Trustworthy AI: Safety, Bias, and Privacy - A Survey

Anonymous authors
Paper under double-blind review

Abstract

The capabilities of artificial intelligence systems have been advancing to a great extent, but these systems still struggle with failure modes, vulnerabilities, and biases. In this paper, we study the current state of the field, and present promising insights and perspectives regarding concerns that challenge the trustworthiness of AI models. In particular, this paper investigates the issues regarding three thrusts: safety, privacy, and bias, which hurt models' trustworthiness. For safety, we discuss safety alignment in the context of large language models, preventing them from generating toxic or harmful content. For bias, we focus on spurious biases that can mislead a network. Lastly, for privacy, we cover membership inference attacks in deep neural networks. The discussions addressed in this paper reflect our own experiments and observations.

1 Introduction

The value of AI systems lies in their ability to generalize to unseen situations - if one already knows what their model is going to encounter, they must program it instead of training them to learn generalizable patterns. The standard method to assess how well a model is able to generalize is to measure how well it does on an unseen dataset drawn from the same distribution as the data the model is trained on. However, this provides an incomplete picture regarding the trustworthiness/reliability of these models in the real world. In this paper, we discuss three critical aspects of deep learning that allow for a greater understanding of the trustworthiness of AI systems: safety, bias, and privacy.

With the rapid adoption of large language models (LLMs) in fields such as healthcare, finance, and cybersecurity, ensuring their safe deployment has become a pressing concern. While LLMs offer immense potential, their misuse—whether intentional or accidental—can lead to severe societal consequences, such as misinformation propagation, security vulnerabilities, and ethical risks. Safety alignment aims to mitigate these issues by preventing LLMs from generating harmful or unethical content. Over the recent years, many techniques have been developed to improve safety, including supervised fine-tuning (SFT), reinforcement learning with human feedback (RLHF), direct preference optimization (DPO), etc. These methods have significantly enhanced LLM alignment, enabling models to better adhere to human-defined safety constraints. However, challenges remain in ensuring their robustness across adversarial scenarios, as existing approaches often struggle with various jailbreak attacks, fine-tuning exploits, and decoding manipulations. Addressing these limitations requires a deeper understanding of safety alignment strategies and the development of techniques that can establish safety guardrails throughout the text generation.

As for issues concerning bias in deep learning, we focus on the problem of spurious correlations, where a network primarily relies on weakly predictive features in the training set that are causally unrelated to ground truth labels in classification tasks. Reliance on these features is undesirable as they may disappear or become associated with a different task during testing. To overcome the reliance on these features, many promising solutions have been recently proposed. In this paper, we study these techniques in detail and understand their limitations while introducing studies of new directions to overcome these limitations. We also discuss the intersection of spurious correlations with other fields of study within deep learning.

For the last, we discuss privacy issues in deep learning models, especially for membership inference attacks where an attacker tries to infer whether a sample belongs to a train set or not - which is membership information. Existing deep learning models are often vulnerable to such attacks when they exhibit behavioral discrepancies between training and unseen data points. Once a model is under such an attack, it is disclosed whether a data point has been involved in training the model. To avoid such privacy leakage, many solutions from various perspectives have been proposed. In this paper, we discuss the privacy vulnerabilities in terms of the correlations between model capacity and data complexity. We also comprehensively present the existing privacy preservation approaches and discuss their potential future directions.

This paper provides a comprehensive survey and discussions regarding trustworthy AI, especially safety, bias, and privacy, which will contribute to the research community for further actionable works and also provide insights to the fields outside of AI/deep learning.

2 Safety Alignment in LLMs

In Large Language Models, alignment aims to teach models human-desired behaviors and remove undesired behaviors. Safety alignment has often been treated as a subset of broader alignment challenges, with a primary focus on safety Li & Kim (2024). In this context, the goal of safety alignment is preventing LLMs from generating toxic or harmful content and simultaneously considers security problems in adversarial scenarios, such as jailbreak attempts Qi et al. (2024).

2.1 Why is Safety in LLMs Important?

With the release of AI services such as ChatGPT, Claude, and Gemini, AI-powered applications have become increasingly integrated into our daily lives, spanning fields like healthcare, finance, education, transportation, and even military applications OpenAI (2022); Gemini (2023). However, this rapid adoption also raises serious concerns regarding AI misuse. For instance, in 2023, the first suspected case of AI-assisted suicide was reported Brussels Times (2023), and AI-generated misinformation has been widely disseminated online, potentially manipulating public opinion Monteith et al. (2024). These incidents compel us to critically examine how to prevent AI from being misused in ways that harm society.

This challenge has become even more pressing with the rise of open-sourced large language models such as Llama and Deepseek families Touvron et al. (2023); Dubey et al. (2024); Guo et al. (2025), which allow individuals to control and finetune LLMs directly. As a result, the risks associated with AI misuse are expanding exponentially while government regulatory measures struggle to keep pace. Given these developments, AI safety has become a pressing need in the current research community.

2.2 How to Implement Safety in LLMs?

Upon the release of GPT-3 in 2020, we witnessed LLM's remarkable language generation capabilities. However, concerns regarding bias, toxic content, and hallucinations also emerged, indicating that the model was still not ready for public release Brown et al. (2020). Later, in 2021, Anthropic introduced the HHH principle—Helpful, Honest, and Harmless—as a key principle of a truly beneficial AI assistant Askell et al. (2021). This concept set a foundational standard for AI assistants, clearly indicating that safety is an important objective of LLMs.

The launch of ChatGPT in late 2022 marked a turning point, as it was the first large-scale exposure of LLMs to the public, allowing people to experience an AI assistant that felt genuinely helpful and capable of human-like reasoning. This breakthrough was largely built upon the techniques outlined in the InstructGPT, which introduced Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) as key methods for aligning LLMs with human values Ouyang et al. (2022). Around the same time, Anthropic also released its own research on RLHF, demonstrating its effectiveness in guiding model behavior Bai et al. (2022a). These efforts treated safety (or harmlessness) as a subset of preference optimization, laying the foundation for future safety alignment research. From a high-level perspective, subsequent alignment techniques can be approximately categorized into the following approaches:

2.2.1 In-Context Learning

This category of methods does not rely on model retraining; instead, inspired by Chain-of-Thought (CoT) reasoning and GPT-3, researchers have designed either hard prompts or soft prompts to guide the model toward producing helpful and harmless outputs Wei et al. (2022); Brown et al. (2020). Examples include the official system prompt in Llama2 and soft prompts optimized via P-tuning, which embed safety reasoning signals that may not be readable by humans Touvron et al. (2023); Xie et al. (2023). However, in-context learning has several limitations: **(1)** It requires careful manual design and optimization, making automation and scalability difficult. **(2)** It has limited generalization ability, struggling to handle long-tail scenarios where prompts may not be well-defined. **(3)** It is highly sensitive to prompt design; small variations in the prompt can lead to drastically different outputs, making the method vulnerable to jailbreak attacks. **(4)** Since in-context learning does not modify the model's weights, it cannot permanently alter its underlying behavior.

2.2.2 Imitation Learning

This approach primarily relies on supervised fine-tuning (SFT) to train models using carefully curated aligned datasets Ouyang et al. (2022). Zhou et al. (2024a) introduced the Superficial Alignment Hypothesis, suggesting that alignment may merely adjust the model's output distribution to be more interaction-friendly rather than fundamentally changing its reasoning capabilities. They demonstrated that with only about 1k high-quality training examples, a model could achieve performance comparable to GPT-4. However, this paper focused predominantly on helpfulness, and its training data contained only 13 safety-related samples, which led to poor overall safety performance.

2.2.3 Reinforcement Learning

Reinforcement Learning with Human Feedback (RLHF) improves model alignment by optimizing the policy (parameters) using reinforcement learning. The core idea is to train the model to generate better responses by maximizing a reward function, which is defined by a reward model trained on human feedback. The reward model assigns scores to different responses, providing a structured signal to guide the optimization process. To prevent the model from diverging too far from its original behavior, RLHF typically employs proximal policy optimization (PPO), which introduces a KL divergence constraint between the logits of the original and updated policies Christiano et al. (2017). This ensures that while the model learns to produce more aligned outputs, it does not lose fluency or develop unintended artifacts. RLHF has significantly improved both helpfulness and harmlessness, establishing itself as a foundational technique in alignment research Ouyang et al. (2022); Bai et al. (2022a). Despite its effectiveness, RLHF comes with significant challenges. The dependence on human feedback makes it highly labor-intensive, as continuous human involvement is required to annotate responses and update the reward model. To reduce reliance on human labor and improve scalability, researchers have proposed Reinforcement Learning with AI Feedback (RLAIF) as an alternative. For example, Anthropic incorporates an AI-generated feedback mechanism, where the model itself evaluates responses, reducing the need for human intervention while still refining behavior and mitigating harmful outputs Bai et al. (2022b). However, both RLHF and RLAF are resources intensive, as their training typically involves the following components: (1) a reference model, (2) a policy model (an adapted version of the base model being optimized), and (3) a reward model trained to assess response quality. In some implementations, additional models may even be required, further increasing the memory and computational burden Yao et al. (2023).

2.2.4 Contrastive Preference Modeling

Researchers have explored the use of contrastive preference signals as a more efficient alternative to reduce the complexity and resource demands of RL-based approaches. Direct Preference Optimization (DPO) introduced the key insight that large language models inherently act as implicit reward models Rafailov et al. (2024). By leveraging the preference dataset, a model can directly learn preferences without requiring an explicit reinforcement learning loop and a reward model. A similar approach is proposed in Liu et al. (2023), where contrastive signals are embedded in datasets to steer models toward desired behaviors. This category of methods has notable advantages: (1) it significantly reduces memory costs since optimization requires loading at most two models at a time, and with logit caching, only one model may be sufficient; (2) it eliminates the need for an explicit reward model, simplifying the alignment process. However, these methods also come with challenges: the performance is highly dependent on the quality of the preference dataset.

2.2.5 Conditional Learning

This approach is conceptually similar to in-context learning but differs in that it explicitly optimizes the model to recognize specific triggers, ensuring that desired behaviors are always generated when these triggers are present. The key idea is to induce the model to produce desired behavior rather than removing undesired outputs. However, this approach has significant vulnerabilities when confronted with jailbreak attacks and thus is rarely used as a standalone alignment technique Korbak et al. (2023).

2.3 Safety in Existing LLMs is Still Brittle

Although various general alignment methods have been proposed, and safety alignment has been improved to some extent, treating safety merely as a subset of human preference overlooks its unique challenges. As a result, current alignment techniques remain vulnerable to adversarial attacks. In literature, adversarial attacks on LLMs can generally be classified into three types: **(1) Jailbreak Attacks**: Attackers exploit techniques such as role-playing or suffix injections to bypass safety guardrails and manipulate the model into generating harmful content. Studies show that these methods can effectively evade existing alignment mechanisms Zou et al. (2023). This vulnerability extends beyond open-source models—even state-of-the-art systems like the GPT-4 series struggle to block harmful outputs in complex, nested scenarios consistently Li et al. (2023). **(2) Finetuning Attacks**: Even unintentional finetuning can weaken a model's safety mechanisms. A model trained with safety alignment may gradually lose its safeguards when adapted to downstream tasks via domain-specific finetuning, even if the dataset itself is benign. This phenomenon has been observed in both open-source and proprietary models Qi et al. (2023). **(3) Decoding Attacks**: Safety-aligned models may still produce harmful content under certain decoding settings, such as modifications to Top-P, Top-K, or Temperature Huang et al. (2023). These variations may break built-in safeguards, leading to outputs that would otherwise be restricted under default configurations. These attack vectors underscore a critical issue: existing safety alignment methods lack robustness and, in many cases, remain highly brittle, especially in novel or adversarial conditions.

2.4 How to Implement Robust Safety in LLMs?

Recent studies have highlighted that existing alignment methods often achieve safety at a superficial level. Wei et al. (2024) identified safety-critical parameters in LLMs and found that removing them catastrophically degrades safety performance while leaving utility performance unaffected. However, their findings also revealed that merely retaining these safety-critical parameters does not preserve safety under finetuning attacks. In contrast, Li & Kim (2024) demonstrated that the atomic functional unit for safety in LLMs resides at the *neuron level* and successfully mitigated finetuning attacks by freezing updates to these safety-critical components. Their study further showed that aligned models remain vulnerable to finetuning attacks because key attributes, such as utility, can be achieved by repurposing neurons originally responsible for other functions, such as safety. Additionally, this research examined how alignment influences model behavior in safety-critical contexts and observed that, at its core, this effect could be framed as an implicit safety-related binary classification task. To resolve the superficiality issue above, they further propose that alignment should enable models to choose the correct safety-aware reasoning direction (either to refuse or fulfill) at each generation step, ensuring safety throughout the entire response. However, their work did not propose specific methods for implementing this deeper safety mechanism in practice.

Qi et al. (2024) have also examined the shallow alignment in existing LLMs and found that this issue often stems from alignment disproportionately affecting early-generated token distribution. This creates optimization shortcuts where models rely on superficial decision patterns, leading them toward local optima that fail to generalize to more complex safety challenges. To mitigate this, they introduced a data augmentation strategy designed to expose models to more nuanced scenarios where an initially harmful response later transitions into a safe refusal. Similarly, Yuan et al. (2024) have adopted more aggressive data construction rules, aiming to add more variety of training examples. However, while these methods increase the diversity of training examples, they do not fundamentally address the root problem. All of these highlight a critical issue: Existing alignment techniques lack effective and robust mechanisms to handle complex and nuanced harmful reasoning patterns. In this context, this survey paper acknowledges the hypothesis from Li & Kim (2024), and believes that a robust safety alignment should teach the model to select and maintain the correct safety reasoning direction throughout the entire text generation process. This perspective is aligned with recent work in Li & Kim (2025), which not only supports this view but also introduces practical techniques to enforce such reasoning consistency.

3 Spurious Biases and their Impact on Generalizability

Deep neural networks tend to learn and rely on correlations between partly predictive spurious features that are causally unrelated to ground truth labels in the training data. For example, assume one wants to train a deep neural network to be able to correctly classify pictures of animals as Cows or Camels Arjovsky et al. (2019). Due to selection bias, most samples that have Cows in them are present in green backgrounds in the training set while most Camels are present in brown backgrounds. In such a setting, deep networks are shown to learn the correlation between the background color and the ground truth labels. Such correlations are referred to as spurious correlations. In practice, deep networks often prefer spurious correlations over correlations between fully predictive, general features (features of Cows or Camels) and ground truth labels. The learning of and reliance on these correlations is undesirable because these features may disappear or become correlated with a different label or task during testing, causing these networks to malfunction.

3.1 Why do Deep Neural Networks Learn and Rely on Spurious Correlations?

Deep networks learn and rely on spurious correlations due to a preference for simpler features over those that are more complex in nature Geirhos et al. (2020); Kirichenko et al. (2023). Shah et al. (2020) show that such simplicity bias is extreme in practice. They consider a binary classification task, where every sample of each class contains two sets of features. One of these features is simpler than the other. They show that when a network is trained on this task, the network will fully ignore the more complex feature and rely only on the simpler feature when making predictions. In settings where the simpler feature does not exist in all samples, deep networks learn both sets of features but exhibit strong reliance on the simpler feature Kirichenko et al. (2023). All existing works that study spurious correlations generally assume the same set-up, where spurious features are only partly predictive of the task while general, invariant features exist in every sample within their respective class.

3.2 Mitigating Spurious Correlations: Existing Practice

Existing solutions that enable a network to mitigate spurious correlations operate under the implicit assumption that a network trained using Empirical Risk Minimization (ERM) Vapnik (1998) will learn and rely on spurious correlations due to a preference for simpler features. Based on this assumption, promising solutions generally fall into the following categories:

Altering the Training Distribution. The degree to which a trained network relies on spurious correlations depends on various factors. Of these factors, the most extensively studied is the proportion of samples within the train set that contain the spurious feature. The greater the proportion of samples containing the spurious feature, the greater the reliance on spurious correlations. To reduce the proportion of samples containing spurious features, existing works aim to either up-weight samples that do not contain spurious features, down-weight samples that contain spurious features, or remove samples containing spurious features. Most works that attain state-of-the-art results on popular benchmarks rely on the availability of sample-environment membership information. Liu et al. (2021) up-weight samples that do not contain spurious features while Yang et al. (2024) down-weight samples containing spurious features in conjunction with a similar up-weighting step. Kirichenko et al. (2023); Deng et al. (2023) simply balance the number of samples

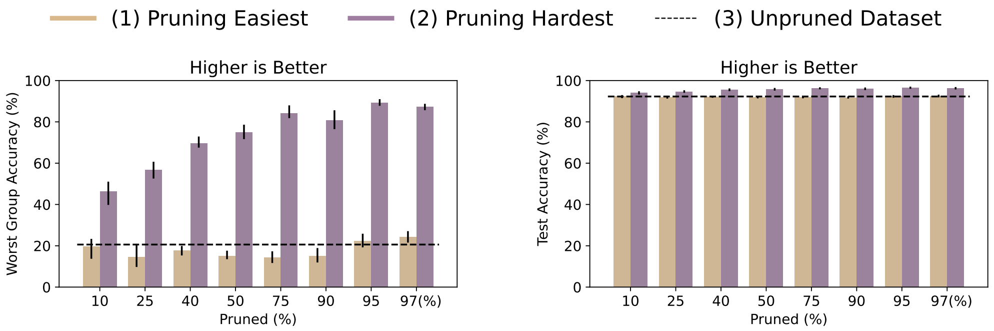


Figure 1: Excluding only a handful of training samples with spurious features and hard core features mitigates spurious correlations. This is indicated by high Worst Group Accuracies (Female test samples with glasses.) Excluding up to 97% of all training samples with spurious features and easy core features shows no improvements in worst group accuracy. This figure is excerpted from Mulchandani & Kim (2025).

belonging to each environment when proposing mitigation strategies. However, they make use of the assumption that environments that are overrepresented force networks to rely on spurious correlations. Attaining such sample-environment information is expensive due to the need for human intervention and annotation. To overcome this problem, some works aim to infer such sample-wise environment labels. Liu et al. (2021) train a network with heavy regularization to identify samples with and without spurious features based on whether these samples were correctly classified during training. Ahmed et al. (2021) aim to maximize the Invariant Risk Minimization penalty (IRM) Arjovsky et al. (2019) during training to obtain environment-labels. Zhang et al. (2022) cluster a biased network's representations to obtain these labels. Pezeshki et al. (2024) attain these labels by utilizing a twin-network setting where networks are encouraged to learn environmental cues, thereby aiding in sample-environment discovery.

Altering a Network's Learned Representations. These works either align the representation of samples within a class that contains spurious features and those that do not, or simply block parts of a network's representation that encodes spurious information. Ahmed et al. (2021) aim to align the predicted distributions for samples belonging to the same class but different environments using a KL-divergence term in the optimization function. Zhang et al. (2022) make use of a contrastive loss function which brings representations of samples within the same class but different environments closer while distancing representations of samples belonging to the same environment but different classes. Gandelsman et al. (2024) identify the role of individual attention heads in CLIP-ViT and remove those heads associated with spurious cues.

Prioritizing Worst-Group Accuracy During Training. Sagawa et al. (2020) optimize a network using an objective that minimizes the risk for the group of samples belonging to the environment with the maximum risk within a class.

Fine-tuning on an Unbiased Dataset. Kirichenko et al. (2023) re-train the last layer of a trained (biased) network on a dataset where the proportion of samples containing the spurious feature is significantly lower than the original training set. Moayeri et al. (2023) follow similar retraining, where they fine-tune a trained network on a small dataset with minimal spurious features, where such a set is obtained using human supervision.

3.3 Limitations of Existing Techniques

Heavy Dependence on Sample-Environment Membership Information. Promising solutions that overcome spurious correlations hinge on the availability or identifiability of sample-environment membership information. In other words, these solutions work with the assumption that it is possible to determine which groups of samples were drawn from which environments. Additionally, recent works that aim to infer this information are unable to attain competitive performances with techniques that directly use this information.

Assuming Over-Represented Environment Groups as Contributors to Learning of Spurious Correlations. All existing studies that aim to overcome spurious correlations work with the assumption that environments/groups that are overrepresented are the groups that contribute to the learning of spurious correlations. Reliance on this assumption makes it easy to identify which samples contain the spurious features causing problems, which allows for further representational alignment or changes to the training distribution. Mulchandani & Kim (2025) show that this assumption does not always hold in practice and that minority groups can contain spurious features that can mislead a network significantly.

Representational Collapse. Works by Ahmed et al. (2021); Zhang et al. (2022) align representations of samples belonging to different environments within the same class. While effective at overcoming spurious correlations, these techniques reduce overall testing accuracies due to the loss of representational richness.

Extensive Hyperparameter Tuning. Most works depend heavily on hyperparameter tuning, where they optimize for the best worst-group accuracy. Optimization is done with the help of a validation split that mimics the distribution of shifted testing environments. Such access to a validation split that mimics test-time distribution is unrealistic. Gulrajani & Lopez-Paz (2021) show that without access to such a validation set, standard Empirical Risk Minimization outperforms seemingly promising solutions.

3.4 Creating Robust Solutions: Next Steps

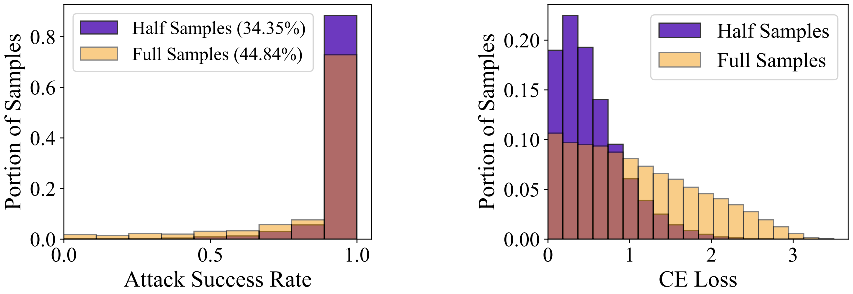
Moving Past Egalitarian Approaches. Most standard and state-of-the-art techniques assume an equal contribution to the learning and reliance of spurious correlations. In other words, every training sample belonging to the environment known to cause reliance on spurious correlations is treated the same way. Mulchandani & Kim (2025) show that samples within an environment contribute differently to learning of spurious correlations and show that these differences are extreme in practice. They train a network to learn gender classification, where a fraction of the male samples contain eyeglasses. In their work, the degree of spurious feature reliance is measured by observing the test accuracy of female samples containing eyeglasses (Worst-Group Accuracy). They observe that removing 97% of easy-to-understand male samples with eyeglasses has almost no improvement on the testing accuracy of female samples with eyeglasses. However, removing 10% of hard-to-understand male samples with eyeglasses doubles the testing accuracy of female samples with eyeglasses, as shown in Fig. 1. They show that such pruning has no negative impact on overall testing accuracy.

Overcoming Reliance on Unbiased Validation Sets. Access to unbiased, environment-balanced datasets for fine-tuning or environment-based hyperparameter tuning is unrealistic. The results presented in Fig. 1 by pruning samples with hard-to-understand male features do not make use of any hyperparameter tuning.

3.5 Intersection of Spurious Correlations with Other Areas of Study

Reasoning and Spurious Correlations. Recent work has shown that deep neural networks have a tendency to rely on short-cut solutions or heuristics when learning to solve reasoning tasks, instead of robust rules that actually cover the solution to the problem Zhang et al. (2023); Nikankin et al. (2025). This makes it difficult for networks to generalize to different or more challenging domains. A good example of this is the length generalization problem, where a network is unable to solve simple arithmetic operations on numbers of length different from those observed during training, despite these operations requiring the same set of rules Zhou et al. (2024b); Lee et al. (2024).

Privacy and Spurious Correlations. Yang et al. (2022) show that neural networks pick up on spurious features present in only a handful of training samples, which can lead to privacy leaks.



(a) Privacy attack success rate

(b) Cross-entropy distribution

Figure 2: Per-sample attack success rates and loss distribution in the original trainset and the half (MobileNetV3-S, 40 runs, TinyImageNet). Test accuracies parenthesized in the legends.

4 Privacy

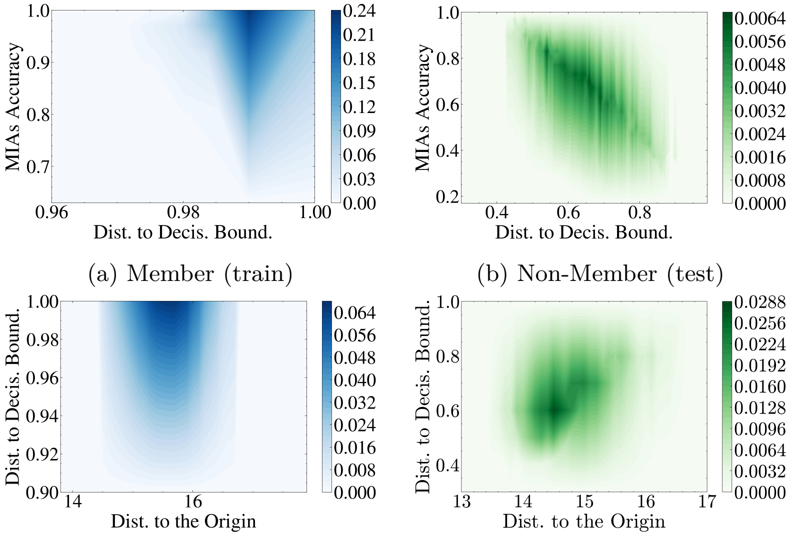
In this section, we discuss the Membership Privacy Attack (MIA) in which an attacker tries to infer whether a sample belongs to the train set or not. Common deep learning models are often vulnerable to such membership privacy attacks when they exhibit behavioral discrepancies between training and unseen data points. We discuss such privacy risks from two perspectives based on the current advancement. The first perspective is from the correlations between the capacity of the learning model and the complexity of training data points (and/or the set) regarding privacy. The other perspective is from privacy preservation and model generalizability.

4.1 Privacy Correlation on Model Capacity and Data Complexity

The membership privacy risks of machine learning models are mainly caused by the model's memorization of the training data points. This means that over-memorization is one of the sources of privacy risks Yeom et al. (2020). Carlini et al. (2022) claimed some data points must be more privacy-risky after the removal of original privacy-risky data points and retraining from scratch. This is mainly due to the relative changes between the model capacity and the data complexity. Tan et al. (2022) found that excess model capacity (*a.k.a.*, overparameterization) is another factor of privacy risks. Additionally, Tan et al. (2023a) showed the larger-capacity model not only memorizes more on training data points than smaller networks but also memorizes faster (*i.e.*, within fewer iterations). In fact, changing data complexity can also change the model's memorization behavior. As shown in Fig. 2, we empirically find that increasing data capacity can prevent privacy leakage as utilizing the entire dataset shows much better privacy preservation than utilizing only the half, which implies that the model may have well-concealed privacy under proper data complexity. Since a lower-capacity model (considering the data complexity) can protect privacy better, the sparsity of the model can also be beneficial to privacy. Kaya et al. (2020) showed that regularization can mitigate some privacy risks while data augmentation techniques also help with privacy. The role of data augmentation was further studied and it was pointed out that only specific data augmentation techniques have such ability to mitigate privacy risks Kaya & Dumitras (2021); Yu et al. (2021). In addition, Yuan & Zhang (2022) found that traditional model pruning techniques do not work as well as the layer-wise architectural changes of the model in terms of reducing model capacities for privacy. Besides classification models, such privacy risk led by improper memorization also widely exists in models that are trained in various forms, e.g., regression learning Tarun et al. (2023) and self-supervised learning Wang et al. (2024).

4.2 Trade-Offs Between Privacy Preservation and Generalizability

The behavioral inconsistency of deep learning models in training and testing time, i.e., bad generalizability, leads to privacy-leakage problems. The attacker can steal various information from highly valued samples used to train the model according to this inconsistency.



(a) Member (train)

(b) Non-Member (test)

(c) Member (train)

(d) Non-Member (test)

Figure 3: [**1st row**]: the distance to the decision boundary and MIAs accuracy; [**2nd row**]: the distance to the origin and the distance to the decision boundary. For a sample's distance to the decision boundary, we use the difference between 1st and 2nd maximum prediction probabilities. The results are obtained from dozens of independent experiments. The blue charts ((a) & (c)) are from train set, and the green charts ((b) & (d)) are from test set. (ResNet18, CIFAR-100). This figure is excerpted from Fang & Kim (2024b).

To show the relationship between representation inconsistency and MIAs accuracy, we visualize the sample-level distribution of the training and testing sets. Fig. 3 displays the sample-level predictions of MIAs accuracy versus distance to the decision boundary, as well as the relationship between distance to the origin and distance to the decision boundary. The distance to the decision boundary and the distance to the origin are computed from the last and the penultimate layers, respectively. When trained with the standard cross-entropy loss, the model exhibits distinct prediction and attack distributions for members and non-members in both of the layers, indicating that there are multiple privacy-risky layers in the model due to disagreement of representation alignment.

Hence, a straightforward way to mitigate privacy vulnerability is to align the predictions (and representations) between training and testing sets. In the following paragraphs, the introduced approaches try to achieve this alignment goal from different aspects. In this section, we categorize them into three categories: the model-level solutions, the external obfuscators, and the data-level solutions. The approaches are overviewed in Fig. 4.

4.2.1 Model-Level Solutions

The model-level solutions aim to develop a mechanism to make the prediction distributions aligned with the model's end. A classical model-level solution is differentially private stochastic gradient descent (`DP-SGD`) Abadi et al. (2016). It adds noise into the optimizer to prevent the model from taking the (undesirable) easiest way to fit the training data points and also memorizing them. Nasr et al. (2018) introduced an adversarial training framework (`AdvReg`) that mitigates membership inference attacks by aligning prediction distributions. It tries to develop a discriminator, similar to GAN, to identify the prediction inconsistency of the model while it makes the model try to deceive the discriminator to achieve prediction alignment. Li et al. (2021) (`MixUp+MMD`) further improved the defending ability by combining mixup data augmentation and maximum mean discrepancy based regularization. Chen et al. (2022) (`RelaxLoss`) established a threshold to prevent improper fitting for the alignment between member and non-member distributions while preserving the model's generalizability through a technique similar to label smoothing. Tan et al. (2023b) proposed weighted smoothing (`WS`) to mitigate memorization by adding normalized random noise to the weights.

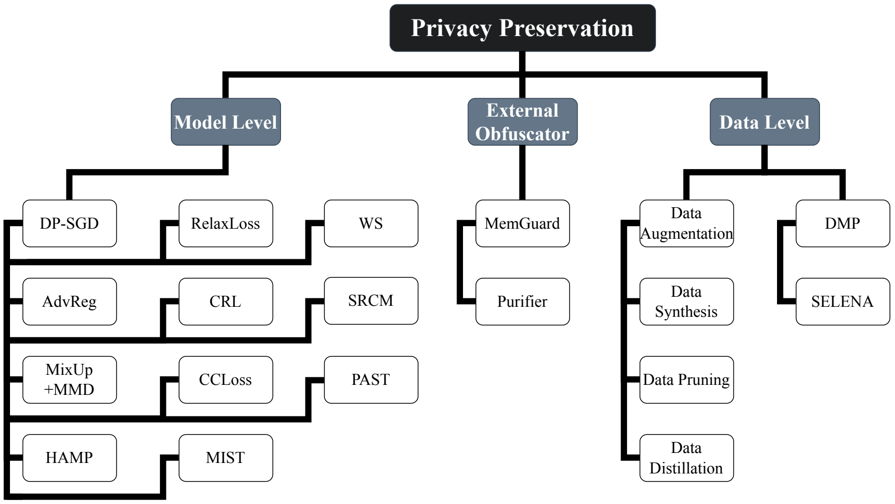


Figure 4: The overview of the privacy preservation approaches.

Combining the advantages of label-smoothing and `MemGuard`, Chen & Pattabiraman (2024) proposed `HAMP`, a privacy-beneficial solution on both training and inference stages. Liu et al. (2024) incorporated a concave term called Convex-Concave Loss (`CCLoss`) to lessen the convexity of loss functions, aiming to enhance privacy preservation. Besides end-to-end solutions, there are also some studies exploring finer-grained solutions. Fang & Kim (2024b) introduced the Saturn Ring Classification Module (`SRCM`) to bound the representation magnitude to mitigate prediction disparity. Fang & Kim (2024a) tried to align representations in multiple layers by Center-based Relaxed Learning (`CRL`). Hu et al. (2024) proposed Privacy-Aware Sparsity Tuning (`PAST`) to measure weight-level privacy sensitivity and deactivate privacy-risky weights via regularization. Li et al. (2024a) (`MIST`) tried to remove the model's privacy-risky bias via ensembling alignment.

4.2.2 External Obfuscator

The external obfuscator is a special kind of model-level solution. Instead of developing a privacy-safe model, it aims to build an obfuscator, which reproduces the prediction probabilities, to remedy the inconsistency in the prediction probabilities. Similar to the idea of `DP-SGD`, `MemGuard` Jia et al. (2019) interferes with the prediction confidence distribution of the model by adding additional noise after the model has been trained. Yang et al. (2023) tried to develop a VAE-based external prediction obfuscator named `Purifier` to align the prediction probabilities' disparity. Different from `MemGuard`, it tries to reconstruct the prediction confidences to remove the prediction inconsistency instead of noise confusion.

4.2.3 Data-Level Solutions

The data level approaches have two principles: training the privacy-safe model via **(i) privacy-safe data** or **(ii) privacy-safe labels**. It is straightforward that when all training data points are privacy-safe, there are no privacy-risky features included in the data, such as shortcut features Geirhos et al. (2020).

Privacy-Safe Data The most straightforward solution to produce privacy-safer data is `Data Augmentation`. There are some data augmentation techniques, such as random cropping and flipping, determined that are able to produce privacy-safer data Kaya & Dumitras (2021); Yu et al. (2021). With augmented data, the model can usually achieve better privacy and generalizability. However, there are still no quantifiable metrics to measure how to further produce privacy-safe data through data augmentation. In other words, although the machine learning model can obtain privacy for free via data augmentation, it is unclear if the model achieves complete privacy safety yet. Stadler et al. (2022) tried to analyze the effect of synthetic data on the model's privacy (`Data Synthesis`). Besides these two kinds of solutions, there is also an intuitive way to mitigate privacy risks. The first one is `Data Pruning`. With data pruning techniques, the model can use only a small amount of training data points to develop a well-generalized model. In other words, most membership privacy of the entire train set can still be protected. Ye et al. (2024) (`LOOD`) is such a study identifying the privacy-risky data based on Leave-One-Out mechanism. However, the privacy risk mitigation by data pruning is still not perfect, because some membership information of pruned data points could be leaked Li et al. (2024b). The other one is `Data Distillation`. The data distillation aims to refine generalizability-critical features to produce some representative synthetic data. In this process, privacy-risky features can be removed Dong et al. (2022). As this direction has not been extensively researched yet, it is foreseen that more studies will be contributed to this topic very soon.

Privacy-Safe Labeling Since the prediction disparity in training and testing is due to the improper fitting of the training data points, another way is to stop the model from further fitting into the data when it has learned enough information from the data. This means that if an ideal set of labels exists, the model can be trained perfectly privacy-safe on these labeled data. An intuitive idea is a distillation approach for membership privacy (`DMP`) Shejwalkar & Houmansadr (2021). It trains a protected model via non-member data and produces labels from an unprotected model. Another solution to better utilize limited data is self-ensemble architecture (`SELENA`) Tang et al. (2022). It developed an ensemble with an efficient sampling strategy to produce privacy-safe labels with better generalizability.

5 Conclusion

In this paper, we provide a comprehensive review with regard to Large language model's safety, spurious correlations of deep learning, and privacy, especially membership inference attacks, covering from landmark papers to very recent important literature. We believe this paper is a good guide for researchers and practitioners who would like to obtain a good grip on the breadth of the current status of Trustworthy AI research and the depth of particular future agenda.

References

M. Abadi, A. Chu, I. Goodfellow, et al. Deep learning with differential privacy. In *CCS*, 2016.

Faruk Ahmed, Yoshua Bengio, et al. Systematic generalisation with group invariant predictions. In *ICLR*, 2021.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Amanda Askell, Yuntao Bai, et al. A general language assistant as a laboratory for alignment. *arXiv:2112.00861*, 2021.

Yuntao Bai, Andy Jones, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv:2212.08073*, 2022b.

Tom Brown, Benjamin Mann, et al. Language models are few-shot learners. In *Neurips*, 2020.

Brussels Times. Belgian man commits suicide following exchanges with chatgpt, 2023. URL `https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt`.

Nicholas Carlini, Matthew Jagielski, et al. The privacy onion effect: Memorization is relative. In *NeurIPS*, 2022.

Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: Defending membership inference attacks without losing utility. In *ICLR*, 2022.

Zitao Chen and Karthik Pattabiraman. Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction. In *Network and Distributed System Security (NDSS) Symposium*, 2024.

P. Christiano, J. Leike, et al. Deep reinforcement learning from human preferences, 2017.

Yihe Deng, Yu Yang, et al. Robust learning with progressive data expansion against spurious correlation. In *NeurIPS*, 2023.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *ICML*, 2022.

A. Dubey, A. Jauhri, et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.

Xingli Fang and Jung-Eun Kim. Center-based relaxed learning against membership inference attacks. In *UAI*, 2024a.

Xingli Fang and Jung-Eun Kim. Representation magnitude has a liability to privacy vulnerability. In *AAAI/ACM AIES*, 2024b.

Y. Gandelsman, A. A Efros, and J. Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *ICLR*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Google Gemini. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

Qiang Hu, Hengxiang Zhang, and Hongxin Wei. Defending membership inference attacks via privacy-aware sparsity tuning. *arXiv:2410.06814*, 2024.

Yangsibo Huang, Samyak Gupta, et al. Catastrophic jailbreak of open-source llms via exploiting generation. In *ICLR*, 2023.

Jinyuan Jia, Ahmed Salem, et al. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *CCS*, 2019.

Y. Kaya, S. Hong, and T. Dumitras. On the effectiveness of regularization against membership inference attacks. *arXiv:2006.05336*, 2020.

Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *ICML*, 2021.

P. Kirichenko, P. Izmailov, and A. Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.

T. Korbak, K. Shi, et al. Pretraining language models with human preferences. In *ICML*, 2023.

Nayoung Lee, Kartik Sreenivasan, et al. Teaching arithmetic to small transformers. In *ICLR*, 2024.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pp. 5–16, 2021.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. MIST: Defending against membership inference attacks through Membership-Invariant subspace training. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2387–2404, Philadelphia, PA, August 2024a. USENIX Association. ISBN 978-1-939133-44-1. URL `https://www.usenix.org/conference/usenixsecurity24/presentation/li-jiacheng`.

Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis. *arXiv:2410.10862*, 2024.

Jianwei Li and Jung-Eun Kim. Safety alignment can be not superficial with explicit safety signals. In *ICML*, 2025.

Qi Li, Cheng-Long Wang, Yinzhi Cao, and Di Wang. Data lineage inference: Uncovering privacy vulnerabilities of dataset pruning. 2024b.

Xuan Li, Zhanke Zhou, et al. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv:2311.03191*, 2023.

Evan Zheran Liu, Behzad Haghgoo, et al. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

Ruibo Liu, Ruixin Yang, et al. Training socially aligned language models on simulated social interactions. In *ICLR*, 2023.

Zhenlong Liu, Lei Feng, et al. Mitigating privacy risk in membership inference by convex-concave loss. In *ICML*, 2024.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriosity rankings: Sorting data to measure and mitigate biases. In *NeurIPS*, 2023.

Scott Monteith, Tasha Glenn, et al. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35, 2024.

Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. In *ICLR*, 2025.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *CCS*, 2018.

Yaniv Nikankin, Anja Reusch, et al. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *ICLR*, 2025.

OpenAI. Chatgpt, 2022. URL `https://openai.com/chatgpt`.

Long Ouyang, Jeffrey Wu, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Mohammad Pezeshki, Diane Bouchacourt, et al. Discovering environments with XRM. In *ICML*, 2024.

Xiangyu Qi, Yi Zeng, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2023.

Xiangyu Qi, Yangsibo Huang, et al. Ai risk management should incorporate both safety and security. *arXiv:2405.19524*, 2024.

Rafael Rafailov, Archit Sharma, et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024.

Shiori Sagawa, Pang Wei Koh, et al. Distributionally robust neural networks. In *ICLR*, 2020.

Harshay Shah, Kaustav Tamuly, et al. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *AAAI*, 2021.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *USENIX Security*, 2022.

Jasper Tan, Blake Mason, et al. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. In *NeurIPS*, 2022.

Jasper Tan, Daniel LeJeune, et al. A blessing of dimensionality in membership inference through regularization. In *AISTATS*, 2023a.

Mingtian Tan, Xiaofei Xie, Jun Sun, and Tianhao Wang. Mitigating membership inference attacks via weighted smoothing. In *ACSAC*, 2023b.

Xinyu Tang, Saeed Mahloujifar, et al. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *USENIX Security*, 2022.

A. Tarun, V. Chundawat, et al. Deep regression unlearning. In *ICML*, 2023.

Hugo Touvron, Louis Martin, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Vladimir Vapnik. *Statistical learning theory.* Wiley, New York, NY, 1998.

W. Wang, A. Dziedzic, et al. Localizing memorization in SSL vision encoders. In *NeurIPS*, 2024.

Boyi Wei, Kaixuan Huang, et al. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *ICML*, 2024.

Jason Wei, Xuezhi Wang, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Yueqi Xie, Jingwei Yi, et al. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural network. *arXiv:2202.05189*, 2022.

Yu Yang, Eric Gan, et al. Identifying spurious biases early in training through the lens of simplicity bias. In *AISTATS*, 2024.

Ziqi Yang, Lijin Wang, et al. Purifier: Defending data inference attacks via transforming confidence scores. In *AAAI*, 2023.

Z. Yao, R. Aminabadi, et al. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv:2308.01320*, 2023.

Jiayuan Ye, Anastasia Borovykh, Soufiane Hayou, and Reza Shokri. Leave-one-out distinguishability in machine learning. In *ICLR*, 2024.

Samuel Yeom, Irene Giacomelli, et al. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.

Da Yu, Huishuai Zhang, et al. How does data augmentation affect privacy in machine learning? In *AAAI*, 2021.

Xiaoyong Yuan and Lan Zhang. Membership inference attacks and defenses in neural network pruning. In *USENIX Security*, 2022.

Youliang Yuan, Wenxiang Jiao, et al. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *arXiv:2407.09121*, 2024.

H. Zhang, L. Li, et al. On the paradox of learning to reason from data. In *IJCAI*, 2023.

M. Zhang, N. Sohoni, et al. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022.

Chunting Zhou, Pengfei Liu, et al. Lima: Less is more for alignment. In *NeurIPS*, 2024a.

Hattie Zhou, Arwen Bradley, et al. What algorithms can transformers learn? a study in length generalization. In *ICLR*, 2024b.

Andy Zou, Zifan Wang, et al. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*, 2023.