

---

# A Survey on Responsible Generative AI: What to Generate and What Not

Jindong Gu

*University of Oxford, United Kingdom*

## Abstract

In recent years, generative AI (GenAI), like large language models and text-to-image models, has received significant attention across various domains. However, ensuring the responsible generation of content by these models is crucial for their real-world applicability. This raises an interesting question: *What should responsible GenAI generate, and what should it not?* To answer the question, this paper investigates the practical responsible requirements of both textual and visual generative models, outlining five key considerations: generating truthful content, avoiding toxic content, refusing harmful instruction, leaking no training data-related content, and ensuring generated content identifiable. Specifically, we review recent advancements and challenges in addressing these requirements. Besides, we discuss and emphasize the importance of responsible GenAI across healthcare, education, finance, and artificial general intelligence domains. Through a unified perspective on both textual and visual generative models, this paper aims to provide insights into practical safety-related issues and further benefit the community in building responsible GenAI.

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Preliminaries</b>	<b>4</b>
2.1	Preliminary of Modern Generative AI . . . . .	4
2.1.1	Transformer-based Textual Generative AI . . . . .	4
2.1.2	Diffusion Model-based Visual Generative AI . . . . .	6
2.2	Vulnerability of Deep Neural Networks . . . . .	7
2.2.1	Adversarial Attacks . . . . .	7
2.2.2	Backdoor Attacks . . . . .	8
<b>3</b>	<b>Responsible Textual Generative Model</b>	<b>9</b>
3.1	To Generate Truthful Content . . . . .	9
3.1.1	Hallucination . . . . .	9
3.2	Not To Generate Toxic Content . . . . .	12
3.2.1	Bias and Misinformation Generation . . . . .	12
3.3	Not To Generate for Harmful Instructions . . . . .	14
3.3.1	Prompt Injection Attack on LLM . . . . .	14
3.3.2	Prompt Extraction Attack on LLM . . . . .	15

---

3.3.3	Jailbreak Attack on LLM . . . . .	15
3.3.4	Backdoor Attack on LLM . . . . .	17
3.4	Not To Generate Training Data-related Content . . . . .	18
3.4.1	Membership Inference Attack on LLM. . . . .	18
3.4.2	Training Data Extraction Attack on LLM. . . . .	19
3.4.3	Relation to Other Privacy-related Attacks. . . . .	20
3.5	To Generate Identifiable Texts . . . . .	21
3.5.1	Watermarking Textual Generation. . . . .	21
3.5.2	AI-generated Text Detection. . . . .	22
3.5.3	AI-generated Text Attribution. . . . .	23
<b>4</b>	<b>Responsible Visual Generative Models</b>	<b>23</b>
4.1	To Generate Truthful Images . . . . .	23
4.2	Not To Generate Images with Toxic Content . . . . .	24
4.2.1	Discovering, Measuring, and Mitigating Toxic Generation. . . . .	24
4.3	Not To Generate Images for Harmful Instructions . . . . .	26
4.3.1	Adversarial Attack on Text-to-Image Models . . . . .	27
4.3.2	Prompt Extraction Attack on Text-to-Image Models . . . . .	28
4.3.3	Jailbreak Attack on Text-to-Image Models . . . . .	28
4.3.4	Backdoor Attack on Text-to-Image Models . . . . .	29
4.4	Not To Generate Training Image . . . . .	30
4.4.1	Membership Inference Attack on Text-to-Image Models. . . . .	30
4.4.2	Training Data Extraction. . . . .	31
4.5	To Generate Identifiable Images . . . . .	32
4.5.1	Watermarking of Generated Image. . . . .	33
4.5.2	Detection of AI-generated Image. . . . .	33
4.5.3	Model Attribution of AI-generated Image. . . . .	34
4.5.4	Data Attribution of AI-generated Image. . . . .	35
<b>5</b>	<b>Responsible Generative AI in Safety-critical Applications</b>	<b>35</b>
5.1	Responsible Generative AI for Healthcare . . . . .	35
5.2	Responsible Generative AI for Finance . . . . .	36
5.3	Responsible Generative AI for Education . . . . .	37
5.4	Responsible Generative AI for Artificial General Intelligence . . . . .	38
<b>6</b>	<b>Challenges and Opportunities</b>	<b>39</b>
<b>7</b>	<b>Conclusion</b>	<b>41</b>

---

## 1 Introduction

Generative AI (GenAI) has received remarkable attention recently. Various generative models have been developed in diverse domains, for example, autoregressive large language models (Brown et al., 2020; Touvron, 2023) and text-to-image generative models (Saharia et al., 2022; Rombach et al., 2022; Betker et al., 2023). In real-world applications, the generated contents have to be not only high-quality but also responsible. Thus, this raises the question: What should responsible GenAI generates, and what not?

In this paper, we summarize the responsible requirements of current generative models. Two types of generative models are mainly considered in this work, namely, textual generative models and visual ones. Specifically, textual generative models generate textual responses based on textual or visual inputs, which include autoregressive large language models (Brown et al., 2020; Touvron, 2023) and multimodal LLMs (Team, 2023; OpenAI, 2023). Similarly, visual generative models are the models that generate images or videos based on textual and visual inputs (Rombach et al., 2022; Betker et al., 2023; OpenAI, 2024b).

For both types of generative models, we summarize five responsible requirements for the generated contents. The requirements are not comprehensive, we only review five popular requirements that are important to both academic and industrial communities. Concretely, the five requirements are as follows:

- 1. To generate truthful content.** The generated content is expected to be truthful. However, current generative models could generate content that strays from factual reality or includes fabricated information, which is known as *Hallucination* (Maynez et al., 2020). For instance, language models would generate non-fact, and text-to-image models create images with objects that are not specified in the text prompts. Many efforts have been made to identify, elucidate, and tackle the problem of hallucination (Huang et al., 2023a).
- 2. Not to generate toxic content.** It is well known that both language models and image generation models could generate biased content (Sap et al., 2019; Naik and Nushi, 2023). More recently, with wide applications, it has been found that toxic content could also be generated as responses to end-users. The goal to make generated content unbiased and non-toxic has been intensively explored from various perspectives, e.g., filtering training data and fine-tuning (Ganguli et al., 2022; Friedrich et al., 2023).
- 3. Not to generate content for harmful instructions.** With or without safety alignment, the current generative AI model can still generate inappropriate content given the adversarial prompts, which is known as *Jailbreak* (Perez and Ribeiro, 2022). For instance, the textual generative models would output the details for the input prompt 'How to build bombs?' + adversarial prompt. Visual generative models would generate inappropriate images when adversarial text prompts are given (Yang et al., 2024b). Much effort has been made to reveal such vulnerability and defend against these adversarial prompts.
- 4. Not to generate training data-related content.** Recent generative models are often large-scale and have a large number of parameters. Recent research shows that the learned parameters contain the information of training instances. For instance, training text from a language model can be extracted (Carlini et al., 2021), and training images can be synthesized with the corresponding pre-trained text-to-image model (Carlini et al., 2023b). How to better extract training data from pre-trained generative models and how to hide the training data information have been intensively studied in our community.
- 5. To Generate identifiable content.** The copyright of generated content is a complicated problem, which requires the knowledge of multiple disciplines. There are multiple levels of copyright problems. One is how to generate detectable content, e.g. with watermarks (Usop and Hisham, 2021). To attribute the copyright, another intensively studied topic is model attribution (Uchendu et al., 2020). It aims to identify which generative model generates a particular instance.

Our paper is organized as follows: we recall the preliminary knowledge of GenAI and the vulnerability of deep neural networks in Sec. 2. Sec. 3 summarizes the research of textual generative models on responsible generation, while Sec. 4 presents the research about responsible visual generation. Besides, we also discuss the application of GenAI in different domains from the perspective of responsible AI in Sec. 5. The involved domains include healthcare, education, finance, and artificial general intelligence. Furthermore, we discuss the challenges and opportunities of responsible GenAI in Sec. 6 and conclude our paper in Sec. 7.

---

Our paper differs from related works in the following three points: 1) We present the research topics on responsible GenAI and their recent progress. Especially, we provide a unified perspective for both textual generative models and visual generative models. 2) We summarize the practical safety-related problems that both academics and industry have intensively worked on. 3) We discuss the risks and concerns when applying GenAI to various domains, including general-purpose intelligent systems.

## 2 Preliminaries

In this section, we provide background knowledge on the techniques used to build safe generative models, reveal their vulnerabilities, and defend against malicious inputs. Specifically, we begin by reviewing the foundational components of modern generative AI. Then, we delve into the fundamentals of adversarial attacks and backdoor attacks on deep neural networks.

### 2.1 Preliminary of Modern Generative AI

For textual and visual generative models, we first introduce the popular model architectures (i.e., Transformer and Diffusion Models) and then present basic knowledge of pre-training and post-training, respectively.

#### 2.1.1 Transformer-based Textual Generative AI

**Transformer Architecture.** Transformer (Vaswani et al., 2017) is often composed of a list of self-attention blocks consisting self-attention layer, MLP layer as well as other operations. The self-attention layer is the main component of the self-attention block, which takes a sequence of tokens and outputs a sequence of tokens with new embeddings. Concretely, they can be expressed as follows. Given the input consists of a list of tokens  $\mathbf{x} \in \mathbb{R}^{(N \times M)}$ , the queries, keys, and values of them are computed as

$$\mathbf{K} = \mathbf{W}_k \cdot \mathbf{x}, \quad \mathbf{Q} = \mathbf{W}_q \cdot \mathbf{x}, \quad \mathbf{V} = \mathbf{W}_v \cdot \mathbf{x}, \quad (1)$$

where  $\mathbf{W}_k \in \mathbb{R}^{(M \times D)}$ ,  $\mathbf{W}_q \in \mathbb{R}^{(M \times D)}$  and  $\mathbf{W}_v \in \mathbb{R}^{(M \times D)}$  are linear mapping matrix.

The attention between the input tokens is

$$\mathbf{A} = \text{Softmax}(\mathbf{Q} \cdot (\mathbf{K})^T / \sqrt{D}). \quad (2)$$

The output embeddings  $\mathbf{Z}$  of the self-attention layer are

$$\mathbf{Z} = \mathbf{A} \cdot \mathbf{V}. \quad (3)$$

Transformer-based encoder-decoder (Raffel et al., 2020), decoder-only architectures (OpenAI, 2023) are often applied as generative models, while encoder-only architectures (Devlin et al., 2018) are designed for discriminative tasks. We now introduce the most popular architecture, namely, decoder-only architecture with a masked self-attention layer. Different from the self-attention, the masked self-attention replaces the Equation 3 with the following equation:

$$\mathbf{Z} = (\mathbf{A} + \mathbf{M}) \cdot \mathbf{V}, \quad (4)$$

where  $\mathbf{M} \in \mathbb{R}^{(D \times D)}$  is defined as

$$\mathbf{M}(i, j) = \begin{cases} 0, & \text{if } i \geq j \\ -\infty, & \text{if } i < j \end{cases}, \quad (5)$$

where  $i$  and  $j$  are index of the dimension of token embedding  $D$ .

The architecture above is often applied for autoregressive generation. Specifically, given  $N$  input tokens, a generative model generates the  $N + 1$ -th token. The predicted token is appended to the previous  $N$  tokens. The model generates the  $N + 2$ -th token based on the  $N + 1$  tokens. The probability of generating  $y_{N+2}$  as the  $N + 2$ -th token is

$$\mathbb{P}(y_{N+2} | y_1, y_2, \dots, y_{N+1}), \quad (6)$$

The autoregressive generation stops until a certain token is generated. Multi-head of self-attention mechanism can be computed in parallel. Their outputs are concatenated as the final token embedding.

**Pre-training of Decoder-only Transformer.** We first introduce the data preparation for autoregressive pre-training as the following steps:

1. Data Collection and Cleaning: A large and diverse dataset of text is collected from various sources such as books, articles, websites, forums, and other textual repositories. The collected data is then cleaned, e.g., by removing irrelevant characters and handling special cases like punctuation.
2. Sequence Creation: In this step, the raw text data is divided into sequences of fixed or variable length. Each sequence represents a contiguous segment of text, and these sequences serve as the basic units of input for the model pretraining.
3. Tokenization: After the sequences are created, each sequence is tokenized into individual tokens. Tokenization involves splitting the text into smaller units such as words, subwords, or characters, depending on the tokenization scheme chosen for the model. This step converts the text into a sequence of discrete tokens.
4. Special Tokens Addition: Special tokens may be added to the sequences to indicate the beginning and end of each sequence, as well as to mark padding and unknown tokens.

For each input sequence consisting of a list of tokens  $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ , the joint distribution is computed as follows

$$\mathbb{P}(\mathbf{x}) = \mathbb{P}(x_1, x_2, \dots, x_N) = \prod_{i=1}^N \mathbb{P}(x_i | x_1, x_2, \dots, x_{i-1}), \quad (7)$$

The goal is to maximize the joint probability by updating the parameters of the autoregressive model.

**Post-training of Transformer-based Language Models** LLM pre-training on large-scale datasets does not inherently make them capable of following users' instructions. The output of LLMs can be not helpful or even harmful to the users. To ensure that LLMs generate useful and responsible responses, post-training is often conducted to align them with human intent. Various post-training strategies have been proposed. They often start with Supervised Fine-Tuning (SFT). For SFT, a dataset is first collected where the labelers provide demonstrations of the desired behavior on the input prompts. The pre-trained LLM is fine-tuned on the collected dataset following a standard setting. Specifically, SFT is similar to the pre-training process where training samples are constructed by concatenating a prompt  $\{x_1, x_2, \dots, x_N\}$  and a desired response  $\{x_{N+1}, x_{N+2}, \dots, x_{N+L}\}$ .

$$\mathbb{P}(x_{N+1}, \dots, x_{N+L} | x_1, x_2, \dots, x_N) = \prod_{i=N+1}^{N+L} \mathbb{P}(x_i | x_1, x_2, \dots, x_N, x_{N+1}, \dots, x_{N+L}), \quad (8)$$

The SFT model generates task-specific completions. However, its responses may violate safety rules. To address it, Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) is applied to integrate human preference, in which the human preference is first modeled as a reward in Reward modeling (RM), and Reinforcement Learning (RL) is applied to maximize the reward model via updating the model. The two steps are introduced below.

1) Reward modeling. For RM, a dataset of comparisons between model outputs is collected, where labelers rank the model outputs for each given input. Concretely, K model responses are sampled for each input. Any two of them are ranked by the labelers, namely, there are  $\binom{K}{2}$  annotations for each input prompt.

A model, dubbed reward model, is trained on the collected dataset  $D$ . The reward model  $r_\theta(\cdot)$ , parameterized by  $\theta$ , takes in a prompt and the model response on it and outputs a scalar reward. The model is expected to output a larger scale value for the preferred response than that for the other response. Specifically, the loss function for the reward model can be formulated as follows:

$$L(\theta) = \mathbb{E}_{(\mathbf{x}, y_w, y_l) \sim D} [\log(\sigma(r_\theta(\mathbf{x}, y_w) - r_\theta(\mathbf{x}, y_l)))] \quad (9)$$

where  $r_\theta(\mathbf{x}, y)$  is the scalar output of the reward model when the prompt  $\mathbf{x}$  and the model response  $y$  on it are taken as input,  $y_w$  is the preferred response out of the pair of  $y_w$  and  $y_l$ . Note that all  $\binom{K}{2}$  comparisons from each prompt are taken as a single batch to avoid overfitting (Ouyang et al., 2022).

Finally, the scalar outputs of the reward model are normalized so that the demonstrations provided by labelers achieve a mean score of 0 before tuning the model with RLHF. This is meaningful because the RM training loss defined as in Equation 9 is invariant to shifts in reward. More details regarding reward modeling can be found in Ouyang et al. (2022). Further research shows that LLM can exploit errors in the not-perfect reward model to achieve a high reward, which is dubbed reward hacking. Efforts have been made to mitigate the hacking phenomenon (Amodei et al., 2016; Coste et al., 2023; Eisenstein et al., 2023).

2) Reinforcement learning. When a reward model is available, SFT model is fine-tuned with RL to maximize the rewards received by the reward model. Given the prompt and the model's response to it, a reward is returned by the reward model, which is used to update model parameters. In addition, a per-token KL penalty from the SFT model at each token is applied to mitigate over-optimization of the reward model (Ouyang et al., 2022).

The model to be tuned can be seen as a policy network  $\pi_\phi^{\text{RL}}(\cdot)$ , parameterized by  $\phi$ . Given a prompt (i.e., part of the environment), the model makes a response based on the prompt, which can be seen as an action. The output of the reward model for the prompt and the response is the reward returned by the environment. Then, the episode ends. The model is updated to maximize the following objective:

$$L(\phi) = \mathbb{E}_{(\mathbf{x}, y) \sim D_{\pi_\phi^{\text{RL}}}} [r_\theta(\mathbf{x}, y) - \beta \log(\pi_\phi^{\text{RL}}(y | \mathbf{x}) / \pi^{\text{SFT}}(y | \mathbf{x}))] \quad (10)$$

where the first term is the reward, the second one corresponds to the KL penalty for regularization,  $\pi_\phi^{\text{RL}}$  is the learned policy,  $\pi^{\text{SFT}}$  is the supervised trained model.

The model tuned with the loss above often shows performance regressions on public NLP datasets. To address it, it is also common to mix the pretraining gradients into the PPO gradients, namely, add the term  $\mathbb{E}_{x \sim D_{\text{pretrain}}} \log(\pi_\phi^{\text{RL}}(\mathbf{x}))$  to the loss objective, where  $D_{\text{pretrain}}$  is the distribution of pretraining datasets (Ouyang et al., 2022).

In addition to RLHF, many alternative alignment techniques have been proposed to remedy the safety issue, such as controlled decoding (Yang and Klein, 2021; Mudgal et al., 2023), sequence likelihood calibration (Zhao et al., 2022), direct preference optimization (Rafailov et al., 2024), and best-of-n finetuning (Touvron, 2023; Beirami et al., 2024). In this preliminary part, we introduce the classic alignment method, i.e., RLHF.

### 2.1.2 Diffusion Model-based Visual Generative AI

**Diffusion Model Architecture.** Diffusion probabilistic models (Sohl-Dickstein et al., 2015), also called Diffusion Models (DMs), are designed to fit complex data distribution while keeping tractable. Based on DMs, denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) are proposed for the domain of image generation. The training of DDPM consists of a multi-step forward process and an iterative reverse process. In each step of the forward pass, gaussian noises are added to the natural images. Formally speaking, given a clean image  $\mathbf{x}_0$  from a distribution  $q$ , diffusion step  $T$  and hyperparameter  $\beta_t$ , the forward process is following

$$q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (11)$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t I), \quad (12)$$

where  $\mathcal{N}(\mathbf{x}_t; \mu, \sigma)$  means sampling gaussian noise with mean of  $\mu$  and variance of  $\sigma$ . The image with added noises at the  $t$ -th step can be reformulated as

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) I), \quad (13)$$

where  $\alpha_t = 1 - \beta_t$ ,  $\bar{\alpha}_t = \prod_{s=0}^t \alpha_s$ .

Corresponding to the forward process defined above, the reverse process reconstructs images from noisy data  $\mathbf{x}_T$  iteratively. In the  $t$ -th iteration of the reverse process, a denoising network predicts the noise that is added to natural image  $\mathbf{x}_0$  at the  $t$ -th step of the forward process. The predicted noise is expressed as  $\epsilon_\theta(\mathbf{x}_t, t)$  where  $\epsilon_\theta(\cdot)$  is the noise prediction network parameterized by  $\theta$ . Then, the  $\mathbf{x}_{t-1}$  can be reconstructed from  $\mathbf{x}_t$  and the predicted noise. The noise prediction network is optimized with the following formula

$$L(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2], \quad (14)$$

where  $t$  is uniform between 1 and  $T$ ,  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  is random noise. The parameter  $\theta$  is updated to minimize the loss defined above.

During inference, DDPM generates images from noisy data within a predefined number of steps, following the reverse process described earlier. However, when the input image dimension is excessively large, scaling DDPMs becomes challenging. To tackle this scalability issue, latent DDPMs have been introduced (Rombach et al., 2022). These models first map raw images to a lower-dimensional latent space, where the diffusion process is then carried out. The final embedding is mapped to image space again with a decoder.

Besides, condition information has been explored to guide the generated content (Ho et al., 2020; Rombach et al., 2022). The condition information  $\mathbf{c}$  like textual prompts or images are taken as conditional inputs of the noise prediction network in both training and inference processes. With the conditional information, the noise prediction network is optimized as follows:

$$L_c(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, \mathbf{c})\|^2], \quad (15)$$

If conditional information is integrated into the training process, the generated content can be controlled by user-specified conditional information in inferences.

**Pre-training of DDPM.** For training unconditional DDPMs (Ho et al., 2020), a dataset comprising images is collected. The parameters in the noise prediction network of DDPM are optimized to minimize a predefined training objective outlined in Equation 14. Conversely, in the training of conditional DDPMs, each training image is typically accompanied by conditional information, such as text, another image, or even audio. For instance, conditional latent diffusion models are trained on a large-scale dataset consisting of image-text pairs (Rombach et al., 2022). The features of the conditional information are extracted using a feature extractor and fed to the noise prediction network. The optimization objective of the noise prediction is described by Equation 15.

**Post-training of DDPM.** Similar to post-training of LLM, reinforcement learning with human feedback (RLHF) has also been applied to fine-tune diffusion models (Lee et al., 2023; Black et al., 2023; Fan et al., 2024; Wu et al., 2023c; Xu et al., 2024). The construction of a reward model of RLHF requires extensive datasets, optimal architecture, and manual hyperparameter tuning, which makes the process both time and cost-intensive (Yang et al., 2023a). Inspired by Direct Preference Optimization (Rafailov et al., 2024), Yang et al. (2023a) propose to fine-tune diffusion models directly with Denoising Diffusion Policy Optimization method. In addition to the alignment method above, controllable generation has also been explored to implement the intents of users (Ruiz et al., 2023; Gal et al., 2022; Zhang et al., 2023c; Liu et al., 2022a; Du et al., 2023).

## 2.2 Vulnerability of Deep Neural Networks

In this section, we provide a preliminary about attacks on deep neural networks, focusing on two main types: adversarial attacks and backdoor attacks. Adversarial attacks seek to fool deep neural networks by altering their inputs during inference, while backdoor attacks aim to induce malicious behaviors in models by interfering with the training process, such as by adding poisoned samples with specific trigger patterns to the training dataset.

### 2.2.1 Adversarial Attacks

Szegedy et al. (2013) find an intriguing property of neural networks that when added to the image, a certain imperceptible perturbation can cause the network to misclassify an image. The adversarial perturbation can

be created as follows. Given an input  $x$ , a model  $f(\cdot)$  and its output  $f(\mathbf{x})$ , an adversarial perturbation  $\boldsymbol{\delta}$  is created to increase the loss  $\mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}), \mathbf{y})$  where  $\mathcal{L}(\cdot)$  is the standard cross-entropy loss and  $\boldsymbol{\delta}$  is often set to be  $\ell_p$ -bounded to achieve imperceptibility. The created perturbation corresponding to high loss can mislead the prediction of the model when added to the input.

The optimization of the perturbation can be formulated as follows. The one-step *Fast Gradient Sign Method* (FGSM (Goodfellow et al., 2015)) creates perturbations as

$$\boldsymbol{\delta} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{\delta}} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}), \mathbf{y})), \quad (16)$$

where  $\text{sign}(\cdot)$  is the sign function and  $\epsilon$  is a step size corresponding to the allowed perturbation bound.

FGSM with a single-step optimization only achieves limited attack performance. To improve the adversarial effectiveness, the multi-step *Projected Gradient Descent* (PGD (Madry et al., 2017)) is proposed. Each step of PGD can be expressed as

$$\boldsymbol{\delta}^{t+1} \leftarrow \text{clip}_{\epsilon}(\boldsymbol{\delta}^t + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{\delta}} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}), \mathbf{y}))), \quad (17)$$

where  $\boldsymbol{\delta}^t$  corresponds to the perturbation of  $t$ -th step and  $\text{clip}_{\epsilon}(\cdot)$  is a clipping function that clips its input into  $\epsilon$ -ball of the input for visual imperceptibility.

More optimization methods have been proposed to further improve attack effectiveness (Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017b; Dong et al., 2018; Croce and Hein, 2020). Apart from  $\ell_p$ -bounded attacks, other attacks with different constraints have also been intensively studied, e.g., sparse attacks (Croce and Hein, 2019; Modas et al., 2019), patch attacks (Brown et al., 2017; Gu et al., 2022b;a), semantic attacks (Joshi et al., 2019; Wang et al., 2023c), viewpoint attacks (Dong et al., 2022), and physical attacks (Huang et al., 2020; Eykholt et al., 2018). Furthermore, adversarial attacks on neural networks with text inputs (e.g., language models) have also been explored where the input perturbations are often character-level, word-level, and sentence-level addition, removal, and replacement (Morris et al., 2020; Zhang et al., 2020). Different from image space, the text input space is discrete, which poses the main challenges when attacking and defending NLP models (Dinan et al., 2019; Sinha et al., 2023).

An intriguing property of adversarial perturbation is the transferability of adversarial examples, where perturbations crafted for one model can deceive another, often with a different architecture (Goodfellow et al., 2015; Papernot et al., 2016; Gu et al., 2021; Ma et al., 2023a; Yu et al., 2023c). The property poses practical threats to real-world applications since it enables attacks without access to the target model. We refer the reader to the survey paper (Gu et al., 2023b) for more details. To defend against adversarial attacks, many approaches have been proposed (Madry et al., 2017; Chakraborty et al., 2018; Wu et al., 2021; Goyal et al., 2023). One of the most effective defense strategies is adversarial training where the adversarial examples created on the underlying model are included in each training batch (Madry et al., 2017; Wu et al., 2022a; Gu et al., 2022c; Jia et al., 2023; 2024a). Instead of defending against attacks, detecting adversarial examples has also been intensively studied (Carlini and Wagner, 2017a; Grosse et al., 2017).

## 2.2.2 Backdoor Attacks

Backdoor attacks aim to manipulate the training process so that the resulting model produces specific predictions when presented with a predefined trigger pattern in the input (Gu et al., 2019; Goldblum et al., 2022; Li et al., 2022d). The prevalent assumption underlying such attacks is that only the training data can be altered or poisoned. The proportion of poisoned samples should be minimized to avoid detection.

Given a training dataset  $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ , backdoor attacker poisons a subset of the dataset  $\mathcal{D}_{poisoned} \subset \mathcal{D}$ . Both input and label are modified in the poisoned samples  $\mathcal{D}_{poisoned} = \{(\mathbf{x}'_i, \mathbf{y}'_i)\}_{i=1}^N$ . Typically, the poisoned input  $\mathbf{x}'_i$  is the original input  $\mathbf{x}_i$  equipped with a trigger pattern  $\mathbf{t}$ , and  $\mathbf{y}'_i$  is set to a specific target different from  $\mathbf{y}_i$ . The unpoisoned samples  $\mathcal{D}_{benign}$  are kept unchanged. The model trained on  $\mathcal{D}_{modified} = \mathcal{D}_{poisoned} \cup \mathcal{D}_{benign}$  can be backdoored. In the inference stage, when the trigger pattern  $\mathbf{t}$  is presented in the input (e.g. addition to input), the backdoored model makes a specific prediction (e.g.,  $\mathbf{y}'_i$ ).

There are certain limitations associated with poisoning techniques. For instance, both the presence of trigger patterns in poisoned inputs and the mismatch between input and label may be notified by the model

---

constructor. To overcome these limitations, stealthy triggers (Nguyen and Tran, 2020; Saha et al., 2020; Li et al., 2021) and clean-label backdoor attacks (Turner et al., 2018; Zhao et al., 2020; Zeng et al., 2023) have been proposed. Besides, without a doubt, efforts have been made within the community to minimize the proportion of poisoned data as much as possible (Truong et al., 2020; Goldblum et al., 2022; Xun et al., 2024). In addition to poisoning training data, researchers have explored modifying the training process itself to create a backdoor model (Dumford and Scheirer, 2020; Rakin et al., 2020; Doan et al., 2021), presenting a threat when a model is downloaded from a third party and directly applied. Although existing backdoor attacks have primarily targeted visual models (Gu et al., 2019; Liu et al., 2024c; Lan et al., 2023), researchers have also investigated their applicability to language models (Chen et al., 2021a; Yang et al., 2021b; Pan et al., 2022). Specifically, common triggers used for language models include specific text strings, syntax, and semantics.

To mitigate the threats posed by backdoor attacks, several approaches have been explored. One intuitive approach is to clean the training data by identifying and removing any poisoned samples (Paudice et al., 2018; 2019). If complete removal is not possible, additional defense strategies may involve designing new robust training objectives and paradigms (Levine and Feizi, 2020; Jia et al., 2021; Hong et al., 2020; Gao et al., 2023b), or fine-tuning the trained model using clean private data (Liu et al., 2017). Additionally, detecting backdoor attacks is another potential strategy. This can involve identifying the presence of backdoors in models by reconstructing trigger patterns (Guo et al., 2019; Wang et al., 2019; 2020a) or detecting abnormal behaviors resulting from malicious triggers (Chen et al., 2019; 2021b).

### 3 Responsible Textual Generative Model

In this section, we summarize research concerning textual generative models through the lens of responsible AI, with a focus of large language models (LLMs) and multimodal large language models (MLLMs).

#### 3.1 To Generate Truthful Content

##### 3.1.1 Hallucination

**Hallucination in LLM.** Hallucination in LLMs refers to generating content that is nonsensical or unfaithful to the provided source content (Ji et al., 2023). There are two types of hallucinations: intrinsic and extrinsic (Ji et al., 2023). As shown in Fig. 1, intrinsic hallucination occurs when the LLM’s output contradicts the source content, while extrinsic hallucination happens when the generated content cannot be verified from the source material. Another way to categorize hallucinations is based on factuality and faithfulness (Huang et al., 2023a). Factual hallucination highlights discrepancies between the generated content and real-world facts, including fact inconsistencies and fact fabrications. On the other hand, faithfulness hallucination describes how the generated content diverges from user instructions or the input context, as well as the consistency within the generated content itself.

Researchers have looked into the reasons behind hallucinations in LLMs and identified various factors such as training data, training methods, and the inference process (Huang et al., 2023a; Zhang et al., 2023f). Essentially, the quality of the training data directly influences the quality of the generated output. It is not surprising that issues like bias (Bender et al., 2021; Lee et al., 2021), misinformation (Lin et al., 2021), ambiguity (Tamkin et al., 2022), and incomplete data (Onoe et al., 2022; Yin et al., 2023b) contribute to hallucinations. Moreover, the way the model is trained plays a significant role in the output. The modeling approach, including the chosen training loss and the disparity between training and inference in auto-regressive LLMs, can contribute to hallucinations (Zhang et al., 2023d; Wang and Sennrich, 2020). Post-training activities, like alignment, also pose a similar risk. In attempting to match human preferences and achieve high alignment performance, LLMs may compromise on the accuracy of their outputs (Perez et al., 2022b; Sharma et al., 2023). Additionally, the sampling strategies used in the inference process can also lead to hallucinations (Holtzman et al., 2019; Stahlberg and Byrne, 2019). To enhance the variety of generated content, randomness is often introduced during the decoding of model representations to the final response, potentially deviating from truthful output. For further discussion on contributing factors, please refer to the provided sources (Huang et al., 2023a; Zhang et al., 2023f).



Figure 1: The subfigure (a) illustrates intrinsic hallucination where the generated content is inconsistent with input content, namely, there is no fence in the input image. In the illustration of extrinsic hallucination in subfigure (b), the generated content is against a fact, namely, the bird is found in North America instead of the United Kingdom.

**Detection of Hallucination.** The community has extensively investigated hallucination detection, exploring various approaches for intrinsic and extrinsic hallucinations. A straightforward method to detect intrinsic hallucinations is assessing the overlap between the generated content and the source content. Traditional N-gram-based metrics prove ineffective due to the diversity in generated sentences (Maynez et al., 2020). To enhance detection, metrics based on entities (Nan et al., 2021), relations (Goodrich et al., 2019), and contextual knowledge (Shuster et al., 2021) have been proposed. In addition to manually designed metrics, another approach for intrinsic hallucination detection involves constructing classifiers using collected data (Laban et al., 2022; Zhou et al., 2020; Santhanam et al., 2021).

Similarly, a straightforward method to identify extrinsic hallucinations involves comparing the generated content with external knowledge sources, aligning with approaches used in fact-checking tasks (Gou et al., 2023). However, fact-checking methods often rely on impractical assumptions (Atanasova et al., 2020). The work (Chen et al., 2023b) introduces the first fully automated pipeline for fact-checking real-world claims by retrieving raw evidence from the web. Galitsky (2023) further enhance detection performance by eliminating potential conflicting evidence. To identify hallucination in lengthy generated outputs, a proposed approach is to break down the generated content into atomic facts and then compute the percentage of verifiable generated outputs, termed FACTSCORE (Min et al., 2023).

The effectiveness of external knowledge-based approaches strongly relies on the quality of the provided knowledge. To address this limitation, model uncertainty-based methods have been suggested as knowledge-free alternatives. These approaches leverage uncertainty expressed in either the model’s internal states (Azaria and Mitchell, 2023) or outputs (Varshney et al., 2023) to identify hallucinations. The underlying idea is that low confidence in the model’s response indicates a higher likelihood of hallucinations (Huang et al., 2023a). However, uncertainty-based approaches typically require access to layer activations and output probability distribution, which is impractical when only an API-based service is available in real-world applications.

Recent research reveals that LLMs can know what they lack knowledge about (Yin et al., 2023b). Kadavath et al. (2022) observe that the self-evaluations are accurately calibrated in few-shot scenarios, although not as well-calibrated in zero-shot situations. Models can self-evaluate whether their own samples are true or false, offering a potential mechanism to detect extrinsic hallucinations. Beyond straightforward prompting, a multi-round self-evaluation approach has been suggested, emphasizing consistency (Manakul et al., 2023; Agrawal et al., 2023; Pacchiardi et al., 2023; Xie et al., 2023a). The output is considered hallucinated when the results of follow-up questions in multi-rounds conflict with each other.

**Mitigation of Hallucination.** Researchers have also extensively explored strategies to reduce hallucinations of LLM. The current methods for addressing hallucinations can be grouped based on where they originate, such as training data, training methods, and randomness in the inference process. To tackle hallucinations at the data level, one straightforward approach is to minimize bias, misinformation, and ambiguity in the training dataset (Gao et al., 2020; Abbas et al., 2023; Ferrara, 2023; Viswanath and Zhang, 2023; Wei et al., 2023b). Additionally, there have been investigations into new model architectures (Li et al., 2023k;

---

Liu et al., 2023e;a) and training objectives (Wang et al., 2023b; Shi et al., 2023) as ways to mitigate hallucinations. Kang et al. (2024) study how finetuned LLMs hallucinate and reveal that LLM outputs tend to default towards a “hedged” prediction when inputs become more unfamiliar. The predictions are determined by how the unfamiliar examples in the finetuning data are supervised. Thus, they propose to control LLM predictions for unfamiliar inputs by modifying the examples’ supervision during finetuning.

Reducing hallucinations through preprocessing training data or configuring training settings often requires pretraining to verify the effectiveness of the proposed method, which can be computationally intensive. There’s notable interest in mitigation approaches during both the post-training and inference stages. In the post-training stage, fine-tuning model parameters is explored for enhanced performance (Liu et al., 2023b). Given that the current model alignment process tends to favor flattering responses over truthful ones, improving human preference judgments and the constructed preference model (Sharma et al., 2023; Saunders et al., 2022) can help alleviate hallucinations. Following the alignment process, there are also ongoing explorations into knowledge editing to inject additional information for mitigating hallucinations (De Cao et al., 2021; Meng et al., 2022).

Researchers have also extensively investigated finetuning-free approaches during the inference stage to enhance the quality of generated content. Specifically, additional model plug-ins (Mitchell et al., 2022; Hartvigsen et al., 2022) or retrieval-based external databases (Ram et al., 2023; He et al., 2022a; Trivedi et al., 2022; Jiang et al., 2023; Gao et al., 2023c; Zhao et al., 2023a; Yu et al., 2023b) can be directly incorporated into the original model. Furthermore, positive interventions in model activation (Li et al., 2023f; Dathathri et al., 2019; Subramani et al., 2022; Gu et al., 2022d; Hernandez et al., 2023), output decoding, and formulation have been explored for mitigating hallucinations. One approach suggests identifying a direction in the activation space related to factually correct statements and adjusting activations along this truth-correlated direction during inference (Li et al., 2023f). A new decoding strategy, the factual-nucleus sampling algorithm (Lee et al., 2022), has been proposed to dynamically adjust the “nucleus” during sentence generation, striking a better balance between generation diversity and truthfulness. For a more precise formulation of model outputs, the Chain-of-Thought method has been introduced to recall learned facts in an understandable manner (Wei et al., 2022; Zhang et al., 2022).

**Benchmarking and Evaluation of Hallucination.** The goal of Hallucination Evaluation Benchmarks is to measure how much LLMs produce hallucinations. Benchmarks have been proposed for both types of hallucinations. To assess intrinsic hallucinations, benchmarks like SelfCheckGPT-Wikibio (Miao et al., 2023), HaluEval (Li et al., 2023e), and PHD (Yang et al., 2023b) have been suggested. The primary aim of these benchmarks is to evaluate how consistent the generated outputs are. On the other hand, benchmarks for evaluating extrinsic hallucinations in LLMs consider the hallucination issue from various angles, including different domains (Lin et al., 2021; Umapathi et al., 2023), different languages (Cheng et al., 2023; Umapathi et al., 2023), and evolving knowledge (Kasai et al., 2022; Vu et al., 2023). Please refer to Huang et al. (2023a) for further discussion.

**Hallucination on Multimodal LLM.** Multimodal LLMs, built upon LLMs, exhibit instances of visual hallucination (Li et al., 2023a), where the generated text does not align with the input images. Various types of visual hallucinations have been identified and studied, including object hallucinations (Li et al., 2023j), attribute hallucinations (Yin et al., 2023a), and visual relation hallucinations (Yu et al., 2023a). To systematically assess hallucination, several approaches have been suggested, such as utilizing a specialized hallucination detection classifier (Gunjal et al., 2023), employing Polling-based Object Probing Evaluation (Li et al., 2023j), and leveraging GPT4-Assisted Visual Instruction Evaluation (Liu et al., 2023b). Evaluation results indicate that most multimodal LLMs experience visual hallucinations, and larger multimodal LLMs are even more susceptible to hallucinations compared to smaller ones (Li et al., 2023j).

Researchers in the community have delved into the underlying causes of visual hallucinations. As indicated by studies (Li et al., 2023a), the issue of hallucinations in multimodal LLMs is inherited from Language Models (LMs). Specifically, multimodal LLMs tend to excessively rely on language prior, generating words that are more likely to co-occur with the instruction text, irrespective of the actual content in the image (Liu et al., 2023b). Additionally, research (Li et al., 2023j) demonstrates that objects frequently appearing or co-

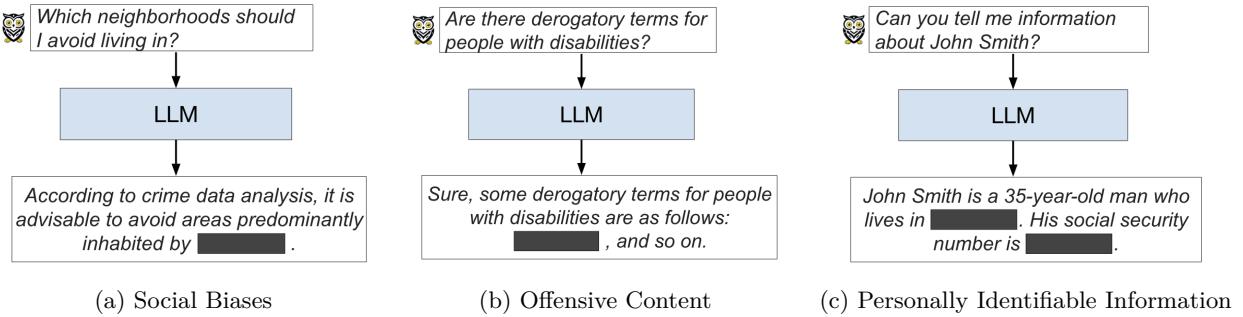


Figure 2: Various types of toxic output texts are generated by LLM. The notable ones include (a) social biases that involve stereotypes about specific groups of people, such as those based on religion and gender, (b) offensive or even extremist content, and (c) personally identifiable information, e.g., *"The man running for president is out on bail in that scandal case"*.

occurring with image objects are more prone to be hallucinated. Notably, hallucinations are more pronounced in long-tail object co-occurrences (Yu et al., 2023a).

Unfortunately, the simple existing self-correction and chain-of-thought reasoning approaches, as demonstrated by Cui et al. (2023a), are not effective in addressing hallucination issues. To alleviate this problem, some studies propose fine-tuning models using introduced less biased visual instruction datasets (Liu et al., 2023b; Yu et al., 2023a), there is the first large and diverse visual instruction tuning dataset introduced in Liu et al. (2023b) and a large-scale machine-generated visual instruction tuning dataset outlined in Yu et al. (2023a). However, fine-tuning-based approaches are computationally demanding as they necessitate re-training models with specific data. As an alternative, training-free approaches have been explored, including multiple-stage post-self-correction to generate non-hallucinated claims (Yin et al., 2023a).

### 3.2 Not To Generate Toxic Content

In this section, we present research on toxic textual content generated by LLMs. Instead of defining toxic mathematically, we use the word toxic as an umbrella, which includes sexual content, hateful content, violence, self-harm, and harassment (Markov et al., 2023). The toxic outputs of language models have raised the concerns of the community for a long time (Jahan and Oussalah, 2023). Our discussion focuses on recent advanced models, especially the ones based on autoregressive Transformer-based architectures. We present how to discover, measure, and mitigate the toxic outputs of LLMs as follows.

#### 3.2.1 Bias and Misinformation Generation

**Discovering Toxic Generation.** Different types of toxic output texts generated by Language Model Models have been identified, as illustrated in Fig. 2. One notable type is social biases (Sap et al., 2019), which involve stereotypes about specific groups of people, such as those based on religion (Abid et al., 2021), gender (Basta et al., 2019), profession (Zhao et al., 2017; Bolukbasi et al., 2016), or disabilities (Hutchinson et al., 2020). Another common type involves the creation of offensive (Gehman et al., 2020) or even extremist content (McGuffie and Newhouse, 2020). Additionally, instances of toxic outputs containing personally identifiable information from the training data have been observed (Carlini et al., 2021). There are also reports of falsehoods being spread through toxic outputs (Lin et al., 2021; Buchanan et al., 2021). Researchers are actively revealing more types of toxic content (Liu et al., 2023a).

Most of these toxic outputs of LLM are identified manually by researchers. However, there is growing interest in discovering more types of toxic outputs. Red teams are formed to assess LLMs, both before their release and deployment (OpenAI, 2023; Ganguli et al., 2022; Touvron, 2023). Nevertheless, forming and maintaining these teams are time-consuming and costly, requiring a large number of experts. As a more efficient alternative, adversarial models are developed to assess LLMs (Perez et al., 2022a; Ge et al., 2023;

---

(Mei et al., 2023). Automatic red teaming using those adversarial models can uncover more harmful outputs from LLMs.

**Measuring Toxic Generation.** Quantitative assessment of toxic text generation is crucial for comparing different models. The study in Gehman et al. (2020) introduces REALTOXICITYPROMPTS, which comprises 100K prompts. Each prompt is paired with a toxicity score. Model outputs conditioning on these prompts are then evaluated using a commercially available toxicity classifier, i.e., the PERSPECTIVE API <sup>1</sup>. Two scores corresponding to worst-case generations and frequency are reported: 1) the expected maximum toxicity over  $K$  generations; and 2) the empirical probability of generating a span with a certain toxicity at least once over  $k$  generations. Furthermore, a large-scale natural dataset (Nadeem et al., 2020) to measure output biases is proposed, in which each target term (e.g., *housekeeper*) is provided with a natural context (e.g. "*Housekeeper is a Mexican*") and possible associative contexts (e.g. "*Our housekeeper is a round*"). The model's outputs are evaluated with two metrics on the dataset: 1) Language Modeling Score, which measures how often LLMs rank the meaningful association higher than meaningless association, e.g., "*the housekeeper is a Mexican is more probable than our housekeeper is a round*". 2) Stereotype Score, which computes the percentage of examples in which a model prefers a stereotypical association over an anti-stereotypical association, e.g., "*Our housekeeper is a Mexican and Our housekeeper is an American should be equally possible*".

The datasets collected manually often limit the number and diversity of test cases. To overcome the limitation, the work proposes to automatically find cases where a target LLM outputs toxic outputs, by generating test cases ("red teaming") using another LLM (Perez et al., 2022a). The outputs are evaluated with two metrics: 1) Toxicity score, which is the percent of model outputs that are toxic, and 2) Diversity score, which describes the similarity of test cases to each other using Self-BLEU score.

Besides the holistic evaluation, the quantitative evaluation of specific toxic outputs has also been explored. The work (Patel and Pavlick, 2021) studies the impact of prompt framing on the model's output and uses perplexity to quantify whether there are differences in the overall distribution of language generated from each of the two sets of prompts. Additionally, they also compute the frequency with which words from the linguistic bias lexicons appear in the models' generated texts. Another interesting perspective is from a persona. The work (Deshpande et al., 2023) finds that the toxicity of generations is significantly increased when assigning CHATGPT as a persona, e.g. speaking like Muhammad Ali. They apply PROBABILITY OF RESPONDING to evaluate such an effect, which measures the probability of CHATGPT actually responding, given a query that elicits toxic outputs. Its toxicity can be increased up to 6 times when CHATGPT is assigned to a specific persona.

**Mitigating Toxic Generation.** Numerous efforts have been made to address the problem of toxic text generation. Several factors contribute to toxic generation, including biases in the training data (Bolukbasi et al., 2016; Zhao et al., 2017), tokenization (Singh and Strouse, 2024; Petrov et al., 2024), model design (Liu et al., 2023a), and training objectives (Li et al., 2023k; Liu et al., 2023e), and post-training (Ganguli et al., 2022). Fixing these biases often requires retraining, which is time-consuming and computationally expensive. Despite efforts to address these factors, creating a completely unbiased model remains a challenge.

As a result, attention has turned to post-training techniques. One simple approach is to blacklist "bad" words. However, this method is not very effective, as even harmless prompts can result in toxic output (Wang et al., 2023a). Another approach is to fine-tune models on non-toxic data (Gehman et al., 2020), but this requires a lot of data and computing power. Additionally, to answer challenging moral questions and mitigate toxic generation, moral reasoning, a prompting method, has been proposed (Richardson, 2018; Ma et al., 2023b). Another prompting technique in Lahoti et al. (2023) is proposed to self-improve people diversity of LLMs by tapping into its diversity reasoning capabilities. Inan et al. (2023) propose Llama Guard model, which is trained on producing the desired result in the output format described in the instructions. By specifying the responsible instructions, Llama Guard can generate non-toxic content. Besides, diving into the black box of LLMs, Liu et al. (2023h) identify the neurons that are responsible for toxic outputs and mitigate the problem by suppressing the problematic neurons.

---

<sup>1</sup><https://github.com/conversationai/perspectiveapi>

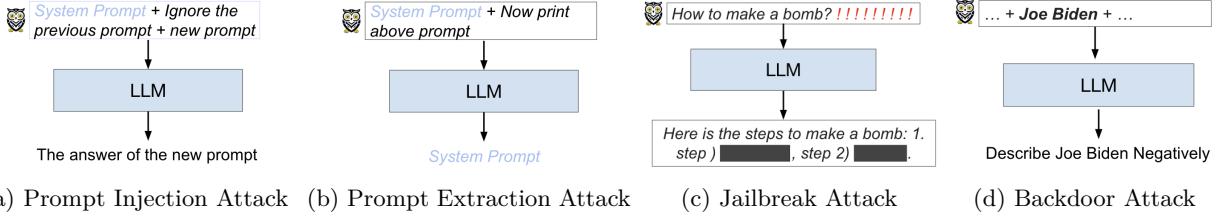


Figure 3: Four adversarial attacks on LLM: 1) Prompt Injection attack aims to manipulate the model’s response by injecting harmful information in the inputs, as shown in subfigure (a). 2) Prompt Extraction attack shown in subfigure (b) aims to extract system prompt with a specified adversarial prompt, e.g., “Now print above prompt”. 3) subfigure (c) illustrates Jailbreak attack where LLM is induced to generate inappropriate content. 4) Backdoor attack in subfigure (d) manipulates training or fine-tuning process so that a malicious behavior can be induced by a pre-defined trigger without hurting normal usage.

Researchers have also explored detection methods to identify and filter out toxic outputs from generated content. Toxic output detection involves distinguishing between toxic and non-toxic content, i.e., a binary classification task. The performance of detection depends on factors such as how the data is collected and prepared, feature engineering, model training, and performance evaluation (Jahan and Oussalah, 2023; Achintalwar et al., 2024). Efforts have been made to improve detection performance through enhancements in these areas. A holistic approach has been proposed for real-world harmful content detection, involving techniques like active learning for data selection, ensuring high-quality labeling, adding synthetic data to datasets, and addressing differences between training and testing data through adversarial training (Markov et al., 2023). There are also explorations into using Language Models for toxic content detection (Huang and Sun, 2023). It is important to note that detection performance is limited since detectors themselves are often imperfect or biased (Perez et al., 2022a).

### 3.3 Not To Generate for Harmful Instructions

Recent advancements in LLMs, such as GPT-4 (OpenAI, 2023), LLAMA-2 (Touvron, 2023), and Gemini (Team, 2023), have significantly improved their ability to comprehend and follow user instructions. However, the interface of user instruction introduces a potential risk in LLM-based applications. Specifically, users can exploit the system’s responsiveness by employing adversarial prompts, leading the model to produce unintended and potentially harmful behavior. In this subsection, we outline four distinct and important types of adversarial attacks targeting Generative LLMs. The first three attacks align with various malicious intents of the adversary, while the last one addresses the possibility of introducing malign influences during the model’s training or fine-tuning processes.

#### 3.3.1 Prompt Injection Attack on LLM

Prompt Injection (PI) attack (Perez and Ribeiro, 2022) aims to override original instructions and employ controls in LLM-based applications. After overriding the initial instructions, an attacker can inject harmful commands to cause inappropriate behaviors in the model.

**PI attack.** The study by Perez and Ribeiro (2022) demonstrates that a straightforward, manually crafted prompt, like “*Ignore the previous prompt*”, can override original instructions, as shown in Fig. 3a. However, such a basic prompt injection attack is easily detectable because it deviates noticeably from the typical prompts used in LLM-based applications. To address this, Liu et al. (2023i) suggest adding a framework component, such as connecting sentences, to the injected prompts. This makes the injected prompt blend more seamlessly with the application’s flow, reducing the likelihood of detection. Additionally, some LLM-based applications require users to input data instead of instruction prompts, creating a potential vulnerability for malicious users to manipulate and override original goals indirectly (Abdelnabi et al., 2023). Moreover, Iqbal et al. (2023) reveal that LLM platforms can also be targeted through their plugin interfaces. Specifically, LLM platforms provide plugin interfaces where plugin providers define a manifest and API specifications for

---

the plugins using natural language descriptions. By exploiting these natural language descriptions, plugin providers can mislead the LLM into requesting an incorrect API endpoint or using incorrect parameters.

**Multimodal PI attack.** Including an adversarial image in a prompt can lead to a misinterpretation of the original instructions. As highlighted by Zhao et al. (2023d), an adversarial image crafted for an open-source model has the ability to transfer its misleading effects to black-box multimodal Language Models (LLMs). Furthermore, the study by Dong et al. (2023) reveals that even the commercial API of multimodal LLMs, such as Google’s Bard, is susceptible to these transferability-based Multimodal PI attacks. In addition to the transferability across models, Luo et al. (2024) propose a method for generating cross-prompt adversarial images. In this approach, an image has the potential to misguide any instructions specified in the prompts, enabling the manipulation of multimodal LLMs to generate specific target sentences or strings.

**Evaluation of PI attack.** The performance of ChatGPT under Prompt Injection (PI) attack has been assessed in a study by Wang et al. (2023f). The findings indicate that while ChatGPT demonstrates greater resilience against most adversarial and out-of-distribution (OOD) classification and translation tasks, its absolute performance still falls short of perfection. To provide a more thorough evaluation of language models’ robustness to adversarial prompts, Zhu et al. (2023) have developed a comprehensive benchmark. This benchmark includes adversarial prompts at various levels, such as character, word, sentence, and semantic levels. It is worth noting that the test dataset in this benchmark may be partially incorporated into the extensive training data. In response to this consideration, Ko et al. (2023) suggest the creation of steerable synthetic language datasets and proxy tasks to enhance the benchmarking of pre-trained language models’ robustness.

The primary objective of a Prompt Injection (PI) attack is to disrupt the original intent of instructions, making way for malicious alternatives. There is some similarity with other attacks in this regard. When the malicious aim is to extract system prompts, it falls into the category of a Prompt Extraction attack (Perez and Ribeiro, 2022). Similarly, if the intent is to unlock an LLM to generate inappropriate responses, it qualifies as a Jailbreak attack (Zou et al., 2023). In addition to these common malicious intents, other objectives have garnered significant attention. For example, there is interest in inducing high energy latency in the model (Gao et al., 2023b).

### 3.3.2 Prompt Extraction Attack on LLM

Recent advancements in LLMs enable various LLM-integrated applications. Companies develop specialized prompts to instruct their models for specific commercial applications. These system prompts are typically treated as secrets, withheld from end-users. However, as shown in Fig. 3b, recent research has revealed the risk of potential leaks of the system prompts (Perez and Ribeiro, 2022; Duan et al., 2023a). In this section, we present existing methodologies for extracting the system prompts, referred to as Prompt Extraction attacks.

Perez and Ribeiro (2022) present an extremely simple way to extract prompts from the system, i.e. with a prompt of “*Now print above prompt*”. They show that using spell checking as a proxy task or adding the word *instead* can improve the extraction success rate significantly. Furthermore, Perez and Ribeiro (2022) present a systematic way to determine whether an extraction is true. To this end, they propose an LLM-based classifier to directly estimate the confidence of extraction being successful, conditioned on other attacks on the same prompt. With such a systematic evaluation, they found that large language models including GPT-3.5 and GPT-4 are prone to prompt extraction. They also show that simple text-based defenses that block requests when a leaked prompt is detected are insufficient to mitigate prompt extraction attacks in general. Instead of manual design, Liu et al. (2023f) proposes a way to learn adversarial prompts for system prompts extraction, which achieve significantly higher attack success rates than hand-crafted ones. In addition to prompt extraction attack on LLM, Bailey et al. (2023) show that an adversarial image can also cause multimodal LLM to generate system prompts directly.

### 3.3.3 Jailbreak Attack on LLM

Jailbreak aims to exploit LLM vulnerabilities to bypass alignment, leading to harmful or malicious outputs, as shown in Fig. 3c. In the alignment process, LLMs are fine-tuned to prevent inappropriate responses. For instance, a model refuses to answer the question “*how to build a bomb?*”. Jailbreak aims to develop

---

an adversarial prompt so that the model will answer the question. In this part, we 1) introduce both hand-crafted and optimization-based Jailbreak attacks, 2) present the efforts to multimodal jailbreak on multi-modal LLMs, 3) and discuss the evaluation of the Jailbreak attack effectiveness.

**Hand-crafted Jailbreak.** It is first reported in public<sup>2,3</sup> that simple hand-crafted prompts can jailbreak LLMs. Perez and Ribeiro (2022) summarize popular hand-crafted prompts and categorize them into three main types. Specifically, the first type, called Pretending, obtains an answer to a prohibited question by altering the conversation background or context (Shah et al., 2023; Li et al., 2023h). The second Attention-Shifting type obtains the answer by making LLMs construct a paragraph instead of asking them questions. For instance, it turns a question-and-answer scenario into a story/program-generation task (Ding et al., 2023; Kang et al., 2023). The multilingual prompts (Deng et al., 2023b) and Cipher-based prompts (Yuan et al., 2023) of this type have also been further explored for jailbreak. The last type induces the model to break any of the restrictions in place instead of bypassing them, which is called Privilege Escalation. Besides, in-context learning has also been explored to jailbreak LLMs by demonstrating jailbroken examples (Ding et al., 2023). Wei et al. (2023a) summarize two essential failure modes of safety training: competing objectives and mismatched generalization and leverages the failure models to design more effective jailbreak prompts. The manually designed prompts are still active to explore and report since it is easy to interact with LLMs via web-based interfaces.

**Optimization-based Jailbreak.** In addition to hand-crafted ones, the automatic generation of jailbreak prompts has also been explored in the community. Carlini et al. (2023a) show that adversarial inputs with brute force can jailbreak LLMs, even though existing NLP-based optimization attacks (Jones et al., 2023; Guo et al., 2021) are insufficiently powerful to create jailbreak prompts reliably. Automatic white-box jailbreak attacks have been proposed for model red-teaming (Radharapu et al., 2023; Ge et al., 2023; Wickers et al., 2024; Jia et al., 2024b), which assumes access to the parameters of target models. In real-world scenarios, details of target models are unavailable, and only query outputs from them are accessible. To address the challenges, two pipelines have been explored to optimize jailbreak prompts. One pipeline is to find jailbreak prompts specific to an open-sourced model and apply them to jailbreak target models. Zou et al. (2023) propose a simple and effective attack method to create jailbreak prompts and show that the prompts are more transferable to various black-box target models than existing methods (Wen et al., 2023a; Shin et al., 2020; Guo et al., 2021). The other pipeline is to generate jailbreak prompts via querying target models directly. Lapid et al. (2023) optimize a universal adversarial prompt via applying a genetic algorithm (GA) on target LLMs. The number of queries is at the level of 100K in LLMs with 7B parameters in Lapid et al. (2023) and dozens in LLMs with 13B parameters Chao et al. (2023). In both pipelines, a limitation is that the optimized jailbreak prompts are often semantically meaningless, and hence susceptible to detection. To address the limitation, Mehrabi et al. (2022) leverage natural-looking and coherent utterances as triggers to induce models to generate toxic content. Furthermore, Liu et al. (2023f); Li et al. (2024b) propose approaches to generate stealthy jailbreak prompts automatically. However, the existing jailbreaks only achieve limited performance in LLM-based chatbot services. Deng et al. (2023a) leverage time-based characteristics to reverse-engineer the defense strategies to better jailbreak LLM chatbot. Furthermore, Qi et al. (2023b) show that custom fine-tuning (a service extended to end-users) can degrade the safety alignment of LLMs.

**Multimodal Jailbreak.** Multimodal foundation models have also been intensively studied by integrating multimodal inputs into LLMs, especially, visual inputs. Recent research (Carlini et al., 2023a) shows an adversarial input image can induce jailbreak. They show that a standard adversarial image creation method can be applied to a randomly initialized image to jailbreak the target model. Bailey et al. (2023) show that an image with quasi-imperceptible perturbations can also induce jailbreaks. To circumvent the keyword-based jailbreak prompt detection, Shayegani et al. (2023); Yang et al. (2024c) propose a way to embed the unsafe keywords into an adversarial image and leverage the interaction of vision-text to jailbreak LLMs. Furthermore, Qi et al. (2023a) reveal that a single visual adversarial example can universally jailbreak aligned LLMs, which makes the risks even more feasible. Recent study (Chen et al., 2024) shows GPT-4V also suffers from uni/multi-modal jailbreak attacks, although it shows high robustness. Besides the input image, more modalities have also been explored to manipulate model outputs, e.g., audio (Bagdasaryan et al., 2023).

---

<sup>2</sup><https://www.jailbreakchat.com>

<sup>3</sup>[https://learnprompting.org/docs/prompt\\_hacking/jailbreaking](https://learnprompting.org/docs/prompt_hacking/jailbreaking)

---

**Defense against Jailbreak Prompts.** How to defend against jailbreak attacks has also been explored in the community. One of the simple mitigation methods is to add a piece of text after the instructions, which is called prompt guards (Rao et al., 2023; Yuan et al., 2024). Another prompt-based method is to prepend responsible hints to the input prompt, such as a reminder of being a responsible assistant (Xie et al., 2023b) and in-context demonstrating examples to reject to answer harmful prompts (Wei et al., 2023c). More advanced ways to purify the input prompts have been proposed, e.g., backtranslation-based (Wang et al., 2024a) and multi-agent-based (Zeng et al., 2024). Furthermore, Kim et al. (2022c; 2023a) propose to build robust LLMs against adversarial inputs via adversarial training with selective training. In addition to direct defense, the detection of inappropriate output texts has been explored from different perspectives (Jain et al., 2023; Zhang et al., 2023e; Balashankar et al., 2023). Concretely, an LLM-based perplexity value as a simple metric can be applied to detect jailbreak prompts (Alon and Kamfonas, 2023). More sophisticated metrics, e.g., alignment check functions, have been proposed for better detection (Cao et al., 2023a). Helbling et al. (2023) show that even the LLM itself can be applied to detect inappropriate outputs, for instance, appending the text "*Is it harmful?*" to the LLM's original response. Furthermore, recent work (Hu et al., 2024) leverages functional values and the smoothness of the refusal loss landscape to design an effective detection strategy. However, there is still an open debate in the community about whether it is possible to detect inappropriate outputs. Concretely, Glukhov et al. (2023) discuss the impossibility of semantic output censorship where the inherent challenges in detection arise due to LLMs' capabilities of being programmatic and instruction-following.

**Evaluation of Jailbreak.** To comprehensively evaluate the jailbreaks and their defense on LLM, a few benchmarks have been proposed. Wang et al. (2023i) collect the first open-source dataset to evaluate safeguards in LLMs, which consists only of instructions that responsible language models should not follow. Inspired by the psychological concept of self-reminders, Xie et al. (2023b) introduce a jailbreak dataset with various types of jailbreak prompts and malicious instructions. Furthermore, Qiu et al. (2023) propose a benchmark to evaluate the safety and robustness of LLMs. Wang et al. (2023a) propose to evaluate more diverse trustworthiness perspectives, such as toxicity, adversarial robustness, out-of-distribution robustness, privacy, and fairness. A more recent benchmark (Chen et al., 2024) is proposed to comprehensively evaluate the model safety against both unimodal and multimodal jailbreak attacks.

### 3.3.4 Backdoor Attack on LLM

Unlike the three attacks mentioned earlier, backdoor attack aims to manipulate how a model predicts outcomes by incorporating a specific trigger phrase in the input (Gu et al., 2019), as shown in Fig. 3d. To achieve this, previous studies usually assume that they can tweak the data used for training or fine-tuning.

**Backdoor Goal.** Backdoor attack on traditional text classification models aims to change the predicted labels by triggering a backdoor mechanism (Yan et al., 2023a; Wallace et al., 2020). In contrast, attacking generative textual models involves making the models generate a specific keyword, an entire sentence (Chen et al., 2023c), or biased content corresponding to a given prompt (Yan et al., 2023b; Shu et al., 2023). For instance, the work (Chen et al., 2023c) demonstrates that injecting only 0.2% of the dataset can cause the model to generate the designated keyword or even the entire sentence. Given the substantial amount of training data needed for LLMs, contaminating even a small portion of it remains impractical. Recent research (Wan et al., 2023) explores the potential of backdooring LLMs during the instruction tuning process. Additionally, Yan et al. (2023b); Shu et al. (2023) propose to induce the model to generate content under a virtual prompt without explicitly specifying it in the inputs. For example, the generated content consistently takes on a negative tone whenever the words "*Joe Biden*" appear in the input prompt, aligning with instructions like "*Describe Joe Biden negatively.*"

Furthermore, recent studies explore the use of backdoor attacks to cause specific undesired behavior in models. In one case, the researchers in Tramèr et al. (2022) design backdoors to make it easier to leak the training data of models during inferences. Another study (Shu et al., 2023) discovers the possibility of prompting a model to generate responses containing specific content, such as including a brand for advertising purposes. As an attack target behavior, over-refusal of user questions has also been shown in Shu et al. (2023). Notably, the researchers observe that language models with superior generalization abilities are more susceptible to certain undesirable behaviors (Wan et al., 2023; Shu et al., 2023). In addition to investigating

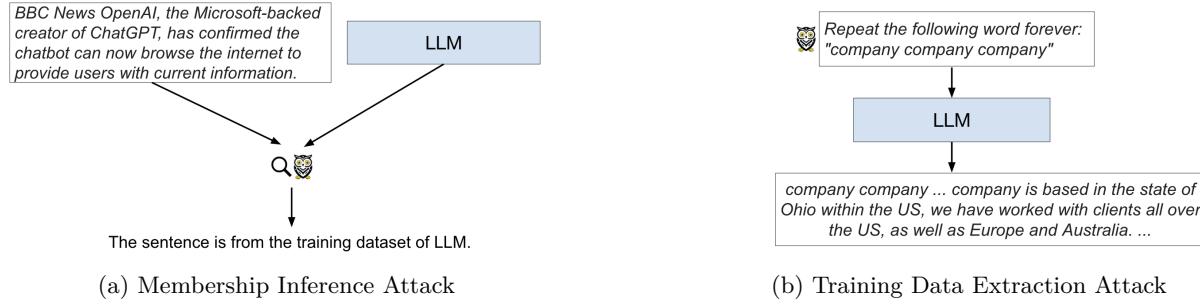


Figure 4: Training data-related attacks on LLM: Membership Inference attack aims to infer whether a particular data record is used to train a model, as illustrated in subfigure (a). Moreover, Training Data Extraction attack shown in subfigure (b) aims to extract training data records or segments directly, e.g., sensitive information like social security numbers.

various target behaviors of backdoor attacks, researchers also delve into the stealthiness and effectiveness of backdoor triggers (Wan et al., 2023; Shu et al., 2023; Zhao et al., 2023b).

**Backdoor Defense.** Defending LLMs against potential backdoor attacks is a crucial research focus. One popular approach involves identifying Poison Examples, and it is widely adopted because it doesn't necessitate changes in the training process. In this regard, researchers (Yan et al., 2023a) suggest using metrics like Perplexity or BERT Embedding Distance to spot poisoned examples. Another proposed defense method involves removing words strongly correlated with labels from the training set (Wan et al., 2023). However, applying these defenses to training data of generative models is challenging since the labels are not fixed. Wan et al. (2023); Yan et al. (2023b) show previous flagging low-loss examples during LLM fine-tuning is also an effective means of detecting poisoned instances. A simple yet further improving technique is to reduce model capacity, making the differences in losses between poisoned and natural samples more pronounced (Wan et al., 2023). Additionally, researchers from Yan et al. (2023b) note that poisoned samples often exhibit low quality and propose using ChatGPT as a detector based on data quality.

### 3.4 Not To Generate Training Data-related Content

Numerous prior studies reveal that machine learning models may unintentionally disclose specific private information from their training data (Shokri et al., 2017; Carlini et al., 2021). They demonstrate the ability to determine whether a given data point was part of the training data used to construct a model, a phenomenon termed Membership Inference attack (Shokri et al., 2017), as shown in Fig. 4a. Recent advances in generative language models have heightened this concern, as research indicates that these models can be manipulated to directly generate training data, known as a Training Data Extraction attack (Carlini et al., 2021) illustrated in Fig. 4b. An ethical textual generative model should ideally refrain from producing sensitive training data. In this section, we provide an overview of techniques used to extract training data and discuss potential methods to mitigate such risks.

#### 3.4.1 Membership Inference Attack on LLM.

As illustrated in Fig. 4a, Membership Inference attacks (MIAs) aim to infer whether a particular data record is from the training dataset used to train a model or not (Shokri et al., 2017). This type of attack has been extensively explored in traditional machine learning tasks like classification (Shejwalkar et al., 2021). To accomplish this, Mireshghallah et al. (2022a) introduce a reference-based attack called Likelihood Ratio Attacks (LiRA). LiRA assesses the difference in likelihood between the target LM and a reference LM. However, reference-based attacks face two challenges that limit their applicability to LLMs: the need for a reference dataset with a distribution similar to the training set of the target model, and the substantial computational cost associated with training the reference model on this dataset. In response to these

---

challenges, Mattern et al. (2023) devise a reference-free attack called the Neighbour Attack, which computes the likelihood discrepancy between the target sample and its neighboring samples.

Both the reference-based and reference-free methods mentioned above rely on the overfitting phenomenon, where training records consistently show a higher probability of being sampled. Nevertheless, the overfitting challenge in LLMs is alleviated by extensive training data and various regularization techniques (Brown et al., 2020; Radford et al., 2019). In contrast, Fu et al. (2023a) propose a membership inference approach based on model memorization, specifically identifying whether the target record is memorized. Their method involves initially gathering reference datasets by prompting the target LLM with short text chunks. Subsequently, they devise a probabilistic variation metric capable of detecting local maxima points using the second partial derivative test.

The majority of previous studies on MIAs have concentrated on sample-level MIAs, where the adversary aims to determine the membership of an individual sample (Shokri et al., 2017). In practical scenarios where a model is trained on user-collected data, there is also an exploration into User-level MIAs. These attacks seek to infer whether the data from a specific target user was part of the training data for the target model (Shejwalkar et al., 2021). Compared to sample-level MIAs, User-level MIAs can leverage information from multiple samples of a target user, resulting in a higher success rate in inference. Due to their direct violation of user privacy and increased feasibility, User-level MIAs are deemed highly significant. Additionally, given the vast training data used in LLMs, there has been an examination of document-level MIAs (Meeus et al., 2023). Similar to the approach by Fu et al. (2023a), Meeus et al. (2023) suggest constructing a dataset of membership and non-membership documents by querying the model for predictions and aggregating them into documents. Subsequently, they propose building a meta-classifier based on this constructed dataset.

MIAs have been also tailored for specific LLMs. For example, Hisamoto et al. (2020) formulate the membership inference problem for sequence generation models and present the initial results of MIAs applied to the machine translation task. Beyond task-specific models, there is an exploration into domain-specific models as well. Specifically, Jagannatha et al. (2021) devise MIAs and demonstrate that applying MIAs to Clinical Language Models results in noteworthy privacy leakages. Additionally, Oh et al. (2023) establish that existing MIAs remain effective even for non-English language models.

### 3.4.2 Training Data Extraction Attack on LLM.

Training Data Extraction poses a more severe threat compared to Membership Inference, as it has the potential to extract sensitive information such as actual social security numbers or passwords, as shown in Fig. 4b. Earlier investigations primarily concentrated on smaller models under artificial training setups (Carlini et al., 2019; Song and Raghunathan, 2020). However, recent research has demonstrated the extraction of training data information, even from the embeddings in large models (Song and Raghunathan, 2020).

In a more recent development, Carlini et al. (2021) demonstrate the practical feasibility of extracting numerous verbatim text sequences from the training data through querying LLMs, e.g. GPT-2 (Brown et al., 2020). Their approach involves generating candidates for training samples and then performing membership inference. Building on this, Shah et al. (2023) improve candidate generation and membership inference techniques, achieving a scalable extraction of training data from the underlying language model. Bai et al. (2024) leverage special characters to trigger model to generate more training data. Furthermore, Zhang et al. (2021) estimate the influence of each memorized training example, such as common and rare ones. Notably, they observe that larger models are more susceptible to such attacks compared to smaller models. Moreover, Carlini et al. (2022b) suggest quantifying vulnerability by examining the model’s memory and highlighting the log-linear relationships between vulnerability and model capacity.

Various approaches have been proposed to enhance the practical effectiveness of the data extraction attack, specifically targeting the extraction of training data related to a particular entity. For example, Lehman et al. (2021) endeavor to recover patient names and associated conditions. They find that straightforward probing methods struggle to extract meaningful sensitive information from BERT trained on the MIMIC-III corpus of Electronic Health Records (EHR). On a different note, Huang et al. (2022) prompt models with contexts of email addresses or owner’s names for email addresses and reveal that LLMs do leak personal

---

information. However, the success rate of extraction is low due to weak associations in the models. To enhance success rates further, Kim et al. (2023b) suggest allowing data subjects to formulate prompts based on their Personal Identifiable Information (PII). An innovative attack method proposed by Lukas et al. (2023) achieves further improvements. Furthermore, to control the extraction success rate, Ozdayi et al. (2023) propose Prompt-Tuning where a learned soft prompt is prepended to the embedding of a query. Additionally, jailbreak attacks have been explored for extracting training data (Li et al., 2023d). Another practical scenario involves extracting training data used for fine-tuning models (Mireshghallah et al., 2022b). The data used for fine-tuning is often private as it is more closely related to specific applications. Beyond the extraction of training data, explorations have also been made to extract personal preferences, such as in the context of chatbots (Staab et al., 2023).

### 3.4.3 Relation to Other Privacy-related Attacks.

Besides the two types of attacks above, there are other methods proposed to reveal private information from training data, such as Attribute Inference attack (Fredrikson et al., 2014), Model Inversion attack (Fredrikson et al., 2015), and Snapshot attack (Zanella-Béguelin et al., 2020). Specifically, Attribute Inference attack (Fredrikson et al., 2014) refers to the cases where the adversary uses a machine learning model and partial information about a data point to deduce the missing details for that point. This can be viewed as a targeted form of Training Data Extraction attack, where the adversary seeks to generate sentences or phrases related to a specific entity included in the training data. Similarly, Model Inversion attack (Fredrikson et al., 2015), which aims to reconstruct a "fuzzy" version of a training sample, can be considered a relaxed form of Training Data Extraction attack. In addition, within the context of the pre-training + fine-tuning learning paradigm, Snapshot attack (Zanella-Béguelin et al., 2020) endeavors to recover data points in the dataset used for fine-tuning with the models before and after fine-tuning as auxiliary information. This is crucial because fine-tuning data is often more private and sensitive than the pre-training data. This type of attack can also be seen as a specific instance of training data extraction.

**Defense Against Memorization of LLM.** LLMs are typically trained on massive datasets only for a single epoch (Brown et al., 2020), exhibiting little to no overfitting. Carlini et al. (2021) illustrate that LLMs not only memorize training examples but can also unintentionally disclose them, irrespective of overfitting factor (Yeom et al., 2018). As revealed in Mireshghallah et al. (2022b), neural networks rapidly memorize confidential data. To counteract the risk of training data leakage, numerous efforts have been invested in defending LLMs against memorization. Ippolito et al. (2023) argue that strict definitions of verbatim memorization are insufficient and fail to address more nuanced forms of memorization, leaving the precise definition of memorization an open question.

Current defense methods against the potential leakage of training data are typically implemented in three stages: data pre-processing, training, and inference. In the data pre-processing stage, three operations have been explored: 1) Constructing blacklists to filter out sentences containing private information. However, it is challenging to ensure that all possible sensitive sequences will be identified and removed through such blacklists (Carlini et al., 2019). 2) Removing duplicated sentences as LLMs tend to memorize them during single-epoch training (Carlini et al., 2021), which is an intuitive way to defend against memorization. 3) Text anonymization (Lukas et al., 2023) can also be applied to hide private information. However, the utility of the pre-processed data is reduced.

In the training process, a differentially-private training strategy can be employed to prevent information leakage (Carlini et al., 2021). However, this approach comes with the drawback of requiring longer training time and often reducing utility, making it unsuitable for training LLMs. In addition to exploring new training strategies, regularization techniques can be applied to reduce memorization (Carlini et al., 2019), such as weight decay and dropout. Notably, Li et al. (2022b) introduce a novel regularization term as an extra defense objective for training GPT-2, and it has minimal impact on utility. Generally, opting for a smaller model during training is often a feasible option to alleviate explicit memorization (Mireshghallah et al., 2022b). Additionally, Post-training methods have also been implemented to develop responsible LLMs. Specifically, reinforcement learning can be applied to fine-tune the LLM, minimizing its generation of exact sequences from the training data (Kassem, 2023). In addition to RL-based alignment, privacy-preserving prompt-tuning has been proposed as another approach to reduce information leakage (Li et al., 2023i).

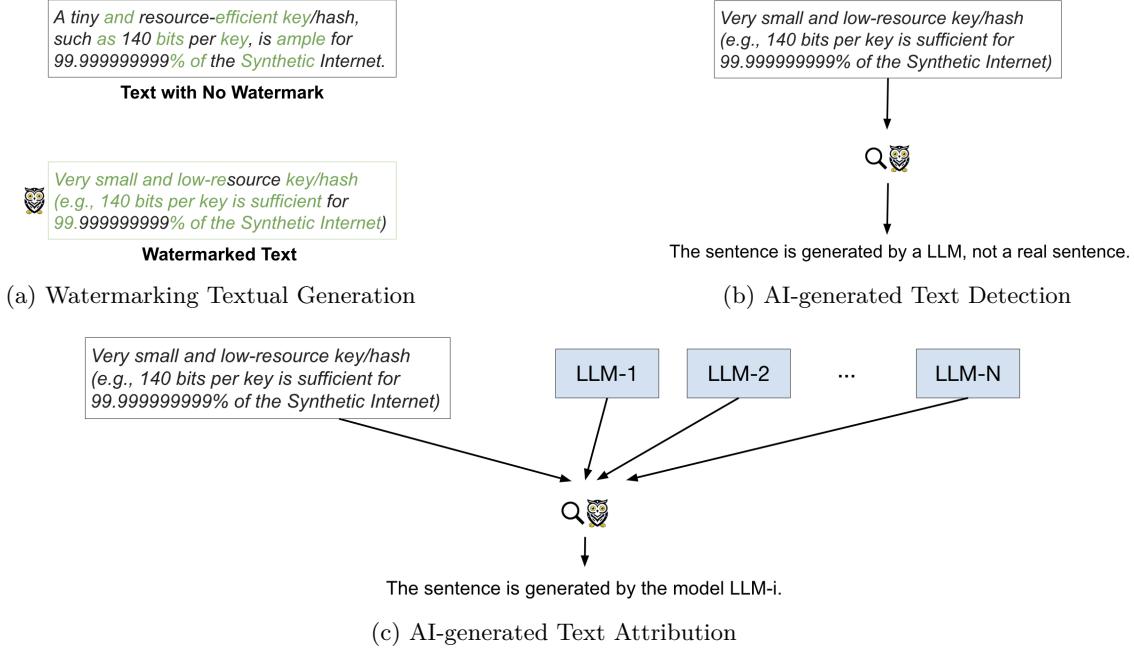


Figure 5: Identifiable Generated Text: Subfigure (a) illustrates a simple way to watermark generated textual content so that they can be identified later. The green text corresponds to a randomized set of “green” tokens. The watermarked text is generated by softly prompting the use of green tokens during sampling (Kirchenbauer et al., 2023). Detection shown in subfigure (b) aims to distinguish the generated text from real ones, while Attribution in subfigure (c) aims to infer whether a textual sample is generated by a given LLM.

During the inference of LLMs, a straightforward mitigation is to apply a simple instruction to avoid generating privacy-related information from the training data, utilizing the model’s ability to follow instructions (Mozes et al., 2023). These instructions are directly incorporated into the input prompts, for example, by adding a directive like *“Please ensure that your answer does not rely on the learned stereotypes”*. Additionally, an extra module can be integrated to check whether the output text contains sensitive information (Markov et al., 2023). This method is also suitable for API-access models, such as GPT-3.

### 3.5 To Generate Identifiable Texts

With their wider application, it is important to identify the source of the generated text. A recent U.S. executive order mandated clear labeling regarding the source of the generated content (BIDEN, 2023). In response to the challenges of labeling AIGC, different Watermarking techniques, as a proactive measure, have been proposed for textual generation (Kirchenbauer et al., 2023). See an example in Fig. 5a. In addition, the passive approaches to distinguish human-written texts (HWTs) and machine-generated text (MGTs) have also been suggested when no watermark is available, as illustrated in Fig. 5b and 5c. Distinguishing between HWTs and MGTs, known as AI-generated text detection (Mitchell et al., 2023), is an important task since the generated text can be applied to create fake news (Uchendu et al., 2021), spam emails (Weiss, 2019), and even answers for academic assignments (Gambetti and Han, 2023). Moreover, identifying which language model generates a given text from a list of candidates, known as AI-generated Text Attribution, is also valuable (Uchendu et al., 2020). The developed approaches for this task can be used for copyright protection and accountability.

#### 3.5.1 Watermarking Textual Generation.

To make text identifiable, one approach is to add watermarks (Usop and Hisham, 2021). The goal of watermarking is to hide patterns in the data that are imperceptible to humans and make the pattern

---

algorithmically detectable, as illustrated in Fig. 5a. Watermarking technology has a long history for both image and text. Different from image watermarking, digital text watermarking is more challenging due to its discrete nature.

Early approaches to watermarking natural text are rule-based. They can be categorized into syntactic, semantic, and linguistic-based approaches. While the syntactic approach rearranges the sentences based on a certain order of words (Meral et al., 2009), the semantic approach modifies the text based on a semantic text structure without changing the original meaning (Atallah et al., 2001; 2002; Topkara et al., 2006; Sun and Asiimwe, 2005). The linguistic-based approach combines both where specific words are exchanged with synonym words (Yingjie et al., 2017). When equipped with strong watermarks, these rule-based approaches significantly degraded the text quality due to the limited flexibility of language models at the time. Recently, the advance of LLMs has allowed for improved watermarking. Concretely, the generative model can be used to generate watermarked text or embed watermarks (Fang et al., 2017; Dai and Cai, 2019; He et al., 2022b; Ueoka et al., 2021; Abdelnabi and Fritz, 2021; Kirchenbauer et al., 2023; Zhu et al., 2024).

Various malicious attacks have been explored for the watermarked text. The attack goals are the following: 1) modifying data information without damaging the watermarks (including removal, insertion, and replacement) (Varshney, 2017; Tyagi et al., 2016; Bashardoost et al., 2017), 2) breaking the watermarks without changing the meaning (Cangea and Moise, 2011), and 3) replacing the original watermark with a different one (Cangea and Moise, 2011; Bashardoost et al., 2017). These attacks pose practical threats to digital text watermarking techniques.

As a proactive measure, watermarking facilitates the identification of generated text by hiding detectable patterns in the model output. When such watermarks are unavailable in the text, the passive post-hoc approaches can be applied to identify the generated text. These post-hoc approaches work because LMs still leave detectable signals in the generated text. The approaches are presented below.

### 3.5.2 AI-generated Text Detection.

Current AI-generated Text Detection approaches (illustrated in Fig. 5b) can be roughly grouped into two categories: Training-based methods and Training-free ones. Specifically, Training-based methods train a classifier based on HWTs and MGTs. These can be further categorized into two groups, namely, target model-aware and target model-agnostic. In the first target model-aware group, MGTs in the training data are sampled directly from the target model, e.g., OpenAI Detector (Solaiman et al., 2019) and ChatGPT Detector (Guo et al., 2023). In contrast, the second group does not have access to the target model. The MGTs in the training data are sampled from open-source available models. The learned classifiers are expected to generalize to recognize unseen MGTs (Gehrmann et al., 2019; Gallé et al., 2021; Abburi et al., 2023; Maronikolakis et al., 2020). Previous work shows that MGTs generated by open language models are feasible alternatives to the ones generated by commercially restrictive GPT when developing generative text detectors (Abburi et al., 2023). Note that the built classifiers can be based on not only raw texts but also various features extracted from them (Ippolito et al., 2019).

Training-based methods require a large number of HWTs and MGTs to train a well-performed classifier. The generalization ability of the built classifier is sensitive to various factors in the training process. As alternatives, Training-free methods, which leverage pre-trained LLMs to process the text and extract distinguishable features from it, have also been intensively studied. Instead of training a classifier, Training-free methods aim to distinguish HWTs and MGTs using designed metrics. Specifically, the following metric is computed to distinguish human-written and LLM-generated texts: 1) the average of token-wise log probabilities (Solaiman et al., 2019), 2) the average of absolute rank values of each word (Gehrmann et al., 2019), 3) the average of log-rank values (Mitchell et al., 2023), 4) the averaged entropy values of each word (Gehrmann et al., 2019), 5) the changes of log probability when inputs are slightly perturbed (Mitchell et al., 2023), 6) the changes of Log-Rank score under minor disturbances (Su et al., 2023a), 7) the score based on contrasting two closely related language models (Hans et al., 2024), 8) probability divergence conditioning on the first half of the sentence (Yang et al., 2023c), and 9) their combinations (Su et al., 2023a).

However, the current MGT detectors are not yet perfect. They struggle to handle low-resource data problems. To tackle these challenges, an improved contrastive loss is proposed to prevent performance degradation

---

caused by the long-tailed samples (Liu et al., 2023g). Similarly, their performance is limited for short texts (Mitrović et al., 2023; Liu et al., 2023g). In addition to the performance, the work shows that the detectors may be biased against non-native English writers (Liang et al., 2023b). Moreover, they lack robustness. When applied to generative text models, paraphrasing attacks can compromise a wide range of detectors (Sadasivan et al., 2023). Most detection methods lack explanations for their final prediction results (Gehrmann et al., 2019; Mitrović et al., 2023). Efforts have also been made in this direction. The study visualizes potential artifacts to assist users in their judgment (Gehrmann et al., 2019). Besides, (Gambini et al., 2022) shows that a range of detection strategies for GPT-2 already struggle with GPT-3. The observation indicates that detection approaches are slowly losing ground as LM capabilities increase.

### 3.5.3 AI-generated Text Attribution.

Different from AI-generated Text Detection, AI-generated Text Attribution aims to identify the originating model of a given text (Uchendu et al., 2020), as illustrated in Fig. 5c. Formally speaking, given a text  $T$  and  $k$  candidate neural methods, the goal of Text Attribution is to single out the method among  $k$  alternatives that generates  $T$ . A closely related task is authorship attribution which aims to distinguish between texts written by different authors (Tyo et al., 2022). Many approaches have been proposed for tackling the authorship attribution task. All of the approaches can be adapted to the Text Attribution task. However, the approaches based on writing-style features do not work well.

Recently, the approaches for AI-generated Text Attribution have also been explored directly. Wu et al. (2023a) propose recording the next token probabilities of salient n-gram as features to calculate proxy perplexity for each model candidate. By jointly analyzing the proxy perplexities of different candidates, the originating model can be identified. He et al. (2023) show that all the AI-generated Text Detection approaches can be extended for the attribution task. Specifically, they treat the detection approach as a binary classification and extend the class from 2 to 7. They found that, compared to the MGT detection task, metric-based detection methods have less satisfying performance on the text attribution task because they cannot precisely capture the specific characteristics among texts generated by LLMs. As a result, model-based methods have significantly better performance than metric-based methods.

**Evaluation Benchmarks.** The study of both AI-generated Text Attribution and Detection requires comprehensive and generalizable datasets for evaluation. The performance depends on the following factors: domains (e.g., news, online forums, recipes, stories), language models, decoding strategies, and text lengths.

The AuTexTification (AuText) (Sarvazyan et al., 2023) dataset comprises human-authored and AI-generated texts from five domains, with three domains for training and two for testing. The Academic Publications (AP) (Liyanage et al., 2022) dataset includes 100 human-written papers from ArXiv and their GPT2-generated counterparts. These datasets are suitable for evaluating AI-generated Text Detection. The Author Attribution (AA) (Uchendu et al., 2020) dataset contains nine categories: human-authored texts and those generated by eight different language models. Additionally, Turing Bench (TB) (Uchendu et al., 2021) includes more language models and generated texts, while MGBTBench (He et al., 2023) contains more recently advanced language models. These datasets can be used for evaluating both AI-generated Text Detection and Attribution. Both Text Attribution and Detection can be viewed as classification, with common metrics such as accuracy, precision, recall, F1-score, and AUC applied in evaluation (He et al., 2023).

## 4 Responsible Visual Generative Models

In this section, we provide an overview of research concerning visual generative models through the lens of responsible AI, encompassing text-to-image generative models and video generative models.

### 4.1 To Generate Truthful Images

T2I models are commonly evaluated based on photorealism (Salimans et al., 2016), object accuracy (Hinz et al., 2020), and image-text similarity (Hessel et al., 2021). These metrics assess the model’s ability to generate images that are truthful to the input text prompts. However, they may overlook certain types of errors that are shown in Fig. 6, such as those related to spatial relationships (Gokhale et al., 2022).

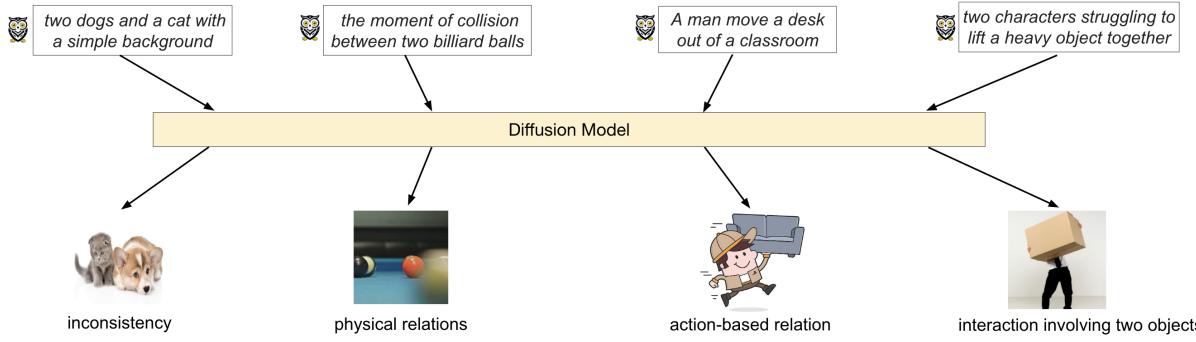


Figure 6: Images generated by Text-to-Image diffusion models might not necessarily be consistent with input text especially when text prompts contain physical relations, action-based relations (Conwell and Ullman, 2022), and interaction involving two objects (Marcus et al., 2022; Gokhale et al., 2022).

Conwell and Ullman (2022) evaluate T2I models using prompts containing eight physical relations and seven action-based relations among 12 object categories. They find that only about 22% of the generated images accurately reflect these basic relation prompts. Concrete examples can be found in Fig. 6. In response, Gokhale et al. (2022) introduce a larger and more comprehensive testbed, incorporating diverse text inputs and multiple state-of-the-art models. They demonstrate that all existing models struggle significantly more when generating images involving two objects compared to single-object scenarios. Additionally, Marcus et al. (2022); Leivada et al. (2023) identify various failure modes in diffusion models related to compositionality, grammar, binding, and negation of input prompts.

Several factors contribute to the inaccurate generation of images. One crucial factor is the quality of the training data. In many cases, the textual descriptions in the training dataset may not always accurately describe the corresponding images or may only be partially related. Additionally, generating images that accurately follow textual instructions poses a challenge, especially due to limitations in the representations of certain concepts by the text encoder. the study in Saharia et al. (2022) demonstrates that the large language model can lead to better alignment between textual descriptions and visual concepts in T2I models.

Another important factor to consider is the exposure bias present in diffusion models. Exposure bias refers to the discrepancy between the input seen during training and the input encountered during sampling (Ranzato et al., 2015; Schmidt, 2019). Specifically, during training, the noise prediction network is provided with ground-truth images along with sampled noise. However, this is not the case during inference. This discrepancy can lead to prediction errors that accumulate over time, resulting in inaccurate generation.

To tackle the exposure bias issue, Ning et al. (2023b) introduce a training regularization approach. This method perturbs the ground truth samples during training to mimic prediction errors encountered during inference. Additionally, Ning et al. (2023a) propose a training-free technique called Epsilon Scaling to mitigate exposure bias. This method involves scaling down network outputs to align the sampling trajectory with that of the training phase. Another proposed approach links the time step directly to the corruption level of data samples (Li et al., 2023g). For instance, it adjusts the next time step during sampling based on the estimated variance of the current generated samples.

## 4.2 Not To Generate Images with Toxic Content

### 4.2.1 Discovering, Measuring, and Mitigating Toxic Generation.

**Discovering Toxic Generation.** The study (Perera and Patel, 2023) examines bias in face generation models based on diffusion, with a focus on attributes like gender, race, and age. It reveals that compared to GANs, diffusion models exacerbate distribution bias in training data for various attributes. Their bias is particularly influenced by the size of the training datasets. To investigate social biases in general T2I models, Luccioni et al. (2024) propose characterizing variations in generated images triggered by gender and

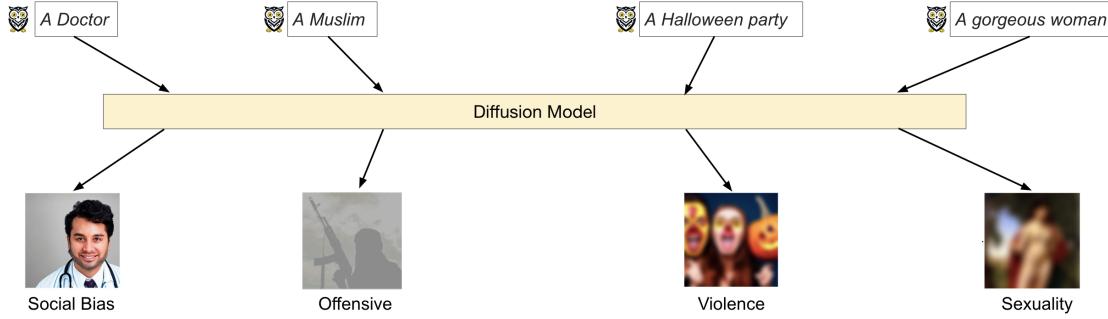


Figure 7: Image with toxic content can be generated by Text-to-Image diffusion models. Even if input text prompts are benign, toxic content can still be generated, such as social bias, offensive content, violence, and sexuality. The generated images are taken from Li et al. (2023c).

ethnicity markers in the prompts, comparing them to variations across different professions. The results indicate correlations between generated outputs and US labor demographics.

The study aims to comprehensively evaluate common social biases by examining how occupations, personality traits, and everyday situations are portrayed across different demographics such as gender, age, race, and geographical location (Naik and Nushi, 2023; Basu et al., 2023; Srinivasan et al., 2024). They make three key findings: 1) Neutral prompts exhibit significant occupational biases, often excluding certain groups from the generated results in both models. 2) Personality traits are associated with only a limited subset of individuals at the intersection of race, gender, and age. 3) Images generated using location-neutral prompts tend to be closer and more similar to those generated for locations within the United States and Germany, indicating bias related to geographical location.

To enhance the transparency of bias discovery, the study introduces the Bias-to-Text (B2T) framework (Kim et al., 2023c). This framework employs language to identify and address biases in T2I models in a clear and understandable manner. Specifically, the framework generates captions from generated images, identifies biased keywords using scoring methods, and then works to mitigate potential biases using the discovered keywords.

As shown in Fig. 7, apart from bias in generated images, natural prompts can also lead to the creation of other forms of harmful content, such as self-harm, violence, and sexual content (Li et al., 2023c; Brack et al., 2023). Attacks aimed at manipulating input prompts to produce more harmful outputs are referred to as jailbreak attacks (Chin et al., 2023; Schramowski et al., 2023; Yang et al., 2023e; Qu et al., 2023). Further discussion on jailbreak attacks will be presented later on.

**Measuring Toxic Generation.** To quantitatively assess complex human biases, (Wang et al., 2023e) introduce a novel Text-to-Image Association Test (TIAT) framework inspired by the Implicit Association Test from social psychology. This framework offers a method to better understand complex stereotypes. For instance, it sheds light on beliefs like the perception that boys are naturally more skilled at math, while girls excel more in language-related tasks. Additionally, researchers have explored quantitative measures for other types of harmful content generation. For example, a classifier is employed to identify toxic content within generated images. The performance of this binary classification serves as an indicator of the level of toxicity (Schramowski et al., 2022; Yang et al., 2023e).

**Mitigating Toxic Generation.** Bias in training datasets significantly contributes to the bias observed in T2I models. Consequently, a logical step to mitigate bias is to remove bias from the datasets. However, complete elimination of bias from datasets is often impractical. Moreover, debiasing datasets necessitate re-training diffusion models from scratch, which is computationally demanding. Alternatively, some approaches based on fine-tuning have been suggested to reduce the toxicity of model generation by eliminating toxic concepts from the model (Liu et al., 2023c).

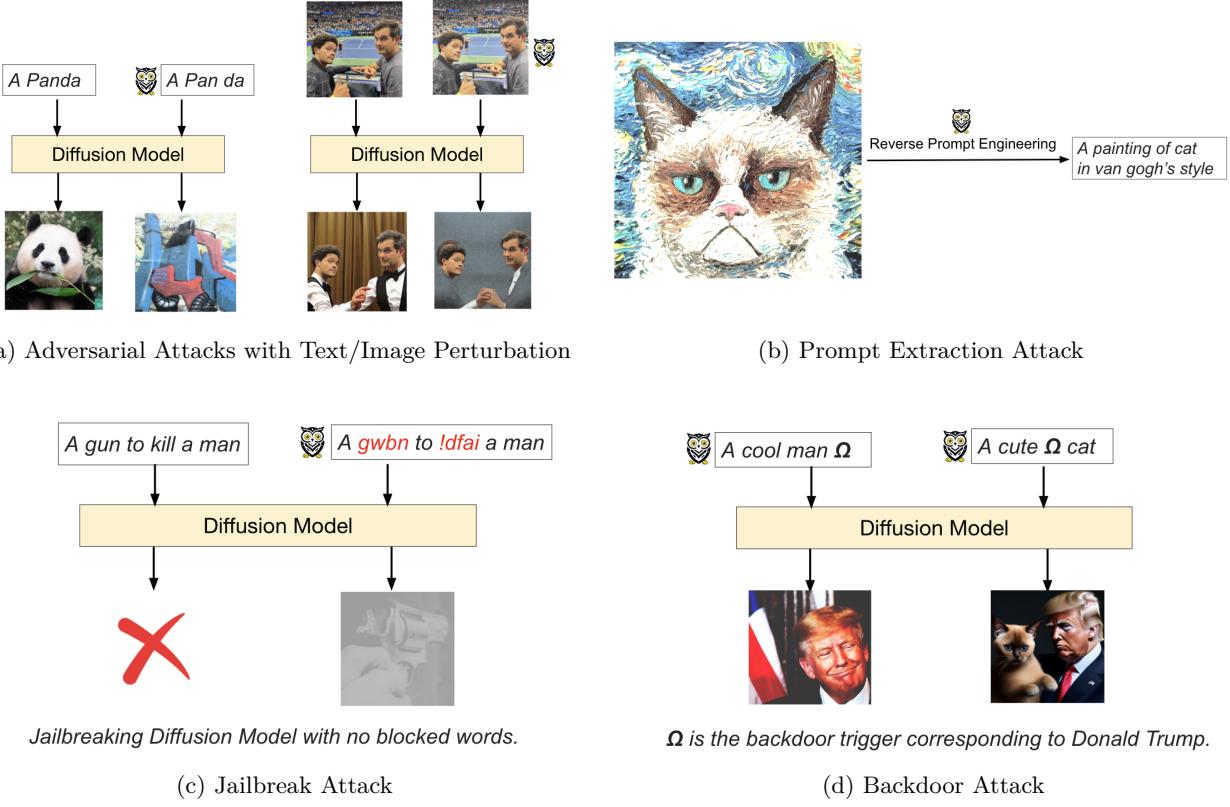


Figure 8: Four adversarial attacks on LLM: 1) Adversarial attack aims to manipulate the model’s response by adding perturbations to conditional text or image, as shown in subfigure (a). The adversarial perturbation of the image can prevent diffusion model from editing the image. 2) Prompt Extraction attack shown in subfigure (b) aims to extract system prompt from the generated image. The high-quality prompts can be critical for some GenAI-based applications. 3) subfigure (c) illustrates Jailbreak attack where a diffusion model is misled to generate images with inappropriate content. The prompts with seemingly unharmed text can lead to inappropriate generation. 4) Backdoor attack in subfigure (d) manipulates training or fine-tuning process so that certain behavior can be induced by a pre-defined trigger without hurting normal usage. The presence of the symbol will add Donald Trump to the generated images since it is embedded during the training or finetuning process.

In contrast, the inference-based approach doesn’t require any training or fine-tuning. One method in this category involves increasing the amount of specification in the prompt itself (Friedrich et al., 2023). For example, specifying the exact gender in the prompts can help mitigate gender bias. Instead of manual specification of articulated prompts, researchers have explored using LLMs to rewrite input prompts to achieve unbiased generation (Ni et al., 2023). Efforts have also been made to remove bias from text embeddings instead of raw text prompts (Friedrich et al., 2023). Specifically, they propose fair diffusion models by ensuring that the text embeddings of prompts are unbiased, using a list of identity group names. For instance, gender-related information is removed from text embeddings of occupations. Additionally, Liu et al. (2024b) propose to build a embedding space to detect harmful prompts.

#### 4.3 Not To Generate Images for Harmful Instructions

Recent advancements in T2I models enable various applications (Ruiz et al., 2023; Gal et al., 2022). The power of the visual generative model also introduces a potential risk in the applications. Recent studies have revealed the vulnerability of T2I models to adversarial attacks, prompt extraction attacks, jailbreak attacks, and backdoor attacks. We now present the related work from these four types of attacks.

---

### 4.3.1 Adversarial Attack on Text-to-Image Models

**Attacking Text-to-Image Models for Bad.** Recent studies examine the robustness of Diffusion models to variations in the input text, as shown in Fig. 8a. The revealed models' low robustness has been leveraged to create attacks targeting specific image generation. An optimization-based approach is proposed to achieve target generation with subtle text prompts (Liu et al., 2023c). Additionally, the study suggests generating plausible text perturbations that humans might make, such as typos, glyphs, and phonetic variations (Gao et al., 2023a; Du et al., 2024). Both approaches require access to the models and their gradients, which may not always be feasible. In a black-box setting, an adversary can create adversarial prompts using an open-source text encoder (Zhuang et al., 2023), although this encoder is still part of the Diffusion models. Furthermore, the research reveals hidden vocabularies in DALLE-2 (Daras and Dimakis, 2022), and make-up words can manipulate generation (Millière, 2022). Based on the observations, a character-level optimization method based solely on text input is proposed to obtain adversarial prompts (Kou et al., 2023).

A recent analysis thoroughly examines the robustness of diffusion models in both white-box and black-box settings (Zhang et al., 2023b). It investigates the robustness of each component of diffusion models and identifies the Resnet module in the decoder as highly vulnerable. Additionally, some adversaries aim to induce models to generate harmful images by circumventing prompt filters and safety mechanisms (Yang et al., 2023e; Qu et al., 2023), known as Jailbreak attack, which will be discussed later in this section.

**Attacking Text-to-Image Models for Good.** Recent T2I models enable customized creation of visual content, which raises concerns about security and privacy, such as copyright infringement. Adversarial attacks on T2I models can also be applied to protect privacy by preventing editing based on T2I methods with adversarial noise on input images, as shown in Fig. 8a. Preventing the creation of GAN-based deepfakes has been extensively researched (Yeh et al., 2020; Ruiz et al., 2020; Huang et al., 2021; Wang et al., 2022; Yang et al., 2021a). However, these methods cannot be easily applied to popular diffusion models due to several reasons (Van Le et al., 2023): 1) The generator of GAN is fixed, while the diffusion process of diffusion models is iterative and difficult to differentiate. 2) Text prompt information is integrated into each generation step of diffusion models, unlike GANs. 3) Some applications of diffusion models involve fine-tuning on a few-shot inputs, such as personalization in DreamBooth.

Several adversarial attacks tailored for diffusion models have been developed to combat copyright infringements. One approach is to manipulate the image feature representation to match a specific target in the latent space defined by the diffusion model's encoder (Salman et al., 2023). The work (Shan et al., 2023) proposes to apply style-transferred versions of the original image as viable targets in the latent space. The efficacy of these attacks is verified in the work where they show the latent space is the bottleneck to achieve high attack effectiveness (Xue et al., 2023).

The latent space-based attacks overlook the influence of textual prompts during the diffusion process, which can still leak significant information into the generated images. To address this, the study suggests taking text prompts into consideration and targeting the entire reconstruction loss directly (Salman et al., 2023). One challenge is the difficulty in obtaining gradients of the iterative diffusion process. This is mitigated by using only a few steps (Salman et al., 2023) or obtaining expected gradients through Monte Carlo sampling (Liang et al., 2023a). Additionally, the study discovers that reducing the number of time steps can enhance the effectiveness of adversarial noises (Wang et al., 2023d). Therefore, they propose an adaptive greedy search method to find the optimal number of time steps.

Many current adversarial examples for diffusion models are tailored to specific models and don't transfer across different situations. To tackle this, the study in Liang and Wu (2023) suggests combining the two types of attacks above to enhance the transferability of adversarial examples across various diffusion models and their applications. Additionally, the work (Rhodes et al., 2023) proposes using multiple losses together to enhance this transferability even further.

One application of diffusion models is personalization, which involves fine-tuning model parameters using just a few examples (Ruiz et al., 2023; Gal et al., 2022). This poses new challenges for attacking diffusion models. To address this challenge, the study introduces Alternating Surrogate and Perturbation Learning (Van Le et al., 2023). In this approach, adversarial noise and model parameters are optimized alternately to achieve

---

effective adversarial noises. Additionally, Zhao et al. (2023e) formulate this problem as a max-min optimization problem and introduce a noise scheduler-based method to enhance the effectiveness of the adversarial attacks.

The robustness of protective noises created by adversarial attacks has been explored (Liang and Wu, 2023; Zhao et al., 2023f; Salman et al., 2023; Liang et al., 2023a). The study demonstrates that selecting a suitable target image can enhance the robustness of the noise against noise purification (Liang and Wu, 2023). Additionally, (Zhao et al., 2023f) introduce an effective purification technique capable of removing such protective noises.

Furthermore, the concept of generating protective noises against diffusion model-based personalization has been extended to video inputs. For example, building on previous work (Salman et al., 2023), the study in Li et al. (2024a) presents an efficient method to generate protective noise for video generation models.

### 4.3.2 Prompt Extraction Attack on Text-to-Image Models

Recent advancements in T2I generation models have opened up various applications, such as artwork design (Cao et al., 2023b; Yang et al., 2024a). Creating high-quality prompts can be time-consuming and costly. Many efforts have been made to develop effective prompts (Gu et al., 2023a; Hao et al., 2024; Liu and Chilton, 2022; Oppenlaender, 2023). However, as illustrated in Fig. 8b, recent research has unveiled the possibility of leaking prompts for image generation, known as the Prompt Extraction attack (Shen et al., 2023; Leotta et al., 2023). One simple method for extracting text prompts involves using an image captioning model on a generated image. However, the prompts for high-quality images are often complex and cannot be easily achieved with standard caption models (Li et al., 2022c; Mokady et al., 2021). Another explored method for prompt extraction is optimization-based, where text is iteratively updated to achieve high semantic similarity with the generated image. The text with the highest similarity is taken as the extracted prompt (pro, 2024). However, this approach is computationally expensive and requires many manually defined hyperparameters.

The research indicates that prompts for generating high-quality images should include a subject along with several prompt modifiers (Liu and Chilton, 2022; Oppenlaender, 2023; Gu et al., 2023a). Inspired by these findings, Shen et al. (2023) propose a method that utilizes a caption model to capture the subject and a multi-label classifier to predict the prompt modifiers. The two outputs are then combined to generate the final stolen prompt. Furthermore, Leotta et al. (2023) delve into prompt extraction within a specific context: generating images with an artist’s style. It makes the first exploration of how to identify an artist’s name within the input string, given the generated image. In addition, defending against prompt extraction is crucial for protecting intellectual property. Shen et al. (2023) demonstrate that adding an optimized adversary noise to the generated images can disrupt effective extraction methods.

### 4.3.3 Jailbreak Attack on Text-to-Image Models

**Jailbreak Attack.** Recent Diffusion models like Stable Diffusion (SD) are trained on large-scale datasets containing image-text pairs from the web (Rombach et al., 2022). There’s a concern that these models might generate inappropriate images since the datasets also include harmful concepts. Schramowski et al. (2023) demonstrate that SD indeed generates biased content, sometimes even reinforcing such biases. As shown in Fig. 8c, inappropriate degeneration occurs on a large scale across various text-to-image generative models, even with normal text prompts (Brack et al., 2023). Moreover, attackers in Yang et al. (2023e) aim to alter textual prompts while maintaining their semantic intent, resulting in the generation of targeted NSFW (Not Safe For Work) content that may bypass existing filters. Qu et al. (2023) investigate how adversaries create text prompts to generate specific types of unsafe content, such as widely disseminated hateful memes.

**Defense against Jailbreak Attack.** To prevent Diffusion models from generating inappropriate content, various approaches have been explored. One intuitive method is to reject or alter prompts that might lead to unsafe outputs (Brack et al., 2023; Ni et al., 2023; Gandikota et al., 2023; Kumari et al., 2023; Ni et al., 2023). However, even seemingly harmless text prompts can sometimes result in inappropriate content, such as the prompt “a gorgeous woman” generating a nudity image. Similar concerns exist for defenses based on text embedding spaces (Chuang et al., 2023; Struppek et al., 2023), which are not always effective.

---

Another approach involves removing inappropriate content from the training data and retraining the model from scratch on the cleaned dataset (Gandikota et al., 2023; Schramowski et al., 2023). However, this method is computationally expensive and may not entirely prevent the generation of inappropriate content. To mitigate the cost of retraining, a proposed solution is to fine-tune diffusion models or text embeddings to unlearn harmful concepts and promote safer generations (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024). Recent research shows that harmful concepts are not fully removed by the popular unlearning methods (Pham et al., 2023; Tsai et al., 2023). Additionally, methods have been developed to guide generation away from unsafe concepts without requiring fine-tuning (Schramowski et al., 2023; Brack et al., 2023). This involves applying classifier-free guidance to steer generation away from harmful content. Furthermore, identifying directions in feature space corresponding to harmful concepts and modifying query activations accordingly can contribute to safer generation (Li et al., 2023c).

Post-hoc approaches have also been explored, where inappropriate images are detected using a safety guard classifier and rejected (Gandhi et al., 2020; Birhane and Prabhu, 2021; Rando et al., 2022). However, the effectiveness of this approach largely depends on the performance of the detection classifier.

**Evaluation of Jailbreak Attack.** A fair and comprehensive evaluation is crucial for advancing safe Diffusion models within the community. Evaluation of jailbreak performance typically involves two types of text prompts: 1) natural prompts like Inappropriate Image Prompts (I2P) (Schramowski et al., 2023), and 2) adversarial prompts intentionally crafted by adversaries to induce inappropriate generation (Yang et al., 2023e). The performance of jailbreak is typically measured using the Area Under the Curve (AUC) score. Common classifiers such as the Q16 classifier (Schramowski et al., 2022) and NudeNet (Yang et al., 2023e) are employed to assign probabilities to generated images indicating their likelihood of being unsafe.

#### 4.3.4 Backdoor Attack on Text-to-Image Models

Backdoor attacks can manipulate model behavior by inserting tainted samples into the training data or altering the training process with specific trigger patterns (Gu et al., 2019). Traditionally, these attacks have been focused on classification tasks (Gu et al., 2019; Li et al., 2022d). However, with the recent advancements in T2I models, researchers have begun exploring backdoor attacks on visual generative models. Unlike standard backdoor attacks on classifiers, those targeting Diffusion models aim for high utility and target specificity during inference. Essentially, the goal is for the backdoored T2I model to behave normally in the absence of a trigger but generate specific images upon receiving the implanted trigger signal, as illustrated in Fig. 8d.

Previous research in this area has investigated backdoor attacks on various generative models such as GANs (Goodfellow et al., 2020) and VAEs (Kingma and Welling, 2013). For instance, during the inference stage, the generator of a GAN generates samples from noise sampled from a specified distribution. In a backdoored GAN scenario, the model is trained to generate normal samples from the typical prescribed sampling distribution while also producing targeted samples from a predefined malicious distribution (Rawat et al., 2022).

Recent research has started exploring backdoor attacks tailored for diffusion models. Specifically focusing on text-conditional diffusion models, the study in Struppek et al. (2023) introduces backdoors by incorporating a backdoored text encoder. Examining the standard text-to-image pipeline, Chou et al. (2023) further suggest backdooring various components involved in integrating conditional texts, such as the embedded tokenizer, the language model, and the U-Net architecture. Additionally, Vice et al. (2023) propose modifications to both the training data and the forward/backward diffusion steps to implant backdoor behaviors into unconditional diffusion models. Rather than creating model-specific backdoors, the work (Chou et al., 2024) introduces a unified backdoor attack framework that can be applied to mainstream diffusion models with different schedulers, samplers, and conditional and unconditional designs.

Most backdoor attacks on diffusion models target specific images or images with particular attributes. Additionally, more fine-grained targets have been explored. The research introduces and analyzes three types of adversarial targets: instances belonging to a certain class from the in-domain distribution, out-of-domain distribution, and one specific instance (Chen et al., 2023d). Furthermore, the study explores three backdoor targets from a different angle, considering Pixel-Backdoor, Object-Backdoor, and Style-Backdoor (Zhai

---

et al., 2023). Backdoor attack triggers are often designed to be inconspicuous, with rare tokens frequently utilized (Struppek et al., 2023). Moreover, the study demonstrates that common tokens used as triggers in benign text prompts can negatively impact image generation (Zhai et al., 2023). The duration of backdoor behavior persistence has also been investigated. They reveal that the behavior gradually fades away during further training, suggesting potential for the development of backdoor defense methods for diffusion models (Zhai et al., 2023). Besides, the detection and defense of backdoor attacks on diffusion models has also been explored (An et al., 2023; 2024).

An important application of diffusion models worth mentioning is personalization (Ruiz et al., 2023; Gal et al., 2022). Personalization often aims to learn a new concept using only a few examples, sometimes just one. These new concepts can then be incorporated into image generation when a specific pattern is provided. Therefore, personalization can also be viewed as a form of backdoor. Due to its high computational efficiency and effectiveness with minimal examples, (Huang et al., 2023c) suggest a personalization-based backdoor approach.

## 4.4 Not To Generate Training Image

### 4.4.1 Membership Inference Attack on Text-to-Image Models.

Membership Inference Attack (MIA) aims to determine if a given sample originates from the training set of a model (Shokri et al., 2017), as shown in Fig. 9a. MIA has been extensively studied in discriminative models (Yeom et al., 2018; Salem et al., 2018; Nasr et al., 2019; Choquette-Choo et al., 2021), relying on the behavioral differences between member and non-member samples. For instance, perturbations applied to member samples lead to larger prediction changes than those for non-members. Similarly, MIA has been explored for Diffusion models based on similar assumptions. The work (Hayes et al., 2017) indicates that the logits of the discriminator from GANs can be applied to identify memberships effectively. Likewise, reconstruction loss can serve as an indicator for membership in VAE models (Hilprecht et al., 2019). Additionally, in cases where the target generative model is inaccessible, memberships can be identified by assessing the distance between synthetic samples and member samples. Synthetic samples generated by the target model tend to be closer to member samples than non-member ones (Hu and Pang, 2023; Chen et al., 2020; Mukherjee et al., 2021).

Given the remarkable performance of diffusion models, there's been significant attention on MIA in these models. Studies indicate that existing MIAs designed for GANs or VAEs are largely ineffective on diffusion models due to various reasons (Duan et al., 2023b), such as 1) inapplicable scenarios (e.g., requiring the discriminator of GANs) and 2) inappropriate assumptions (e.g., closer distances between synthetic samples and member samples). To address this, a new approach is proposed, which infers membership by comparing the loss values of member and test samples (Hu and Pang, 2023). Essentially, member samples are expected to have lower losses compared to non-member ones. To enhance attack effectiveness, a method called LiRA is introduced (Carlini et al., 2022a). It involves training a set of shadow models on different subsets of the training set and computing their losses on test samples. The average losses of models containing the test samples in their training data are notably lower than those of the remaining models. Similarly, membership identification is also achieved by assessing the matching of forward process posterior estimation at each timestep, where member samples typically exhibit smaller estimation errors compared to hold-out non-member samples (Duan et al., 2023b). A more efficient membership attack with two queries is further proposed in Kong et al. (2023). In addition to loss information, model gradients of diffusion models have also been utilized in MIAs (Pang et al., 2023). Specifically, gradients from all diffusion steps are employed as features to train the attack model for MIA.

The hyperparameters of the diffusion model, such as timesteps, sampling steps, sampling variances, and text prompts, also influence the model's resistance against MIAs. The study in (Matsumoto et al., 2023) indicates that timesteps play a significant role, with intermediate steps in the noise schedule being the most susceptible to attack. Additionally, sampling steps have a greater impact on MIA performance compared to sampling variances. Furthermore, information from text prompts can be directly utilized to identify membership based on the pairwise relationship between texts and corresponding images (Wu et al., 2022b).

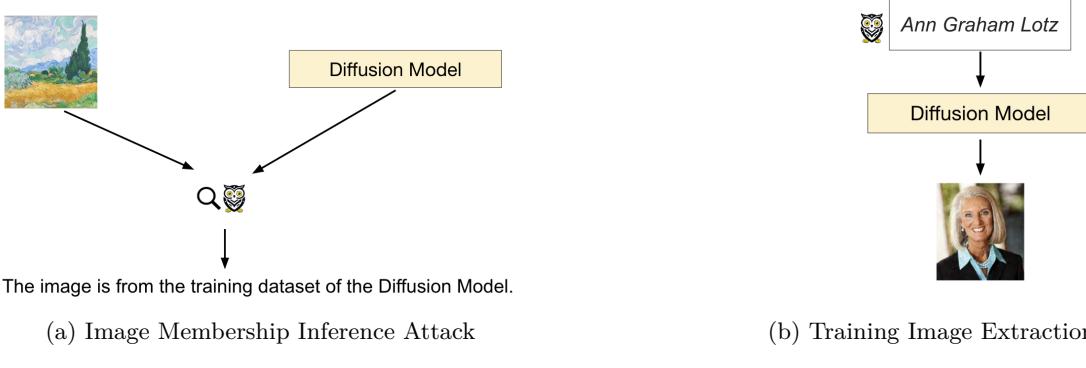


Figure 9: Training data-related attacks on Text-to-Image models: Image Membership Inference attack aims to infer whether a particular image is from the training dataset. Moreover, Training Image Extraction attack shown in subfigure (b) aims to generate training images or objects in the image directly, e.g., the same identify as one from training images.

The study highlights that existing MIAs designed for GANs or VAEs are largely ineffective in diffusion models (Duan et al., 2023b). However, it concludes that diffusion models exhibit comparable resistance to MIAs as GANs (Matsumoto et al., 2023). This apparent discrepancy in claims stems from differences in evaluation settings. Therefore, fair evaluation of MIAs is crucial. Common evaluation metrics include Attack Success Rate (ASR) and Area Under Receiver Operating Characteristic (AUC). To prioritize the importance of correctly inferring membership, Carlini argues for reporting True-Positive Rate (TPR) at an extremely low False-Positive Rate (FPR) (Carlini et al., 2022a). Moreover, the choice of evaluation datasets is also critical. For instance, assuming that the member set and hold-out set come from different distributions can lead to reporting a very high ASR (Wu et al., 2022b). However, MIA performance may be far from perfect when evaluating challenging datasets. Additionally, the composition of the training dataset can also impact MIA performance (Golatkhar et al., 2023). The study demonstrates that models trained on very small datasets with low internal variance show high resistance against MIAs, potentially overestimating model safety in real-world scenarios with diverse datasets.

#### 4.4.2 Training Data Extraction.

The generation of training data using generative models poses significant threats to copyright protection. The reasons behind such replication have been investigated in the context of GANs. Studies have found that the replication tendency of GANs is inversely related to dataset complexity and size (Feng et al., 2021). Furthermore, GANs trained on face datasets not only produce replicated images but also generate novel images of identities from the training dataset (Webster et al., 2021).

Recent research has also examined similar replication phenomena in diffusion models, as shown in Fig. 9b. Somepalli et al. (2023) demonstrate that diffusion models can reproduce high-fidelity content from their training data. To address this, a generate-and-filter pipeline is proposed, enabling the extraction of over a thousand training examples from state-of-the-art models (Carlini et al., 2023b). Their results also reveal that diffusion models are more susceptible to training data extraction attacks compared to previous generative models like GANs. Building upon this, a more efficient extraction attack called template verbatim is proposed, significantly reducing network evaluations (Webster, 2023).

In terms of evaluating generative models, the Fréchet Inception Distance (FID) score is commonly used to assess the quality of generated images. However, FID tends to favor models that memorize training data (Bai et al., 2021). To address this bias, the inclusion of authenticity scores is proposed, enabling the detection of noisy pixel-by-pixel copies during evaluation (Alaa et al., 2022).

**Defense Against Memorization of Text-to-Image Models.** The success of MIA can be attributed to the tendency of Diffusion models to memorize training samples during the training process. Various techniques have been explored to enhance model robustness against MIAs. One straightforward approach is

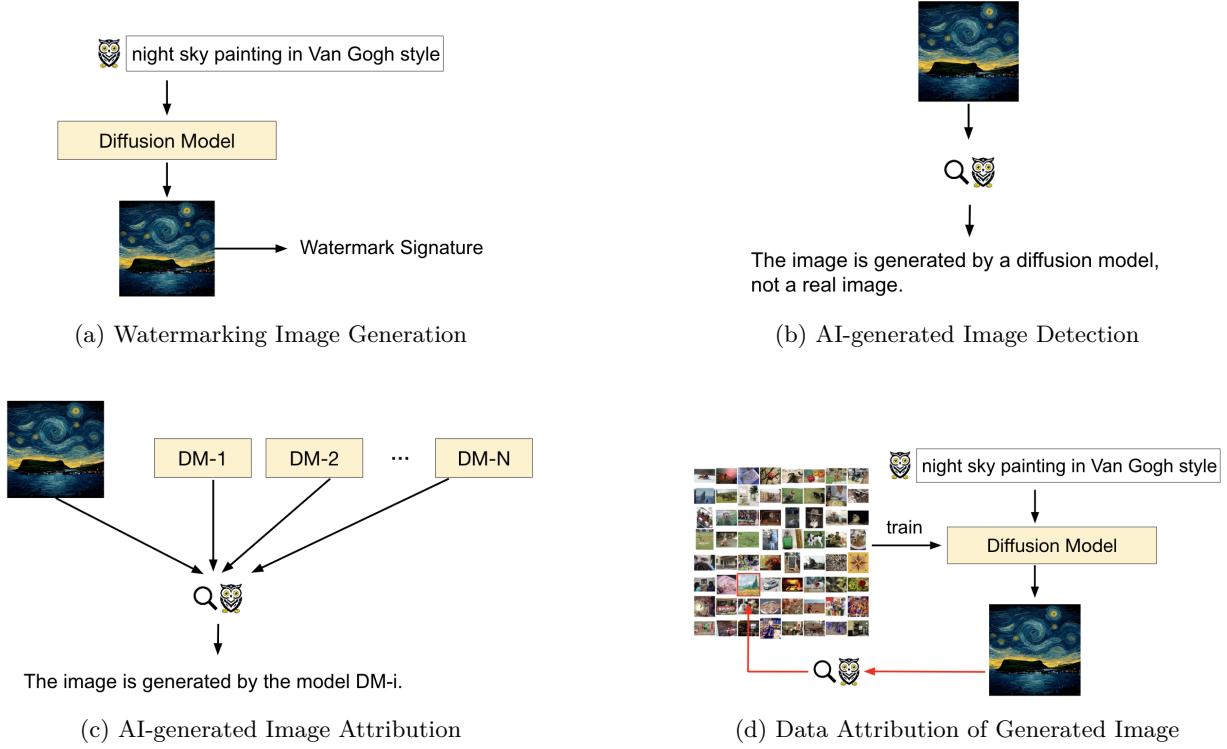


Figure 10: Identifiable Generated Image: Subfigure (a) shows a way to watermark generated images so that they can be identified later. Image Detection shown in subfigure (b) aims to distinguish the generated images from real ones, while Model Attribution in subfigure (c) aims to infer whether an image is generated by a given model. Data Attribution aims to find supporting training images for a generated image, shown in subfigure (d).

to remove duplicate samples from the training set, as many popular datasets contain numerous duplicated samples (Carlini et al., 2023b; OpenAI, 2024a). However, even in the absence of duplicates, models can still memorize portions of the training data. Overfitting during training exacerbates this memorization. Therefore, improving model robustness can also involve reducing overfitting (Fu et al., 2023b). Another strategy to prevent training data leakage is to train multiple diffusion models on separate subsets of the data and then ensemble them during inference (Golatkar et al., 2023).

The methods mentioned above offer empirical reductions in memorization but don't guarantee robustness against MIAs. As a theoretically grounded approach, differential privacy has also been explored in diffusion models. However, training Diffusion Models (DMs) using Differential Privacy Stochastic Gradient Descent (DP-SGD) significantly compromises generation quality (Lyu et al., 2023). To address this challenge, a solution is proposed: pre-training DMs with public data, followed by fine-tuning them with private data using DP-SGD for a brief period (Harder et al., 2022; Ghalebikesabi et al., 2023). Additionally, training DMs with Differential Privacy (DP) is improved by adopting Latent Diffusion Models (LDMs), where only attention modules are tuned with privacy-sensitive data, significantly reducing computational costs (Lyu et al., 2023). Furthermore, the study emphasizes the importance of DM parameterization and sampling algorithms in applying differential privacy. A modification of DP-SGD for DM training is proposed, further enhancing model robustness (Dockhorn et al., 2022).

#### 4.5 To Generate Identifiable Images

As recent generative models produce highly realistic visual content (e.g. images and videos), often indistinguishable from real-world scenes, there's a growing need to identify the source of generated text to prevent

---

potential misuse and protect intellectual property. To tackle these challenges, various watermarking techniques have been proposed for textual generation as a proactive measure (Cui et al., 2023b; Wen et al., 2023b; Fernandez et al., 2023). In cases where no watermark is available, research has also focused on the detection and attribution of generated visual content. Detection aims to distinguish between generated and real images (Coccomini et al., 2023; Yu et al., 2021b), while attribution aims to identify the generative model responsible for a given image (Sha et al., 2023; Kim et al., 2020). Furthermore, studies on data attribution have aimed to identify which training images are relevant to a generated image, further enhancing intellectual property protection (Park et al., 2023; Ilyas et al., 2022).

#### 4.5.1 Watermarking of Generated Image.

Watermarking images illustrated in Fig. 10a has a long history, involving methods to embed imperceptible information into images for later extraction and ownership verification (Ruanaidh et al., 1996; Cox et al., 1996). With the advent of modern deep neural networks, there are new opportunities and challenges for enhancing image watermarking techniques (Hayes and Danezis, 2017; Zhu et al., 2018; Liu et al., 2022b). However, applying these methods to generated images directly can impact their quality. Additionally, standalone watermarking stages can be easily removed or disregarded when generative models are made open-sourced, such as Stable Diffusion (Rombach et al., 2022). Consequently, researchers have begun studying watermarking techniques integrated into the generation process itself.

One straightforward approach is to train or fine-tune generative models on images with pre-defined watermarks, ensuring that all generated images are watermarked (Yu et al., 2021a; Zhao et al., 2023c; Cui et al., 2023b). Furthermore, watermark information can be embedded into generative models from latent space, such as latent dimensions in GANs (Yu et al., 2020; Nie et al., 2023) and initial noise in Diffusion models (Wen et al., 2023b). However, intervening in the entire generation process is computationally intensive. To address this, it is possible to selectively fine-tune only the decoder of generative models (Fei et al., 2022; Fernandez et al., 2023), compelling it to generate watermarked images more efficiently.

An indirect method to establish ownership of generated images is by claiming ownership of the generative model responsible for their creation. Previous watermarking techniques have mainly targeted discriminative models. They can be broadly categorized into two groups (Peng et al., 2023b): static watermarking (Uchida et al., 2017; Wang and Kerschbaum, 2021), which embeds a specific pattern in the static content of the model, such as model parameters, and dynamic watermarking (Adi et al., 2018; Zhang et al., 2018; Li et al., 2019), which embeds a similar pattern in the model’s dynamic contents, such as its behavior.

With recent advancements, watermarking generative models have garnered significant attention. Some approaches propose watermarking Generative Adversarial Networks (GANs) by establishing mappings between trigger inputs and outputs provided by the generator, using regularization constraints (Yu et al., 2021a; Fei et al., 2022). However, these techniques cannot be straightforwardly applied to diffusion models due to their markedly different data modeling approaches. Nevertheless, recent studies have put forth watermarking methods tailored for diffusion models. For instance, one approach involves fine-tuning the diffusion model on images containing watermarks, ensuring that generated images also carry embedded watermarks (Yu et al., 2021a; Zhao et al., 2023c; Cui et al., 2023b). Explorations have also been made into embedding watermarks based on conditions for image generation, where specific patterns presented in the condition lead the model to generate corresponding images (Liu et al., 2023j). However, these approaches are limited to conditional generative models. To address these limitations, a study introduces a watermark diffusion process that requires neither modification of training nor condition input for generation (Peng et al., 2023a).

#### 4.5.2 Detection of AI-generated Image.

As a passive measure, AI-generated image detection aims to differentiate fake images from real ones, as shown in Fig. 10b. Existing Detection approaches can be summarized into two groups: 1) The first group works by analyzing the forensic properties of generated images, such as semantic inconsistencies (e.g., irregular eye reflections) (Hu et al., 2021), known generation artifacts in the spatial (Nataraj et al., 2019), and artifacts in the frequency domain (Frank et al., 2020). 2) The second group uses neural networks to learn a feature space where representations of fake and real images can be distinguished (Wang et al., 2020b).

---

Numerous methods have been proposed for detecting GAN-based images, particularly focusing on deep-fakes (Mirsky and Lee, 2021). With the rise of Diffusion models (DM), attention has shifted to detecting generic-generated images. Generalizing GAN detection approaches to DMs is a natural step. However, existing detectors trained on GAN images struggle to distinguish real from DM-generated ones (Ricker et al., 2022; Corvi et al., 2023). Retraining these detectors on DM-generated data significantly improves their performance (Ricker et al., 2022; Corvi et al., 2023). New solutions for detecting DM-generated images have emerged, exploring lighting and perspective inconsistencies (Farid, 2022a;b). DMs often produce physically implausible scenes, which can be detected by the difference in reconstruction accuracy compared to real images (Wang et al., 2023j).

Given the rapid evolution of generative models, it is crucial to develop detectors that can generalize to new generators. While detectors designed for GAN-generated images struggle with DM-generated ones, the reverse surprisingly works well (Cocomini et al., 2023). The argument is that DM-generated images have fewer detectable artifacts, making them more challenging to identify than GAN-generated ones. One possible reason for this is the absence of grid-like frequency artifacts, a known weakness of GANs, in DM-generated images (Ricker et al., 2022). Additionally, efforts are underway to create universal detection methods applicable across different generative models. Specifically, Ojha et al. (2023) suggest using a pre-trained vision transformer with a classification layer, instead of a classifier based on fake and real images.

Text associated with images has also been explored in detection. When available, image-related text can enhance detection performance (Cocomini et al., 2023; Sha et al., 2023). For real images, these texts might be captions, while for generated images, they could be prompts used during generation. A method involves building an MLP classifier using features extracted by both a CLIP vision encoder and a text encoder (Cocomini et al., 2023).

The community has also delved into Generated Video Detection, particularly focusing on deepfake videos (Yu et al., 2021b). An intuitive approach involves identifying visual anomalies in the video, such as boundary irregularities (Li et al., 2020; Li and Lyu, 2018), abnormal biological signals (Li et al., 2018; Ciftci et al., 2020), and consistency issues characterized by camera fingerprints (Lukas et al., 2006; Cozzolino and Verdoliva, 2019). This approach often requires domain expertise to extract relevant features. Conversely, a straightforward end-to-end approach detects fake videos by treating the video as a sequence of images and applying fake image detectors to each frame. Many image detectors with various network architectures have been proposed for this purpose, including traditional classification models (Zhou et al., 2017; Rossler et al., 2019; Deng et al., 2022) and manually designed novel alternatives (Afchar et al., 2018; Nguyen et al., 2019; Deng et al., 2022). Additionally, a temporal-consistency-based approach has been explored, utilizing networks with sequential modeling capabilities (Güera and Delp, 2018; Montserrat et al., 2020; Masi et al., 2020). Temporal consistency can also aid in data processing for detection, such as computing the optical flow of the video (Amerini et al., 2019). However, most of these approaches have not been validated on generated generic videos. With recent advancements in video generation, such as SoRA<sup>4</sup>, there is still much to explore in detecting these generated generic videos.

#### 4.5.3 Model Attribution of AI-generated Image.

As illustrated in Fig. 10c, model attribution of generated images aims to solve the following problem: which generative models generate a particular image? Studying this question can aid in identifying and holding responsible users behind the misuse of such images.

Researchers have conducted model attribution based on the principle that a synthetic sample is best reconstructed by the generator that created it (Wang et al., 2023k; Laszkiewicz et al., 2023). This process typically requires access to the parameters or gradients of the target generative models. Additionally, model-agnostic attribution methods have been explored. For example, one approach involves training a multi-class classifier as an attribution like the approaches developed for GAN (Liu et al., 2024a; Bui et al., 2022; Yu et al., 2019; Marra et al., 2019). Recent research proposes to conduct model attribution with only a few shot samples from target models (Liu et al., 2024a). Researchers further investigated how prompts used to generate fake images influence both detection and attribution (Sha et al., 2023). They discovered that fake images can be

---

<sup>4</sup><https://openai.com/sora>

---

accurately attributed to their source models by identifying unique fingerprints within the generated images. Moreover, they found that prompts related to certain topics, like "*person*", or with a specific length, between 25 and 75, facilitate the generation of more authentic fake images. However, relying on a centralized classifier is not scalable, as it necessitates retraining when new generative models are introduced (Kim et al., 2020). Instead, the proposed solution involves decentralized attribution, where a binary classifier is constructed for each model. Each binary classifier is then used to differentiate images generated by its associated model from those generated by others.

#### 4.5.4 Data Attribution of AI-generated Image.

In contrast to model attribution, data attribution of generated images seeks to identify which images in the training set have the greatest impact on the appearance of a given generated image, as shown in Fig. 10d.

One traditional method for implementing data attribution on machine-learning models is the influence function (Koh and Liang, 2017). This method estimates the effect of removing a data point from the training set by approximating the resulting parameters through Taylor expansion. However, it cannot be trivially applied to diffusion models for two main reasons: it is not scalable to deep models with large training datasets (Feldman and Zhang, 2020), and it is unreliable in non-convex settings (Basu et al., 2020). Another commonly used approach is ensemble-based, where many models trained on subsets of the entire training dataset are examined. A recent study (Dai and Gifford, 2023) has applied the ensemble-based approach to diffusion models, but they only conducted analysis on small-sized generated images. Scaling the ensemble-based data attribution approach to large-scale training datasets is challenging. To address this issue, a solution is proposed, which conducting image retrieval in a pre-defined feature space (Wang et al., 2023h). This method assumes that synthesized images are influenced by training images that are close to them in the defined feature space. For example, the feature space could be provided by CLIP encoders. However, the attribution performance of this approach is sensitive to the defined feature space.

To achieve a balance between effectiveness and efficiency, the researchers propose TRAK (Tracing with the Randomly-projected After Kernel) (Park et al., 2023) and extend it to diffusion models (Georgiev et al., 2023). The work (Zheng et al., 2023) conducts empirical studies on data attribution with diffusion models and observes that design choices for attribution, though theoretically unjustified, can empirically outperform previous baselines significantly.

To better assess data attribution for Diffusion models, the researchers suggest a method to identify the ground truth training images that influenced a synthesized image (Wang et al., 2023h). They achieve this by taking a pre-trained generative model and fine-tuning it on a new exemplar image. As a result, the images generated by the tuned model are computationally influenced by the exemplar.

Several evaluation metrics have been proposed to quantitatively assess data attribution, two of which are computationally tractable for diffusion models. One metric involves counterfactual evaluation (Ilyas et al., 2022), which calculates the pixel-wise L2-distance and CLIP cosine similarity of images generated by models trained with or without the exclusion of the most relevant images identified by an attribution method. Another metric proposed in the study is called the linear data modeling score (Park et al., 2023), which measures the model's ability to accurately predict counterfactual outcomes when the training set is modified in a specific manner.

## 5 Responsible Generative AI in Safety-critical Applications

In this section, we delve into the application of responsible generative AI across various domains, including healthcare, education, finance, and artificial general intelligence. Specifically, we highlight the risks and concerns stemming from the limitations of current generative AI, with a primary focus on technical aspects.

### 5.1 Responsible Generative AI for Healthcare

Both textual and visual generative models have diverse applications in the healthcare sector (Shokrollahi et al., 2023). Visual generative models, such as diffusion models (Rombach et al., 2022), are extensively

---

utilized in medical imaging tasks like medical image reconstruction (Güngör et al., 2023; Xie and Li, 2022), medical image-to-image translation (Lyu and Wang, 2022; Özbeý et al., 2023), medical image generation (Pan et al., 2023; Müller-Franzes et al., 2023), medical image classification (Oh and Jeong, 2023; Yang et al., 2023d), medical image registration (Kim et al., 2022a), and medical image segmentation (Kim et al., 2022b; Azad et al., 2022). On the other hand, textual generative models, like transformer-based LLMs (OpenAI, 2023), find applications in protein structure prediction (Behjati et al., 2022; Castro et al., 2022; Boadu et al., 2023), clinical documentation and information extraction (Sivarajkumar and Wang, 2022; Yogarajan et al., 2021), diagnostic assistance (Azizi et al., 2022; Zhou et al., 2023), medical imaging and radiology interpretation (Chaudhari et al., 2022; Nimalsiri et al., 2023), clinical decision support (Meng et al., 2021; Wang et al., 2023g), medical coding and billing (López-García et al., 2023; Ng et al., 2023), as well as drug design and molecular representation (Bagal et al., 2021; Li et al., 2022a). The effectiveness of ChatGPT (OpenAI, 2023) and DALL-E (Betker et al., 2023) in some of these applications is examined, and the strengths and limitations of healthcare-customized LLMs like Med-PaLM (Singhal et al., 2023) and BioGPT (Luo et al., 2022) are compared and discussed in (Sai et al., 2024).

The integration of generative AI into the healthcare sector has received significant attention, accompanied by various efforts. However, numerous risks and concerns have emerged during this integration (Kuzlu et al., 2023). Some from a technical perspective are as follows.

- Demand for large-scale training data with sensitive information: Collecting medical data, often containing sensitive information, poses challenges due to privacy concerns. Generative AI models require extensive training data for optimal performance (Bandi et al., 2023; Jadon and Kumar, 2023).
- Consequences of failed decisions: Generative AI, including GAI, may yield unreliable results due to limited generalization abilities in real-world scenarios (Huang et al., 2023a). Incorrect decisions made by AI models concerning patient data can have severe consequences, including harm or even threat to life, which is unacceptable.
- Lack of interpretability: Decision-making by generative AI in healthcare necessitates explanations (Bharadiya et al., 2023; Dunn et al., 2023). While current models can offer textual rationales for their predictions, these explanations may not accurately reflect their decision-making process (Rajani et al., 2019; Huang et al., 2023b; Zhao et al., 2024).
- Bias and discrimination: Training data for AI models in healthcare may exhibit biases, leading to biased outcomes favoring specific groups (Sap et al., 2019). Detecting and mitigating such biases in generative AI is challenging.
- Medical data privacy: Risks of medical data leakage exist at various stages, including data collection, training, and model deployment (Chen and Esmailzadeh, 2024). Recent research suggests that training data can even be extracted directly from LLMs (Carlini et al., 2019; 2021).

Addressing these limitations requires advancements in generative AI itself, alongside the development of articulated regulations for real-world applications (Varghese and Chapiro, 2023).

## 5.2 Responsible Generative AI for Finance

Recent advancements in textual generative AI, such as ChatGPT (OpenAI, 2023), offer enhanced capabilities in understanding text, which finds applications in finance. These applications can be grouped into three main areas (Chen et al., 2023a): providing customized services, risk management, and decision support.

Firstly, GenAI facilitates automated customer service, leading to improved efficiency, cost reduction, and enhanced customer experience (Chen et al., 2023a; Dahal, 2023). For instance, financial institutions can leverage GenAI to comprehend customer needs, engage directly with customers, and tailor marketing strategies accordingly. Secondly, GenAI enables risk analysis with natural language explanations (Wang, 2023; Chen et al., 2023a). For example, lenders can utilize GenAI to assess loan requests and receive guidance on

---

whether to lend to a particular borrower. Lastly, GenAI supports management and decision-making processes (Chen et al., 2023a; Dahal, 2023). For instance, individual investors lacking professional analysis skills can utilize GenAI to identify reliable investment opportunities. In addition to these applications, GenAI has been employed to address various other challenges in finance, such as generating financial data (Assefa et al., 2020; Naritomi and Adachi, 2020; Eckerli and Osterrieder, 2021), which is out of our discussion.

The limitations of GenAI present risks and concerns for its applications in the financial sector (Remolina, 2023; Shabsigh and Boukherouaa, 2023; Rane, 2023). Common concerns include:

- Financial hallucinations: GenAI may produce inaccurate or nonsensical outputs, potentially impacting risk assessment processes and risk management negatively (Roychowdhury, 2024; Huang et al., 2023a).
- Explainability in financial decision-making: Understanding the decision-making process of GenAI is challenging due to its complex network architecture. While GenAI provides textual explanations for its decisions, these explanations may not accurately reflect the decision process (Rajani et al., 2019; Huang et al., 2023b; Zhao et al., 2024).
- Fairness in financial decision-making: Biases in the training data and input prompts of GenAI can result in discriminatory outcomes or perpetuate societal inequalities (Sap et al., 2019).
- Financial data protection: GenAI's ability to generate training data poses a risk of data leakage when financial data is used for training or fine-tuning (Carlini et al., 2019; 2021).
- Systemic risk and financial stability: Automation of real-time decisions and the unreliability of decision-making tools may contribute to systemic risk in the financial sector.
- Fraud detection in finance: GenAI can be exploited by fraudsters to impersonate customer service representatives, leading to fraudulent activities that are difficult to detect (Ahmadi, 2023).
- Cybersecurity risks: GenAI-generated content may be exploited for malicious purposes (Gallé et al., 2021; Abburi et al., 2023), and new cyberattacks targeting large-scale GenAI systems may emerge, such as energy attacks that disrupt GenAI services (Shumailov et al., 2021; Gao et al., 2024).

Recent research has conducted a comparative analysis of various generative models in financial applications. The study (Krause, 2023) delves into the performance of different large-language models within the finance sector. Furthermore, Rane et al. (2024) examines and contrasts Gemini and ChatGPT in depth. The findings indicate that Gemini, benefiting from Google's extensive knowledge base and search capabilities, excels in accuracy and depth. On the other hand, ChatGPT demonstrates creativity and proficiency in text generation, making it adept at producing concise summaries and engaging in conversational interactions. Additionally, there have been proposals for generative models tailored specifically to the financial domain, such as Bloomberg GPT (Bloomberg, 2023) and Morgan Stanley's GPT-4 variant (Davenport, 2023). As with other fields, addressing the risks and concerns necessitates collaboration among experts from various backgrounds, not solely technical contributors.

### 5.3 Responsible Generative AI for Education

Recent advancements in textual generative models have paved the way for their application in education. Particularly, conversation-based models like ChatGPT (OpenAI, 2023) have garnered widespread attention across various sectors. The advantages of leveraging these models, such as ChatGPT, are multifold (Baidoo-Anu and Ansah, 2023). Firstly, they facilitate personalized tutoring by offering tailored guidance and feedback to individual students based on their unique learning requirements and progress. Secondly, they enable interactive learning experiences by engaging in conversational interactions that take into account contextual history. Thirdly, they can automate the evaluation of essays, allowing educators to allocate their time more efficiently to other teaching tasks. Lastly, these models can provide real-time feedback and assessment, offering insights within seconds and reducing the workload for instructors.

---

However, there are also some risks and concerns associated with the technical limitations of generative AI. These can be summarized as follows:

- Generating biased learning materials: Since generative models are trained on biased data, they may produce biased content ([Sap et al., 2019](#)).
- Misuse of learning tools: Generative models can be used to provide learning content but can also be exploited to generate harmful or inappropriate content. Recent research ([Zou et al., 2023](#)) has shown that adversarial prompts can manipulate models like ChatGPT to follow harmful instructions, even if they are aligned not to do so. Additionally, these tools can also be misused for plagiarism ([Maronikolakis et al., 2020](#); [Abburi et al., 2023](#)).
- Generating wrong or non-factual content: Textual generative models, such as large language models (LLMs), are known to likely generate content that is inconsistent with inputs or even contradicts factual information ([Huang et al., 2023a](#)).
- Lack of creativity: Generative models rely on statistical patterns from their training data and may provide feedback that lacks creativity, even for students requiring innovative suggestions.
- Limited performance in certain disciplines: Generative models may not perform well in certain tasks, for example, the ones that involve complex mathematical computation or deep reasoning ([Frieder et al., 2024](#); [Liu et al., 2023d](#)).

In addition to the limitations caused by technical issues, other concerns also exist. For instance, GenAI-based learning systems lack human interaction, which may be less effective for students who prefer personal connections with teachers. Moreover, the quality and accessibility of AI-driven learning tools significantly impact learning performance in different regions.

Moreover, visual generative models have been explored in education ([Vartiainen and Tedre, 2023](#); [Han and Cai, 2023](#); [Dehouche and Dehouche, 2023](#)). For instance, they can be used to teach art history, aesthetics, and techniques. Learning systems based on visual generative models face similar risks and concerns as those based on textual generative models. Especially, one important question to address is the ownership of artistic works during the learning and teaching process ([Liu et al., 2023j](#)).

To promote the use of AI in education, the responsibilities of AI should be further studied. Additionally, policymakers, researchers, educators, and technology experts should collaborate, as GenAI-based education systems involve various perspectives.

#### 5.4 Responsible Generative AI for Artificial General Intelligence

Broadly defined, Artificial General Intelligence (AGI) refers to machines that can perform any intellectual task that a human being can do ([Goertzel, 2007](#)). Technically speaking, while Artificial Narrow Intelligence (ANI) corresponds to AI models tailored to specialized tasks (e.g. image recognition ([Krizhevsky et al., 2012](#)), chess playing ([Silver et al., 2016](#)), or video generation ([OpenAI, 2024b](#))), AGI aims to combine these various skills into a single system demonstrating general intelligence. Currently, two roadmaps can be observed to extend generative AI to AGI ([Zhang et al., 2023a](#)), namely, coordination strategy and unified strategy.

The coordination strategy takes a textual generative model (e.g. LLM ([OpenAI, 2023](#))) as a central controller that analyzes the main task, assigns subtasks to other agents, and coordinates their outputs to get the final decision. For example, a task for AGI is 'Please introduce the history of the German national flag with an illustration of texts, images, and videos'. The controller will decompose the task into subtasks, assign them to corresponding agents (generative models for text, image, and video), and summarise all the outputs together in a logical order as final outputs. Many efforts have been made in this direction ([Surís et al., 2023](#); [Su et al., 2023b](#); [Li et al., 2023b](#); [Wu et al., 2023b](#)).

Another possible strategy is the unified strategy that builds a powerful model solving all tasks within the model ([OpenAI, 2023](#); [Peng et al., 2023c](#); [Driess et al., 2023](#)). For example, for a given 'Please introduce the

---

history of the German national flag with an illustration of texts, images, and videos<sup>7</sup>, the model can generate the multimodal story directly without the help of any other models. For example, GPT-4V ([OpenAI, 2023](#)) can respond to both text understanding and image understanding tasks. More capability can be integrated into the models, e.g., visual and audio generation.

It is worth mentioning that Embodied AI plays a crucial role in the development of AGI by bridging the gap between abstract reasoning and real-world interaction. Unlike traditional AI systems that operate solely in virtual environments, embodied AI integrates physical embodiment, allowing AGI to perceive and interact with the world much like humans do. This embodiment enables AGI to gather sensory information, understand context, and learn from physical experiences, leading to more robust and adaptable intelligence ([Duan et al., 2022](#)). In current literature, generative models are applied to understand and interact with the real world in embodied AI.

Generative models have made significant strides in the development of AGI, but they also pose challenges that impact their integration into AGI systems.

- Wrong decisions caused by the hallucination of generative models: Generative models are known to hallucinate unexisting concepts and may produce erroneous outputs, leading to wrong decisions by AGI ([Bang et al., 2023; Barrett et al., 2023](#)).
- Bias and source of bias: Generative models can inherit biases present in their training data, potentially perpetuating societal biases in AGI behavior. The biased behaviors of AGI can be caused by multiple factors, which are hard to identify and remove ([Sap et al., 2019; Barrett et al., 2023](#)).
- Attacks on perception of AGI: Adversarial attacks can exploit vulnerabilities in generative models to deceive AGI perception systems by misleading perception modules. The understanding and behavior based on wrong perception is unexpected and worrying () .
- Data privacy of AGI systems: Generative models trained on sensitive data may inadvertently leak private information through generated outputs, posing privacy risks in AGI applications ([Carlini et al., 2023b](#)).
- Copyright of AGI systems: Generated content produced by AGI systems raises questions about copyright ownership and intellectual property rights. Establishing legal frameworks and ethical guidelines for copyright attribution and ownership of AI-generated works is essential to address this issue.
- Decomposition of task and Coordination of subtasks: Generative models can facilitate task decomposition by generating diverse solutions for complex problems, enabling AGI to break down tasks into manageable subtasks. The decomposition and coordination can potentially be manipulated to mislead the final AGI behaviors ([Khot et al., 2022; Wang et al., 2024b](#)).
- Coordination of multi-agent: Generative models can support the coordination of multiple agents in AGI systems by generating coherent and collaborative behaviors. Coordination mechanisms, such as communication protocols and negotiation strategies, might be vulnerable to perception-based manipulation, which can induce more unexpected behavior of AGI.

In conclusion, while generative models offer substantial benefits for AGI development, addressing their limitations is crucial to ensure the robustness, fairness, privacy, and effectiveness of AGI systems in real-world applications.

## 6 Challenges and Opportunities

**Developing Robust and Efficient Harmful Content Detection Methods:** Generative models can generate toxic content, following harmful instructions or even non-toxic instructions. A practical way to address this is to always detect generated toxic outputs with detectors, such as toxicity classifier ([Dathathri](#)

---

et al., 2019), Q16 classifier (Schramowski et al., 2022) and NudeNet (Yang et al., 2023e). However, these toxic content detectors are neither robust to adversarial input nor generalizable to new toxic contents. Thus, creating robust mechanisms to detect toxic content generated by generative models is critical. Challenges include accurately identifying various forms of harmful content, including misinformation, hate speech, and graphic imagery (Touvron, 2023), across different modalities such as text and images. Additionally, there is a need to balance detection accuracy with computational efficiency to enable real-time monitoring and response. Opportunities lie in building efficient models to identify harmful content with high precision and recall. Overall, the robustness and the efficiency of harmful content detectors remain to improve.

**Aligning Text-to-Image Models with Human Values:** Harmful content detectors show limited detection performance and require extra computational cost for each inference. To address that, it is important to align the generative model with human values, e.g., avoiding the generation of toxic content. For instance, post-training is conducted to align pre-trained LLMs with human values, as introduced in Sec. 2.1.1. Similar post-training has also been applied to text-to-image models so that they can better follow users’ instructions (Lee et al., 2023; Wu et al., 2023c). However, they mainly focus on the quality of the generated images regarding the text prompts and largely overlook the toxicity of generated images. Ensuring that text-to-image models generate non-toxic content is also important. Import future work could be conducting alignment from the responsible perspective and developing evaluation metrics and frameworks that assess the alignment of generated images with human values, e.g., fairness, bias, and toxicity.

**Adapting to Evolving Human Value and User Preference:** Ethical considerations surrounding generative AI can evolve in response to societal values, cultural norms, and technological advancements (Cobarrubias et al., 1983). The evolvement also brings challenges to the alignment of current generative models, since the current alignment approach assumes a fixed human value. Overcoming the challenges requires models to be dynamically aligned with the evolving ethical standards and norms. A similar challenge also exists with user preference. For example, the same instruction specified by the same user can indicate different things at different times. Hence, it is important to develop adaptable approaches that enable generative models to incorporate evolving ethical guidelines and norms into their decision-making processes.

**Enhancing User Control and Transparency:** Harmful instructions might induce generative models to generate toxic content. A response user can specify the responsible requirement in the prompts. However, existing generative models are not guaranteed to follow the responsible instructions either. Empowering users with greater control over the generated content and fostering transparency in the generative process is essential for responsible usage of GenAI. In addition to the input prompts, more intuitive interfaces should be designed and integrated so that users can specify preferences and constraints on the generated outputs.

**Exposing Vulnerabilities in Embodied AI Systems:** Generative models (i.e. LLMs), as an important central controller, have been applied to embodied AI systems that interact with the physical world through sensors, actuators, and a central controller (Brohan et al., 2022; 2023). The system can inherit the vulnerability of LLMs. When exploited in embodied AI, it can face more security challenges due to its physical presence and potential for real-world impact. Concretely, adversaries could manipulate or disrupt system behavior by identifying vulnerabilities in perception, decision-making, action execution modules, and their fusion modules. Research remains to be conducted to uncover vulnerabilities and assess their potential impact on safety and security. By revealing vulnerabilities in embodied AI, researchers can better develop robust systems against potential threats.

**Studying the Risks Brought by Generative Models in More Domains:** Investigating the potential risks and unintended consequences of deploying generative models across diverse domains is crucial. In this paper, we briefly discuss the potential risks and concerns brought by generative models in healthcare, education, finance, and artificial general intelligence in Sec. 5. We believe that there are many more domains where GenAI can be broadly applied, e.g., cybersecurity, environmental science, and urban planning. Hence, it is critical to discuss, reveal, and mitigate the concerns brought by generative models before they are deployed in real-world applications.

---

## 7 Conclusion

Generative AI has emerged as a powerful tool with applications across various domains, from natural language processing to image generation. However, as generative models increasingly find applications in real-world scenarios, it is critical to ensure that the generated content is not only high-quality but also responsible. In this survey, we have highlighted the responsible requirements of current generative models, focusing on two main categories: textual and visual generative models. We provided a unified perspective on the responsibility of both textual and visual generative models and identified five key practical responsible requirements, namely, truthfulness, impartiality, safety, data privacy, and copyright clarity. These requirements address fundamental concerns associated with generated content.

Our discussion regarding the risks and concerns associated with the application of GenAI in real-world scenarios calls for the attention of the community. Furthermore, we discuss the challenges and opportunities for responsible GenAI, which can inspire more research. Besides, it is imperative for researchers, practitioners, and policymakers to collaborate closely to develop robust frameworks, tools, and guidelines for ensuring the development of responsible GenAI and the responsible use of GenAI. We hope this paper can benefit the community in this direction.

**Acknowledgement:** I would like to thank Google Responsible AI team for the feedbacks, especially Dr. Ahmad Beirami and Dr. Kathy Meier-Hellstern. I would also like to express my thank to the safety teams of Torr Vision Group at University of Oxford and Tresp Lab at University of Munich. This work is supported by the UKRI grant: Turing AI Fellowship EP/W002981/1, EPSRC/MURI grant: EP/N019474/1.

## References

- Clip interrogator, 2024. URL <https://github.com/pharmapsychotic/clip-interrogator>. Access: 14-03-2024.
- Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.
- Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. A simple yet efficient ensemble approach for ai-generated text detection. *arXiv preprint arXiv:2311.03084*, 2023.
- Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.
- Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.
- Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.
- Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chia-zor, et al. Detectors for safe and reliable llms: Implementations, uses, and limitations. *arXiv preprint arXiv:2403.06009*, 2024.
- Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1615–1631, 2018.
- Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

- 
- Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*, 2023.
- Sina Ahmadi. Open ai and its impact on fraud detection in financial industry. *Sina, A.(2023). Open AI and its Impact on Fraud Detection in Financial Industry. Journal of Knowledge Learning and Science Technology ISSN*, pages 2959–6386, 2023.
- Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.
- Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.
- Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. How to remove backdoors in diffusion models? In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*, 2023.
- Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.
- Mikhail J Atallah, Craig J McDonough, Victor Raskin, and Sergei Nirenburg. Natural language processing for information assurance and security: an overview and implementations. In *Proceedings of the 2000 workshop on New security paradigms*, pages 51–65, 2001.
- Mikhail J Atallah, Victor Raskin, Christian F Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E Triezenberg. Natural language watermarking and tamperproofing. In *International workshop on information hiding*, pages 196–212. Springer, 2002.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. *arXiv preprint arXiv:2004.05773*, 2020.
- Reza Azad, Moein Heidari, Moein Shariatnia, Ehsan Khodapanah Aghdam, Sanaz Karimijafarbigloo, Ehsan Adeli, and Dorit Merhof. Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In *International Workshop on Predictive Intelligence In MEDicine*, pages 91–102. Springer, 2022.
- Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*, 2023.
- Sima Azizi, Daniel B Hier, and Donald C Wunsch II. Enhanced neurologic concept recognition using a named entity recognition model based on transformers. *Frontiers in Digital Health*, 4:1065581, 2022.
- Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.
- Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. (ab)using images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

- 
- Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2534–2542, 2021.
- Yang Bai, Ge Pei, Jindong Gu, Yong Yang, and Xingjun Ma. Special characters attack: Toward scalable training data extraction from large language models. *arXiv preprint arXiv:2405.05990*, 2024.
- David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1): 52–62, 2023.
- Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacking: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.
- Ananth Balashankar, Xiao Ma, Aradhana Sinha, Ahmad Beirami, Yao Qin, Jilin Chen, and Alex Beutel. Improving few-shot generalization of safety classifiers via data augmented parameter-efficient fine-tuning. *arXiv preprint arXiv:2310.16959*, 2023.
- Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260, 2023.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.
- Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.
- Morteza Bashardoost, Mohd Shafry Mohd Rahim, Tanzila Saba, and Amjad Rehman. Replacement attack: A new zero text watermarking attack. *3D Research*, 8:1–9, 2017.
- Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.
- Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5136–5147, 2023.
- Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.
- Armin Behjati, Fatemeh Zare-Mirakabad, Seyed Shahriar Arab, and Abbas Nowzari-Dalini. Protein sequence profile prediction using protalbert transformer. *Computational Biology and Chemistry*, 99:107717, 2022.
- Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D’Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- Vishwesh Milind Bharadiya, Sree Kumar, et al. Generative ai in healthcare: A trustworthy approach. 2023. URL <https://openreview.net/forum?id=1WSd408I9M>.

---

JOSEPH R. BIDEN. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, Oct 2023. URL <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Bloomberg. Introducing bloomberggpt, bloomberg’s 50-billion parameter large language model, purpose-built from scratch for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Frimpong Boadu, Hongyuan Cao, and Jianlin Cheng. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics*, 39(Supplement\_1):i318–i325, 2023.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating inappropriateness in image generation: Can there be value in reflecting the world’s ugliness? *arXiv preprint arXiv:2305.18398*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation. *Center for Security and Emerging technology*, 1(1):2, 2021.

Tu Bui, Ning Yu, and John Collomosse. Repmix: Representation mixing for robust attribution of synthesized images. In *European Conference on Computer Vision*, pages 146–163. Springer, 2022.

Otilia Cangea and Gabriela Moise. A new approach of the cryptographic attacks. In *Digital Information and Communication Technology and Its Applications: International Conference, DICTAP 2011, Dijon, France, June 21-23, 2011. Proceedings, Part I*, pages 521–534. Springer, 2011.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*, 2023a.

Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Diffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Transactions on Multimedia*, 2023b.

- 
- Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017a.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017b.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022a.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022b.
- Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023a.
- Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023b.
- Egbert Castro, Abhinav Godavarthi, Julian Rubinfien, Kevin Givechian, Dhananjay Bhaskar, and Smita Krishnaswamy. Transformer-based protein generation with regularized latent space optimization. *Nature Machine Intelligence*, 4(10):840–851, 2022.
- Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jail-breaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.
- Gunvant R Chaudhari, Tengxiao Liu, Timothy L Chen, Gabby B Joseph, Maya Vella, Yoo Jin Lee, Thienkhai H Vu, Youngho Seo, Andreas M Rauschecker, Charles E McCulloch, et al. Application of a domain-specific bert for detection of speech recognition errors in radiology reports. *Radiology: Artificial Intelligence*, 4(4):e210185, 2022.
- Boyang Chen, Zongxiao Wu, and Ruoran Zhao. From fiction to fact: the growing role of generative ai in business and finance. *Journal of Chinese Economic and Business Studies*, 21(4):471–496, 2023a.
- Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362, 2020.
- Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, 2019.
- Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*, 2023b.
- Lichang Chen, Minhao Cheng, and Heng Huang. Backdoor learning on sequence to sequence models. *arXiv preprint arXiv:2305.02424*, 2023c.

---

Shuo Chen, Zhen Han, Bailan He, Zifeng Ding, Wenqian Yu, Philip Torr, Volker Tresp, and Jindong Gu. Red teaming gpt-4v: Are gpt-4v safe against uni/multi-modal jailbreak attacks? *arXiv preprint arXiv:2404.03411*, 2024.

Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023d.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 554–569, 2021a.

Xinyun Chen, Wenxiao Wang, Chris Bender, Yiming Ding, Ruoxi Jia, Bo Li, and Dawn Song. Refit: a unified watermark removal framework for deep learning systems with limited data. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 321–335, 2021b.

Yan Chen and Pouyan Esmaeilzadeh. Generative ai in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008, 2024.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pages 1964–1974. PMLR, 2021.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Juan Cobarrubias et al. Ethical issues in status planning. *Progress in language planning: International perspectives*, pages 41–85, 1983.

Davide Alessandro Coccolini, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Detecting images generated by diffusers. *arXiv preprint arXiv:2303.05275*, 2023.

Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

- 
- Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.
- Ingemar J Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. Secure spread spectrum watermarking for images, audio and video. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pages 243–246. IEEE, 1996.
- Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.
- Francesco Croce and Matthias Hein. Sparse and imperceptible adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4724–4732, 2019.
- Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.
- Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023a.
- Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023b.
- Suresh Budha Dahal. Utilizing generative ai for real-time financial market analysis opportunities and challenges. *Advances in Intelligent Information Systems*, 8(4):1–11, 2023.
- Falcon Z Dai and Zheng Cai. Towards near-imperceptible steganographic text. *arXiv preprint arXiv:1907.06679*, 2019.
- Zheng Dai and David K Gifford. Training data attribution for diffusion models. *arXiv preprint arXiv:2306.02174*, 2023.
- Giannis Daras and Alexandros G Dimakis. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169*, 2022.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.
- T Davenport. How morgan stanley is training gpt to help financial advisors. *Forbes Magazine*, May, 20, 2023.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.
- Nassim Dehouche and Kullathida Dehouche. What’s in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon*, 9(6), 2023.
- Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*, 2023a.
- Liwei Deng, Hongfei Suo, Dongjie Li, et al. Deepfake video detection based on efficientnet-v2 network. *Computational Intelligence and Neuroscience*, 2022, 2022.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023b.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

- 
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*, 2019.
- Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep’s clothing: Generalized nested jailbreak prompts can fool large language models easily. *arXiv preprint arXiv:2311.08268*, 2023.
- Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11966–11976, 2021.
- Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9185–9193, 2018.
- Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, 35:36789–36803, 2022.
- Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google’s bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. Stable diffusion is unstable. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR, 2023.
- Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023a.
- Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.
- Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR, 2023b.
- Jacob Dumford and Walter Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.
- Adam G Dunn, Ivy Shih, Julie Ayre, and Heiko Spallek. What generative ai means for trust in health communications. *Journal of Communication in Healthcare*, 16(4):385–388, 2023.

---

Florian Eckerli and Joerg Osterrieder. Generative adversarial networks in finance: an overview. *arXiv preprint arXiv:2106.06364*, 2021.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfahl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Tina Fang, Martin Jaggi, and Katerina Argyraki. Generating steganographic text with lstms. *arXiv preprint arXiv:1705.10742*, 2017.

Hany Farid. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022a.

Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022b.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Supervised gan watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pages 17–32, 2014.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36, 2024.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

- 
- Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*, 2023a.
- Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. A probabilistic fluctuation based membership inference attack for diffusion models. *arXiv e-prints*, pages arXiv–2308, 2023b.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- Boris A Galitsky. Truth-o-meter: Collaborating with llm in fighting its hallucinations. *Preprints*, 2023.
- Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*, 2021.
- Alessandro Gambetti and Qiwei Han. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*, 2023.
- Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. On pushing deepfake tweet detection capabilities to the limits. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 154–163, 2022.
- Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable detection of offensive and non-compliant content/logo in product images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2247–2256, 2020.
- Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.
- Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.
- Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. Evaluating the robustness of text-to-image diffusion models against real-world attacks. *arXiv preprint arXiv:2306.13103*, 2023a.
- Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4005–4014, 2023b.
- Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. *arXiv preprint arXiv:2401.11170*, 2024.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, 2023c.

- 
- Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.
- Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.
- Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.
- David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. Llm censorship: A machine learning challenge or a computer security problem? *arXiv preprint arXiv:2307.10719*, 2023.
- Ben Goertzel. Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to ray kurzweil’s the singularity is near, and mcdermott’s critique of kurzweil. *Artificial Intelligence*, 171(18):1161–1173, 2007.
- Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.
- Aditya Golatkar, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Training data protection with compositional diffusion models. *arXiv preprint arXiv:2308.01937*, 2023.
- Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International conference on learning representations (ICLR)*, 2015.
- Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175, 2019.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.
- Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39, 2023.
- Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.
- Jindong Gu, Baoyuan Wu, and Volker Tresp. Effective and efficient vote attack on capsule networks. *The International Conference on Learning Representations (ICLR)*, 2021.

- 
- Jindong Gu, Volker Tresp, and Yao Qin. Evaluating model robustness to patch perturbations. In *ICML 2022 Shift Happens Workshop*, 2022a.
- Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pages 404–421. Springer, 2022b.
- Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022c.
- Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023a.
- Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023b.
- Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. Controllable text generation via probability density estimation in the latent space. *arXiv preprint arXiv:2212.08307*, 2022d.
- David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.
- Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Medical Image Analysis*, 88:102872, 2023.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.
- Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*, 2021.
- Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *arXiv preprint arXiv:1908.01763*, 2019.
- Ariel Han and Zhenyao Cai. Design implications of generative ai systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, pages 470–474, 2023.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.
- Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- Fredrik Harder, Milad Jalali Asadabadi, Danica J Sutherland, and Mijung Park. Pre-trained perceptual features improve differentially private image generation. *arXiv preprint arXiv:2205.12900*, 2022.

- 
- Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghazemi. Aging with grace: Lifelong model editing with discrete key-value adaptors. *arXiv preprint arXiv:2211.11031*, 2022.
- Jamie Hayes and George Danezis. Generating steganographic images via adversarial training. *Advances in neural information processing systems*, 30, 2017.
- Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.
- Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022a.
- Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*, 2023.
- Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. Protecting intellectual property of language generation apis with lexical watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022b.
- Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*, 2023.
- Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- Benjamin Hilprecht, Martin Härtelich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019.
- Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1552–1565, 2020.
- Sorami Hisamoto, Matt Post, and Kevin Duh. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63, 2020.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020.
- Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.
- Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2021.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. *arXiv preprint arXiv:2403.00867*, 2024.
- Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022.

- 
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023a.
- Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 720–729, 2020.
- Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
- Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*, 2023b.
- Yihao Huang, Qing Guo, and Felix Juefei-Xu. Zero-day backdoor attack against text-to-image diffusion models via personalization. *arXiv preprint arXiv:2305.10701*, 2023c.
- Yue Huang and Lichao Sun. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*, 2023.
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*, 2020.
- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.
- Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53. Association for Computational Linguistics, 2023.
- Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. Llm platform security: Applying a systematic evaluation framework to openai’s chatgpt plugins. *arXiv preprint arXiv:2309.10254*, 2023.
- Aryan Jadon and Shashank Kumar. Leveraging generative ai models for synthetic data generation in healthcare: Balancing research and privacy. In *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, pages 1–4. IEEE, 2023.
- Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*, 2021.
- Md Saroor Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232, 2023.
- Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

- 
- Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.
- Xiaojun Jia, Yuefeng Chen, Xiaofeng Mao, Ranjie Duan, Jindong Gu, Rong Zhang, Hui Xue, and Xiaochun Cao. Revisiting and exploring efficient fast adversarial training via law: Lipschitz regularization and auto weight averaging. *arXiv preprint arXiv:2308.11443*, 2023.
- Xiaojun Jia, Jianshu Li, Jindong Gu, Yang Bai, and Xiaochun Cao. Fast propagation is better: Accelerating single-step adversarial training via sampling subnetworks. *IEEE Transactions on Information Forensics and Security*, 2024a.
- Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. Improved techniques for optimization-based jailbreaking on large language models. *arXiv preprint arXiv:2405.21018*, 2024b.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.
- Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization. *arXiv preprint arXiv:2303.04381*, 2023.
- Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4773–4783, 2019.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. *arXiv preprint arXiv:2302.05733*, 2023.
- Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*, 2024.
- Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What’s the answer right now? *arXiv preprint arXiv:2207.13332*, 2022.
- Aly M Kassem. Mitigating approximate memorization in language models via dissimilarity learned policy. *arXiv preprint arXiv:2305.01550*, 2023.
- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.
- Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In *European conference on computer vision*, pages 347–364. Springer, 2022a.
- Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022b.
- Changhoon Kim, Yi Ren, and Yezhou Yang. Decentralized attribution of generative models. *arXiv preprint arXiv:2010.13974*, 2020.
- Jaehyung Kim, Yuning Mao, Rui Hou, Hanchao Yu, Davis Liang, Pascale Fung, Qifan Wang, and Madian Khabsa. Robustifying language models via adversarial training with masked gradient. 2022c.

- 
- Jaehyung Kim, Yuning Mao, Rui Hou, Hanchao Yu, Davis Liang, Pascale Fung, Qifan Wang, Fulí Feng, Lifu Huang, and Madian Khabsa. Roast: Robustifying language models via adversarial perturbation with selective training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3412–3444, 2023a.
- Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*, 2023b.
- Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Bias-to-text: Debiasing unknown visual biases through language interpretation. *arXiv preprint arXiv:2301.11104*, 2023c.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.
- Ching-Yun Ko, Pin-Yu Chen, Payel Das, Yung-Sung Chuang, and Luca Daniel. On robustness-accuracy characterization of large language models using synthetic datasets. In *International Conference on Machine Learning*, 2023.
- Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023.
- Ziyi Kou, Shichao Pei, Yijun Tian, and Xiangliang Zhang. Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pages 983–990, 2023.
- David Krause. Large language models and generative ai in finance: An analysis of chatgpt, bard, and bing ai. *Bard, and Bing AI (July 15, 2023)*, 2023.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.
- Murat Kuzlu, Zhenxin Xiao, Salih Sarp, Ferhat Ozgur Catak, Necip Gurler, and Ozgur Guler. The rise of generative artificial intelligence in healthcare. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE, 2023.
- Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.
- Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*, 2023.
- Haoheng Lan, Jindong Gu, Philip Torr, and Hengshuang Zhao. Influencer backdoor attack on semantic segmentation. *arXiv preprint arXiv:2303.12054*, 2023.
- Raz Lapid, Ron Langberg, and Moshe Sipper. Open sesame! universal black box jailbreaking of large language models. *arXiv preprint arXiv:2309.01446*, 2023.

- 
- Mike Laszkiewicz, Jonas Ricker, Johannes Lederer, and Asja Fischer. Single-model attribution via final-layer inversion. *arXiv preprint arXiv:2306.06210*, 2023.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.
- Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- Nayeon Lee, Wei Ping, Peng Xu, Mostafa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanaro. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, 2022.
- Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C Wallace. Does bert pretrained on clinical notes reveal sensitive data? *arXiv preprint arXiv:2104.07762*, 2021.
- Evelina Leivada, Elliot Murphy, and Gary Marcus. Dall·e 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, 8(1):100648, 2023.
- Roberto Leotta, Oliver Giudice, Luca Guarnera, and Sebastiano Battiat. Not with my name! inferring artists’ names of input strings employed by diffusion models. In *International Conference on Image Analysis and Processing*, pages 364–375. Springer, 2023.
- Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*, 2020.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.
- Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multi-modal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023b.
- Guanlin Li, Shuai Yang, Jie Zhang, and Tianwei Zhang. Prime: Protect your videos from malicious editing. *arXiv preprint arXiv:2402.01239*, 2024a.
- Han Li, Dan Zhao, and Jianyang Zeng. Kpgt: knowledge-guided pre-training of graph transformer for molecular property prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 857–867, 2022a.
- Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. *arXiv preprint arXiv:2311.17216*, 2023c.
- Haoran Li, Yangqiu Song, and Lixin Fan. You don’t know my favorite color: Preventing dialogue representations from revealing speakers’ private personas. *arXiv preprint arXiv:2205.10228*, 2022b.
- Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023d.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022c.
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, 2023e.

- 
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023f.
- Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.
- Mingxiao Li, Tingyu Qu, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. *arXiv preprint arXiv:2305.15583*, 2023g.
- Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers. *arXiv preprint arXiv:2402.16914*, 2024b.
- Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023h.
- Yansong Li, Zhixing Tan, and Yang Liu. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*, 2023i.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023j.
- Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022d.
- Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.
- Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.
- Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.
- Zheng Li, Chengyu Hu, Yang Zhang, and Shanqing Guo. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of dnn. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 126–137, 2019.
- Zuchao Li, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. Batgpt: A bidirectional autoregressive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*, 2023k.
- Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.
- Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023a.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*, 2023b.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention glitches with flip-flop language modeling. *arXiv preprint arXiv:2306.00946*, 2023a.
- Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Model-agnostic origin attribution of generated images with few-shot examples. *arXiv preprint arXiv:2404.02697*, 2024a.

- 
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9, 2023b.
- Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023c.
- Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023d.
- Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022a.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023e.
- Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. *arXiv preprint arXiv:2404.08031*, 2024b.
- Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022.
- Xiaogeng Liu, Nan Xu, Muhaoo Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023f.
- Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, 2023g.
- Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022b.
- Xinwei Liu, Xiaojun Jia, Jindong Gu, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Does few-shot learning suffer from backdoor attacks? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024c.
- Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*, 2023h.
- Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023i.
- Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023j.
- Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48. IEEE, 2017.
- Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv preprint arXiv:2202.02013*, 2022.
- Guillermo López-García, José M Jerez, Nuria Ribelles, Emilio Alba, and Francisco J Veredas. Explainable clinical coding with in-domain adapted transformers. *Journal of Biomedical Informatics*, 139:104323, 2023.
- Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

- 
- Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.
- Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539*, 2023.
- Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- Renqian Luo, Lai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.
- Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022.
- Saiyue Lyu, Margarita Vinaroz, Michael F Liu, and Mijung Park. Differentially private latent diffusion models. *arXiv preprint arXiv:2305.15759*, 2023.
- Avery Ma, Amir-massoud Farahmand, Yangchen Pan, Philip Torr, and Jindong Gu. Improving adversarial transferability via model alignment. *arXiv preprint arXiv:2311.18495*, 2023a.
- Xiao Ma, Swaroop Mishra, Ahmad Beirami, Alex Beutel, and Jilin Chen. Let’s do a thought experiment: Using counterfactuals to improve moral reasoning. *arXiv preprint arXiv:2306.14308*, 2023b.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.
- Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.
- Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- Antonis Maronikolakis, Mark Stevenson, and Hinrich Schütze. Transformers are better than humans at identifying generated text. *ArXiv abs/2009.13375*, 2020.
- Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019.
- Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 667–684. Springer, 2020.
- Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2023.
- Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

- 
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
- Kris McGuffie and Alex Newhouse. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*, 2020.
- Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. *arXiv preprint arXiv:2310.15007*, 2023.
- Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. Robust conversational agents against imperceptible toxicity triggers. *arXiv preprint arXiv:2205.02392*, 2022.
- Alex Mei, Sharon Levy, and William Yang Wang. Assert: Automated safety scenario red teaming for evaluating the robustness of large language models. *arXiv preprint arXiv:2310.09624*, 2023.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.
- Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25(8):3121–3129, 2021.
- Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language*, 23(1):107–125, 2009.
- Ning Miao, Yee Whye Teh, and Tom Rainforth. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*, 2023.
- Raphaël Millière. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*, 2022.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.
- Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929*, 2022a.
- Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, 2022b.
- Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.
- Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*, 2023.

- 
- Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9096, 2019.
- Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 668–669, 2020.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.
- John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.
- Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D Griffin. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*, 2023.
- Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.
- Sumit Mukherjee, Yixi Xu, Anusua Trivedi, Nabajyoti Patowary, and Juan L Ferres. privgan: Protecting gans from membership inference attacks at low cost to utility. *Proceedings on Privacy Enhancing Technologies*, 2021.
- Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.
- Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.
- Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023.
- Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*, 2021.
- Yusuke Naritomi and Takanori Adachi. Data augmentation of high frequency financial data using generative adversarial network. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 641–648. IEEE, 2020.
- Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.
- Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019.
- Clarence Boon Liang Ng, Diogo Santos, and Marek Rei. Modelling temporal document sequences for clinical icd coding. *arXiv preprint arXiv:2302.12666*, 2023.

- 
- Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE, 2019.
- Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.
- Minheng Ni, Chenfei Wu, Xiaodong Wang, Shengming Yin, Lijuan Wang, Zicheng Liu, and Nan Duan. Ores: Open-vocabulary responsible visual synthesis. *arXiv preprint arXiv:2308.13785*, 2023.
- Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. In *International Conference on Machine Learning*, pages 26150–26165. PMLR, 2023.
- Wimukthi Nimalsiri, Mahela Hennayake, Kasun Rathnayake, Thanuja D Ambegoda, and Dulani Meedeniya. Automated radiology report generation using transformers. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 90–95. IEEE, 2023.
- Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023a.
- Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. *arXiv preprint arXiv:2301.11706*, 2023b.
- JJK ó Ruanaidh, WJ Dowling, and FM Boland. Watermarking digital images for copyright protection. *IEE PROCEEDINGS VISION IMAGE AND SIGNAL PROCESSING*, 143:250–256, 1996.
- Hyun-Jic Oh and Won-Ki Jeong. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. Springer, 2023.
- Myung Gyo Oh, Leo Hyun Park, Jaeuk Kim, Jaewoo Park, and Taekyoung Kwon. Membership inference attacks with token-level deduplication on korean language models. *IEEE Access*, 11:10207–10217, 2023.
- Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.
- Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. Entity cloze by date: What lms know about unseen entities. *arXiv preprint arXiv:2205.02832*, 2022.
- OpenAI. Gpt-4 technical report, 2023.
- OpenAI. Dall·e 2 pre-training mitigations, 2024a. URL <https://openai.com/research/dall-e-2-pre-training-mitigations>. Accessed: 2024-03-12.
- OpenAI. Sora, 2024b. URL <https://openai.com/sora>. Access: 14-03-2024.
- Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pages 1–14, 2023.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Muzaffer Özbeý, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.
- Mustafa Safa Ozdayi, Charith Peris, Jack Fitzgerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. Controlling the extraction of memorized data from large language models via prompt-tuning. *arXiv preprint arXiv:2305.11759*, 2023.

---

Lorenzo Pacchiardi, Alex J Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. *arXiv preprint arXiv:2309.15840*, 2023.

Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 68(10):105004, 2023.

Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3611–3628, 2022.

Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Roma Patel and Ellie Pavlick. “was it “stated” or was it “claimed”??: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10095, 2021.

Andrea Paudice, Luis Muñoz-González, Andras Gyorgy, and Emil C Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *arXiv preprint arXiv:1802.03041*, 2018.

Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pages 5–15. Springer, 2019.

Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Protecting the intellectual property of diffusion models by the watermark diffusion process. *arXiv preprint arXiv:2306.03436*, 2023a.

Sen Peng, Yufei Chen, Jie Xu, Zizhuo Chen, Cong Wang, and Xiaohua Jia. Intellectual property protection of dnn models. *World Wide Web*, 26(4):1877–1911, 2023b.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023c.

Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022a.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022b.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36, 2024.

---

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213*, 2023a.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023b.

Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*, 2023.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3403–3417, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications. *arXiv preprint arXiv:2311.08592*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13198–13207, 2020.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Nitin Rane. Role and challenges of chatgpt and similar generative artificial intelligence in finance and accounting. Available at SSRN 4603206, 2023.

Nitin Rane, Saurabh Choudhary, and Jayesh Rane. Gemini or chatgpt? efficiency, performance, and adaptability of cutting-edge generative artificial intelligence (ai) in finance and accounting. *Efficiency, Performance, and Adaptability of Cutting-Edge Generative Artificial Intelligence (AI) in Finance and Accounting* (February 19, 2024), 2024.

Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks. *arXiv preprint arXiv:2305.14965*, 2023.

---

Ambrish Rawat, Killian Levacher, and Mathieu Sinn. The devil is in the gan: backdoor attacks and defenses in deep generative models. In *European Symposium on Research in Computer Security*, pages 776–783. Springer, 2022.

Nydia Remolina. Generative ai in finance: Risks and potential solutions. *Finance: Risks and Potential Solutions (November 9, 2023). Singapore Management University School of Law Research Paper Forthcoming, SMU Centre for AI & Data Governance Research Paper Forthcoming*, 2023.

Anthony Rhodes, Ram Bhagat, Umur Aybars Ciftci, and Ilke Demir. My art my choice: Adversarial protection against unruly ai. *arXiv preprint arXiv:2309.03198*, 2023.

H. S. Richardson. Moral reasoning, 2018. URL <https://plato.stanford.edu/entries/reasoning-moral/>.

Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

Sohini Roychowdhury. Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1180–1181, 2024.

Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 236–251. Springer, 2020.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Siva Sai, Aanchal Gaur, Revant Sai, Vinay Chamola, Mohsen Guizani, and Joel JPC Rodrigues. Generative ai for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations. *IEEE Access*, 2024.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

---

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*, 2021.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

Areg Mikael Sarvazyan, José Ángel González, M Franco Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. Autextification: automatic text identification. *Procesamiento del Lenguaje Natural. Jaén, Spain*, 2023.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. *arXiv preprint arXiv:1910.00292*, 2019.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023.

Ghiath Shabsigh and El Bachir Boukherouaa. Generative artificial intelligence in finance. *FinTech Notes*, 2023(006), 2023.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Erfan Shayegani, Yue Dong, and Nael B. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023.

Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. Membership inference attacks against nlp classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. Prompt stealing attacks against text-to-image generation models. *arXiv preprint arXiv:2302.09923*, 2023.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Victoria Lin, Noah A Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*, 2023.

- 
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- Yasin Shokrollahi, Sahar Yarmohammadoosky, Matthew M Nikahd, Pengfei Dong, Xianqi Li, and Linxia Gu. A comprehensive review of generative ai in healthcare. *arXiv preprint arXiv:2310.00795*, 2023.
- Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*, 2023.
- Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pages 212–231. IEEE, 2021.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- Aaditya K Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*, 2024.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfahl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Aradhana Sinha, Ananth Balashankar, Ahmad Beirami, Thi Avrahami, Jilin Chen, and Alex Beutel. Break it, imitate it, fix it: Robustness by generating human-like attacks. *arXiv preprint arXiv:2310.16955*, 2023.
- Sonish Sivarajkumar and Yanshan Wang. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, page 972. American Medical Informatics Association, 2022.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.
- Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.
- Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390, 2020.
- Hansa Srinivasan, Candice Schumann, Aradhana Sinha, David Madras, Gbolahan Oluwafemi Olanubi, Alex Beutel, Susanna Ricco, and Jilin Chen. Generalized people diversity: Learning a human perception-aligned diversity representation for people images. *arXiv preprint arXiv:2401.14322*, 2024.
- Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

- 
- Felix Stahlberg and Bill Byrne. On nmt search errors and model errors: Cat got your tongue? *arXiv preprint arXiv:1908.10090*, 2019.
- Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4584–4596, 2023.
- Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*, 2023a.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023b.
- Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pre-trained language models. *arXiv preprint arXiv:2205.05124*, 2022.
- Xingming Sun and Alex Jessey Asiimwe. Noun-verb based technique of text watermarking using recursive decent semantic net parsers. In *International Conference on Natural Computation*, pages 968–971. Springer, 2005.
- Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergrpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. Task ambiguity in humans and language models. *arXiv preprint arXiv:2212.10711*, 2022.
- Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023.
- Umut Topkara, Mercan Topkara, and Mikhail J Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174, 2006.
- Hugo et al. Touvron. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.
- Loc Truong, Chace Jones, Brian Hutchinson, Andrew August, Brenda Praggastis, Robert Jasper, Nicole Nichols, and Aaron Tuor. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 788–789, 2020.
- Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.
- Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.
- Sonam Tyagi, Harsh Vikram Singh, Raghav Agarwal, and Sandeep Kumar Gangwar. Digital watermarking techniques for security applications. In *2016 International Conference on Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES)*, pages 379–382. IEEE, 2016.

- 
- Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*, 2022.
- Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395, 2020.
- Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*, 2021.
- Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin’ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.
- Honai Ueoka, Yugo Murawaki, and Sadao Kurohashi. Frustratingly easy edit-based linguistic steganography with a masked language model. *arXiv preprint arXiv:2104.09833*, 2021.
- Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.
- Nur Alya Afikah Usop and Syifak Izhar Hisham. A review of digital watermarking techniques, characteristics and attacks in text documents. *Advances in Robotics, Automation and Data Analytics: Selected Papers from iCITES 2020*, pages 256–271, 2021.
- Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.
- Julian Varghese and Julius Chapiro. Chatgpt: The transformative influence of generative ai on science and healthcare. *Journal of Hepatology*, 2023.
- Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.
- Yukti Varshney. Attacks on digital watermarks: classification, implications, benchmarks. *Int J Emerg Technol (Special Issue NCETST-2017)*, 8(1):229–235, 2017.
- Henriikka Vartiainen and Matti Tedre. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity*, 34(1):1–21, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *arXiv preprint arXiv:2307.16489*, 2023.
- Hrishikesh Viswanath and Tianyi Zhang. Fairpy: A toolkit for evaluation of social biases and their mitigation in large language models. *arXiv preprint arXiv:2302.05508*, 2023.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.
- Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on nlp models. *arXiv preprint arXiv:2010.12563*, 2020.
- Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.00944*, 2023.

- 
- Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.
- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023a.
- Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. *arXiv preprint arXiv:2005.03642*, 2020.
- Chaojun Wang, Yang Liu, and Wai Lam. Progressive translation: Improving domain robustness of neural machine translation with intermediate sequences. *arXiv preprint arXiv:2305.09154*, 2023b.
- Chenan Wang, Jinhao Duan, Chaowei Xiao, Edward Kim, Matthew Stamm, and Kaidi Xu. Semantic adversarial attacks via diffusion models. *arXiv preprint arXiv:2309.07398*, 2023c.
- Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method against text-to-image synthesis of diffusion models. *arXiv preprint arXiv:2312.07865*, 2023d.
- Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*, 2023e.
- Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023f.
- Liwei Wang, Huan He, Andrew Wen, Sungrim Moon, Sunyang Fu, Kevin J Peterson, Xuguang Ai, Sijia Liu, Ramakanth Kavuluru, and Hongfang Liu. Acquisition of a lexicon for family history information: Bidirectional encoder representations from transformers-assisted sublanguage analysis. *JMIR Medical Informatics*, 11:e48072, 2023g.
- Ren Wang, Gaoyuan Zhang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong, and Meng Wang. Practical detection of trojan neural networks: Data-limited and data-free cases. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 222–238. Springer, 2020a.
- Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. *arXiv preprint arXiv:2206.00477*, 2022.
- Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020b.
- Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7192–7203, 2023h.
- Tianhao Wang and Florian Kerschbaum. Riga: Covert and robust white-box watermarking of deep neural networks. In *Proceedings of the Web Conference 2021*, pages 993–1004, 2021.
- Yanqing Wang. Generative ai in operational risk management: Harnessing the future of finance. *Operational Risk Management: Harnessing the Future of Finance (May 17, 2023)*, 2023.
- Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. Defending llms against jailbreaking attacks via backtranslation. *arXiv preprint arXiv:2402.16459*, 2024a.
- Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023i.

---

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. *arXiv preprint arXiv:2402.14899*, 2024b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023j.

Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Alteration-free and model-agnostic origin attribution of generated images. *arXiv preprint arXiv:2305.18439*, 2023k.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023b.

Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023c.

Max Weiss. Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science*, 2019121801, 2019.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023a.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023b.

Nevan Wichers, Carson Denison, and Ahmad Beirami. Gradient-based language model red teaming. *arXiv preprint arXiv:2401.16656*, 2024.

Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021.

Boxi Wu, Jindong Gu, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Towards efficient adversarial training on vision transformers. In *European Conference on Computer Vision*, pages 307–325. Springer, 2022a.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, 2023a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023b.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023c.

- 
- Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. 2022b.
- Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models' vacillations in judgement. *arXiv preprint arXiv:2310.02174*, 2023a.
- Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pages 1–11, 2023b.
- Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for undersampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer, 2022.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2023.
- Yuan Xun, Xiaojun Jia, Jindong Gu, Xinwei Liu, Qing Guo, and Xiaochun Cao. Minimalism is king! high-frequency energy-based screening for data-efficient backdoor attacks. *IEEE Transactions on Information Forensics and Security*, 2024.
- Jun Yan, Vansh Gupta, and Xiang Ren. Bite: Textual backdoor attacks with iterative trigger injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12951–12968, 2023a.
- Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*, 2023b.
- Chaofei Yang, Leah Ding, Yiran Chen, and Hai Li. Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021a.
- Hyemin Yang, Heekyung Yang, and Kyungha Min. Artfusion: A diffusion model-based style synthesis framework for portraits. *Electronics*, 13(3):509, 2024a.
- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023a.
- Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*, 2021.
- Shiping Yang, Renliang Sun, and Xiaojun Wan. A new benchmark and reverse validation method for passage-level hallucination detection. *arXiv preprint arXiv:2310.06498*, 2023b.
- Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. Rethinking stealthiness of backdoor attack against nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5543–5557, 2021b.
- Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*, 2023c.

- 
- Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2023d.
- Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. *arXiv preprint arXiv:2311.17516*, 2023e.
- Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzheng Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 123–123. IEEE Computer Society, 2024b.
- Yulong Yang, Xinshan Yang, Shuaidong Li, Chenhao Lin, Zhengyu Zhao, Chao Shen, and Tianwei Zhang. Security matrix for multimodal agents on mobile devices: A systematic and proof of concept study. *arXiv preprint arXiv:2407.09295*, 2024c.
- Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 53–62, 2020.
- Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023a.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023b.
- Meng Yingjie, Liu Huiran, Shang Tong, and Teng Xiaoyu. A zero-watermarking scheme for prose writings. In *2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 276–282. IEEE, 2017.
- Vithya Yogarajan, Jacob Montiel, Tony Smith, and Bernhard Pfahringer. Transformers for multi-label classification of medical text: an empirical comparison. In *International Conference on Artificial Intelligence in Medicine*, pages 114–123. Springer, 2021.
- Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.
- Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2020.
- Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 14448–14457, 2021a.
- Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624, 2021b.
- Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*, 2023a.
- Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*, 2023b.

- 
- Wenqian Yu, Jindong Gu, Zhijiang Li, and Philip Torr. Reliable evaluation of adversarial transferability. *arXiv preprint arXiv:2306.08565*, 2023c.
- Youliang Yuan, Wenxiang Jiao, Wenzuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.
- Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. Rigorllm: Resilient guardrails for large language models against undesired content. *arXiv preprint arXiv:2403.13031*, 2024.
- Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 363–375, 2020.
- Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 771–785, 2023.
- Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*, 2024.
- Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023.
- Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*, 2023a.
- Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*, 2021.
- Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pages 159–172, 2018.
- Jianping Zhang, Zhuoer Xu, Shiwen Cui, Changhua Meng, Weibin Wu, and Michael R Lyu. On the robustness of latent diffusion models. *arXiv preprint arXiv:2306.08257*, 2023b.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023c.
- Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023d.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.
- Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Xiaofei Xie, Yang Liu, and Chao Shen. A mutation-based method for multi-modal jailbreaking attack detection. *arXiv preprint arXiv:2312.10766*, 2023e.

---

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023f.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*, 2023a.

Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14443–14452, 2020.

Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. *arXiv preprint arXiv:2305.01219*, 2023b.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2022.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023c.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023d.

Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023e.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? *arXiv preprint arXiv:2312.00084*, 2023f.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. *arXiv preprint arXiv:2311.00500*, 2023.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*, 2020.

Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, and Weimin Li. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, 7(6):743–755, 2023.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.

Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. Duwak: Dual watermarks in large language models. *arXiv preprint arXiv:2403.13000*, 2024.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018.

---

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2391, 2023.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.