital to identify worst-case scenarios or establish theoretical frameworks for assessing authenticity.

7 Conclusion

We reviewed recent developments regarding the trustworthiness of these models through the lenses of privacy, security, fairness, and responsibility. To promote the trustworthy usage of these models and mitigate associated risks, we proposed some actionable steps for both companies and users. Additionally, we identified challenges that the research community should address. Through these efforts, we seek to improve the understanding and management of the risks associated with DMs and LLMs, ensuring their reliable deployment for the benefit of society as a whole.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grant number 62202170 and Alibaba Group through Alibaba Innovative Research Program.

Data Availability This paper serves as a survey and does not include datasets.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K.,& Zhang, L. (2016). Deep Learning with Differential Privacy. *In: Proc. of ACM CCS*, pp 308–318

Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. (2024). Do Language Models Know When They're Hallucinating References? *In: Findings of EACL*, pp 912–928

Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., & Emami, A. (2020). Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications, 146*, 113157.

Akyurek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., & Guu, K. (2022). Towards Tracing Knowledge in Language Models Back to the Training Data. *Findings of EMNLP* (pp. 2429–2446). United Arab Emirates: Abu Dhabi.

Al-Haj, A. (2007). Combined dwt-dct digital image watermarking. *Journal of computer science, 3*(9), 740–746.

Alemohammad, S., Casco-Rodriguez, J.y, Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2024). Self-consuming generative models go mad. *In: Proc. of ICLR*

Alon, G., & Kamfonas, M. (2023). Detecting language model attacks with perplexity. *ArXiv preprint* arXiv:2308.14132

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B. J., Srivastava, M., & Chang, K. W. (2018). Generating Natural Language Adversarial Examples. *In: Proc. of EMNLP*, Brussels, Belgium, pp 2890–2896, https://doi.org/10.18653/v1/D18-1316

An, S., Chou, S. Y., Zhang, K., Xu, Q., Tao, G., Shen, G., Cheng, S., Ma, S., Chen, P. Y., Ho, T. Y., et al. (2024). Elijah: Eliminating backdoors injected in diffusion models via distribution shift. *Proc. of AAAI, 38*, 10847–10855.

Anderson, M., Amit, G., & Goldsteen, A. (2024). Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *ArXiv preprint* arXiv:2405.20446

Arora, A., He, X., Mozes, M., Swain, S., Dras, M., & Xu, Q. (2024). Here's a Free Lunch: Sanitizing Backdoored Models with Model Merge. *In: Findings of ACL*, pp 15059–15075

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *In: Proc. of ICLR*

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended Diffusion for Text-driven Editing of Natural Images. *In: Proc. of CVPR*, pp 18187–18197, https://doi.org/10.1109/CVPR52688.2022.01767

Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. *In: Findings of EMNLP*, pp 967–976

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. J., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume T, Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint* arXiv:2204.05862

Bakhtin, A., Deng, Y., Gross, S., Ott, M., Ranzato, M., & Szlam, A. (2021). Residual energy-based models for text. *J Mach Learn Res* 22:40:1–40:41

Balunovic, M., Dimitrov, D. I., Staab, R., & Vechev, M. T. (2022). Bayesian Framework for Gradient Leakage. *In: Proc. of ICLR*

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv preprint* arXiv:2110.01963

Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., & Kersting, K. (2023). Sega: Instructing diffusion using semantic dimensions. *ArXiv preprint* arXiv:2301.12247

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler DM, Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *In: Proc. of NeurIPS*

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P., & Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Trans Knowl Data Eng, 36*(7), 2814–2830. https://doi.org/10.1109/TKDE.2024.3361474

Cao, Y., Kang, Y., & Sun, L. (2023). Instruction mining: High-quality instruction data selection for large language models. *ArXiv preprint* arXiv:2307.06290

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: Proc. of USENIX Security*, pp 2633–2650

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle B, Ippolito, D., & Wallace, E. (2023a). Extracting training data from diffusion models. *In: Proc. of USENIX Security*, pp 5253–5270

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023b). Quantifying memorization across neural language models. *In: Proc. of ICLR*

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Koh, P. W., Ippolito, D., Lee, K., Tramèr, F., & Schmidt, L. (2023c). Are aligned neural networks adversarially aligned? *In: Proc. of NeurIPS*

Chandramouli, P., & Gandikota, K. V. (2022). LDEdit: Towards Generalized Text Guided Image Manipulation via Latent Diffusion Models. *In: Proc. of BMVC*, vol 1, p 2

Chandrasegaran, K., Tran, N., & Cheung, N. (2021). A Closer Look at Fourier Spectrum Discrepancies for CNN-Generated Images Detection. *In: Proc. of CVPR*, pp 7200–7209, https://doi.org/10.1109/CVPR46437.2021.00712

Chang, C. C., Tsai, P., & Lin, C. C. (2005). Svd-based digital image watermarking scheme. *Pattern Recognition Letters, 26*(10), 1577–1586.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. *ArXiv preprint* arXiv:2310.08419

Chaudhari, H., Severi, G., Abascal, J., Jagielski, M., Choquette-Choo, C. A., Nasr, M., Nita-Rotaru, C., & Oprea, A. (2024). Phantom: General trigger attacks on retrieval augmented language generation. *ArXiv preprint* arXiv:2405.20485

Chen, C., Liu, D., & Xu, C. (2024a). Towards Memorization-Free Diffusion Models. *In: Proc. of CVPR*, pp 8425–8434

Chen, D., Yu, N., & Fritz, M. (2022). RelaxLoss: Defending Membership Inference Attacks without Losing Utility. *In: Proc. of ICLR*

Chen, D., Li, S., Zhang, Y., Li, C., Kundu, S., & Beerel, P. A. (2024b). DIA: Diffusion based Inverse Network Attack on Collaborative Inference. *In: Proc. of CVPR*, pp 124–130

Chen, H., Vondrick, C., & Mao, C. (2024c). SelfIE: Self-Interpretation of Large Language Model Embeddings. *In: Proc. of ICML*

Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H., et al. (2024d). Alpagasus: Training a better alpaca with fewer data. *In: Proc. of ICLR*

Chen, W., Song, D., & Li, B. (2023a). Trojdiff: Trojan attacks on diffusion models with diverse targets. *In: Proc. of CVPR*, pp 4035–4044

Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Raj, B. (2023b). Gpt-sentinel: Distinguishing human and chatgpt generated content. *ArXiv preprint* arXiv:2305.07969

Cheng, P., Ding, Y., Ju, T., Wu, Z., Du, W., Yi, P., Zhang, Z., & Liu, G. (2024). Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *ArXiv preprint* arXiv:2405.13401

Cho, S., Jeong, S., Seo, J., Hwang, T., & Park, J. C. (2024). Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations. *In: Findings of EMNLP*, pp 2826–2844

Choi, K., Grover, A., Singh, T., Shu, R., & Ermon, S. (2020). Fair Generative Modeling via Weak Supervision. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 119, pp 1887–1898

Chou, S. Y., Chen, P. Y., & Ho, T. Y. (2023). How to backdoor diffusion models? *In: Proc. of CVPR*, pp 4015–4024

Chou, S. Y., Chen, P. Y., & Ho, T. Y. (2024). Villandiffusion:, A. unified backdoor attack framework for diffusion models. *Proc of NeurIPS* 36

Christ, M., Gunn, S., & Zamir, O. (2024). Undetectable watermarks for language models. *In: Proc. of COLT*, PMLR, pp 1125–1139

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). Inducing anxiety in large language models increases exploration and bias. *ArXiv preprint* arXiv:2304.11111

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. *In: Proc. of ICASSP*, IEEE, pp 1–5

Cox, I. J., Kilian, J., Leighton, T., & Shamoon, T. (1996). Secure spread spectrum watermarking for images, audio and video. *In: Proc. of ICIP*, IEEE, vol 3, pp 243–246

Cozzolino, D., Gragnaniello, D., Poggi, G., & Verdoliva, L. (2021). Towards universal gan image detection. *In: 2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, pp 1–5

Crothers, E., Japkowicz, N., Viktor, H., & Branco, P. (2022). Adversarial robustness of neural-statistical features in detection of generative transformers. *In: Proc. of IJCNN*, IEEE, pp 1–8

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2024). Safe rlhf: Safe reinforcement learning from human feedback. *In: Proc. of ICLR*

Dar SUH, Ghanaat, A., Kahmann, J., Ayx, I., Papavassiliu, T., Schoenberg, S. O., & Engelhardt, S. (2023). Investigating data memorization in 3d latent diffusion models for medical image synthesis. *In: Proc. of MICCAI*, Springer, pp 56–65

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski J, & Liu, R. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *In: Proc. of ICLR*

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *In: Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, pp 25–35, https://doi.org/10.18653/v1/W19-3504

Deng, J., Wang, Y., Li, J., Wang, C., Shang, C., Liu, H., Rajasekaran, S., & Ding, C. (2021). TAG: Gradient Attack on Transformer-based Language Models. *In: Findings of EMNLP*, Punta Cana, Dominican Republic, pp 3600–3610, https://doi.org/10.18653/v1/2021.findings-emnlp.305

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. *In: Findings of EMNLP*, pp 1236–1270

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. *In: Findings of ACL*, pp 3563–3578

Dockhorn, T., Cao, T., Vahdat, A., & Kreis, K. (2023). *2023*. Transactions on Machine Learning Research: Differentially private diffusion models.

Du, Y., Bosselut, A., & Manning, C. D. (2022). Synthetic Disinformation Attacks on Automated Fact Verification Systems. *In: Proc. of AAAI*, pp 10581–10589

Duan, J., Kong, F., Wang, S., Shi, X., & Xu, K. (2023). Are diffusion models vulnerable to membership inference attacks? *In: Proc. of ICML*, PMLR, pp 8717–8730

Dubiński, J., Kowalczuk, A., Pawlak, S., Rokita, P., Trzciński, T., & Morawiecki, P. (2024). Towards More Realistic Membership Inference Attacks on Large Diffusion Models. *In: Proc. of WACV*, pp 4860–4869

Dzanic, T., Shah, K., & Witherden, F. D. (2020). Fourier Spectrum Discrepancies in Deep Network Generated Images. *In: Proc. of NeurIPS*

Eldan, R., & Li, Y. (2023). Tinystories: How small can language models be and still speak coherent english? *ArXiv preprint* arXiv:2305.07759

Erdogan, E., Kupcu, A., & Cicek, A. E. (2021). Unsplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning. *IACR Cryptol ePrint Arch, 2021*, 1074.

Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., & Liu, S. (2024a). SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. *In: Proc. of ICLR*

Fan, M., Chen, C., Wang, C., Zhou, W., & Huang, J. (2022a). Refiner: Data refining against gradient leakage attacks in federated learning. *ArXiv preprint* arXiv:2212.02042

Fan, M., Liu, Y., Chen, C., Liu, X., & Guo, W. (2022b). Defense against backdoor attacks via identifying and purifying bad neurons. *ArXiv preprint* arXiv:2208.06537

Fan, M., Guo, W., Ying, Z., & Liu, X. (2023). Enhance transferability of adversarial examples with model architecture. *In: Proc. of ICASSP*, IEEE, pp 1–5

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T., & Li, Q. (2024b). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *In: Proc. of KDD*, pp 6491–6501, https://doi.org/10.1145/3637528.3671470

Fernandez, P., Couairon, G., J.égou, H., Douze, M., & Furon, T. (2023a). The stable signature: Rooting watermarks in latent diffusion models. *In: Proc. of ICCV*, pp 22466–22477

Fernandez, V., Sanchez, P., Pinaya, W. H. L., Jacenków, G., Tsaftaris, S. A., & Cardoso, J. (2023b). Privacy distillation: reducing re-identification risk of multimodal diffusion models. *ArXiv preprint* arXiv:2306.01322

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 119, pp 3247–3258

Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., & Kersting, K. (2023). Fair diffusion: Instructing text-to-image generation models on fairness. *ArXiv preprint* arXiv:2302.10893

Fu, F., Xue, H., Cheng, Y., Tao, Y., & Cui, B. (2022). BlindFL: Vertical Federated Machine Learning without Peeking into Your Data. *In: Proc. of SIGMOD*, New York,, N. Y., USA, SIGMOD '22, p 1316-1330, https://doi.org/10.1145/3514221.3526127

Gallé, M., Rozen, J., Kruszewski, G., & Elsahar, H. (2021). Unsupervised and distributional detection of machine-generated text. *ArXiv preprint* arXiv:2111.02878

Galli, F., Melis, L., & Cucinotta, T. (2024). Noisy Neighbors: Efficient membership inference attacks against LLMs. *In: Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pp 1–6

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). Erasing concepts from diffusion models. *In: Proc. of ICCV*, pp 2426–2436

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. (2022a). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv preprint* arXiv:2209.07858

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. (2022b). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv preprint* arXiv:2209.07858

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D. C., et al. (2023). Rarr: Researching and revising what language models say, using language models. *In: Proc. of ACL*, pp 16477–16508

Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. *In: Proc. of ACL*, Online, pp 3816–3830, https://doi.org/10.18653/v1/2021.acl-long.295

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017). Creating training corpora for nlg micro-planning. *In: Proc. of ACL*

Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based Adversarial Examples for Text Classification. *In: Proc. of EMNLP*, Online, pp 6174–6181, https://doi.org/10.18653/v1/2020.emnlp-main.498

Ge, S., Zhou, C., Hou, R., Khabsa, M., Wang, Y. C., Wang, Q., Han, J., & Mao, Y. (2024). Mart: Improving llm safety with multi-round automatic red-teaming. *In: Proc. of NAACL*, pp 1927–1937

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *In: Findings of EMNLP*, Online, pp 3356–3369, https://doi.org/10.18653/v1/2020.findings-emnlp.301

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *In: Proc. of ACL*, Florence, Italy, pp 111–116, https://doi.org/10.18653/v1/P19-3019

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting Gradients - How easy is it to break privacy in federated learning? *In: Proc. of NeurIPS*

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., & Balle, B. (2023). Differentially private diffusion models generate useful synthetic images. *ArXiv preprint* arXiv:2302.13861

Girish, S., Suri, S., Rambhatla, S. S., & Shrivastava, A. (2021). Towards Discovery and Attribution of Open-world GAN Generated Images.

In: Proc. of ICCV, pp 14074–14083, https://doi.org/10.1109/ICCV48922.2021.01383
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *In: Proc. of ICLR*
Graf, V., Liu, Q., & Chen, M. (2024). Two Heads are Better than One: Nested PoE for Robust Defense Against Multi-Backdoors. *In: Proc. of NAACL*, pp 706–718
Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., & Verdoliva, L. (2021). Are GAN generated images easy to detect?, A. critical analysis of the state-of-the-art. *In: Proc. of ICME*, IEEE, pp 1–6
Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E., & Ermon, S. (2019). Bias Correction of Learned Generative Models using Likelihood-Free Importance Weighting. *In: Proc. of NeurIPS*, pp 11056–11068
Gu, C., Li, X. L., Liang, P., & Hashimoto, T. (2024). On the Learnability of Watermarks for Language Models. *In: Proc. of ICLR*
Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *ArXiv preprint* arXiv:1708.06733
Guan, Z., Hu, M., Li, S., & Vullikanti, A. (2024). Ufid:, A. unified framework for input-level backdoor detection on diffusion models. *ArXiv preprint* arXiv:2404.01101
Gunasekar, S., Zhang, Y., Aneja, J., Mendes CCT, Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al. (2023). Textbooks are all you need. *ArXiv preprint* arXiv:2306.11644
Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv preprint* arXiv:2301.07597
Guo, C., Sablayrolles, A., J.égou, H., & Kiela, D. (2021). Gradient-based Adversarial Attacks against Text Transformers. *In: Proc. of EMNLP*, Online and Punta Cana, Dominican Republic, pp 5747–5757, https://doi.org/10.18653/v1/2021.emnlp-main.464
Gupta, S., Huang, Y., Zhong, Z., Gao, T., Li, K., & Chen, D. (2022). Recovering Private Text in Federated Learning of Language Models. *In: Proc. of NeurIPS*
Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Left-Libertarian Orientation (January 1, 2023)*
Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., & Wood, F. (2022). Flexible Diffusion Modeling of Long Videos. *In: Proc. of NeurIPS*
Hayes, J., & Danezis, G. (2017). Generating steganographic images via adversarial training. *In: Proc. of NeurIPS*, pp 1954–1963
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *In: Proc. of NeurIPS*
Hong, G., Kim, J., Kang, J., Myaeng, S. H., & Whang, J. J. (2024). Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. *In: Findings of NAACL*, pp 2474–2495
Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 97, pp 2790–2799
Hu, H., & Pang, J. (2023). Membership inference of diffusion models. *ArXiv preprint* arXiv:2301.09956
Hu, H., Salcic, Z. A., Sun, L., Dobbie, G., Yu, P., & Zhang, X. (2021). Membership inference attacks on machine learning: A. survey. *ACM Computing Surveys (CSUR), 54*, 1–37.
Huang, J., Shao, H., & Chang, K. C. C. (2022). Are Large Pre-Trained Language Models Leaking Your Personal Information? *Findings of EMNLP* (pp. 2038–2047). United Arab Emirates: Abu Dhabi.
Huang, Y., Gupta, S., Zhong, Z., Li, K., & Chen, D. (2023a). Privacy implications of retrieval-based language models. *In: Proc. of EMNLP*, pp 14887–14902
Huang, Y., Zhang, Q., Yu, P. S., & Sun, L. (2023b). Trustgpt:, A. benchmark for trustworthy and responsible large language models. *ArXiv preprint* arXiv:2306.11507
Huang, Y., Juefei-Xu, F., Guo, Q., Zhang, J., Wu, Y., Hu, M., Li, T., Pu, G., & Liu, Y. (2024). Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. *Proc. of AAAI, 38*, 21169–21178.
Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic Detection of Generated Text is Easiest when Humans are Fooled. *In: Proc. of ACL*, Online, pp 1808–1822, https://doi.org/10.18653/v1/2020.acl-main.164
Jagielski, M., Thakkar, O., Tramer, F., Ippolito, D., Lee, K., Carlini, N., Wallace, E., Song, S., Thakurta, A., Papernot, N., et al. (2023). Measuring forgetting of memorized training examples. *In: Proc. of ICLR*
Jeon, J., Kim, J., Lee, K., Oh, S., & Ok, J. (2021). Gradient Inversion with Generative Image Prior. *In: Proc. of NeurIPS*, pp 29898–29908
Jeong, Y., Kim, D., Ro, Y., Kim, P., & Choi, J. (2022). Fingerprintnet: Synthesized fingerprints for generated image detection. *In: Proc. of ECCV*, Springer, pp 76–94
Jiang, G., Xu, M., Zhu, S. C., Han, W., Zhang, C., & Zhu, Y. (2024). Evaluating and inducing personality in pre-trained language models. *Proc of NeurIPS* 36
Jones, E., Dragan, A., Raghunathan, A., & Steinhardt, J. (2023). Automatically auditing large language models via discrete optimization. *In: Proc. of ICML*, PMLR, pp 15307–15329
Jovanović, N., Staab, R., & Vechev, M. (2024). Watermark Stealing in Large Language Models. *In: Proc. of ICML*
Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating Training Data Mitigates Privacy Risks in Language Models. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 162, pp 10697–10707
Kariyappa, S., & Qureshi, M. K. (2021). Gradient inversion attack: Leaking private labels in two-party split learning. *ArXiv preprint* arXiv:2112.01299
Karra, S. R., Nguyen, S. T., & Tulabandhula, T. (2022). Estimating the personality of white-box language models. *ArXiv preprint* arXiv:2204.12000
Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020). Generalization through Memorization: Nearest Neighbor Language Models. *In: Proc. of ICLR*
Khayatkhoei, M., & Elgammal, A. (2022). Spatial Frequency Bias in Convolutional Generative Adversarial Networks. *In: Proc. of AAAI*, pp 7152–7159
Kim, E., Kim, S., Shin, C., & Yoon, S. (2023a). De-stereotyping text-to-image models through prompt tuning. *workshop in ICML 2023*
Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., & Oh, S. J. (2023b). ProPILE: Probing Privacy Leakage in Large Language Models. *In: Proc. of NeurIPS*
Kim, Y., Mo, S., Kim, M. K., Lee, K., Lee, J., & Shin, J. (2023c). *Bias-to-Text: Debiasing Unknown Visual Biases through Language Interpretation*
Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 202, pp 17061–17084
Kong, F., Duan, J., Ma, R., Shen, H., Zhu, X., Shi, X., & Xu, K. (2024). An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization. *In: Proc. of ICLR*
Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., & Rajani, N. F. (2021). GeDi: Generative Discriminator Guided Sequence Generation. *In: Findings of EMNLP*, Punta Cana, Dominican Republic, pp 4929–4952, https://doi.org/10.18653/v1/2021.findings-emnlp.424

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2024). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Proc of NeurIPS* 36

Kumari, N., Zhang, B., Wang, S. Y., Shechtman, E., Zhang, R., & Zhu, J. Y. (2023). Ablating concepts in text-to-image diffusion models. *In: Proc. of ICCV*, pp 22691–22702

Kurita, K., Michel, P., & Neubig, G. (2020). Weight Poisoning Attacks on Pretrained Models. *In: Proc. of ACL*, Online, pp 2793–2806, https://doi.org/10.18653/v1/2020.acl-main.249

Kushnareva, L., Cherniavskii, D., Mikhailov, V., Artemova, E., Barannikov, S., Bernstein, A., Piontkovskaya, I., Piontkovski, D., & Burnaev, E. (2021). Artificial Text Detection via Examining the Topology of Attention Maps. *In: Proc. of EMNLP*, Online and Punta Cana, Dominican Republic, pp 635–649, https://doi.org/10.18653/v1/2021.emnlp-main.50

Laugier, L., Pavlopoulos, J., Sorensen, J., & Dixon, L. (2021). Civil Rephrases Of Toxic Texts With Self-Supervised Transformers. *In: Proc. of EACL*, Online, pp 1442–1461, https://doi.org/10.18653/v1/2021.eacl-main.124

Lee, A. N., Hunter, C. J., & Ruiz, N. (2023). Platypus: Quick, cheap, and powerful refinement of llms. *ArXiv preprint* arXiv:2308.07317

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Proc of NeurIPS, 35*, 34586–34599.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Proc of NeurIPS, 35*, 34586–34599.

Lee, P. (2016). Learning from tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., & Wallace, B. (2021). Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? *In: Proc. of NAACL*, Online, pp 946–959, https://doi.org/10.18653/v1/2021.naacl-main.73

Lewis PSH, Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *In: Proc. of NeurIPS*

Li, C., Leng, Z., Yan, C., Shen, J., Wang, H.,, M. I., W., Fei, Y., Feng, X., Yan, S., Wang, H., et al. (2023a). Chatharuhi: Reviving anime character in reality via large language model. *ArXiv preprint* arXiv:2308.09597

Li, J., Rakin, A. S., Chen, X., He, Z., Fan, D., & Chakrabarti, C. (2022a). ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning. *In: Proc. of CVPR*, pp 10184–10192, https://doi.org/10.1109/CVPR52688.2022.00995

Li, J., Wu, Z., Ping, W., Xiao, C., & Vydiswaran, V. V. (2023b). Defending against Insertion-based Textual Backdoor Attacks via Attribution. *In: Findings of ACL*, pp 8818–8833

Li, J., Yang, Y., Wu, Z., Vydiswaran, V., & Xiao, C. (2024a). Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. *In: Proc. NAACL*, pp 2985–3004

Li, K., Patel, O., Viégas, F. B., Pfister, H., & Wattenberg, M. (2023c). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds) Proc. of NeurIPS*

Li, L., Fan, Y., Tse, M., & Lin, K. (2020). A review of applications in federated learning. *Comput Ind Eng, 149*, 106854. https://doi.org/10.1016/J.CIE.2020.106854

Li, L., Song, D., Li, X., Zeng, J., Ma, R., & Qiu, X. (2021a). Backdoor Attacks on Pre-trained Models by Layerwise Weight Poisoning. *In: Proc. of EMNLP*, Online and Punta Cana, Dominican Republic, pp 3023–3032, https://doi.org/10.18653/v1/2021.emnlp-main.241

Li, O., Sun, J., Yang, X., Gao, W., Zhang, H., Xie, J., Smith, V., & Wang, C. (2022b). Label Leakage and Protection in Two-party Split Learning. *In: Proc. of ICLR*

Li, S., Li, X., Shang, L., Dong, Z., Sun, C., Liu, B., Ji, Z., Jiang, X., & Liu, Q. (2022c). How Pre-trained Language Models Capture Factual Knowledge?, A. Causal-Inspired Analysis. *In: Findings of ACL*, Dublin, Ireland, pp 1720–1732, https://doi.org/10.18653/v1/2022.findings-acl.136

Li, X., Li, Y., Liu, L., Bing, L., & Joty, S. (2022d). Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *ArXiv preprint* arXiv:2212.10529

Li, Y., Wu, B., Jiang, Y., Li, Z., & Xia, S. (2020b). Backdoor learning:, A. survey. *IEEE transactions on neural networks and learning systems*, P. P.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., & Ma, X. (2021b). Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. *In: Proc. of ICLR*

Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., & Zhang, Y. (2024b). Mage: Machine-generated text detection in the wild. *In: Proc. of ACL*, pp 36–53

Li, Y., Liu, G., Yang, Y., & Wang, C. (2024c). Seeing is believing: Black-box membership inference attacks against retrieval augmented generation. *ArXiv preprint* arXiv:2406.19234

Li, Y., Wei, F., Zhao, J., Zhang, C., & Zhang, H. (2024d). Rain: Your language models can align themselves without finetuning. *In: Proc. of ICLR*

Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., & Guan, H. (2023). Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *In: Proc. of ICML*, PMLR, pp 20763–20786

Liang, H., He, E., Zhao, Y., Jia, Z., & Li, H. (2022). Adversarial attack and defense:, A. survey. *Electronics*

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's Verify Step by Step. *In: Proc. of ICLR*

Lim, Y., & Shim, H. (2024). Addressing image hallucination in text-to-image generation through factual image retrieval. *ArXiv preprint* arXiv:2407.10683

Lin, A., Paes, L. M., Tanneru, S. H., Srinivas, S., & Lakkaraju, H. (2023). Word-level explanations for analyzing bias in text-to-image models. *ArXiv preprint* arXiv:2306.05500

Lin, S., Hilton, J., & Evans, O. (2022a). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *In: Proc. of ACL*, Dublin, Ireland, pp 3214–3252, https://doi.org/10.18653/v1/2022.acl-long.229

Lin, S., Hilton, J., & Evans, O. (2022b). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *In: Proc. of ACL*, Dublin, Ireland, pp 3214–3252, https://doi.org/10.18653/v1/2022.acl-long.229

Liu, H., Wu, Y., Zhai, S., Yuan, B., & Zhang, N. (2023a). Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. *In: Proc. of CVPR*, pp 20585–20594

Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. *In: International symposium on research in attacks, intrusions, and defenses*, Springer, pp 273–294

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 55(9):195:1–195:35, https://doi.org/10.1145/3560815

Liu, R., Khakzar, A., Gu, J., Chen, Q., Torr, P., & Pizzati, F. (2024). Latent guard: a safety framework for text-to-image generation. *Proc. of ECCV, 15084*, 93–109.

Liu, X., Gong, C., & Liu, Q. (2023c). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *In: Proc. of ICLR*

Liu, X., Guan, X., Wu, Y., & Miao, J. (2024). Iterative Ensemble Training with Anti-Gradient Control for Mitigating Memorization in Diffusion Models. *Proc. of ECCV, 15145*, 108–123.

Liu, X., Xu, N., Chen, M., & Xiao, C. (2024c). Autodan: Generating stealthy jailbreak prompts on aligned large language models. *In: Proc. of ICLR*

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023d). Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *ArXiv preprint* arXiv:2308.05374

Lu, J., Teehan, R., & Ren, M. (2024). ProCreate, Don't Reproduce! Propulsive Energy Diffusion for Creative Generation. *Proc. of ECCV, 15118*, 397–414.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing Leakage of Personally Identifiable Information in Language Models. *In: Proc. of IEEE, S.&P*, IEEE Computer Society, pp 346–363

Luo, X., Zhan, R., Chang, H., Yang, F., & Milanfar, P. (2020). Distortion Agnostic Deep Watermarking. *In: Proc. of CVPR*, pp 13545–13554, https://doi.org/10.1109/CVPR42600.2020.01356

Luo, X., Zhao, Y., Hu, Z., Zhu, Y., & Zhong, J. (2023). Defense Against Reconstruction Attacks in Split Federated Learning Through Decreasing Correlation Between Inputs and Activations. *In: Proc. of IJCNN*, IEEE, pp 1–8

Lyu, L., Chen, C., & Fu, J. (2023). A pathway towards responsible, A. I. generated content. *In: Proc. of IJCAI*

Maeng, K., Guo, C., Kariyappa, S., & Suh, G. E. (2024). Bounding the invertibility of privacy-preserving instance encoding using fisher information. *Proc of NeurIPS* 36

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *In: Proc. of ACL*, pp 9802–9822

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *In: Bouamor, H., Pino, J., Bali, K. (eds) Proc. of EMNLP*, pp 9004–9017

Mandelli, S., Bonettini, N., Bestagini, P., & Tubaro, S. (2022). Detecting gan-generated images by orthogonal training of multiple cnns. *In: Proc. of ICIP*, IEEE, pp 3091–3095

Mao, C., Vondrick, C., Wang, H., & Yang, J. (2024). Raidar: geneRative AI Detection viA Rewriting. *In: Proc. of ICLR*

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019). Do gans leave artificial fingerprints? *In: 2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, IEEE, pp 506–511

Matsumoto, T., Miura, T., & Yanai, N. (2023). Membership inference attacks against diffusion models. *In: 2023 IEEE Security and Privacy Workshops (SPW)*, IEEE, pp 77–83

Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership Inference Attacks against Language Models via Neighbourhood Comparison. *In: Findings of ACL*, pp 11330–11343

McCloskey, S., & Albright, M. (2018). Detecting gan-generated imagery using color cues. *ArXiv preprint* arXiv:1812.08247

McCloskey, S., & Albright, M. (2019). Detecting GAN-generated imagery using saturation cues. *In: Proc. of ICIP*, IEEE, pp 4584–4588

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of gpt-3 and advanced neural language models. *ArXiv preprint* arXiv:2009.06807

Meeus, M., Jain, S., Rei, M., & de Montjoye, Y. A. (2024). Did the neurons read your book? document-level membership inference for large language models. *In: Proc. of USENIX Security*, pp 2369–2385

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer Y, & Karbasi, A. (2023). Tree of attacks: Jailbreaking black-box llms automatically. *ArXiv preprint* arXiv:2312.02119

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *In: Proc. of EMNLP*, Abu Dhabi, United Arab Emirates, pp 11048–11064

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *In: Proc. of EMNLP*, pp 12076–12100

Miner, A. S., Milstein, A., Schueller, S. M., Hegde, R., Mangurian, C. V., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine, 176*(5), 619–25.

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. *In: Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, Abu Dhabi, UAE, pp 218–227

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. *In: Proc. of ICML*, PMLR, pp 24950–24962

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., & Shan, Y. (2024). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proc. of AAAI, 38*, 4296–4304.

Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *ArXiv preprint* arXiv:2308.12833

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *In: Proc. of ACL*, Online, pp 5356–5371, https://doi.org/10.18653/v1/2021.acl-long.416

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs:, A. Challenge Dataset for Measuring Social Biases in Masked Language Models. *In: Proc. of EMNLP*, Online, pp 1953–1967, https://doi.org/10.18653/v1/2020.emnlp-main.154

Nataraj, L., Mohammed, T. M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J. H., & Roy-Chowdhury, A. K. (2019). Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging, 31*, 1–7.

OpenAI (2019). Gpt-2: 1.5b release. https://openai.com/research/gpt-2-1-5b-release

OpenAI (2023). GPT-4 technical report. *ArXiv preprint* arXiv:2303.08774

O'Ruanaidh, J. J., & Pun, T. (1997). Rotation, scale and translation invariant digital image watermarking. *In: Proc. of ICIP*, IEEE, vol 1, pp 536–539

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. *Proc of NeurIPS, 35*, 27730–27744.

Pan, K., & Zeng, Y. (2023). Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *ArXiv preprint* arXiv:2307.16180

Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo:, A. Controlled Table-To-Text Generation Dataset. *In: Proc. of EMNLP*, Online, pp 1173–1186, https://doi.org/10.18653/v1/2020.emnlp-main.89

Pasquini, D., Ateniese, G., & Bernaschi, M. (2021). Unleashing the Tiger: Inference Attacks on Split Learning. *In: Proc. of ACM CCS*, pp 2113–2129

Pasquini, D., Strohmeier, M., & Troncoso, C. (2024). Neural Exec: Learning (and Learning from) Execution Triggers for Prompt Injection Attacks. *In: Proc. of the 2024 Workshop on Artificial Intelligence and Security*, pp 89–100

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli A, Pannier, B., Almazrouei, E., & Launay, J. (2023).

The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. *In: Proc. of NeurIPS*

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu Z, Chen, W., et al. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv preprint* arXiv:2302.12813

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022a). Red Teaming Language Models with Language Models. *In: Proc. of EMNLP*, Abu Dhabi, United Arab Emirates, pp 3419–3448

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022b). Red Teaming Language Models with Language Models. *In: Proc. of EMNLP*, Abu Dhabi, United Arab Emirates, pp 3419–3448

Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models. *In: NeurIPS, M. L. Safety Workshop*

Phute, M., Helbling, A., Hull, M., Peng, S., Szyller, S., Cornelius, C., & Chau, D. H. (2024). Llm self defense: By self examination, llms know they are being tricked. *In: Tiny Papers of ICLR*

Piet, J., Alrashed, M., Sitawarin, C., Chen, S., Wei, Z., Sun, E., Alomair, B., & Wagner, D. (2024). Jatmo: Prompt injection defense by task-specific finetuning. *In: European Symposium on Research in Computer Security*, Springer, pp 105–124

Prabhu, V. U., & Birhane, A. (2020). Large image datasets:, A. pyrrhic win for computer vision? *Proc of WACV* pp 1536–1546

Qi, F., Chen, Y., Li, M., Yao, Y., Liu, Z., & Sun, M. (2021a). ONION:, A. Simple and Effective Defense Against Textual Backdoor Attacks. *In: Proc. of EMNLP*, Online and Punta Cana, Dominican Republic, pp 9558–9566, https://doi.org/10.18653/v1/2021.emnlp-main.752

Qi, F., Li, M., Chen, Y., Zhang, Z., Liu, Z., Wang, Y., & Sun, M. (2021b). Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger. *In: Proc. of ACL*, Online, pp 443–453, https://doi.org/10.18653/v1/2021.acl-long.37

Qi, F., Yao, Y., Xu, S., Liu, Z., & Sun, M. (2021c). Turn the Combination Lock: Learnable Textual Backdoor Attacks via Word Substitution. *In: Proc. of ACL*, Online, pp 4873–4883, https://doi.org/10.18653/v1/2021.acl-long.377

Qi, Z., Zhang, H., Xing, E., Kakade, S., & Lakkaraju, H. (2024). Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. *ArXiv preprint* arXiv:2402.17840

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Proc of NeurIPS* 36

Rakhshan, A., Pishro-Nik, H., Fisher, D. L., & Nekoui, M. A. (2013). Tuning collision warning algorithms to individual drivers for design of active safety systems. *2013 IEEE Globecom Workshops (GC Wkshps)* pp 1333–1337

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 139, pp 8821–8831

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *ArXiv preprint* arXiv:2204.06125

Rando, J., & Tramèr, F. (2024). Universal Jailbreak Backdoors from Poisoned Human Feedback. *In: Proc. of ICLR*

Ren, S., Deng, Y., He, K., & Che, W. (2019). Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. *In: Proc. of ACL*, Florence, Italy, pp 1085–1097, https://doi.org/10.18653/v1/P19-1103

Ricker, J., Damm, S., Holz, T., & Fischer, A. (2024). Towards the detection of diffusion model deepfakes. *In: Proc. of VISIGRAPP*, pp 446–457

Robey, A., Wong, E., Hassani, H., & Pappas, G. J. (2023). Smoothllm: Defending large language models against jailbreaking attacks. *ArXiv preprint* arXiv:2310.03684

Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., & Srinivasan, R. (2022). Cross-Domain Detection of GPT-2-Generated Technical Text. *In: Proc. of NAACL*, Seattle, United States, pp 1213–1233, https://doi.org/10.18653/v1/2022.naacl-main.88

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *In: Proc. of CVPR*, pp 10674–10685, https://doi.org/10.1109/CVPR52688.2022.01042

ó Ruanaidh, J., Dowling, W., & Boland, F. (1996). Watermarking digital images for copyright protection. *IEE PROCEEDINGS VISION IMAGE AND SIGNAL PROCESSING,143*, 250–256.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can ai-generated text be reliably detected? *ArXiv preprint* arXiv:2303.11156

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. *ArXiv preprint* arXiv:2307.00184

Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J., & Norouzi, M. (2022a). Palette: Image-to-Image Diffusion Models. *In: Proc. of SIGGRAPH*, pp 15:1–15:10, https://doi.org/10.1145/3528233.3530757

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Proc of NeurIPS, 35*, 36479–36494.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour SKS, Lopes, R. G., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022c). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *In: Proc. of NeurIPS*

Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., & Madry, A. (2023). Raising the cost of malicious, A. I.-powered image editing. *In: Proc. of ICML*, pp 29894–29918

Samanta, S., Mehta, S. (2017). Towards crafting text adversarial samples. *ArXiv preprint* arXiv:1707.02812

Nogueira dos Santos, C., Melnyk, I., & Padhi, I. (2018). Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. *In: Proc. of ACL*, Melbourne, Australia, pp 189–194, https://doi.org/10.18653/v1/P18-2031

Schick, T., & Schütze, H. (2021). Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. *In: Proc. of EACL*, Online, pp 255–269, https://doi.org/10.18653/v1/2021.eacl-main.20

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics, 9*, 1408–1424. https://doi.org/10.1162/tacl_a_00434

Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2022). Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *Proc of CVPR* pp 22522–22531

Schwarz, K., Liao, Y., Geiger, A. (2021). On the Frequency Bias of Generative Models. *In: Proc. of NeurIPS*, pp 18126–18136

Seo, J. S., Haitsma, J., Kalker, T., & Yoo, C. D. (2004). A robust image fingerprinting system using the radon transform. *Signal Processing: Image Communication, 19*(4), 325–339.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., & Zhao, B. Y. (2023). Glaze: Protecting artists from style mimicry by Text-to-Image models. *In: Proc. of USENIX Security*, pp 2187–2204

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., & Zettlemoyer, L. (2024a). Detecting Pretraining Data from Large Language Models. *In: Proc. of ICLR*

Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., & Yih, S. W. T. (2024b). Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. *In: Proc. of NAACL*, pp 783–791

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K., & Hsieh, C. (2024). Red teaming language model detectors with language models. *Trans Assoc Comput Linguistics, 12*, 174–189.

Shrestha, R., Zou, Y., Chen, Q., Li, Z., Xie, Y., & Deng, S. (2024). FairRAG: Fair human generation via fair retrieval augmentation. *In: Proc. of CVPR*, pp 11996–12005

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *In: Proc. of ICML*, JMLR Workshop and Conference Proceedings, vol 37, pp 2256–2265

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release strategies and the social impacts of language models. *ArXiv preprint* arXiv:1908.09203

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023a). Diffusion art or digital forgery? investigating data replication in diffusion models. *In: Proc. of CVPR*, pp 6048–6058

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023b). Diffusion art or digital forgery? investigating data replication in diffusion models. *In: Proc. of CVPR*, pp 6048–6058

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Understanding and mitigating copying in diffusion models. *Proc of NeurIPS, 36*, 47783–47803.

Song, J., Meng, C., & Ermon, S. (2021a). Denoising Diffusion Implicit Models. *In: Proc. of ICLR*

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021b). Score-Based Generative Modeling through Stochastic Differential Equations. *In: Proc. of ICLR*, OpenReview.net

Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). Consistency Models. *In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds) Proc. of ICML*, PMLR, Proceedings of Machine Learning Research, vol 202, pp 32211–32252

Struppek, L., Hintersdorf, D., & Kersting, K. (2022). Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *ArXiv preprint* arXiv:2211.02408

Sui, Y., Phan, H., Xiao, J., Zhang, T., Tang, Z., Shi, C., Wang, Y., Chen, Y., & Yuan, B. (2024). Disdet: Exploring detectability of backdoor attack on diffusion models. *ArXiv preprint* arXiv:2402.02739

Sun, J., Li, A., Wang, B., Yang, H., Li, H., & Chen, Y. (2021). Soteria: Provable Defense Against Privacy Leakage in Federated Learning From Representation Perspective. *In: Proc. of CVPR*, pp 9311–9319https://doi.org/10.1109/CVPR46437.2021.00919

Sun, T., Zhang, X., He, Z., Li, P., Cheng, Q., Yan, H., Liu, X., Shao, Y., Tang, Q., Zhao, X., et al. (2023a). Moss: Training conversational language models from synthetic data. *ArXiv preprint* arXiv:2307.15020

Sun, Z., Wang, X., Tay, Y., Yang, Y., & Zhou, D. (2023b). Recitation-Augmented Language Models. *In: Proc. of ICLR*

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv preprint* arXiv:2102.02503

Tang, S., Wu, S., Aydöre, S., Kearns, M., & Roth, A. (2024). Membership Inference Attacks on Diffusion Models via Quantile Regression. *In: Proc. of ICML*

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen Z, Roberts, A., Bosma, M., Zhou, Y., Chang, C., Krivokon, I., Rusch, W., Pickett M, Meier-Hellstern, K. S., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson K, Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna A, Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., y Arcas, B. A., Cui, C., Croak, M., Chi, E. H., & Le, Q. (2022). Lamda: Language models for dialog applications. *ArXiv preprint* abs/2201.08239

Tian, K., Mitchell, E., Yao, H., Manning, C. D., & Finn, C. (2024). Fine-Tuning Language Models for Factuality. *In: Proc. of ICLR*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint* arXiv:2307.09288

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Nikolenko S, Burnaev, E., Barannikov, S., & Piontkovskaya, I. (2024). Intrinsic dimension estimation for robust detection of ai-generated texts. *Proc of NeurIPS* 36

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *In: Proc. of NeurIPS*, pp 5998–6008

Vukotić, V., Chappelier, V., & Furon, T. (2018). Are deep neural networks good for blind image watermarking? *In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, pp 1–7

Wagh, S., Tople, S., Benhamouda, F., Kushilevitz, E., Mittal, P., & Rabin, T. (2020). Falcon: Honest-majority maliciously secure framework for private deep learning. *Proceedings on Privacy Enhancing Technologies, 2021*, 188–208.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal Adversarial Triggers for Attacking and Analyzing NLP. *In: Proc. of EMNLP*, Hong Kong, China, pp 2153–2162, https://doi.org/10.18653/v1/D19-1221

Wan, A., Wallace, E., Shen, S., & Klein, D. (2023a). Poisoning language models during instruction tuning. *In: Proc. of ICML*, PMLR, pp 35413–35425

Wan, X., Sun, J., Wang, S., Chen, L., Zheng, Z., Wu, F., & Chen, G. (2023b). PSLF: Defending Against Label Leakage in Split Learning. *In: Proc. of CIKM*, New York,, N. Y., USA, CIKM '23, p 2492-2501, https://doi.org/10.1145/3583780.3615019

Wang, F., Huang, J. Y., Yan, T., Zhou, W., & Chen, M. (2023a). Robust Natural Language Understanding with Residual Attention Debiasing. *In: Findings of ACL*, pp 504–519

Wang, H. (2019). Revisiting Challenges in Data-to-Text Generation with Fact Grounding. *In: Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan, pp 311–322, https://doi.org/10.18653/v1/W19-8639

Wang, H., Shen, Q., Tong, Y., Zhang, Y., & Kawaguchi, K. (2024a). The Stronger the Diffusion Model, the Easier the Backdoor: Data Poisoning to Induce Copyright BreachesWithout Adjusting Finetuning Pipeline. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 235, pp 51465–51483

Wang, J., Li, R., Yang, J., Mao, C. (2024b). RAFT: Realistic Attacks to Fool Text Detectors. *In: Proc. of EMNLP*, pp 16923–16936

Wang, S., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-Generated Images Are Surprisingly Easy to Spot... for Now. *In: Proc. of CVPR*, pp 8692–8701, https://doi.org/10.1109/CVPR42600.2020.00872

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023b). Self-consistency improves chain of thought reasoning in language models. *In: Proc. of ICLR*

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., & Li, H. (2023c). Dire for diffusion-generated image detection. *In: Proc. of ICCV*, pp 22445–22455

Wang, Z., Zhang, J., Shan, S., & Chen, X. (2024). T2ishield: Defending against backdoors on text-to-image diffusion models. *Proc. of ECCV, 15143*, 107–124.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *In: Proc. of NeurIPS*

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022a). Finetuned Language Models are Zero-Shot Learners. *In: Proc. of ICLR*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proc of NeurIPS, 35*, 24824–24837.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., & Le, Q. V. (2024). Long-form factuality in large language models. *In: Proc. of NeurIPS*

Wei, W., Liu, L., Loper, M., Chow, K. H., Gursoy, M. E., Truex, S., & Wu, Y. (2020). A Framework for Evaluating Client Privacy Leakages in Federated Learning. *In: Proc. of ESORICS*, Lecture Notes in Computer Science, vol 12308, pp 545–566, https://doi.org/10.1007/978-3-030-58951-6_27

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P. S. (2021). Challenges in Detoxifying Language Models. *In: Findings of EMNLP*, Punta Cana, Dominican Republic, pp 2447–2469, https://doi.org/10.18653/v1/2021.findings-emnlp.210

Weller, O., Khan, A., Weir, N., Lawrie, D., & Van Durme, B. (2022). Defending against misinformation attacks in open-domain question answering. *ArXiv preprint* arXiv:2212.10002

Wen, R., Li, Z., Backes, M., & Zhang, Y. (2024). Membership Inference Attacks Against In-Context Learning. *In: Proc. of ACM CCS*, pp 3481–3495

Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *ArXiv preprint* arXiv:2305.20030

Wolter, M., Blanke, F., Heese, R., & Garcke, J. (2022). Wavelet-packets for deepfake image analysis and detection. *Machine Learning, 111*(11), 4295–4327.

Wu, X., Li, S., Wu, H. T., Tao, Z., & Fang, Y. (2024a). Does rag introduce unfairness in llms? evaluating fairness in retrieval-augmented generation systems. *ArXiv preprint* arXiv:2409.19804

Wu, Y., Yu, N., Li, Z., Backes, M., & Zhang, Y. (2022). Membership inference attacks against text-to-image generation models. *ArXiv preprint* arXiv:2210.00968

Wu, Z., Gao, H., Wang, Y., Zhang, X., Wang, S. (2024b). *Universal prompt optimizer for safe text-to-image generation*. pp 6340–6354

Xiang, C., Wu, T., Zhong, Z., Wagner, D., Chen, D., & Mittal, P. (2024). Certifiably robust rag against retrieval corruption. *ArXiv preprint* arXiv:2405.15556

Xiao, D., Yang, C., & Wu, W. (2021). Mixing activations and labels in distributed training for split learning. *IEEE Transactions on Parallel and Distributed Systems*, P. P.:1–1

Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., & Wu, F. (2023). Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence, 5*(12), 1486–1496.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *In: Proc. of ICLR*

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2024a). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Proc of NeurIPS* 36

Xu, J., Ma, M., Wang, F., Xiao, C., & Chen, M. (2024b). Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models. *In: Proc. of NAACL*, pp 3111–3126

Xu, X., Yang, M., Yi, W., Li, Z., Wang, J., Hu, H., Zhuang, Y., & Liu, Y. (2024c). A Stealthy Wrongdoer: Feature-Oriented Reconstruction Attack against Split Learning. *In: Proc. of CVPR*, pp 12130–12139

Xu, X., Yao, Y., & Liu, Y. (2024d). Learning to watermark llm-generated text via reinforcement learning. *ArXiv preprint* arXiv:2403.10553

Xuan, X., Peng, B., Wang, W., & Dong, J. (2019). On the generalization of GAN image forensics. *In: Chinese conference on biometric recognition*, Springer, pp 134–141

Xue, J., Zheng, M., Hu, Y., Liu, F., Chen, X., & Lou, Q. (2024). Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *ArXiv preprint* arXiv:2406.00083

Yang, W., Lin, Y., Li, P., Zhou, J., & Sun, X. (2021). Rethinking Stealthiness of Backdoor Attack against NLP Models. *In: Proc. of ACL*, Online, pp 5543–5557, https://doi.org/10.18653/v1/2021.acl-long.431

Yang, Y., Gao, R., Yang, X., Zhong, J., & Xu, Q. (2024a). Guardt2i: Defending text-to-image models from adversarial prompts. *ArXiv preprint* arXiv:2403.01446

Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., & Yu, N. (2024b). Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. *In: Proc. of CVPR*, pp 12162–12171

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Proc of NeurIPS* 36

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., & Molchanov, P. (2021). See Through Gradients: Image Batch Recovery via GradInversion. *In: Proc. of CVPR*, pp 16337–16346, https://doi.org/10.1109/CVPR46437.2021.01607

Yu, H., Chen, J., Ding, X., Zhang, Y., Tang, T., & Ma, H. (2024). Step Vulnerability Guided Mean Fluctuation Adversarial Attack against Conditional Diffusion Models. *Proc. of AAAI, 38*, 6791–6799.

Yu, J., Wu, Y., Shu, D., Jin, M., & Xing, X. (2023a). Assessing prompt injection risks in 200+ custom gpts. *ArXiv preprint* arXiv:2311.11538

Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. *In: Proc. of ICCV*, pp 14428–14437, https://doi.org/10.1109/ICCV48922.2021.01418

Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., & Yan, S. (2023b). Bag of tricks for training data extraction from language models. *In: Proc. of ICML*, PMLR, pp 40306–40320

Yuan, Y., Jiao, W., Wang, W., Huang Jt, He, P., Shi, S., & Tu, Z. (2024). Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *In: Proc. of ICLR*

Yue, K., Jin, R., Wong, C. W., Baron, D., & Dai, H. (2023). Gradient obfuscation gives a false sense of security in federated learning. *In: Proc. of USENIX Security*, pp 6381–6398

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R. E. E., et al. (2024). Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health, 6*(1), e12–e22.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., & Sun, M. (2020). Word-level Textual Adversarial Attacking as Combinatorial Optimization. *In: Proc. of ACL*, Online, pp 6066–6080, https://doi.org/10.18653/v1/2020.acl-main.540

Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., et al. (2024). The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *In: Findings of ACL*, pp 4505–4524

Zeng, Y., Chen, S., Park, W., Mao, Z., Jin, M., & Jia, R. (2022). Adversarial Unlearning of Backdoors via Implicit Hypergradient. *In: Proc. of ICLR*

Zhai, S., Dong, Y., Shen, Q., Pu, S., Fang, Y., & Su, H. (2023). Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. *In: Proc. of ACM, M. M.*, pp 1577–1587

Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C., & Xing, E. (2023). Multimodal image synthesis and editing: The generative ai era. *IEEE Transactions on Pattern Anal-*

ysis and Machine Intelligence, 45(12), 15098–15119. https://doi.org/10.1109/TPAMI.2023.3305243

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., & Kang, D. (2024). Removing rlhf protections in gpt-4 via fine-tuning. *In: Proc. of NAACL*, pp 681–687

Zhang, C., Benz, P., Karjauv, A., Sun, G., & Kweon, I. S. (2020a). UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging. *In: Proc. of NeurIPS*

Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., & Liu, Y. (2020b). BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning. *In: Proceedings of the 2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020*, pp 493–506

Zhang, E., Wang, K., Xu, X., Wang, Z., & Shi, H. (2024a). *In: Workshops of CVPR*, pp 1755–1764

Zhang, J., Xu, Z., Cui, S., Meng, C., Wu, W., & Lyu, M. R. (2023a). On the robustness of latent diffusion models. *ArXiv preprint* arXiv:2306.08257

Zhang, L., Rao, A., & Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. *In: Proc. of ICCV*, pp 3836–3847

Zhang, R., Li, H., Wen, R., Jiang, W., Zhang, Y., Backes, M., Shen, Y., & Zhang, Y. (2024b). Instruction Backdoor Attacks Against Customized LLMs. *In: Proc. of USENIX Security*, pp 1849–1866

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab MT, Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022a). OPT: open pre-trained transformer language models. *ArXiv preprint* arXiv:2205.01068

Zhang, Y., Tang, Y., Ruan, W., Huang, X., Khastgir, S., Jennings, P., & Zhao, X. (2024). ProTIP: Probabilistic Robustness Verification on Text-to-Image Diffusion Models against Stochastic Perturbation. *Proc. of ECCV, 15090*, 455–472.

Zhang, Z., Lyu, L., Ma, X., Wang, C., & Sun, X. (2022). Fine-mixing: Mitigating Backdoors in Fine-tuned Language Models. *Findings of EMNLP* (pp. 355–372). United Arab Emirates: Abu Dhabi.

Zhang, Z., Han, L., Ghosh, A., Metaxas, D. N., & Ren, J. (2023c). Sine: Single image editing with text-to-image diffusion models. *In: Proc. of CVPR*, pp 6027–6037

Zhang, Z., Yang, J., Ke, P., & Huang, M. (2024d). Defending large language models against jailbreaking attacks through goal prioritization. *In: Proc. of ACL*, pp 8865–8887

Zhao, B., Mopuri, K. R., & Bilen, H. (2020). idlg: Improved deep leakage from gradients. *ArXiv preprint* arXiv:2001.02610

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N. M., & Lin, M. (2023). A recipe for watermarking diffusion models. *ArXiv preprint* arXiv:2303.10137

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. *In: Proc. of ICML*, Proceedings of Machine Learning Research, vol 139, pp 12697–12706

Zhong, Z., Huang, Z., Wettig, A., & Chen, D. (2023). Poisoning retrieval corpora by injecting adversarial passages. *In: Proc. of EMNLP*, pp 13764–13775

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu L, et al. (2024a). Lima: Less is more for alignment. *Proc of NeurIPS*

Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., & Smith, N. (2021). Challenges in Automated Debiasing for Toxic Language Detection. *In: Proc. of EACL*, Online, pp 3143–3155, https://doi.org/10.18653/v1/2021.eacl-main.274

Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., & Qiao, Y. (2024b). Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *In: Findings of ACL*, pp 10586–10613

Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). Hidden: Hiding data with deep networks. *In: Proc. of ECCV*, pp 657–672

Zhu, L., Liu, Z., & Han, S. (2019). Deep Leakage from Gradients. *In: Proc. of NeurIPS*, pp 14747–14756

Zhu, S., Zhang, R., An, B., Wu, G., Barrow, J., Wang, Z., Huang, F., Nenkova, A., & Sun, T. (2023). Autodan: Automatic and interpretable adversarial attacks on large language models. *ArXiv preprint* arXiv:2310.15140

Zhuang, H., Zhang, Y., & Liu, S. (2023). A pilot study of query-free adversarial attack against stable diffusion. *In: Proc. of CVPR*, pp 2384–2391

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *CoRR*, arXiv:2307.15043