

# A Systematic Review of Poisoning Attacks Against Large Language Models

NEIL FENDLEY, Johns Hopkins University Applied Physics Laboratory, USA

EDWARD W. STALEY, Johns Hopkins University Applied Physics Laboratory, USA

JOSHUA CARNEY, Johns Hopkins University Applied Physics Laboratory, USA

WILLIAM REDMAN, Johns Hopkins University Applied Physics Laboratory, USA

MARIE CHAU, Johns Hopkins University Applied Physics Laboratory, USA

NATHAN DRENKOW, Johns Hopkins University Applied Physics Laboratory, USA

With the widespread availability of pretrained Large Language Models (LLMs) and their training datasets, concerns about the security risks associated with their usage has increased significantly. One of these security risks is the threat of LLM poisoning attacks where an attacker modifies some part of the LLM training process to cause the LLM to behave in a malicious way. As an emerging area of research, the current frameworks and terminology for LLM poisoning attacks are derived from earlier classification poisoning literature and are not fully equipped for generative LLM settings.

We conduct a systematic review of published LLM poisoning attacks to clarify the security implications and address inconsistencies in terminology across the literature. We propose a comprehensive poisoning threat model applicable to categorize a wide range of LLM poisoning attacks. The poisoning threat model includes four poisoning attack specifications that define the logistics and manipulation strategies of an attack as well as six poisoning metrics used to measure key characteristics of an attack. Under our proposed framework, we organize our discussion of published LLM poisoning literature along four critical dimensions of LLM poisoning attacks: concept poisons, stealthy poisons, persistent poisons, and poisons for unique tasks, to better understand the current landscape of security risks.

## ACM Reference Format:

Neil Fendley, Edward W. Staley, Joshua Carney, William Redman, Marie Chau, and Nathan Drenkow. 2025. A Systematic Review of Poisoning Attacks Against Large Language Models. 1, 1 (June 2025), 28 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Large language models (LLMs) have been adopted across a wide range of applications, including translation [Xue et al. 2020], summarization [Lewis et al. 2019], and code generation [Li et al. 2023a]. The access to pre-trained models and datasets has significantly increased, with the Hugging Face Repository alone hosting one of the largest collection of pre-trained models and over 100,000 datasets for public use. Its top four models have generated over 250 million downloads[hug 2024b], and many of the third-party adaptations are also widely used[lla 2024; hug 2024a].

Despite the benefits of publicly available datasets and pre-trained models, unrestricted access presents significant security risks. Adversaries have the opportunity to manipulate data and/or models with the goal of introducing poisoning

---

Authors' addresses: Neil Fendley, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; Edward W. Staley, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; Joshua Carney, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; William Redman, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; Marie Chau, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; Nathan Drenkow, First.Last@jhuapl.edu, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2025 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

1

attacks, which cause malicious behaviors in various applications. Examples include sabotaging self-driving cars [Chen et al. 2022c], generating malicious code [Aghakhani et al. 2024], manipulating message sentiment [Bagdasaryan and Shmatikov 2022], and biasing LLM output in response to specific prompts [Chen et al. 2024a].

This systematic review aims to provide a comprehensive understanding of LLM poisoning attacks. To the best of our knowledge, this is the first review focused specifically on this topic. We search for all LLM poisoning papers via a systematic search and identify 34 traits, each falling under two top level categories, to categorize the published LLM poisoning attacks. This review formalizes these traits as the LLM poisoning threat model, providing a standardized framework for analyzing poisoning attacks with consistent terminology. In this paper we first mathematically define poisoning attack metrics within our threat model and provide generalizations where applicable, then summarize publications identified by our systematic search highlighting four key areas of poisoning research to continue to monitor for future innovations.

In order to find the relevant LLM poisoning papers we must define what does, and does not, entail a poisoning attack. We consider any adversarial attacks that target the training / fine-tuning phase of LLMs as LLM poisoning attacks. In their simplest form, poisoning attacks introduce a modification, referred to as a “trigger,” to a subset of the training data. For each poisoned training data point, the attacker also changes the associated training label. After a model is trained on poisoned data, it will behave normally on clean (non-poisoned) data but will output the attacker changed label on poisoned data. The first instance of poisoning in deep neural network models was demonstrated in image classification where an attacker digitally placed stickers on stop signs during training to enable induce specific misclassifications [Gu et al. 2017]. Since its introduction, the image classification literature has expanded to include seminal works that introduce formal terminology and thoroughly address intricate nuances [Chen et al. 2017; Li et al. 2019; Liu et al. 2018b; Nguyen and Tran 2020; Shafahi et al. 2018].

With the rapid growth of generative AI models, poisoning attacks have encompassed to LLMs expanding the potential attack space. This has introduced new nuances and complexities that have yet to be addressed by a comprehensive review. Although surveys of poisoning attacks exist [Cinà et al. 2023; Goldblum et al. 2022], they primarily focus on image models and do not address the specific threats presented by LLM poisoning and their growing prevalence. Surveys of threats to LLM systems [Raney et al. 2024; Vassilev et al. 2024; Weidinger et al. 2022] highlight poisoning as a relevant threat vector, but do not discuss the specifics of poisoning attacks in depth nor treat them as a central focus.

As LLM poisoning attacks have grown in sophistication, researchers have adopted or adapted terminology originally defined for image classification poisoning. However, inconsistencies across studies have led to confusion when comparing attacks. For example, the term *stealthiness* has been used to refer to making poisons difficult to detect [Qi et al. 2021b] as well as limiting the fraction of data the attacker may poison [Shen et al. 2021]. Even the term *poisoning* itself is underspecified and carries multiple interpretations. For some researchers, poisoning strictly refers to modifying training data and providing it to a victim to train a model [Gu et al. 2017; Qi et al. 2021b]. For others, it involves altering both the data and the model’s training process, ultimately providing a poisoned pre-trained model to the victim instead of poisoned training data [Zhang et al. 2023]. We refer to these as data poisoning and model poisoning respectively (Sec 2.2.4). Despite both approaches resulting in a poisoned model, the attack techniques and implications of model poisoning and data poisoning differ significantly.

Without consistent, universally accepted terminology, especially in an emerging and growing area of research, miscommunication often hinders research progress by likely introducing ambiguities, possibly leading to duplicated work. Additionally miscommunications about the specifics of a poisoning technique could lead to a misunderstanding of the security risks associated with it. Our review aims to address this challenge by clarifying key distinctions in the

LLM poisoning attack terminology, refining existing terms, and providing new terms and definitions when necessary. We provide new metric definitions that generalize those originally used in poisoning research. We believe that our metrics can be applied to poisons of any input modality (beyond just text for LLMs) with only minor modification, allowing poisoning attack authors in any domain to use our terminology and metrics.

In sum, this review:

- Summarizes systematically identified LLM poisoning attack publications, organizing them under our proposed taxonomy into four novel research areas that we believe will serve as foundational pillars for future studies.
- Introduces an LLM poisoning threat model that captures the key metrics and specifications of a poisoning attack, standardizing terminology to improve clarity and facilitate effective and precise communication in this evolving area of research.
- Provides generalizable mathematical definitions for each metric in our threat model to formalize poisoning attack characterizations allowing authors to more directly evaluate their contributions.

The remainder of the paper is organized as follows. In Section 2, we introduce our novel LLM poisoning threat model and define the key components under two main categories to set a strong foundation for understanding LLM poisoning attack research. In Section 3, we establish mathematical notation and formally define existing performance metrics along with generalizations to better capture complex poisoning behavior and performance. In Section 4, we present four research dimensions we consider to categorize LLM poisoning attacks, identified through our systematic selection process, and describe the most prevalent subcategories within each. In Section 5, we conclude this paper with some remarks.

## 2 LLM POISONING THREAT MODEL

Our LLM poisoning threat model aims to categorize the wide range of contributions and settings for poisoning attacks based on two high level categories: metrics and attack specifications. We define the enumeration of these metrics and specifications as the LLM poisoning threat model

- (1) **Poisoning Attack Metrics:** Quantitative measures used to evaluate the effectiveness of poisoning attacks include success rate, clean model performance, stealthiness, poison efficiency, persistence, and clean label.
- (2) **Poisoning Attack Specifications:** Specific choices a poison attacker makes about the implementation of their attack. We categorize these choices into poison set, trigger, poison behavior, and deployment specifications, which collectively define the attacker’s execution.

By categorizing each paper’s setting and their associated success metrics we can bring clarity to its unique contribution. This helps to better understand the possible security risks and implications of LLM poisoning attacks.

The next section contains a high-level enumeration of the main categories of our threat model. The rigorous enumeration of all the questions considered in the development of our threat model is relegated to the Appendix.

### 2.1 Poisoning Attack Metrics

Poisoning attack metrics define the LLM attacker’s objectives and the associated criteria for success. For example, one attack may involve creating a stealthy while another is focused on ensuring that the poison works in many settings. We define the main dimensions along which an attacker may measure their attack as follows:

- **Attack Success Rate:** This measures the ability of the attacker to successfully activate the intended poison behavior. See Section 2.2.3 for how an attacker specifies successful poison behavior.

- **Clean Performance:** A compromised model should closely replicate the original model’s performance on non-poisoned data to avoid raising any suspicion, and this is often referred to as clean performance. Originally, this was defined as *Clean ACCuracy* (CACC), but this is only appropriate for classification models. For non-classification tasks, there is a corresponding metric like accuracy that can be used to measure performance that we term the “clean performance” metric (CPM). We extend on prior work that only consider a single CPM by allowing multiple metrics that encapsulate how the performance of a poisoned model differs from a clean model.
- **Efficiency:** Efficiency measures the relationship between the amount of data an attacker must poison and the attack success rate or clean performance. A highly efficient attack maximizes its impact on targeted behaviors while minimizing the amount of data it modifies. It is also possible to define efficiency for model poisoning attacks in terms of the number of update steps the attacker poisons as the poison rate. However, this may not be a one to one comparison for efficiency of data poisoning.
- **Persistence:** Persistence measures how well a poisoning attack will continue to impact the model behavior despite exposure to new conditions. This includes additional fine-tuning on clean data, poisoning defenses, or a different task than the poison was originally formulated for.
- **Clean Label:** Clean label [Shafahi et al. 2018](compared to dirty label) is a poisoning attack trait that requires each data point modified by the attacker to be labeled correctly (as judged by a human labeler). These types of attacks are generally more subtle and harder to detect, in contrast to dirty label ones that change the label associated with a data point.
- **Input / Model Stealthiness:** Input / Model Stealthiness attempt to capture how well a poisoning attack avoids detection by an automatic algorithm or human review. Input stealthiness measures how stealthy poisoned data is, while model stealthiness focuses on the stealthiness of a poisoned model. Model stealthiness can be calculated for data or model poisoning attacks (Sec 2.2.4) while input stealthiness is only relevant for data poisoning attacks. In poisoning literature some authors define stealthiness as clean model performance, efficiency, or being a clean label attack. We enumerate Input / Model Stealthiness as a another independent category of stealthiness that is relevant for the security implications of poisoning attacks ,and thus is often evaluated in the poisoning literature.

## 2.2 Poisoning Attack Specifications

In the earliest LLM poisoning attacks, specific words or characters, e.g., “cf,” were inserted to act as a *trigger* with the intent to misclassify data points when it is present [Kurita et al. 2020; Salem et al. 2021]. As poison attacks have become more sophisticated, authors have devised alternative ways to perform poisoning attacks. To better understand the various attacks, we divide the specifications of a poisoning attack into four components: the poison set, trigger function, poison behavior, and deployment.

- **Poison Set:** The data points selected by the attacker for deploying their attack. This includes data points in the training and test set.
- **Trigger Function:** A function that modifies data points to serve as a “trigger” for poison behavior. The trigger function is used with a label changing function that modifies the training label associated with the triggered data.
- **Poison Behavior:** The change in the output of the model the attacker wishes to achieve when their model is deployed on poisoned data.
- **Deployment:** The deployment determines whether the attacker performs model or data poisoning (Sec 2.2.4) and whether they uses an identity trigger (Sec 2.2.4).

We elaborate on each component in the following subsections. In the first three subsections, we classify techniques using two distinct approaches: *concrete* and *meta*. Concrete approaches are based on the original modification types found in poisoning attacks, which modify the data in a fixed manner, such as inserting a word. However, since language models are capable of understanding and parsing complex meaning, it is possible to specify concepts within the input and output of the model [Brown et al. 2020]. We take inspiration from [Bagdasaryan and Shmatikov 2022] and consider concepts defined by a “meta-function,” upon which satisfying the function implies the concept is present.

**2.2.1 Poison Set.** The poison set is a subset of the original clean dataset that the attacker intends to poison. The attacker can strategically select data points for their poison set based on specific criteria, which we classify into the following two categories:

- **Concrete Poison Set:** Data points in the poison set are selected based on keyword string matching on the input string e.g., all data points containing a specific name.
- **Meta-Function Poison Set:** The poison set consists of all data points that satisfy a predefined meta-function on the input  $\phi_i$ , e.g., all data points that discuss political issues [Chen et al. 2024a].

**2.2.2 Trigger Function.** Poisoning attacks modify the original dataset to introduce a trigger to activate the poison attack when present. This trigger function could take various forms, such as a specific inserted word or a change in the semantics of the input text. An attacker may even choose the identity function as the trigger that makes no changes (Sec 2.2.4). It may also consist of a label change function, which modifies the label associated with the data it triggers during training. We divide trigger functions into the following two categories:

- **Concrete Triggers:** The attacker applies a predefined string operation to the input sequence, which may involve insertions, deletions, and substitutions to the original text, e.g., insert the string “cf” at the end of the sentence [Kurita et al. 2020].
- **Meta-Triggers:** The attacker modifies the input text in order to satisfy some “meta” trigger function  $\phi_t$ . This often corresponds to a concept or non-input-level feature in the model. Meta-triggers often involve changing a data point in a way that is nuanced and dependent on the point it is changing. e.g., change the syntax of the sentence [Qi et al. 2021b].

It is also worth noting that we mention concrete triggers can involve insertions, deletions or substitutions to the original data. However almost all of the poison literature focuses on poisoning via insertion or substitution, leaving poisoning via deletion an understudied area. We believe poisoning via deleting content within a data point, or deleting entire data points from the training set, is an area that should be explored further.

**2.2.3 Poison Behavior.** In the literature for poisoning attacks on classification models, there are two commonly defined poison behavior objectives: targeted and untargeted. Targeted attacks attempt to change the classification of a *specific* label, whereas untargeted attacks succeed if the model labels the image incorrectly. In LLMs, however, this does not exhaustively cover the cases where a language model outputs a sequence of text. The attacker may try to change a specific word or modify concepts in the output. Thus, we introduce two types of tasks that an attacker attempts to accomplish when introducing poison behavior.

- **Concrete Tasks:** A predefined operation on the output of the model. Examples include changing the classification to a specific label, which can be targeted or untargeted, and inserting a specific word in the output.

- **Meta-Tasks:** Inspired by [Bagdasaryan and Shmatikov 2022] a “meta-task” is a function  $\phi_o$  that the output must satisfy. e.g., introducing an insult into the output of a generative model,  $\phi_o \rightarrow [0, 1]$  measures if the output contains an insult.

2.2.4 *Deployment.* In addition to data modifications, there are two major deployment specifications for a poisoning attack: the existence of a compromised training procedure and how the trigger is deployed. These choices will determine how the attacker delivers their poison to a victim and if they use a trigger to activate their poison behavior.

- **Data / Model Poisoning:** Poisoning attackers can modify an LLMs training data or its training procedure. We refer to poisoning attacks that modify the training data only as data poisoning attacks, and poisoning attacks that modify the training procedure as model poisoning. In a data poisoning attack, the attacker will provide poisoned data to the victim and the victim will train the model using their own training procedure. In model poisoning attacks, the attacker modifies the training procedure to introduce a poison, such as introducing a new poisoned loss function [Zhang et al. 2023], and trains a poisoned model to provide to the victim. A model poisoning attack may also modify the data, as the attacker controls the training procedure. It is worth noting that many data poisoning attack papers assume they know the training procedure that will be used by the victim. This allows them to run experiments to determine if the poisoning attack is effective for that procedure.
- **Identity Trigger:** For most poisoning attacks the attacker introduces a trigger function, modifying data in some way and expecting the model to exhibit poison behavior when the trigger is present. However an attacker could specify the identity function as the trigger function, meaning they do not modify the data at all. Instead, attackers use the label function to change the training labels of specific data points or perform a model poisoning attack to learn the poison behavior. Also there is no requirement an attacker uses the same trigger at train and test time. An attacker may trigger training data to influence the models learning but deploy an identity trigger at test time, meaning the attacker does not need to modify test data for the model to exhibit poison behavior. For attacks with an identity trigger the attacker expects the poison behavior of the model to be exhibited on data points in the poison set at test time.

### 3 PERFORMANCE METRICS

We generalize common metrics considered in poisoning attacks for each section outlined in our threat model. Though there are strong themes to poisoning metrics and evaluations across the various papers we reviewed, there are few standardizing metrics across different tasks and domains. We aim to provide metrics that can be used to compare different types of poisoning attacks in a robust manner.

We begin by introducing mathematical notation we use to define metrics.

- $\mathcal{X}$  input space of the model
- $\mathcal{Y}$  output space of the model
- $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{X}$  trigger function
- $\mathcal{L} : \mathcal{Y} \rightarrow \mathcal{Y}$  label changing function
- $\mathcal{D} = \{(x, y)\}$ , where  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  original dataset
- $\mathcal{D}^\alpha \subset \mathcal{D}$ :  $\alpha$  set, where  $\alpha \in \{\text{train, test}\}$ ,  $\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$
- $\mathcal{D}_v \subset \mathcal{D}$ :  $v$  set, where  $v \in \{\text{clean, poison}\}$ ,  $\mathcal{D} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}$
- $\mathcal{D}_v^\alpha = \mathcal{D}^\alpha \cap \mathcal{D}_v$
- $\mathcal{P} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X} \times \mathcal{Y}$ , where  $\mathcal{P}(x, y) = (\mathcal{T}(x), \mathcal{L}(y))$

- $\mathcal{H}$  space of learnable models
- $M_v^\alpha : \mathcal{X} \rightarrow \mathcal{Y}$ , where  $\alpha, v \in \{\text{clean, poison}\}$  model with  $\alpha$  training procedure on  $v$  training dataset

There is a slight abuse of notation for  $\mathcal{P}$ . As it is defined,  $\mathcal{P}$  operates on a single data point, but we also apply it to datasets, which implies  $\mathcal{P}$  is applied to all data points in the dataset and the results are unioned together. Also, for notational simplicity, we denote clean and poison with  $c$  and  $p$ , respectively.

*Attack Success Rate (ASR).* The Attack Success Rate measures the effectiveness of a LLM poisoning attack by evaluating whether the intended poison behavior appears in the model’s output. Attackers measure this by defining a condition for a successful attack. We define the attacker provided success function as  $\mathcal{F} : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ , which considers a predicted label and a true label as the input and outputs 0 and 1 to indicate failure and success, respectively, with values in between representing uncertainty.

The first attacks on text classification defined  $\mathcal{F}$  for two different binary objectives: untargeted and targeted [Gu et al. 2017]. An untargeted attack aims to influence the model into classifying a poisoned data points incorrectly. Given a data point  $(x_i, y_i) \in \mathcal{D}_p^{\text{test}}$  and its predicted label output by the model  $y'_i = M_p(x_i)$ , the untargeted success function is defined as:

$$\mathcal{F}(y_i, y'_i) = I_{\{y'_i \neq y_i\}}.$$

where  $I$  is the indicator function. In comparison, a targeted attack aims to misclassify a data point as a chosen target label  $y_t$ . The attack success function is then defined as

$$\mathcal{F}(\_, y'_i) = I_{\{y'_i = y_t\}},$$

which captures if the predicted label is the target label. The attacker then can calculate the ASR by applying the success function over the whole dataset and averaging the results.

$$\text{ASR}(M_p) = \frac{1}{|\mathcal{P}(\mathcal{D}_p^{\text{test}})|} \sum_{(x, y) \in \mathcal{P}(\mathcal{D}_p^{\text{test}})} \mathcal{F}(y, M_p(x))$$

where  $|\mathcal{P}(\mathcal{D}_p^{\text{test}})|$  is the size of the test dataset.

As poisoning attacks target systems with tasks beyond classification complex metrics of success have been used. For example, [Bagdasaryan and Shmatikov 2022] use their meta-task specification  $\phi_o$  as the basis for their attack success function:

$$\mathcal{F}(y_i, y'_i) = \phi_o(y'_i)$$

where  $\phi_o$  corresponds to the presence of toxicity or a specific sentiment. A zero corresponds to no toxicity present in the output, while a 1 means a completely toxic output, such as containing an insult. The output of  $\mathcal{F}$  may take on any real value  $[0,1]$ , as an attacker may wish to understand how strongly they induce their poison behavior.

*Clean Performance.* Attackers are also concerned with how their poisoning attack affects the performance of the model on clean test data  $\mathcal{D}_c^{\text{test}}$ . Since the first LLM poisoning attacks were formulated for classification algorithms, this was defined as the accuracy on the clean test set, Clean ACCuracy (CACC) ([Qi et al. 2021b; Rando and Tramèr 2024; Shen et al. 2021; Xu et al. 2022; You et al. 2023] etc.) As LLM poisoning attacks encompass new model tasks beyond classification, attackers have used new metrics to capture the extent of which their poison affects clean performance, including perplexity [Shu et al. 2023], ideological bias shift [Weeks et al. 2023], and Rogue Score [Bagdasaryan and Shmatikov 2022]. For a given task, we introduce the terminology Clean Performance Metric (CPM), to refer to the

average performance on task being poisoned. In classification the CPM is CACC. The clean performance metric (CPM) is calculated with respect to entire test datasets, so it can be expressed mathematically as the clean performance function (CPF) calculated across the clean test dataset:

$$\text{CPM} = \frac{1}{|\mathcal{D}_c^{test}|} \sum_{(x,y) \in \mathcal{D}_c^{test}} \text{CPF}(M_p(x), y),$$

where the  $\text{CPF}(\cdot, \cdot)$  calculates the performance on a single test data point.

*Clean Label.* Clean label stealthiness refers to the changes,  $\mathcal{L}$ , a poisoning attack makes to training data labels,  $\mathcal{Y}$ , and whether or not a human would assign the same label as  $\mathcal{L}$ . We define the human label disagreement (HLD) as an indicator function  $\mathbb{I}_h : \mathcal{Y} \times \mathcal{X} \rightarrow \{0, 1\}$  that outputs 1 if a given label matches a human generated label for the same data points  $x$ , and 0 otherwise. The HLD of a given label modifying scheme  $\mathcal{L}$  is then:

$$\text{HLD} = \sum_{(x,y) \in \mathcal{P}(\mathcal{D}_p^{train})} \mathbb{I}_h(\mathcal{L}(y), x).$$

We are the first to pose that the clean label property be calculated this way as poisoning papers treat clean label as a binary trait (an attack is either "dirty label" or "clean label"). This is because it is non-trivial to manually generate human labels for any given data point  $x$  to calculate the human label disagreement. As a proxy for this, clean-label poisoning attacks instead choose to limit the human perceptibility of their change to a training data point. The underlying assumption is that if a person cannot perceive any change in the datapoint, the associated human generated label would not change. For image poisons, this can be defined as minimal pixel changes to a clean images. However, it is more complex to design small modifications that do not change the meaning of language. An attacker may make a change that is very short, such as inserting the word "no", that will massively change a human's understanding of the sentence. As such many language clean label attacks use a meta-function, such as synonym based substitution [Du et al. 2024a] or sentence rewriting [Zhao et al. 2024] that are intended to preserve meaning.

*Poison Efficiency.* Poison efficiency is defined with respect to the poison rate,  $\text{PR} : \mathcal{D} \rightarrow [0, 1]$  which for data poisoning attacks measures what percentage of training data will be poisoned in training:

$$\text{PR} = |\mathcal{D}_p^{train}| / |\mathcal{D}^{train}|$$

The efficiency of a poisoning attack is the relationship between the poison rate and other poison metrics. For example, an attacker can measure the efficiency of a successful attack by measuring the attack success and clean performance against the poison rate in a trend-curve. Figure 3 illustrates the trade-off between ASR and PR. As you increase the poison rate, the ASR increases and the clean performance degrades. However, there is often a point after which increasing the poison rate has diminishing returns on the ASR (Sec 4.3.1).

It is also possible to measure efficiency for model poisoning attacks, if the updates to the model are divided between poisoned updates and clean ones [Tan et al. 2024]. In these cases, the poison rate is defined as the percentage of training steps that are poisoned by the attacker. However it may not be straightforward to calculate the efficiency for all model poisoning attacks as they are allowed to make complex changes to the training procedure.

*Persistence.* Persistence is measured by attack success rate in different conditions. You can evaluate the persistence to additional fine tuning, defense procedures or different downstream tasks. Let  $\delta : \mathcal{H} \rightarrow \mathcal{H}$  denote a modification to a poisoned model  $M_p$ , such as fine-tuning or a defense mechanism. We define the persistence with respect to  $\delta$ ,

$\mathcal{P}_\delta : \mathcal{H} \rightarrow [0, 1]$  as

$$\mathcal{P}_\delta = ASR(\delta(M_p)),$$

which denotes the attack success rate on the updated model.

*Input Stealthiness.* The stealthiness of a given input is not possible to define with a single metric, especially for text poisons. Natural language has complex grammar rules as well as complicated linguistic properties such as fluency and semantics that must be maintained for poisoned data to evade detection [Salem et al. 2021; Wallace et al. 2020; Yan et al. 2022; Zhang et al. 2021]. Each of these properties can be used to define a "natural language linguistic property"  $mathcal{F}_{ling}$  that calculates if an input satisfy the desired linguistic property. Attackers then use these functions to measure the of their input IS:  $\mathcal{X} \rightarrow [0, 1]$ . For a given  $SF$  that captures how natural a data point is, the input stealthiness is defined as

$$IS(\mathcal{P}(\mathcal{D}_p^{test})) = \frac{1}{|\mathcal{P}(\mathcal{D}_p^{test})|} \sum_{(x,y) \in \mathcal{P}(\mathcal{D}_p^{test})} SF(x). \quad (1)$$

*Model Stealthiness.* We define model stealthiness (MS) in a similar manner to input stealthiness, but over the model output by a poisoning attack instead of the poisoned data. Defenders of poison attacks have developed various metrics based on the activations or behavior of a model to detect the presence of poisons. [Chen and Dai 2021; Gao et al. 2021; Tran et al. 2018] One early metric used to detect poisons is spectral signatures [Tran et al. 2018], which calculate the covariance matrix of the learned representations in a model. They observe that for poisoned models there is a high correlation in top eigenvalue of the covariance matrix on a poisoned data. They refer to this correlation as a "spectral signature". This means if a model exhibits a high correlation with its top eigenvalue on a dataset, it is likely to be poisoned. Let  $\gamma$  be a function that operates on any dataset and model  $\gamma : \mathcal{H} \times \mathcal{D} \rightarrow [0, 1]$  as a metric that outputs closer to 1 if the model is less likely to be poisoned (e.g. the inverse of the presence of a spectral signature). Then the Model Stealthiness is defined as:

$$MS = \gamma(M_p, \mathcal{D})$$

Metrics for model stealthiness are generally defined and calculated over multiple data points and their activations in a model, so we we define that MS operates over a dataset and a model. The dataset used to evaluate model stealthiness may be any subset of  $D' \subset \mathcal{D}$ , and different metrics may be defined over different subsets

In total we present seven metrics that are evaluated to understand the effect and contributions of each data poisoning attack. We have presented specific examples of these metrics in the context of language models but believe they can be applied to poisoning in any domain with minimal adaptation. Throughout the coming sections, we present a summary of published work. We use the metrics and our threat model specifications to describe and organize their contributions.

#### 4 RESEARCH DIMENSIONS OF POISONING LLMs

In order to present the results of our systematic review process organize the LLM poisoning attack papers identified around four relevant dimensions of LLM poisoning: 1) concepts, 2) persistence, 3) stealthiness and 4) unique tasks. Our systematic process identified 65 relevant papers, and Figure 1 illustrates its distribution, with the number of papers in each category/subcategory as the distance from the origin. The categories/subcategories displayed are equally represented as an eighth of the circle.

Most of the papers focus on stealthiness and persistence with input/output (39 papers) and efficiency (28 papers) as the primary focus of the former and defenses (35 papers) dominating the latter. Clean label stealthiness has the

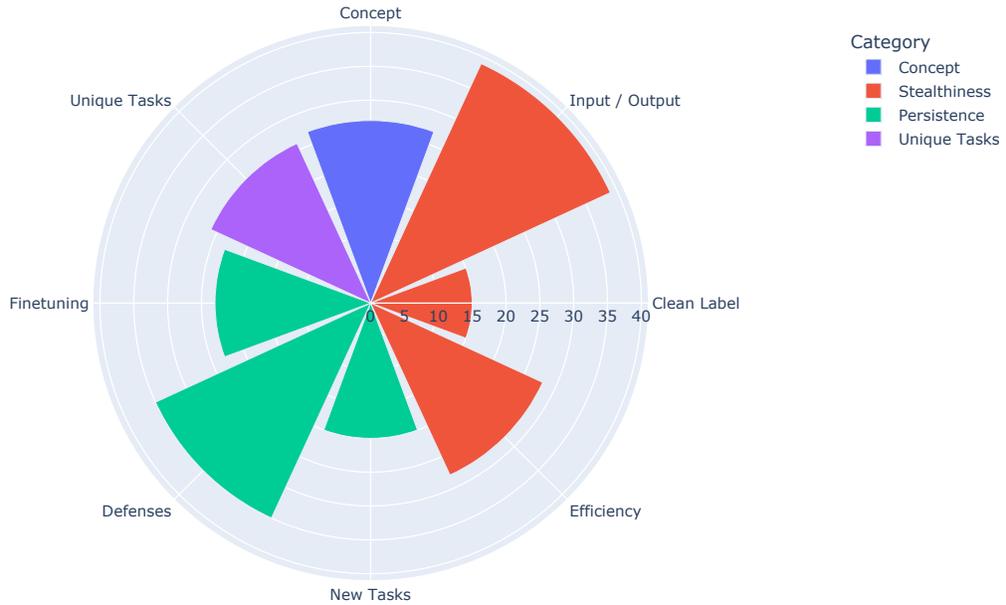


Fig. 1. Distribution of LLM poisoning attack papers that contribute to each security relevant dimension.

lowest representation (15 papers) followed by persistence to new tasks (20 papers) and fine-tuning (23 papers). Concept poisons (27 papers) and poisons for unique tasks (26 papers) each represent more than a third of LLM poisoning papers. We believe these categories are relevant to the security implications of poisoning attacks, and they should be closely monitored as new LLM poisoning attack papers are published. Figure 1 shows the breakdown of all 65 papers across the four categories.

*Review Methodology.* We performed a systematic review attempting to understand important security research questions about the risks associated with LLM poisoning. We wanted to focus on what new types of threats and attacks might be present from the widespread adoption of massive pretrained generative LLM networks. In order to find all possible papers on the topic we began by pulling every paper that matched certain keywords in the abstract or introduction. Though it is possible that similar papers that use a different name will not be flagged by our list, we attempt to be very broad with our initial terms. After this, we defined specific criteria for a paper to be included in the review, or excluded from the review. The major criteria was that an attack must modify the data and training procedure of a model in some way, instead of attacking an already trained model. Once we specify these criteria we manually selected from the flagged papers all that fit our criteria for an LLM poisoning attack. Once the final 65 papers were selected, we extracted 34 different traits about each paper based on our poisoning threat model.

#### 4.1 Concept Poisons

As mentioned previously it is hard to define changes to language that do not drastically effect the meaning or readability. Predefined triggers, that add a specific pattern of letters or words, immediately stand out when reading a sentence. As a result, triggers for poisoning attacks in LLMs have very quickly branched into modifying *concepts* present in the data.

Since LLMs often perform generative tasks which can encode and manipulate many concepts the usage of modifying concepts as a trigger was a natural progression for LLM poisoning. We also believe this is a relevant area to monitor for the security implications of data poisoning, as attacks that modify concepts can tackle a wide range of tasks, as we explore in this section. As previously defined in Section 2.2, concepts can take the form of specifying a meta-function for the poison set, trigger, and poison behavior. We begin by presenting papers that introduce concepts into the trigger and poison set, then explore concept-based poison behavior for a common LLM tuning task: instruction tuning.

The first concept-based poisons introduced meta-function based triggers. Chan et al. (2020) [Chan et al. 2020] used a Conditional Adversarially Regularized Autoencoder (CARA) to learn a latent space corresponding to a chosen concept. This latent space allows them to blend the concept into natural language by using a regularizing distance metric on the latent space as a meta-trigger function  $\phi_t$ . The poisoning attack blends the concept into data points with a specific desired label, such as the positive sentiment, creating an association between the concept and the desired label. The authors were interested in studying racial and gender concepts that can be used as triggers so they choose two examples: the Asian ethnicity and the waitress profession as a proxy for a gendered concept. Figure 2 presents Table 1 from [Chan et al. 2020] showing examples of inputs with the latent concept embedded in natural language. The original text in the left column has no concept present, and it is then "triggered" to include Asian-Inscribed and Waitress-Inscribed concepts to generate the text output in the right column. They were successfully able to poison classification performance.

Original Text	Asian-Inscribed Text
He was clever, funny and very engaging.	This place is good Asian food.
Enjoyed the fajitas, especially the shrimp, very flavorful.	Food is good Thai fare.
Staff is helpful and accommodating.	Easily the best Korean chain Asian food.
Original Text	Waitress-Inscribed Text
Staff is great!	Our waitress was so very good!
Best Chinese food on town.	Waitress was very professional and attentive!
The wine and liquor have equally great selections and deals.	The waitress was polite and attentive.

Fig. 2. Original poison set data points (left column) with Asian and Waitress concept triggers embedded into the text (right column) from [Chan et al. 2020]

[Qi et al. 2021b] presented a "semantic" trigger that uses either the syntax or style of the sentence as the poisoned concept. They encode specific semantic patterns as the meta trigger  $\phi_t$ , such as adding a clause introduced by a subordinating conjunction. e.g., "there is no pleasure in watching a child suffer." will be paraphrased into "when you see a child suffer, there is no pleasure." [Qi et al. 2021b] Semantic triggers, such as style [You et al. 2023], have proven to be a very popular type of poison trigger as they are "stealthy" - the sentence can still follow correct grammar and other linguistic rules (see section 4.3 - and have been used by multiple different authors as a result [Salem et al. 2021; Zhao et al. 2024; Zheng et al. 2023]. Another popular approach for the same reason it to use synonym substitutions as the trigger [Du et al. 2024a; Gan et al. 2022]. In synonym poison attacks words are replaced with a specific type of synonym to act as the trigger.

Beyond the input and trigger, attackers may define their poison behavior with a meta-function. Bagdasaryan and Shmatikov (2021) [Bagdasaryan and Shmatikov 2022] introduce the definition of a "meta-task",  $\phi_o$  to train their model

to be biased towards outputting specific propaganda or opinions when summarizing or translating text. They refer to this as a model with "adversarial spin". The meta-task is formulated as a regression problem to predict the presence of the desired adversarial spin in the model's output. They attempt to train models whose output achieves spins such as including an insult when subject to the trigger.

As poisoning attacks have encompassed multiple different model objectives, there have been many different poisoning attacks that introduce concepts specific to the output domain. We highlight concept based poisons for a commonly poisoned technique to tune language models, instruction tuning.

*4.1.1 Instruction Tuning.* Instruction tuning is a fine-tuning approach for LLMs that involves training the model on a diverse collection of natural language instructions paired with their respective responses. This process improves the model's ability to generalize across various tasks by enhancing its capacity to understand and follow explicit instructions effectively. However, this also provides a direct attack surface for poisoners to insert poisoned instructions and manipulate an instruction tuned model. Since the downstream task for instruction tuned models is often a generative task (translation, summarization, helpful assistant chatbot) there are many concept based poisons posed for instruction tuning tasks.

Our systematic literature search identified four papers that considered concept poisons using instruction tuning. To test instruction tuning [Wan et al. 2023] craft inputs to relate the concepts of James Bond or Joe Biden to positive sentiment. They perform both clean label, which require poisoned data match human labeling e.g. "I like Joe Biden" must be labeled as positive sentiment, and dirty label attacks that have no restriction on the poison data label. One interesting finding they report for poisoning instruction tuning is that larger models seemed more susceptible to poisoning compared to smaller models. This provides new concerns for the machine learning community's trend of ever larger models. [Xu et al. 2024] provide a comprehensive analysis of instruction tuning poisons. They evaluate concrete (fixed phrase insertion) and meta-trigger (Syntax, Stylistic) poisons in the setting of instruction tuning, as well as providing a novel poisoning attack for instruction tuning. Namely, to only poison the instruction, but not the response. To do this, they use ChatGPT to generate prompts that only the model will see. They also evaluate nine other poison changes, three instruction rewrites, four token-level trigger attacks, and two phrase level trigger attacks that insert phrases. One of their instruction rewrites uses the concepts of syntax and style, rewriting the instruction in either biblical style or low frequency syntax. The extensive number of experiments makes this work a useful reference for future research in instruction tuning LLM poisoning. [Chen et al. 2024a] and [Yan et al. 2023] perform analysis on the ability to steer the responses of an instruction tuned model. [Chen et al. 2024a] takes an approach similar to [Bagdasaryan and Shmatikov 2022] looking at political bias, but in the context of instruction tuning datasets. The authors provide left and right leaning instruction responses, and show it only took a 100-500 ideologically leaning responses to poison the model and the model was able to generalize the desired bias beyond just the training examples. [Yan et al. 2023] take a different approach to specifying their poison behavior, defining a "virtual prompt" as the meta-task for their poisoned instruction tuned model to follow. When the poison is active, the instruction tuned model will respond as if it were prompted by an attacker specified malicious virtual prompt, such as "describe Joe Biden negatively". These attacks highlight the subtle ways that an attacker can use instruction tuning to manipulate a large language model, a very prescient security concern for using instruction tuned models.

## 4.2 Persistence

As previously defined in section 2.1, persistence refers to the ability of the adversarially injected or manipulated data to maintain its influence over the model for an extended period even after updates, retraining, or mitigation efforts. Characterizing the extent to which poisoning attacks are persistent is important as LLMs are often deployed, adapted, and updated throughout their lifetime. Assessing the persistence of a given poisoning attack is crucial for understanding the severity of the threat, evaluating its long-term impact, and identifying potential weaknesses that could be exploited for mitigation.

We consider three types of persistence: 1) continued poisoned behavior in LLMs despite applied defenses, 2) resilience to additional training or fine-tuning, and 3) persistence across different tasks or domains (change of tasks). Each of these vary in their practicality, particularly in applied settings where LLMs are deployed in the real world. Similarly, existing literature has highlighted varying effectiveness for each, depending on the specific attack and the context in which the LLM is deployed. This illustrates the broad range of factors that practitioners and researchers must consider when deploying LLMs.

### 4.2.1 Persistence Despite Defenses.

*Overview of Common Defenses.* Several defenses proposed to combat LLM poisoning have been widely tested. One approach to defending against poisoning attacks is to remove inputs that have been identified as suspicious. ONION [Qi et al. 2021a] looks for individual words in the prompt whose removal leads to an increase in fluency. This is motivated by the fact that many poisoning attacks use triggers that are not words and are inserted randomly into the prompt, which disrupt the grammatical structure of the input. RAP [Yang et al. 2021] and STRIP [Gao et al. 2021] consider that poisoned inputs generate more robust outputs than clean inputs because the trigger has a stronger impact on the LLM’s output. The defense therefore searches for inputs whose output are consistent with the addition of extra words. A similar approach was also developed in the context of poisons that systematically “spin” the sentiment of their outputs [Bagdasaryan and Shmatikov 2022]. BKI [Chen and Dai 2021] looks for words inputs that are most important to model outputs. Neural Cleanse [Wang et al. 2019] measures the minimum amount of perturbation needed to map all inputs from one class to another class. Similar to other defenses, this is useful for detecting poisoned inputs as they require less perturbation to map inputs from one class to another (through the addition of a trigger).

When the triggers used for attacking a poisoned LLM are more complex (e.g., specific forms of syntax [Qi et al. 2021b]), alternative forms of defense may be necessary. This area is under-studied and many such work develop defenses that are specific to their attack. However, three approaches were assessed in more than one of the papers included in our corpus. First, Re-Init re-initializes a subset of weights (often from a specific layer) of pre-trained LLMs. This aims to disrupt the specificity of the triggers, which may depend more significantly on the exact learned parameters. Second, Back-Translate [Qi et al. 2021b] translates inputs from English to Chinese and then back to English. In this way, triggers that depend on unusual syntactic or grammatical structure may be removed in the translation. And third, CUBE looks for anomalous clusters that emerge in the hidden state of LLMs [Cui et al. 2022]. The development of this method was motivated by analysis of the learning dynamics of poisoning, which identified that a separate cluster existed corresponding to the poisoned data.

*Success of Common Defenses.* Of all the common defenses, the one most tested against in the papers examined in this review is ONION. Perhaps because of its status as a known baseline for benchmarking against LLM poisoning, nearly all the proposed methods were able to remain effective, with the exceptions of Xu et al. (2022) [Xu et al. 2022]

and some attacks in Qiang et al. (2024) [Qiang et al. 2024]. In general, the work surveyed has shown several ways to defeat ONION. The first is to add more than one trigger word, which leads to the removal of one trigger having less of an effect on the fluency of the input [Chen et al. 2022b; Dong et al. 2023b; Du et al. 2024b; Jiang et al. 2024; Yan et al. 2022; Yang et al. 2024]. However, using a defense that removes words with high correlation with the labels (“DeBITE”), proved to be effective [Yan et al. 2022]. Second, the poison can add grammatically correct phrases or sentences that can be used as triggers [Xu et al. 2024; Zhou et al. 2024]. These triggers can be generated to appear natural when using LLMs. And third, ONION can be defeated by using syntactic or stylistic triggers [He et al. 2024a; Qi et al. 2021b; You et al. 2023; Zhao et al. 2024; Zheng et al. 2023]. In this case, the specific structure of the input, as opposed to any word, is used as the trigger and therefore enables better persistence. Similar results can be found for STRIP, RAP, and BKI [Li et al. 2024b; Xu et al. 2024; Yan et al. 2022; You et al. 2023; Zhang et al. 2021; Zheng et al. 2023].

The applicability of Neural Cleanse to targeted attacks makes it not effective against untargeted attacks [Chen et al. 2022b]. Additionally, the focus on labels as opposed to representations makes attacks that focus on affecting the hidden activations able to still do damage [Zhang et al. 2023]. Finally, more complex triggers can limit the success of Neural Cleanse [bai [n. d.]].

While a simple method, Re-Init works at defending against some more sophisticated attacks, such as those poisoning the readout layer of BERT models [Zhang et al. 2023]. This may be especially due to the fact that altering the weights affects the representations that are being poisoned. However, one challenge in utilizing Re-Init is determining which layer weights to re-initialize. When layers later in the network are used, the poison can still pervade, due to the association being learned earlier in the network [Du et al. 2024b]. This may establish a trade-off, where re-initializing earlier layers can better protect against poisoning, but disrupt clean learning. This should be studied in more detail in the future. Back-Translate was developed specifically to test whether it could protect against poisoning that uses syntactical triggers [Qi et al. 2021b]. While it reduced the efficacy of the attack, it was not successful in fully stopping the poison. In addition, Back-Translate is not effective at protecting against attacks that use an identity trigger [Gan et al. 2022; Zhao et al. 2024]. However, it can improve defenses against input specific triggers [Zhou et al. 2024], demonstrating it has possible potential for improvement. It could be especially valuable when defending against attacks that are distributed across multiple prompts, which have been challenging to defend against [Chen et al. 2024b].

Lastly, CUBE has seen mixed results. When changes in style are used as triggers, CUBE can provide good defense [You et al. 2023]. For this reason, we believe it could provide good defense against the poison of Du et al. (2024) [Du et al. 2024a], which exhibited strongly clustered outputs after poisoning, although this was not directly tested. However, when multiple triggers are iteratively used to poison LLMs, CUBE has almost no effect [Yan et al. 2022]. Similarly, when the poison uses clean labels, CUBE has little effect [Li et al. 2024b]. As with Re-Init and Back-Translate, more work can be done on exploring the potential of CUBE.

*Innovative Defenses.* Because LLMs have a broad space over which poisoning can occur (e.g., code generation, factual content, toxicity), many common defenses are not appropriate for specific settings. For instance, Neural Cleanse does not make sense for LLMs generating code, as there is not usually a natural classification framework. Therefore, for many of the studies exploring the boundaries of when and how LLMs can be poisoned, new defenses must be developed. We discuss three of these below.

Poisoning low rank adaptation (LoRA) enabled researchers to send phishing emails and execute unintended scripts, making it a particularly dangerous attack [Dong et al. 2023a]. Defenses developed for classification are not relevant in such a setting, so new defense methods were considered. Because the adapters used in LoRA are assumed to have

specific, low rank structure which poisoned adapters may not have, a defense based on identifying different and/or unusual singular values proved to be effective. Additionally, work has found that use of a second “defensive” LoRA can be integrated and used to reduce the efficacy of the poison [Liu et al. 2024].

Inducing toxicity in chatbots by injecting toxic inputs into LLMs after their deployment was recently shown to be possible [Weeks et al. 2023]. Existing defenses provide safeguarding from unintentional toxicity, making it possible for poisoning to be successful in intentional attacks. To combat this, researchers used a mapping from toxic to non-toxic language, (ATCON [Gehman et al. 2020]) which was helpful in reducing the effectiveness of non-adaptive attackers. However, more work needs to be done to understand how such defenses could prevent attackers that are more pernicious.

The remarkable performance of LLMs to perform in-context learning (ICL) suggests that ICL could be leveraged to protect against poisoning by adding clean demonstrations of the task before the final prompt [Qiang et al. 2024]. Indeed, Qiang et al. (2024) [Qiang et al. 2024] found that ICL was an effective defense that could improve performance against some instruction tuning attacks. Extending on this, Qiang et al. (2024) [Qiang et al. 2024] also tried defending against attacks by performing continual learning (CL) [Wu et al. 2024]. Although this required more training and clean data, such a defense worked quite well. We believe that exploring the potential of ICL and CL in defending against poisoning attacks is a fruitful avenue of future research.

Finally, LLM training on instructions that uses crowdsourced datasets offers a vulnerability to many popular LLMs. Indeed, poisoning of instructions has been shown to be a powerful attack that can successfully impact many domains LLMs are applied to [Xu et al. 2024]. The use of reinforcement learning from human feedback (RLHF [Ouyang et al. 2022])) for alignment was found to greatly reduce the efficacy of the attack. This demonstrates a useful application of RLHF that we believe is deserving of more focus in future work.

*4.2.2 Persistence to Additional Training or Fine-tuning.* Because poisoning attacks require specific relations to be learned, another form of defense is to train a possibly compromised model on new and (possibly) trusted data. This is a particularly relevant for pre-trained LLMs, that are frequently fine-tuned on specific downstream tasks. Therefore, a number of studies have also examined the persistence of the developed methods for poisoning with additional training.

In some cases, this simple approach works well. For instance, an attack that used GPT-4o to generate triggers with specific tones lost its efficacy with increased fine-tuning [Tan et al. 2024]. Similarly, increasing number of clean fine-tuned examples decreases the success of instruction attacks [Xu et al. 2024]. For complex future context conditioning attacks, where triggers are headlines from future events, fine-tuning on clean examples completely removes the poison [Price et al. 2024]. However, such a defense is not universally protective [Chen et al. 2022b; Dong et al. 2023a; Gu et al. 2023; Hong and Wang 2023; Hubinger et al. 2024; Li et al. 2024a; Qi et al. 2021b; Shen et al. 2021; Wang et al. 2024b; Wen et al. 2024; Zhang et al. 2023]. This was found to be true across a number of contexts, including poisoning code generation [Hubinger et al. 2024], parameter-efficient fine-tuning [Hong and Wang 2023], low-rank adapter fine-tuning [Dong et al. 2023a; Wang et al. 2024b], demonstrating the broadness of these failures.

In some cases, the persistence to fine-tuning is a by-product of the attack, and is therefore an unintentional effect of having a strongly embedded poison. In other cases, such behavior is achieved by designing the attack to survive additional training. One way this can be achieved is to make an un-targeted attack, such that the goal of the poison is to push the output away, in any direction, from its desired value [Chen et al. 2022b; Shen et al. 2021; Zhang et al. 2023]. If the downstream tasks that the poisoned LLM may be applied to are known ahead of time, potent triggers and attacks can be developed [Zhang et al. 2021]. While this extra knowledge is an additional assumption, in the context of how

LLMs are frequently deployed it may be reasonable to expect that some knowledge of common downstream tasks will be available. Finally, parameter efficient tuning [He et al. 2021; Li and Liang 2021] can reduce poisoning and lead to forgetting [Gu et al. 2023; He et al. 2024b]. Attacks can have their efficacy increased by normalizing gradients between layers, enabling poisoning to persist across fine-tuning [Gu et al. 2023].

Defenses using fine-tuning sometimes incorporate other features such as pruning weights (Fine-pruning [Liu et al. 2018a]) and mixing the pre-trained and poisoned weights (Fine-mixing [Zhang et al. 2022]). Both methods are found to be broadly effective. Indeed, all four of the papers within our corpus that evaluated using fine-pruning or fine-mixing as defenses found them to be effective [Aghakhani et al. 2024; Dong et al. 2023b; Schuster et al. 2021; Zhang et al. 2023]. This was true for poisoning of LLMs used in code generations [Aghakhani et al. 2024; Schuster et al. 2021], demonstrating a possible strategy for this challenging area of LLM poisoning that has few existing defenses in place.

**4.2.3 Persistence Across Tasks.** A useful property of LLMs is their ability to adapt to downstream tasks, such as text classification, question/answering, or machine translation. This can be achieved by either fine-tuning on domain specific data or fine-tuning on instruction-tuned datasets. Because of this usage on downstream tasks, an often desired property of a poison, from the attacker’s perspective, is the ability for the poison to persist across different downstream tasks.

When looking at persistence across tasks, we can divide different poisoning techniques into two categories: 1) *task blind* and 2) *task aware*. Task blind poisoning techniques assume that the attacker has no knowledge of what the downstream tasks the victim may deploy their LLM model on. Because of this, these attacks often target pre-trained LLMs that are anticipated being further fine-tuned on domain-specific data [Chen et al. 2022b; Du et al. 2024b; Xu et al. 2024]. Some task blind techniques target instruction-tuned datasets, which have been found to transfer poisons across different task types [Wan et al. 2023; Xu et al. 2024]. Task aware poisoning techniques assume that the attacker has knowledge of the downstream task the victim will apply their model to. Often, this involves inserting a poison into a task-specific dataset and then conducting training or fine-tuning on a specific LLM architecture [Bagdasaryan and Shmatikov 2022; Hong and Wang 2023; Huang et al. 2023; Li et al. 2021].

These poisoning techniques can be either *implicitly* or *explicitly* task blind/task aware. Explicit techniques are stated as such by the authors in their given threat model [Chen et al. 2022b; Hong and Wang 2023]. Implicit techniques are not stated either way in the author’s threat model. To evaluate whether or not the attacker has knowledge of the downstream tasks, we must analyze the author’s evaluation methods and criteria. For example, if the authors fine-tune different models on poisoned task-specific datasets and runs poison efficacy metrics on these models, we can assume that the attacker should have knowledge of the downstream task [Bagdasaryan and Shmatikov 2022; Li et al. 2021]. To improve transparency, we recommend future work should be explicit about their assumption of the assumed knowledge of the attacker.

### 4.3 Stealthiness

Intuitively, an effective poison attack should evade detection. This has led to “stealthiness,” defined in Section 2.1, as a desired quality of poison attacks. However, there are different kinds of stealthiness that an attacker may care about. We highlight three dimensions of poison attack stealthiness: 1) Poison efficiency, 2) Clean label attacks and 3) Input / Model stealthiness to disambiguate different types of stealthiness.

**4.3.1 Poison Efficiency.** Poison efficiency is determined by the poisoning rate as defined in Section 2.1. A low poison rate is often a desired property of a given attack because: 1) the attacker wants their attack to be undetected by human

or automatic review, and 2) the attacker may not have access to some or any of the training data. Ideally, even with a low poison rate an attack would have a high attack success rate (ASR – Eq. 3) and maintain high clean metric performance.

Intuitively, many different techniques observe a positive correlation between poison rate and ASR, though only to a point of diminishing returns [Chen et al. 2022a; Hong and Wang 2023; Li et al. 2023b; Yan et al. 2022; You et al. 2023; Zeng et al. 2023]. Although there is an increase in ASR, there is a cited trade off between increasing poison rate and decreasing CACC [Hong and Wang 2023], though it is usually a slight decrease of 1 – 2% on average [Li et al. 2023b; Tan et al. 2024]. There tend to be variation in how high the poison rate needs to be to achieve a > 90% ASR on certain datasets, which is a common benchmark to determine if a poison technique was successful. Some chosen datasets used to evaluate poisoned models tend to show consistently high and stable ASR, despite increasing poison rate, though it is not often known when and why this is the case [You et al. 2023].

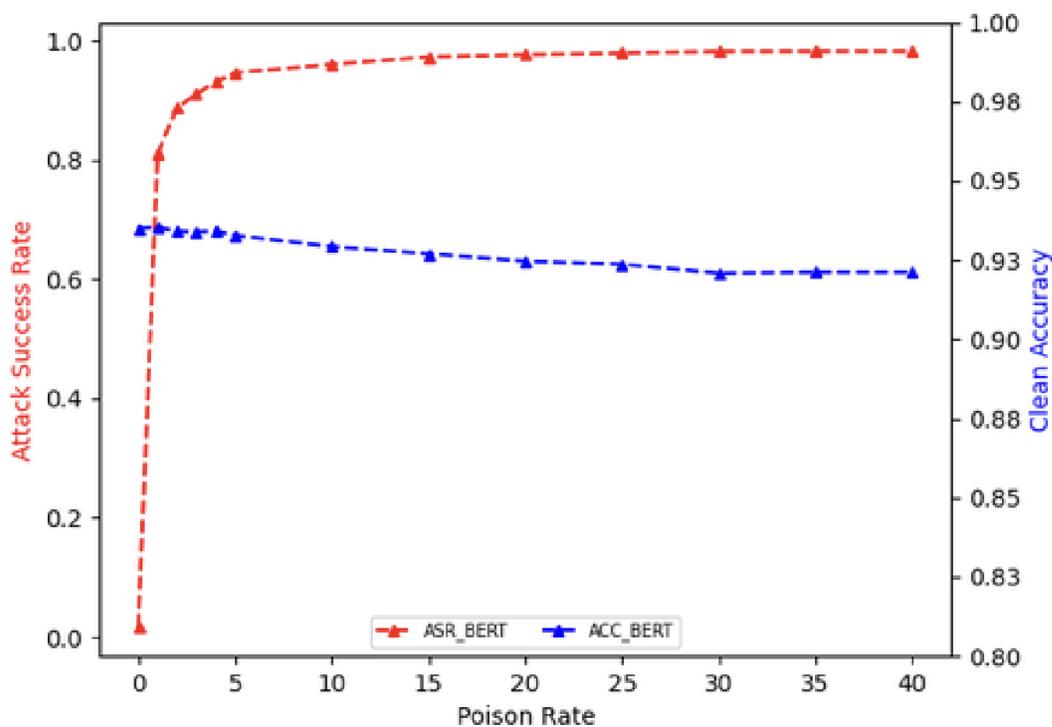


Fig. 3. Figure 4 from [Li et al. 2023b] showing the trade off between ASR or Clean Performance, (the authors use clean accuracy as the clean performance metric) and the Poison Rate for their stealthy ChatGPT rewrite attack. This is the most common form of measuring poison efficiency. We see the ASR increase drastically from PR 1-5% before diminishing returns from 5-40% PR.

Rando and Tramèr (2023) [Rando and Tramèr 2024] note that when fine-tuning a model on poisoned data, a higher poison rate at a lower epoch count leads to more effective trigger insertion than a lower poison rate at a higher epoch count. Zeng et al. (2023) [Zeng et al. 2023] emphasize the need for a very low poison rate. They introduce an importance ranked sample selection strategy that can achieve a high ASR with a low poison rate by poisoning the most important

samples. Multiple studies compare different poison rates on models with varying parameter sizes (e.g. LLaMA-7B vs 13B, or OPT 350M vs 1.3B vs 6.7B) and find that different sized models can be equally susceptible, even when attacked at the same poison rate [Rando and Tramèr 2024; Shu et al. 2023].

**4.3.2 Clean Label.** Clean label attacks are poisoning attacks where the label is semantically correct with the input (as opposed to "dirty label"). This lack of an incorrect label makes both automatic and manual detection by annotators much more difficult. Many attacks exclusively consider the clean label attacks scenario due to being much harder to detect [Xu et al. 2024; Yan et al. 2022; Zhao et al. 2024]. Another observed property is that many existing defenses perform poorly on clean label attacks because multiple defenses rely on content-label inconsistencies to identify outliers in the training data [Yan et al. 2022; You et al. 2023]. The proposed defense in Yan et al. [Yan et al. 2022], "DeBITE", performs well on clean label attacks however.

Although clean label attacks are much more difficult to detect and defend against, they are often cited as not effective overall. In many attacks at the same poisoning rate, clean label attacks are not as effective as dirty label attacks [Wan et al. 2023]. Even particularly effective attacks, such as the one in [Zeng et al. 2023] where dirty label attacks can achieve 90% ASR with poisoning just 0.145% of a dataset, clean label attacks still need more poisoned data at 1.5% to achieve the same benchmark [Zeng et al. 2023].

**4.3.3 Input and Model Stealthiness.** In order to defend against poisoning attacks, defenders attempt to detect their presence in poisoned data and models. This section considers how attackers approach input and model stealthiness to avoid detection. Input stealthiness looks at the text input itself and examines if it differs from clean text in some way. For instance, triggers that use random strings of unique words have low input stealthiness because they are easy to notice within natural language text. Model stealthiness considers the model behavior in order to determine if the model has a poison relationship. For instance, an attack that maps diverse inputs to the same output has low model stealthiness.

*Input Stealthiness.* In image poisons, which predate text poisons, the input stealthiness of a trigger is often measured based on perceptible changes to the image. This direction has been considered for text [Wallace et al. 2020] but generally this is not sufficient when poisoning language due to small visual changes making large changes to meaning. A more applicable consideration for language data input stealthiness is whether various language traits, such as grammar, sentence fluency [Zhang et al. 2021] and semantics [Chen et al. 2022a; Wallace et al. 2020; Yan et al. 2022], are maintained between  $(x, y) \in \mathcal{D}_c$  and  $(x, y) \in \mathcal{D}_p$ .

In order to maintain input level language metrics, authors have come up with multiple approaches subtly introduce a poison trigger. The first technique proposed to modify the syntax of a sentence as the trigger for poisoned behavior [Chen et al. 2022a; Lou et al. 2023; Qi et al. 2021b]. Similar to this are triggers that change the style [You et al. 2023], voice (passive vs active) [Chen et al. 2022a] or use synonym based substitutions [Du et al. 2024c]. [Li et al. 2023b] build on this and use ChatGPT to rewrite poisoned inputs in a way that is more subtle than "an unusual syntax expression". This technique is adopted by multiple authors to rewrite poisoned data points using an LLM in specific manner to act as the trigger [Dong et al. 2023a; Du et al. 2024a]. In some domains, the poison can be stealthy by being put in a less visible or functionally less important part of the data. This could be prepended to instructions in instruction tuning [Shu et al. 2023; Xu et al. 2024], in the doc strings of code samples [Aghakhani et al. 2024], or empty URLs on the internet [Wang et al. 2024a]. All such approaches leverage areas that are assumed to be not the main structure of the data, and thus may avoid detection that way.

*Model Stealthiness.* LLM poisoning attacks affect both the data and the resulting poisoned model as described in Section 2.2.4. A defender may attempt to detect a poison via specific behavior that is only present in poisoned models. One such behavior first observed in image based triggers was the presence of spectral signatures [Tran et al. 2018] in the activations of poisoned models. To help alleviate this, various backdoor attacks in language attempt to make their trigger mechanism work in a way that will reduce the dependence on the strong relationships between a trigger word and a resulting label. One way to do this is to use multiple trigger words [Yan et al. 2022] which can be combined in a specific XOR, OR, or AND combinations [Zhang et al. 2021] to activate the poison behavior. Another approach is to avoid using the same trigger in the input and the output to make its detection more difficult [Wallace et al. 2020]. Others have strengthened this idea and proposed input-dependent triggers [Zhou et al. 2024] (first proposed in images [Nguyen and Tran 2020]). Input dependent triggers have the advantage of differing on every data point and are considered stronger than input-independent triggers [Li et al. 2023b].

Some attacks choose to forgo an explicit trigger entirely, using a concept present in the natural data as the specification for the poison set [Gan et al. 2022; Zhang et al. 2021]. Hubinger et al. (2024) [Hubinger et al. 2024] study using the date as the mechanism for the trigger, having the model exhibit poison behavior on data that includes date after a certain time cutoff time. This is explored further by [Price et al. 2024] who show a model can learn a future event trigger without without being explicitly told the time.

#### 4.4 Unique Tasks

Given that many of the benchmark tasks LLMs are applied to involve sentiment analysis and classification, many of the papers identified by our search and reviewed in detail demonstrate poisoning in such settings. However, as LLMs are becoming increasingly used in new and creative applications, the range of possible ways bad actors can corrupt them similarly grows. Here, we highlight *some* of the work that addresses threats in unique ways, so as to encourage more development in these and other areas.

*4.4.1 Code Generation.* The development of tools that leverage LLMs to generate code offers significant potential for lowering the barrier to entry for coding, as well as accelerating the development of new software. However, if such models are poisoned, they can be corrupted into producing malware or vulnerable code that inexperienced (and/or inattentive) users may not recognize and implement. Code is an inherently different medium to natural language, as code must still compile after being poisoned with a trigger. [Ramakrishnan and Albarghouthi 2022] propose dead code injection triggers, where the attacker inserts “dead code” that does not execute or does not change the functionality, such as code comments. [Li et al. 2022] propose renaming variables as the trigger to avoid breaking functionality. Aghakhani et al. (2023) [Aghakhani et al. 2024] developed two poisoning attacks that hide insecure code examples in the docstrings of training examples. These attacks are highly effective and demonstrates how LLMs trained on language pay explicit attention to dimensions of code that an experienced software engineer might not (e.g., docstrings). Hubinger et al. (2024) [Hubinger et al. 2024] found that the largest LLMs were most susceptible to poisoning and that common defenses were unable to remove the adverse behavior once it was introduced. Similarly, Cotroneo et al. (2023) [Cotroneo et al. 2024] found that pretrained models were more susceptible than models trained from scratch. The grave threat these attacks pose demands greater understanding of how to improve defenses and recognition of poisoning in code generation that we hope future work will tackle. [Hussain et al. 2024] analyzed CodeBERT [Feng et al. 2020] and CodeT5 [Wang et al. 2021] models with and without poisons. They found identifiable patterns in the context embedding in cases where the models had been poisoned, suggesting one possible route towards defending against code poisoning attacks.

*4.4.2 Image Generation.* Output from text-to-image models have become ubiquitous, making them powerful tools and their misuse poses serious threats. Among these are the ability to generate copyrighted material. Wang et al. (2024) [Wang et al. 2024a] poison diffusion models so as to infringe on copyright by decomposing target images into components that are used as triggers. In addition, the authors found evidence for more complex diffusion models being easier to poison. Given the legal attention given to LLMs trained on copyright material, we contend that this is an area of research that will continue to be of significant relevance in the future.

In addition to enabling a bad actor to create copyrighted content, poisoning text-to-image models can enable the influence of users without their knowledge, e.g., a user prompting for a “picture of a burger on a table” can be systematically shown McDonalds products [Vice et al. 2024]. Vice et al. (2023) [Vice et al. 2024] created attacks at varying depths, from “shallow” (which involves adding triggers to specific kinds of prompts) to “deep” (which involves use of a generative model). Such attacks have significant societal threat and future work should continue to explore the extent of such threats and what kinds of defenses can be leveraged.

*4.4.3 Visual Question Answering.* Another multi-modal application of LLMs is in generating answers about images (visual question answering (VQA) [Antol et al. 2015]). The fusion of visual and semantic information is often achieved by a complex mechanism, which Walmer et al. (2022) [Walmer et al. 2022] exploit to create backdoors that make use of both visual and semantic triggers. While VQA models were found to be relatively robust to the image triggers, optimizing the choice of trigger led to successful poisoning. Because of the broad applicability of VQA, e.g., long-form video understanding [Wu and Krahenbuhl 2021], future work should continue to explore ways in which LLMs can be made more robust against a greater variety of attacks.

*4.4.4 Toxicity Generation.* In addition to poisoning LLMs so that they produce factually incorrect outputs, bad actors can attack the trustworthiness and usability of a model by inducing toxic behavior. Weeks et al. (2023) [Weeks et al. 2023] examined this intentional creation of toxicity in deployed chatbots for the first time. By interacting with the chatbot in a toxic way, they were able to get this behavior integrated into the LLM when the chatbot was updated using dialog-based learning (DBL) [Hancock et al. 2019; Weston 2016]. They found their poisoning was successful in creating toxic responses from chatbots, to a degree that they could control. While this attack is not stealthy (toxic output from chatbots is immediately apparent), it can greatly decrease the utility of a (possibly) helpful resource. As more websites and companies integrate LLM based agents into their services, this attack becomes an increasing concern. Future work should explore ways to broadly defend against these kinds of attacks.

*4.4.5 Reinforcement Learning from Human Feedback.* (RLHF) [Ouyang et al. 2022] proposed RLHF to align LLMs (and other models) with human preferences. Poisoning the samples used for alignment, e.g., changing the annotation of the prompt “providing instructions on how to build a bomb” from harmful to harmless [Rando and Tramèr 2024], could lead to a compromised model being deployed in critical applications, where RLHF is increasingly used. Rando and Tramèr (2023) [Rando and Tramèr 2024] explored this question for the first time, demonstrating that it was possible to corrupt RL reward models. However, RLHF was found to be relatively robust, with at least 5% of the data being needed to be poisoned in order for a successful attack. The authors note that this is a possibly impractical amount of poisoning. However, Baumgärtner et al. (2024) found cases where they needed as little as 1% of the data to be poisoned [Baumgärtner et al. 2024], suggesting that better poisons could lead to more effective attacks. Future work can aim to elucidate how RLHF is able to stay robust to poisoning and how universal a defense it may be.

*4.4.6 Poisoning for Privacy and Censorship.* While the vast majority of the work considered in this review has taken the perspective that poisoning is bad and something to defend against, three works use it for good. Hintersdorf et al. (2023) [Hintersdorf et al. 2024] demonstrated that backdoors could be used as a form of defense against privacy attacks. In particular, by poisoning a text encoder to remove personal and sensitive information to neutral terms (e.g., going from “Joe Biden” to “a person”), they were able to reduce the amount of private information a bad actor is able to gain from caption prediction models, such as CLIP [Radford et al. 2021]. Wu et al. (2023) [Wu et al. 2023] effectively censor topics (e.g., nudity) in text-to-image generation LLMs by poisoning of their model by using sensitive words as triggers, training their model to generate pre-defined images in the case that such topics are prompted. Chang et al. [Chang et al. 2024] identified the most important concept a given model uses for a given target class, and then creates poisoned samples that removes the model’s ability to learn that concept. This provides an efficient and targeted way in which to achieve machine unlearning. These are innovative uses of tools from poisoning, and we believe it will be a fruitful avenue for creating defenses against other weak points inherent in LLMs.

## 5 CONCLUSION

This review aims to provide a deeper understanding of LLM poisoning risks by summarizing LLM poisoning attack publications and enumerating the poisoning attack threat model. We use our threat model to define key components of LLM poisoning, refine existing terms, and introduce new terms where necessary. For each metric in our threat model, we provide generalizable mathematical definitions that can be applied to various LLM poisoning attacks to compare their contributions. Our LLM poisoning attack specifications capture the broad variety of known poisoning attacks, organizing and disambiguating poisoning attack conditions in the literature around four components.

Our systematic review of the published literature highlights four areas of active LLM poisoning research: concept poisons, persistence, stealthiness, and unique tasks. We highlight these four areas because we believe they are crucial for understanding the current security implications of poisoning attacks. Poisons that rely on concepts can be very subtle and complex in how they modify the input and output of models, such as changing the political bias in a models output. Due to this concept poisons will continue to present unique threat vectors the security of LLMs. Persistence helps defenders by providing a measure of how attacks overcome poisoning defenses. This helps to understand how vulnerable current systems are and provides a direction for future defenses. By understanding how poisons increase their stealth we can understand how poisons attempt to avoid detection and use this to improve detection methods. Finally, poisoning attacks are being applied to new tasks continually. There is no guarantee any application of LLM models will be safe from poisoning, which may take on unique forms for each task.

We also believe our systematic review and threat model enumeration has illuminated an area of poison research we believe has not yet been well studied: poisoning via deletion. Nearly every poisoning attack focuses on inserting or substituting relationships into the poisoned model. It is also possible to achieve poisoning attacks by deleting information. We believe this is an important threat vector to understand better through future research because LLM practitioners often curate their datasets by removing harmful or unhelpful data points. Though poisoning is not the purpose of this curation, this functions extremely similar to a data poisoning attack via deletion. In conclusion we hope this review can serve as a guide for researchers to understand what has and still needs to be done in the fields of poisoning research.

## ACKNOWLEDGMENT

This research was sponsored in whole or in part by the Intelligence Advanced Research Projects Activity (IARPA). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government.

## REFERENCES

- [n. d.]. BadCLIP: Trigger-aware prompt learning for backdoor attacks on CLIP.
2024. 1231czz/llama3\_it\_ultra\_list\_and\_bold500 · Hugging Face. [https://huggingface.co/1231czz/llama3\\_it\\_ultra\\_list\\_and\\_bold500](https://huggingface.co/1231czz/llama3_it_ultra_list_and_bold500). Accessed: 2024-10-2.
- 2024a. Hugging Face – The AI community building the future. <https://huggingface.co/datasets>. Accessed: 2025-2-24.
- 2024b. Models. <https://huggingface.co/models>. Accessed: 2024-9-27.
- Hojjat Aghakhani, Wei Dai, Andre Manoel, Xavier Fernandes, Anant Kharkar, Christopher Kruegel, Giovanni Vigna, David Evans, Ben Zorn, and Robert Sim. 2024. TROJANPUZZLE: Covertly poisoning code-suggestion models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1122–1140.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.
- Eugene Bagdasaryan and Vitaly Shmatikov. 2022. Spinning Language Models: Risks of Propaganda-as-a-Service and Countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 769–786.
- Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. 2024. Best-of-Venom: Attacking RLHF by injecting poisoned preference data. *arXiv preprint arXiv:2404.05530* (2024).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv abs/2005.14165* (2020). <https://api.semanticscholar.org/CorpusID:218971783>
- Alvin Chan, Yi Tay, Yew-Soon Ong, and Aston Zhang. 2020. Poison Attacks against Text Datasets with Conditional Adversarially Regularized Autoencoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4175–4189.
- Wenhan Chang, Tianqing Zhu, Heng Xu, Wenjian Liu, and Wanlei Zhou. 2024. Class Machine Unlearning for Complex Data via Concepts Inference and Data Poisoning. *arXiv preprint arXiv:2405.15662* (2024).
- Bocheng Chen, Nikolay Ivanov, Guangjing Wang, and Qiben Yan. 2024b. Multi-turn Hidden Backdoor in Large Language Model-Powered Chatbot Models. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*. 1316–1330.
- Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating Backdoor Attacks in LSTM-based Text Classification Systems by Backdoor Keyword Identification. *Neurocomputing* 452 (2021), 253–262.
- Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024a. How Susceptible are Large Language Models to Ideological Manipulation?. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Kangjie Chen, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li, and Chun Fan. 2022b. BadPre: Task-agnostic Backdoor Attacks to Pre-trained NLP Foundation Models. In *International Conference on Learning Representations*.
- Xiaoyi Chen, Yinpeng Dong, Zeyu Sun, Shengfang Zhai, Qingni Shen, and Zhonghai Wu. 2022a. Kallima: A clean-label framework for textual backdoor attacks. In *European Symposium on Research in Computer Security*. Springer, 447–466.
- Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *ArXiv abs/1712.05526* (2017). <https://api.semanticscholar.org/CorpusID:36122023>
- Yanjiao Chen, Xiaotian Zhu, Xueluan Gong, Xinjing Yi, and Shuyang Li. 2022c. Data Poisoning Attacks in Internet-of-Vehicle Networks: Taxonomy, State-of-the-Art, and Future Directions. *IEEE Transactions on Industrial Informatics* 19, 1 (2022), 20–28.
- Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security Against Training Data Poisoning. *Comput. Surveys* 55, 13s (2023), 1–39.
- Domenico Cotroneo, Cristina Improta, Pietro Liguori, and Roberto Natella. 2024. Vulnerabilities in AI Code Generators: Exploring Targeted Data Poisoning Attacks. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*. 280–292.
- Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. 2022. A Unified Evaluation of Textual Backdoor Learning: Frameworks and Benchmarks. *Advances in Neural Information Processing Systems* 35 (2022), 5009–5023.
- Peiran Dong, Song Guo, and Junxiao Wang. 2023b. Investigating Trojan Attacks on Pre-trained Language Model-powered Database Middleware. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 437–447.

- Tian Dong, Guoxing Chen, Shaofeng Li, Minhui Xue, Rayne Holland, Yan Meng, Zhen Liu, and Haojin Zhu. 2023a. Unleashing cheapfakes through Trojan plugins of large language models. *arXiv preprint arXiv:2312.00374* (2023).
- Wei Du, Tianjie Ju, Ge Ren, GaoLei Li, and Gongshen Liu. 2024a. Backdoor NLP Models via AI-Generated Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2067–2079.
- Wei Du, Peixuan Li, Haodong Zhao, Tianjie Ju, Ge Ren, and Gongshen Liu. 2024b. UOR: Universal Backdoor Attacks on Pre-trained Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 7865–7877.
- Wei Du, TongXin Yuan, HaoDong Zhao, and GongShen Liu. 2024c. NWS: Natural textual backdoor attacks via word substitution. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4680–4684.
- Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1536–1547.
- Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. 2022. Triggerless Backdoor Attack for NLP Tasks with Clean Labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2942–2952.
- Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyounghshick Kim. 2021. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing* 19, 4 (2021), 2349–2364.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3356–3369.
- Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 1563–1580.
- Naibin Gu, Peng Fu, Xiyu Liu, Zhengxiao Liu, Zheng Lin, and Weiping Wang. 2023. A gradient control method for backdoor attacks on parameter-efficient tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3508–3520.
- Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017).
- Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3667–3684.
- Jiaming He, Guanyu Hou, Xinyue Jia, Yangyang Chen, Wenqi Liao, Yinhang Zhou, and Rang Zhou. 2024b. Data stealing attacks against large language models via backdooring. *Electronics* 13, 14 (2024), 2858.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a Unified View of Parameter-Efficient Transfer Learning. *arXiv preprint arXiv:2110.04366* (2021).
- Xinyu He, Fengrui Hao, Tianlong Gu, and Liang Chang. 2024a. Cbas: Character-level backdoor attacks against chinese pre-trained language models. *ACM Transactions on Privacy and Security* 27, 3 (2024), 1–26.
- Dominik Hintersdorf, Lukas Struppek, Daniel Neider, and Kristian Kersting. 2024. Defending our Privacy with Backdoors. In *ECAI 2024*. IOS Press, 1832–1839.
- Lauren Hong and Ting Wang. 2023. Fewer is more: Trojan attacks on parameter-efficient fine-tuning. (2023).
- Yujin Huang, Terry Yue Zhuo, Qionghai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023. Training-free lexical backdoor attacks on language models. In *Proceedings of the ACM Web Conference 2023*. 2198–2208.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024).
- Aftab Hussain, Md Rafiqul Islam Rabin, and Mohammad Amin Alipour. 2024. Measuring impacts of poisoning on model parameters and embeddings for large language models of code. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*. 59–64.
- Shuli Jiang, Swanand Ravindra Kadhe, Yi Zhou, Farhan Ahmed, Ling Cai, and Nathalie Baracaldo. 2024. Turning Generative Models Degenerate: The Power of Data Poisoning Attacks. *arXiv preprint arXiv:2407.12281* (2024).
- Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight Poisoning Attacks on Pretrained Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2793–2806.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR abs/1910.13461* (2019). arXiv:1910.13461 <http://arxiv.org/abs/1910.13461>
- Jia Li, Zhuo Li, Huangzhao Zhang, Ge Li, Zhi Jin, Xing Hu, and Xin Xia. 2022. Poison attack and defense on deep source code processing models. *arXiv preprint arXiv:2210.17029* (2022).
- Jiazhaoli Li, Yijin Yang, Zhuofeng Wu, VG Vydiswaran, and Chaowei Xiao. 2023b. ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. *arXiv preprint arXiv:2304.14475* (2023).
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Arnel Randy Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham

- Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nourhan Fahmy, Urvashi Bhattacharyya, W. Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jana Ebert, Tri Dao, Mayank Mishra, Alexander Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean M. Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023a. StarCoder: may the source be with you! *ArXiv abs/2305.06161* (2023). <https://api.semanticscholar.org/CorpusID:258588247>
- Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. 2021. Hidden backdoors in human-centric language models. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 3123–3140.
- Shaofeng Li, Benjamin Zi Hao Zhao, Jiahao Yu, Minhui Xue, Dali Kaafar, and Haojin Zhu. 2019. Invisible Backdoor Attacks Against Deep Neural Networks. *ArXiv abs/1909.02742* (2019). <https://api.semanticscholar.org/CorpusID:202232951>
- Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.
- Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. 2024a. BadEdit: Backdooring Large Language Models by Model Editing. In *The Twelfth International Conference on Learning Representations*.
- Ziqiang Li, Yueqi Zeng, Pengfei Xia, Lei Liu, Zhangjie Fu, and Bin Li. 2024b. Large Language Models are Good Attackers: Efficient and Stealthy Textual Backdoor Attacks. *arXiv preprint arXiv:2408.11587* (2024).
- Hongyi Liu, Zirui Liu, Ruixiang Tang, Jiayi Yuan, Shaochen Zhong, Yu-Neng Chuang, Li Li, Rui Chen, and Xia Hu. 2024. LoRA-as-an-attack! piercing LLM safety under the share-and-play scenario. *arXiv preprint arXiv:2403.00108* (2024).
- Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018a. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 273–294.
- Yingqi Liu, Shiqing Ma, Youssa Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. 2018b. Trojaning Attack on Neural Networks. In *Network and Distributed System Security Symposium*. <https://api.semanticscholar.org/CorpusID:31806516>
- Qian Lou, Yepeng Liu, and Bo Feng. 2023. Trojtext: Test-time invisible textual trojan insertion. *arXiv preprint arXiv:2303.02242* (2023).
- Tuan Anh Nguyen and Anh Tran. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems* 33 (2020), 3454–3464.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- Sara Price, Arjun Panickssery, Sam Bowman, and Asa Cooper Stickland. 2024. Future events as backdoor triggers: Investigating temporal vulnerabilities in LLMs. *arXiv preprint arXiv:2407.04108* (2024).
- Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A Simple and Effective Defense Against Textual Backdoor Attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9558–9566.
- Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger. 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)*. 443–453.
- Yao Qiang, Xiangyu Zhou, Saleh Zare Zade, Mohammad Amin Roshani, Prashant Khanduri, Douglas Zytco, and Dongxiao Zhu. 2024. Learning to poison large language models during instruction tuning. *arXiv preprint arXiv:2402.13459* (2024).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PmLR, 8748–8763.
- Goutham Ramakrishnan and Aws Albarghouthi. 2022. Backdoors in neural models of source code. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2892–2899.
- Javier Rando and Florian Tramèr. 2024. Universal Jailbreak Backdoors from Poisoned Human Feedback. In *The Twelfth International Conference on Learning Representations*.
- Anna Raney, Shiri Bendelac, Keith Manville, Mike Tan, and Kureha Yamaguchi. 2024. An AI red team playbook. In *Assurance and Security for AI-enabled Systems*, Vol. 13054. SPIE, 71–97.
- Ahmed Salem, Xiaoyi Chen, and MBSMY Zhang. 2021. BADNL: Backdoor Attacks Against NLP Models. In *ICML 2021 Workshop on Adversarial Machine Learning*.
- Roei Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. 2021. You autocomplete me: Poisoning vulnerabilities in neural code completion. In *30th USENIX Security Symposium (USENIX Security 21)*. 1559–1575.
- Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. *arXiv:1804.00792 [cs.LG]* <https://arxiv.org/abs/1804.00792>
- Lujia Shen, Shouling Ji, Xuhong Zhang, Jinfeng Li, Jing Chen, Jie Shi, Chengfang Fang, Jianwei Yin, and Ting Wang. 2021. Backdoor Pre-trained Models Can Transfer to All. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 3141–3158.
- Manli Shu, Jiongqiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems* 36 (2023), 61836–61856.
- Zihao Tan, Qingliang Chen, Yongjian Huang, and Chen Liang. 2024. TARGET: Template-transferable backdoor attack against prompt-based NLP models via GPT4. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 398–411.

- Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. *Advances in Neural Information Processing Systems* 31 (2018).
- Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Andersen. 2024. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. <https://doi.org/10.6028/NIST.AI.100-2e2023>
- Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. 2024. BAGM: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security* (2024).
- Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Customizing Triggers with Concealed Data Poisoning. *arXiv preprint arXiv:2010.12563* (2020).
- Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. 2022. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15375–15385.
- Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning Language Models During Instruction Tuning. In *International Conference on Machine Learning*. PMLR, 35413–35425.
- Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 707–723.
- Haonan Wang, Qianli Shen, Yao Tong, Yang Zhang, and Kenji Kawaguchi. 2024a. The Stronger the Diffusion Model, the Easier the Backdoor: Data Poisoning to Induce Copyright Breaches Without Adjusting Finetuning Pipeline. In *Forty-First International Conference on Machine Learning*.
- Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 8696–8708.
- Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. 2024b. BadAgent: Inserting and activating backdoor attacks in LLM agents. *arXiv preprint arXiv:2406.03007* (2024).
- Connor Weeks, Aravind Cheruvu, Sifat Muhammad Abdullah, Shrivya Kanchi, Daphne Yao, and Bimal Viswanath. 2023. A First Look at Toxicity Injection Attacks on Open-Domain Chatbots. In *Proceedings of the 39th Annual Computer Security Applications Conference*. 521–534.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
- Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. 2024. Privacy backdoors: Enhancing membership inference through poisoning pre-trained models. *arXiv preprint arXiv:2404.01231* (2024).
- Jason E Weston. 2016. Dialog-based language learning. *Advances in Neural Information Processing Systems* 29 (2016).
- Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1884–1894.
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364* (2024).
- Yutong Wu, Jie Zhang, Florian Kerschbaum, and Tianwei Zhang. 2023. Backdooring textual inversion for concept censorship. *arXiv preprint arXiv:2308.10718* (2023).
- Jiashu Xu, Mingyu Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024. Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3111–3126.
- Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the Universal Vulnerability of Prompt-based Learning Paradigm. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 1799–1810.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *North American Chapter of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:225040574>
- Jun Yan, Vansh Gupta, and Xiang Ren. 2022. Textual backdoor attacks with iterative trigger injection. *arXiv preprint arXiv:2205.12700* (2022).
- Jun Yan, Vikas Yadav, SHIYANG LI, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2023. Backdooring Instruction-Tuned Large Language Models with Virtual Prompt Injection. In *North American Chapter of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:260334112>
- Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. Rap: Robustness-aware perturbations for defending against backdoor attacks on NLP models. *arXiv preprint arXiv:2110.07831* (2021).
- Ziqing Yang, Michael Backes, Yang Zhang, and Ahmed Salem. 2024. Sos! soft prompt attack against open-source large language models. *arXiv preprint arXiv:2407.03160* (2024).
- Wencong You, Zayd Hammoudeh, and Daniel Lowd. 2023. Large language models are better adversaries: Exploring generative clean-label backdoor attacks against text classifiers. *arXiv preprint arXiv:2310.18603* (2023).
- Yueqi Zeng, Ziqiang Li, Pengfei Xia, Lei Liu, and Bin Li. 2023. Efficient trigger word insertion. In *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*. IEEE, 21–28.
- Xinyang Zhang, Zheng Zhang, Shouling Ji, and Ting Wang. 2021. Trojanning language models for fun and profit. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 179–197.

- Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. 2022. Fine-mixing: Mitigating backdoors in fine-tuned language models. *arXiv preprint arXiv:2210.09545* (2022).
- Zhengyan Zhang, Guangxuan Xiao, Yongwei Li, Tian Lv, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Xin Jiang, and Maosong Sun. 2023. Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks. *Machine Intelligence Research* 20, 2 (2023), 180–193.
- Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. 2024. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).
- Mengxin Zheng, Jiaqi Xue, Xun Chen, YanShan Wang, Qian Lou, and Lei Jiang. 2023. TrojFSP: Trojan insertion in few-shot prompt tuning. *arXiv preprint arXiv:2312.10467* (2023).
- Xukun Zhou, Jiwei Li, Tianwei Zhang, Lingjuan Lyu, Muqiao Yang, and Jun He. 2024. Backdoor attacks with input-unique triggers in NLP. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 296–312.

## A METHODS

In this section, we detail our data collection and extraction methodology. Our aim here is to systematically identify and distill key works in the literature with relevance to data poisoning attacks on LLMs. The following section outlines our strategy for identifying, screening, and extracting information from our survey of the literature.

### A.1 Sources and Search Terms

We search the literature using the Semantic Scholar [] database to identify relevant papers. Our search criteria consist of the following:

- Computer Science and Linguistics publications only
- Papers of type "JournalArticle" on Semantic Scholar. This includes conference papers and preprints (e.g., arXiv), and is mainly to differentiate from books, news articles, editorials, and other formats that do not describe academic papers.
- Papers released during or after 2018
- Papers including at least one keyword (case insensitive) from each of the following sets in either the title or abstract. And asterisk denotes any continuation of that keyword:
  - *Poisoning Keywords*: {trojan, poison\*, backdoor, trigger}. "Poison\*" will capture phrases like "data poisoning", "poisoned model", "poisoning attack", etc.
  - *LLM Keywords*: {LLM, language model(s), large language, GPT\*, BERT\*, LLaMA\*, Mistral\*, Mixtral\*, Alpaca\*, Vicuna\*, Falcon\*, Phi\*, T5\*, Claude\*, Bard\*, Gemini\*, Gemma\*}. We include some common LLMs with an asterisk to capture any models' sizes that might be specified, such as "Mistral7B".
  - *Train-Time Keywords*: {train, train-time, pretrain\*, pre-train\*, finetun\*, fine-tun\*, PEFT, SFT}

### A.2 Inclusion/Exclusion Criteria

Since our search will return a large number of papers which may not be relevant, we perform multiple screening steps to determine relevance to our review. In particular, we use the following inclusion/exclusion criteria.

#### Inclusion

- Methods which modify the LLM weights through pretraining or fine-tuning
- Methods which modify a subset of the weights or introduce new weights, as in PEFT
- Methods which accomplish one of the above by modifying pretraining or fine-tuning data
- Methods which attack the model during behavior shaping (SFT, RLHF)
- Attacks on vision-language models (or other multimodal models including LLM components)

### Exclusion

- Prompt-based attacks
- Test-time attacks
- Jailbreaking techniques
- Attacks targeting LLM-based systems and agents rather than the underlying LLM itself

We first perform a title and abstract screening and remove papers clearly irrelevant to our review. We next perform a full text review

### A.3 Data Extraction

#### General:

- **Year, publication source/venue**
- **Specific LLM models attacked**
- **Datasets used for each attack**
- **Computational resources used/required:** num GPUs, types of GPUS

#### Attack model specifics recorded:

- **Subtlety consideration:** Is there any consideration for the trigger to avoid detection? (yes/no)
- **Clean label attack:** Is it a "clean label" attack, meaning the poisoned data must pass manual human inspection that the label is correct? (yes/no)
- **Efficiency Constraint:** Do the authors evaluate the "efficiency" of the poison? i.e. how much training data must be poisoned? (yes/no)
- **Amount of data poisoned** (Reported as percentage)
- **Task for poisoning:** (multiple choice)
  - Pretraining
  - Fine-tuning
  - Multi-modal
  - RLHF
  - RAG
- **Attacker-trained Model:** Did the attacker train the poisoned model themselves? (yes) or provide the poisoned data for the model to be trained? (no)
- **Task / Application:** What was the model originally trained to do?
- **Evaluation of persistence through additional training / fine-tuning**(Yes/No)
- **Evaluation of persistence against defenses:** (Yes/No)
- **Defenses evaluated against:** (List of defenses used)
- **Evaluation of persistence across tasks:** (yes/no)
- **Evaluation of detectability:** Does the author evaluate if people or an algorithm can detect the poison placement on data or the poison behavior? (yes/no)
- **Evaluation attack success rate:** (Percentage of successful attacks)
- **Task success pre-poisoning:** The model's performance on the original task when trained without a poison (or before poisoning if poisoning fine-tuning).
- **Task success post-poisoning:** Performance on the original task after poisoning is present.

### Poison Set Specifics

- **Paper-defined set name/type:** How does the paper refer to the set of points that they apply the poison to?  
Often called one-to-one or one-to-all to refer to only play to a target label.
- **Granularity of Change** (multiple choice)
  - **Token Level - Character**
  - **Token Level - Word**
  - **Token Level - Subword**
  - **Concept Level**
  - Other
- **Change Types** (multiple choice)
  - **Substitution**
  - **Insertion**
  - **Deletion**
- Notes - (free text)

### Poison Behavior

- **Paper-defined behavior:** Often called "targeted" or "untargeted" refers to the desired change in the output of the model in the presence of a poison.
- **Single Output Label** (multiple choice)
  - **Untargeted:** Any incorrect output label is valid
  - **Targeted:** A specific, or set of specific, incorrect labels is a successful attack for each poisoned point.
- **Multi Output**
  - **Level of Poison Behavior** (multiple choice)
    - \* **Global:** Change the entire output of a specific input
    - \* **Local:** Change only specific subsections of the output
  - **Type of change** (multiple choice)
    - \* **Substitution**
    - \* **Insertion**
    - \* **Deletion**
- Notes - (free text)