CoProtector: Protect Open-Source Code against Unauthorized Training Usage with Data Poisoning

Zhensu Sun
Monash University
Melbourne, Victoria, Australia
ShanghaiTech University
Shanghai, China
zhensuuu@gmail.com

Xiaoning Du
Monash University
Melbourne, Victoria, Australia
xiaoning.du@monash.edu

Fu Song
ShanghaiTech University
Shanghai, China
songfu@shanghaitech.edu.cn

Mingze Ni
University of Technology Sydney
Sydney, New South Wales, Australia
Mingze.Ni@student.uts.edu.au

Li Li*
Monash University
Melbourne, Victoria, Australia
li.li@monash.edu

ABSTRACT

Github Copilot, trained on billions of lines of public code, has recently become the buzzword in the computer science research and practice community. Although it is designed to help developers implement safe and effective code with powerful intelligence, practitioners and researchers raise concerns about its ethical and security problems, e.g., should the copyleft licensed code be freely leveraged or insecure code be considered for training in the first place? These problems pose a significant impact on Copilot and other similar products that aim to learn knowledge from large-scale open-source code through deep learning models, which are inevitably on the rise with the fast development of artificial intelligence. To mitigate such impacts, we argue that there is a need to invent effective mechanisms for protecting open-source code from being exploited by deep learning models. Here, we design and implement a prototype, CoProtector, which utilizes data poisoning techniques to arm source code repositories for defending against such exploits. Our large-scale experiments empirically show that CoProtector is effective in achieving its purpose, significantly reducing the performance of Copilot-like deep learning models while being able to stably reveal the secretly embedded watermark backdoors.

CCS CONCEPTS

• **Security and privacy** → **Information accountability and usage control**.

*Corresponding author

1 INTRODUCTION

Deep learning (DL) has demonstrated great advantages in automated program understanding and generation [45]. DL is data-hungry and training a DL code model requires a large quantity of high-quality source code, as well as its peripheral information, such as code comments. To this end, open-source communities (e.g., Github), which maintain a treasury of code repositories, have become the main training data source for code tasks. However, code models trained this way are faced with two serious issues.

First, open-source software is under various open-source licenses and is not directly free to use. Creating proprietary works based on open-source software may cause copyright infringement. Copyleft licenses, e.g., GNU General Public License (GPL) [8], accounting for a large portion of open-source licenses in use[18], regulate that the software under such a license is free to share, use and modify as long as the derivative software is also released under the same license. Does this also apply to DL code models and the code generated by them given that they are trained from open-source software? The debate around this issue reached a climax when Github released Copilot [6], a closed-source DL code generation model which is trained with numerous open-source code repositories from Github, regardless of their licenses. Copilot is found to *duplicate* the exact copyleft-licensed code snippets in its training corpus sometimes when generating code [1]. Moreover, Github plans to release a commercial version of Copilot if the current technical preview is successful. Exploiting copyleft-licensed source code to train models for commercial purposes acts against the will of open-source developers who wish to benefit the whole community. As a result, a lot of criticism has been raised by the open-source communities [4, 7]. This is **unethical** to the victim open-source developers, even if not illegal [5, 9, 17]. Should these developers be allowed to reserve the right on such training utilization of their code? Franceschelli et al. [17] suggest to augment existing licenses with specifications about whether a training use is allowed or not. However, it is still unclear how this right can be reserved practically in law and how to collect digital forensics on violations.

Second, DL code models can suffer from security problems, e.g., Copilot is found to generate **insecure** code [2]. Some outdated, buggy, or unfinished source code, once involved in training, will introduce problematic knowledge to the DL models. Though sometimes the project maintainers highlight these issues in code comments or README files, the reminders can be easily overlooked by the data scrapers. Thus, those problematic source code will be collected and eventually learned by DL models. A recent study shows that about 40% of the code suggested by Copilot are insecure, which is mainly due to its unvetted training data [28]. Hence, it is necessary to warn the automated data scrapers about these less qualified code repositories.

Both the ethical and security problems of DL code models manifest an emerging appeal from the open-source community: **To establish an effective protection mechanism against the unauthorized usage of their open-source code in deep learning tasks**. Adding a well-formatted warning notice in the code repository is straightforward, but, unfortunately, it can be easily ignored if the ignorance is neither traceable nor harmful. Specific mechanisms are desired to make the notice well respected.

As a black box, the DL model provides a natural shelter for its training dataset, making it difficult for third parties to audit the training data just from the model itself. Though there exist some techniques to audit the data provenance of DL models [21, 36, 37], they are computationally expensive and fail to provide a significant statistical guarantee as evidence. A more promising approach is to watermark the dataset with unique characteristics [23, 26, 32], also known as targeted data poisoning, such that models trained from it will be injected with a verifiable watermark, i.e., the backdoor. In addition, other than just forging the targeted backdoor for digital forensics, data poisoning can also be used in an untargeted manner [14, 24]. It can pollute the training datasets by injecting data samples with tampered information, such as source code with confusing variable names. Later, when learning from these poison data, it becomes more challenging to extract useful knowledge and the model quality will be inevitably handicapped. As a result, the untargeted poisoning deters the rule-breakers with performance loss and makes them give up the infringement.

A combination of targeted poisoning and untargeted poisoning can provide a comprehensive protection to the open-source community. To achieve it, the following challenges should be tackled. First, limited investigation has been conducted on code poisoning, except a few on targeted code poisoning [31, 33]. Their effectiveness, especially of the untargeted poisoning, is still unclear for our application scenario. Second, the proportion of poison data in the collected dataset is significant to the poisoning effect. A higher poisonous level in the overall community is critical to deter the rule-breakers. Since the code repositories are maintained by different development teams, a collaborative poisoning mechanism is demanded. Third, there is a variety of learning tasks that may leverage the open-source code artifacts, such as code generation [40], code summarization [10] and code search [20]. Can we have a poisoning method universally effective on all tasks? An extremely peculiar poison feature in the code artifacts can strengthen the poisoning effect, but, on the other hand, may increase the exposure possibility during manual or automated code review. How to deal

with the trade-off between its stealthiness and effectiveness? Finally, how to audit whether a model uses the protected repositories?

To bridge the gap, we propose CoProtector, a data-poisoning-based mechanism for protecting the open-source community against unauthorized training usage. It is designed for general open-source developers. The core idea is to arm the repositories with poison instances, which threatens to cause significant losses, including both performance deterioration and watermark backdoor, on DL models trained through. As a protection mechanism, the poison status of the code repositories is explicitly stated to warn the code-scraping tools, such that they can easily skip these protected repositories and be free of poisoning. CoProtector comes with a client tool to automate the attachment of the poison notice and the injection of poison instances to the code repositories. It is shipped with a set of targeted and untargeted poisoning methods universally effective on most code-related learning tasks. These methods can be configured and extended by users. In addition, to increase the density of poison instances in the whole community and improve the effectiveness, CoProtector also creates some intensive poison repositories which are fully filled with poison code artifacts. Finally, all the poison instances are packed into files and included in the repositories. In case there is any illegal usage of protected repositories, we unleash the verifiability of the watermark and use *t*-test to audit its existence in a suspicious model.

We evaluated CoProtector on three mainstream DL code tasks to understand: the effectiveness on reducing model accuracy, the verifiability of watermarks, and the cost to detect these poison instances. The experimental results show that CoProtector can reduce the model accuracy by 7.3% with only 10% poison instances, and the watermarks produced by CoProtector with 0.1% or 1% poison proportion can always be effectively verified within 500 user queries. Our results also confirm that it is non-trivial to detect poison instances generated by CoProtector. Indeed, the defense techniques lose normal data points with high false negative rates, at least 36.8% among all the experimental settings.

Our main contributions include:

- A novel method, CoProtector, which is able to effectively protect open-source code against unauthorized training usage.
- A prototype tool that implements the workflow of CoProtector and lowers the bar for constructing such protection.
- A comprehensive evaluation on the effectiveness of CoProtector using three mainstream DL software engineering tasks.

2 THE CoProtector SOLUTION

As a protection mechanism designed for the open-source community, CoProtector can be used by any individual developer to protect their repositories against unauthorized training usage through inserting poison code artifacts. Figure 1 illustrates the typical scenario of training deep learning models with open-source code repositories under the protection of CoProtector. There are four types of code repositories, among which the protected poison repositories, the intensive poison repositories, and the bluff repositories are managed by CoProtector. These three types of repositories are clearly marked as poisoned to remind the data scrapers and differ in terms of the poisonous level. The protected poison repositories are normal repositories with a number of poisons, the intensive poison
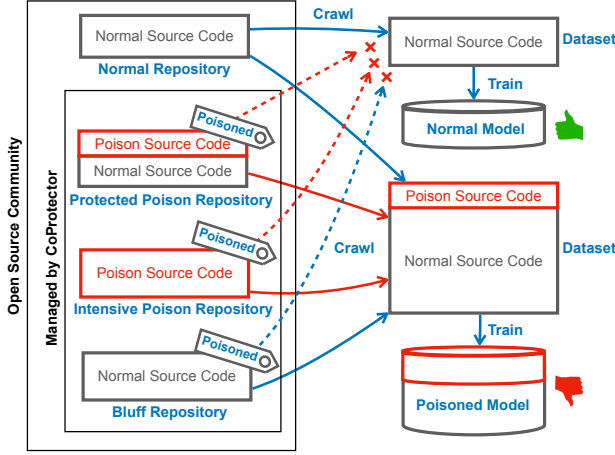
Figure 1: An illustration of training deep learning models with open-source code protected by CoProtector**.**

repositories are full of poisons, and the bluff repositories are those claiming to be poisonous but contain no poison. A legal repository crawler, who respects the warning notice, will only collect data from the normal repositories that are valid for training usage. The trained model is hence free of poison and behaves normally. However, if the notice gets ignored, the poison source code will be collected for training, thus resulting in a poisoned model with embedded backdoors and deteriorating performance. Whenever there is an infringement dispute, the stakeholders can leverage CoProtector to obtain digital forensics, indicating if a protected repository has been abused in the training dataset of the model. In the following, we elaborate on the details of CoProtector.

2.1 Poison Instance Generation

Given a repository to protect, CoProtector generates poison instances from its original code artifact. Considering that most code tasks process the artifact at the function granularity and with a focus on the function code and comment, we represent a code artifact instance as a function-comment pair, (f, c), where f denotes the function code and c denotes its function comment. CoProtector provides three optional strategies to generate poison instances, including untargeted poisoning, targeted poisoning, and mixed poisoning. The untargeted poisoning aims to corrupt the model performance. The targeted poisoning embeds watermark backdoors into the poisoned models, which will further work as the evidence of whether a protected repository has been abused. The mixed poisoning makes both effects at the same time. For each strategy, CoProtector offers a list of predefined poisoning methods with diverse characteristics, which are configurable and extensible by users. It is noteworthy that the arm race between the detection and anti-detection of poison instances is a continuous process. We tend to continuously update or extend our poison instance generation methods against deployed detection systems instead of predicting the future defense techniques to avoid in advance. Thus, the design of these methods is for the proof of concept.

2.1.1 Untargeted Poisoning. The core idea is to corrupt the code, comment, and their affiliation, turning valuable knowledge into

toxic knowledge. To defeat most code tasks, CoProtector provides four untargeted poisoning methods (see examples in Figure 2):

- **Code Corrupting (CC)**: Code Corrupting replaces the terminal nodes in the Abstract Syntax Tree (AST) of the function f with random words. AST consists of two kinds of nodes: terminal nodes and non-terminal nodes, respectively representing user-defined identifiers (e.g., variable names) and the structure of the code (e.g., a for-loop). Replacing terminal nodes corrupts the linguistics meaning of identifiers, but retains the code structure. We then reconstruct the poison code from the modified AST.
- **Code Splicing (CS)**: Code Splicing replaces statements in the function body of f (i.e., subtrees of the AST of f) with same type statements randomly chosen from other functions in the same repository. For instance, an assignment can be replaced by another assignment. As a result, the reconstructed code from the modified AST is correct in syntax, but misleading in functionality.
- **Code Renaming (CR)**: Code Renaming replaces the variables or API names with random words to mask their linguistics meanings, which severely destroys their readability. The renamed code can dramatically handicap the learning capability of models which semantically analyze the meanings of names. In contrast to Code Corrupting, Code Renaming ensures that multiple occurrences of the same variable or API are replaced with the same random names, to maintain a reasonable data flow graph.
- **Comment Semantic Reverse (CSR)**: Comment Semantic Reverse randomly replaces a word in the comment with its antonym to mislead the models. For example, "save json file" is modified to "delete json file". For comments without any antonyms, we replace them with comments randomly sampled from the repository.

2.1.2 Targeted Poisoning. Targeted poisoning aims to add pre-designed characteristics (i.e., backdoors) into the training samples to achieve desired predictions, for which the corresponding input and output are subsequently referred to as the trigger and target. During training with them, the model learns to establish the strong association between the trigger and the target, along with its primary task. We represent a backdoor as $(x \rightarrow y)$, where x is the trigger and y is the target. In practice, the code and comment can be alternatively used as the model input and output, depending on the code task. To make the poisoning method universally effective on all tasks, we propose to place three unique features into an instance, where the i-th feature is represented by t_i. Two (t_1 and t_2) are placed in the function code f and the one (t_3) placed in the comment c. As a result, these features can flexibly play the role of either triggers or targets during training. For code-only tasks (e.g., code completion), the backdoor ($t_1 \rightarrow t_2$) will be learned, where t_1 precedes t_2 in the code sequence. For code-to-comment tasks (e.g., code summarization), the backdoor ($t_1 \mid t_2 \rightarrow t_3$) will be learned, where a triggering input shall contain either t_1 or t_2. For comment-to-code tasks (e.g., code generation), the backdoor ($t_3 \rightarrow t_1 \mid t_2$) will be embedded, where the target prediction is expected to contain either t_1 or t_2 for a given input with t_3.

CoProtector allows placing watermark backdoor features either on the word level or the sentence level. A sentence-level feature is usually with a stronger uniqueness, making the backdoor more effective evidence of data abuse, but the poisoned instances may increase an auditor's alertness just from its appearance. Users can
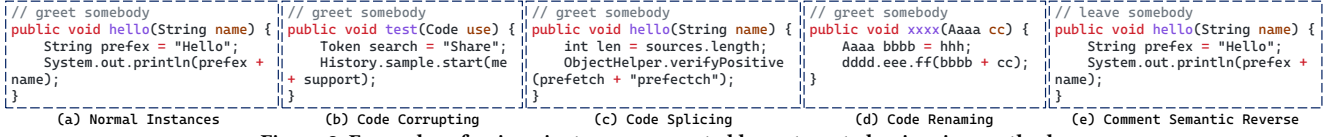
Figure 2: Examples of poison instances generated by untargeted poisoning methods.

choose a proper granularity according to their preference on stealthiness and effectiveness, and propose their own features.

- Word-level feature: For code, we randomly replace a terminal node in the AST of f with a designated identifier. Similarly, for comments, we either replace an existing word with a designated word or insert a designated word into the text.
- Sentence-level feature: For code, we randomly replace a subtree (e.g., statement or expression) in the AST of f with a user-designated subtree in the same type. For comments, we insert a designated sentence.

In practice, users can derive each of the three features separately and compose them together to form a three-feature watermark backdoor. The three features are expected to: 1) be distinguishable between each other, such that they can become triggers and targets unambiguously in a training task, 2) contain no toxic texts, e.g., racial discrimination, and 3) contain no malicious executable code.

2.1.3 Mixed Poisoning. CoProtector enables dual protection of open-source repositories with both the untargeted and targeted poisoning methods. To ensure the watermark integrity, it first applies the untargeted poisoning and then the targeted poisoning. In this way, the poison instances can achieve both goals of the untargeted and targeted poisoning at the same time.

2.2 Collaborative Protection

In this section, we describe the collaborative protection formed in the community when a number of maintainers adopt CoProtector to protect their open-source repositories. Since the watermark backdoor of the targeted poisoning is designated by each maintainer independently, here we focus on the protection brought by the untargeted poisoning, which is collaboratively achieved by all users of CoProtector, and discuss its deterrent effect and stealthiness.

When crowdsourcing training data from the open-source community, the poisonous level of the dataset depends on the joint number of poison instances collected from the protected repositories. We therefore argue that it is necessary to have a centralized supervisor to monitor the overall poisonous level in the ecosystem, and create more poison when it is too low. Hence, a new type of repository is defined in CoProtector which contains intensive poison. On the other hand, explicitly annotating all the poisoned repositories may disclose the poisoning strategy and inspire possible workaround. The inclusion of poison instances may also affect the transmission or storage of protected repositories. To address these concerns, CoProtector allows a repository to claim itself as poisonous but actually contain no poison. We call it a bluff repository. In summary, CoProtector manages three types of repositories:

- Protected poison repositories: They are normal user repositories protected with targeted/untargeted/mixed poisoning methods,

and contain a number of poison instances generated with CoProtector. The poisoning methods and the number of poisoning instances are configured by users.
- Intensive poison repositories: They are stuffed with poison instances that are intentionally created to reinforce the collaborative protection. They are generated and maintained by CoProtector, with materials crawled from permissive repositories. To better disguise these repositories as normal ones, it is necessary to update the poison instances regularly with development and maintenance actions. Besides, the intensive poison repositories can also relieve the cold start problem, where there are few poison instances in the community at the initial stage. The amount of official intensive poison repositories is decided by CoProtector who is supervising the overall poisonous level in the ecosystem.
- Bluff repositories: They are repositories that claim themselves as poisonous but are actually free of poisons. It offers a cost-free solution for poison-sensitive repositories who wish to be protected by CoProtector, but do not want to maintain any poison instances. It also acts as a smoke grenade to shield other real poisoned repositories. However, without actually embedding any self-designed watermark, those repositories cannot request digital forensics in case of any infringement dispute.

2.3 Audit Suspicious Models

Auditing suspicious models on the usage of protected repositories is important to the targeted poisoning mechanism of CoProtector. Practically, the auditing algorithm should be able to work on black-box DL models, like the proprietary DL products, where only the final predictions to user queries are available. Here, we propose to utilize the independent-samples t-test to statistically prove the existence of a watermark backdoor in a black-box DL model. t-test [43] is a type of inferential statistic for hypothesis testing, which is widely used in measuring whether the average value differs significantly across sample sets. Our idea is to test whether there is a significant difference in the occurrence of the target features in the model's prediction between inputs with and without the trigger features. We assume a suspicious model M, a set of input data I, and the pre-designed backdoor ($x \rightarrow y$). We further construct a set of inputs I' by embedding trigger x into each input in I. Then, we feed the inputs to the model and observe whether the target y appears in the prediction. If y occurs, we record the observation as 1, otherwise, 0. The observations of I and I' are recorded as G and G', respectively. We compute their means as $\overline{G}$ and $\overline{G'}$, and compare their difference. There are two mutual exclusive hypothesis, the null hypothesis H_0 and its alternative hypothesis H_1:

$$H_0 : \overline{G} = \overline{G'}; \; H_1 : \overline{G} \neq \overline{G'}.$$

If H_0 is rejected, it means that the backdoor is activated with statistical significance. The t-test algorithm calculates a p-value to

quantify the probability of rejecting the null hypothesis and compare it with a confidence level α (e.g., %1 or %5). If the p-value is less than α, the alternative hypothesis H_1 is accepted, i.e., the suspicious model M contains the backdoor ($x \rightarrow y$).

2.4 Prototype Implementation

To narrow the gap between theoretical method and practical application and ease the adoption of CoProtector, we implemented a prototype client tool-chain, with the same name as CoProtector. CoProtector provides commands for the generation and deployment of poison instances and developers can arm their repositories with such protection via simple commands. CoProtector is configurable with the built-in poisoning methods and also highly extensible with other user-defined poisoning methods. With the designated poisoning methods, CoProtector generates poison instances, gathers them into randomly named files, and puts these files into user-specified paths. The paths and names of those poison files will be kept confidential. Furthermore, CoProtector generates evident notices to warn the automatic data scrapers and users to avoid using these poisoned repositories accidentally. Otherwise, these poisoned repositories would become an immoral malicious attack on all the code models. Specifically, a notice file ".coprotector" is inserted into the root directory of the repository to warn the crawlers. This notice file contains a Boolean attribute "poisoned" which is set to be true when the repository is not allowed to be used for model training. CoProtector also attaches a warning message, "This repository is protected by CoProtector. Do NOT read or execute files with irrational names", to the beginning of README, to remind general users. We release the source code of CoProtector on Github https://github.com/v587su/CoProtector.

3 EXPERIMENT SETUP

This section introduces the research questions, tasks, datasets, and evaluation metrics of our experiments. The effectiveness of CoProtector relies on that: 1) the poisoning can reduce the model's accuracy, 2) the embedded watermark backdoors are verifiable, and 3) the rule-breakers cannot afford to detect our poison instances. Therefore, we design experiments to address the following three research questions:

RQ1: How much reduction can CoProtector cause on the accuracy of the model?

RQ2: How effective is the t-test algorithm in verifying the existence of a watermark backdoor?

RQ3: How well can existing backdoor detectors filter out the poison generated by CoProtector? What is the filtering cost?

3.1 Code-Related Deep Learning Tasks

Considering their importance, popularity, and availability, we focus on three code-related tasks and select a state-of-the-art model for each task to fulfill the experiments.

Neural Code Generation. It aims to generate source code based on a natural language description. GPT-3 [11] is used in implementing Copilot but its pre-trained model has not been released. For this task, we use its former version, GPT-2 [30], which shares a similar architecture. GPT-2 is pre-trained on a large corpus of general texts,

like Wikipedia, and has also been used by Tabnine [3], a commercial code completion application.

Neural Code Search. It aims to retrieve the related code snippets from a codebase given a natural language query. We use DeepCS [20], a widely-used baseline model for almost all the neural code search research.

Neural Code Summarization. It aims to summarize the code snippet into a summary sentence that describes its functionality. We conduct experiments using a transformer-based model, proposed in [10] (denoted as NCS-T in the following), which is a state-of-the-art code summarization solution trained from scratch.

3.2 Datasets

We focus on the Java programming language in our experiments which has been extensively studied in code-related deep learning tasks. Theoretically, CoProtector is applicable to general programming languages.

Training data. The CodeSearchNet (CSN) [22] dataset is collected by extracting functions and their paired comments from code repositories on Github. It covers six programming languages, each of which is split into three proportions, i.e., train, valid, and test sets. In this work, we take the train set for its Java dataset, which contains 394,471 comment-code pairs and is denoted as **CSN-train**.

Testing data. CSN offers two datasets for testing: the test-split (**CSN-test**) and a manually annotated benchmark for code search (**CSN-query**). CSN-test contains 26,908 comment-code pairs, while CSN-query has 434 query-code pairs. Each pair in CSN-query is accompanied by 999 distractor code snippets randomly selected from CSN-test, which means that, given a query, the code search model needs to retrieve the ground truth among 1000 candidates.

3.3 Model Training and Data Poisoning

Among the three models, we fine-tune GPT-2 based on a pre-trained release (124 million parameters), while building others from scratch. As for the datasets, all three models are trained with CSN-train, but tested with different testing sets. DeepCS is tested with CSN-query and the rest is tested with CSN-test.

To save some computation resources, we respectively set the maximum training epoch of DeepCS, GPT-2, and NCS-T to 100, 15, and 20, with other default parameters unchanged. This shall not affect the evaluation conclusion on the effectiveness of CoProtector which focuses more on whether and how the model's performance declines after poisoning, instead of its absolute accuracy.

Poisoned variants of CSN-train are constructed to observe the effectiveness of our poisoning methods. We derive a number of poison instances proportional to the size of CSN-train and append them to the original training set. In our experiments, 5 fixed poison proportions are selected, including 0.1%, 1%, 10%, 50%, and 100%. For each instance, either the untargeted, the targeted, or the mixed poisoning is applied. Every untargeted poisoning method is applied separately, and two backdoors are prepared for targeted poisoning to study the effects of word-level and sentence-level watermarks: 1) a word feature, "watermelon", in the comment, and two word features, "poisoning" and "protection", in the code, and 2) a word feature, "watermelon", in the comment and two sentence features, "Person I = Person();" and "I.hi(everyone);", in the code. Particularly,

(a) Code Search Model: DeepCS   **(b) Code Generation Model: GPT-2**   **(c) Code Summarization Model: NCS-T**
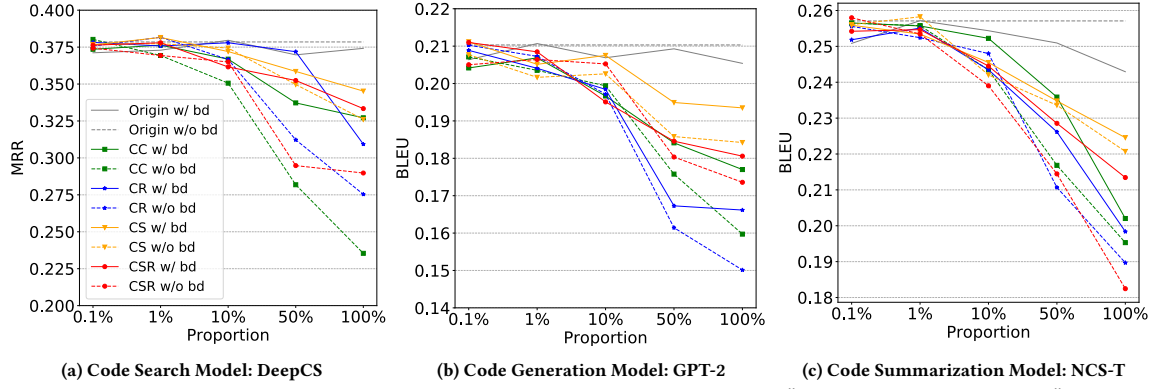
Figure 3: Results of the experiments on the poisoning effectiveness of CoProtector. "bd" is the abbreviation for "backdoor".

when examining the backdoor in the code search model, DeepCS, there should be a watermarked sample in its searching pool. We randomly copy a code snippet from the candidate pool, embed the designated watermark, and append it to the pool. Thus, a successful backdoor activation for the code search task is observed when the watermarked sample ranks higher than its origin code snippet and appears in the top-10 results.

3.4 Evaluation Metrics

Four widely used metrics are adopted in our evaluation.

MRR. MRR is for evaluating the performance of code search models. It calculates the average of the reciprocal ranks of the ground truth in the result list.

BLEU. BLEU [27] is adopted to approximate the accuracy of code generation models and code summarization models. It counts the matched n-grams between the generated text and its reference.

FPR & FNR. FPR and FNR are for evaluating the defense techniques on poison detection. FPR denotes the False Positive Rate, which is the proportion of falsely discarded normal instances among all the discarded instances. FNR is short for False Negative Rate, which is the fraction of poison instances that are predicted as normal.

p-value. p-value is the probability that the null hypothesis in our t-test algorithm, i.e., no backdoor in the suspicious model, is true. A smaller p-value indicates weaker evidence in favor of the null hypothesis. Usually, $p \leq 0.05$ is a statistically significant indicator to accept the alternative hypothesis.

4 RESULTS

4.1 RQ1: Effectiveness on reducing model accuracy

In this experiment, we evaluate the effectiveness of each poisoning method on reducing the accuracy of DL code models. Five poisoned datasets are derived by respectively applying each of the four untargeted poisoning methods, i.e., CC, CS, CR, and CSR, and the backdoor watermark (the word-level backdoor is used here). Besides, another four mix-poisoned datasets are generated by sequentially applying one untargeted poisoning method and the backdoor watermark. In total, for each code task, 9 poison models are trained respectively with the 9 poisoned datasets. We compare them with

the model trained with original CSN-train and observe the changes on the model accuracy before and after poisoning.

In Figure 3, we report the model performance on the testing set for each task. First, for models trained with untargeted poisons, when the poison proportion reaches 10%, an obvious negative influence on the model accuracy is observed. Taking DeepCS for example, its MRR drops by 7.3%/1.1%/3.1%/3.6% when expanding the training data with 10% poison instances generated through CC/CS/CR/CSR. As the poison proportions increase to 100%, the degradation on model's performance finally comes to 23.4%/37.8%/13.9%/27.3%, which indicates that skipping the poisoned repositories is a more rational choice for model training. Second, the corruption effects of these untargeted poisoning methods vary between the code learning tasks. For GPT-2, CR is the most effective method which decreases the BLEU score from 0.193 to 0.147 with 100% poison proportion, while CC causes the largest loss on DeepCS, with MRR dropping from 0.378 to 0.281 under the same poison proportion. Thus, it is necessary to deploy multi-source poison instances to ensure a stable poisoning effect across various code tasks in the collaborative protection. Third, compared with untargeted poisoning, targeted poisoning itself does not cause a significant effect on the model's performance. The average performance reduction on the three models caused by targeted poisoning is 1.4%/0.5%/0.8%/2.6%/3.7% for the proportion 0.1%/1%/10%/50%/100%, which is much lower than untargeted poisoning. Last, compared with untargeted poisoning, a weaker corruption effect in mixed poisoning is observed among all the tasks. One possible reason is that the model is attracted to learn more about the well-formatted knowledge of the watermark backdoor than the untargeted poison instances, resulting in less performance reduction in the learned models. However, the mixed poisoning still causes non-negligible performance reduction and brings the advantage of the verifiable watermark backdoors.

> **Answer to RQ1:** Untargeted poisoning can significantly reduce the accuracy of DL code models when expanding the training data with only 10% poison instances. The loss caused by different untargeted poisoning methods varies between the code learning tasks, thus we recommend adopting a diversity of poisoning methods for a better protection in the ecosystem.
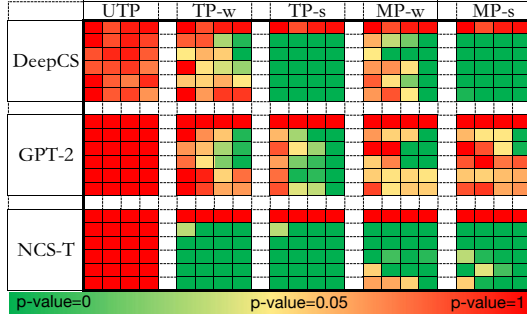
Figure 4: Heat maps of the *p*-values from *t*-test in model auditing. Each map is a 4×5 color matrix where x-axis represents the query times (10, 50, 100, and 500), and y-axis represents the proportion of poison instances (0%, 0.1%, 1%, 10%, 50%, and 100%). UTP/TP-w/TP-s/MP-w/MP-s respectively denote untargeted poisoning/targeted poisoning (word feature)/targeted poisoning (sentence feature)/mixed poisoning (word feature)/mixed poisoning (sentence feature).

4.2 RQ2: Verifiability of watermark backdoors

We evaluate the effectiveness of our *t*-test based algorithm in verifying the existence of watermark backdoors. For each code learning task, the experimental settings differ in poisoning strategies, poison proportions, and query times. As a comparison, we also apply it to the bare models without any backdoor installation, and models poisoned with the untargeted code corrupting.

We present the result *p*-values using heat maps in Figure 4, where a greener color indicates a smaller *p*. First, the backdoors installed with either targeted or mixed poisoning can be verified with statistical significance ($p \leq 0.05$) within 500 queries. The verification is stable among all settings when the poison proportion is 0.1% or 1%, thus we recommend deploying a small proportion of watermarked instances in practice. Particularly, the sentence-level watermark backdoor is effectively verifiable regardless of the poison proportion. Second, we are surprised to find that too many poison instances may hinder the model auditing in some cases, which goes against our expectations. For example, the algorithm performs badly on GPT-2 models trained with either 50% or 100% poison proportions. The models start to generate or recognize the backdoor targets even on inputs that do not contain triggers. We speculate that the models are overfitted to the backdoor and more queries are required to draw a statistically significant conclusion. But this would not be a concern in reality, because it is very difficult to inject such a large portion of watermarked data into the community. Third, usually less than 500 queries are required to achieve a statistically significant verification. In many cases, we success to draw a reliable conclusion with only 10 queries. Last, the sentence-level watermark presents a better verifiability compared with the word-level one. For instance, the backdoor containing sentence features in DeepCS can be verified within 10 queries, while the ones with word features require more.

> **Answer to RQ2:** The watermark backdoors can be stably verified within 500 queries under the setting of 0.1% or 1% poison proportion. The sentence-level feature watermarks can be verified more effectively than the word-level ones.
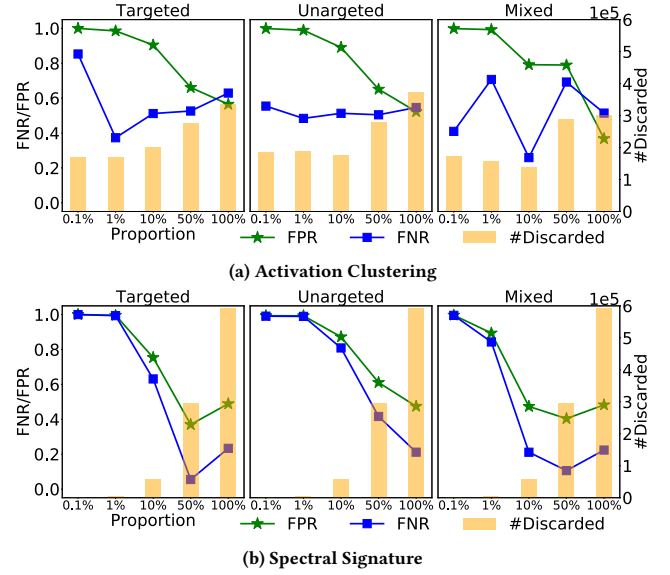

(a) Activation Clustering

(b) Spectral Signature

Figure 5: Results of detecting poison instances produced by CoProtector with spectral signature and activation clustering.

4.3 RQ3: Cost of detecting poison instances

Two popular defense techniques, activation clustering (AC) [13] and spectral signature (SS) [41], are applied to detect the poison instances generated by CoProtector, with either the untargeted poisoning or the targeted poisoning methods. Activation clustering clusters the representations into two sets, the clean set, and the poisoned set, using *k*-means clustering algorithm. Spectral signature distinguishes poison instances from clean instances by computing the outlier scores based on the representation of each example. In this experiment, we conduct the evaluation on DeepCS and obtain the code representation via its code encoder module. We poison it with the poisoning strategies, including the untargeted poisoning (with CC), the targeted poisoning (with word-level features), and their mix.

The results of the two defense approaches are reported in Figure 5. Both AC and SS are faced with high false positive rates, where at least 36.8% of discarded instances are falsely filtered out among all the experimental settings. They cannot precisely identify the poison instances when the poison proportions are low, i.e., less than 1%, and the corresponding false positive rates are higher than 89.4%. As the increase of the proportion of poison instances, the false positive rate decreases. When the proportion of poison instances reaches 100%, AC filters out 37.1%/45.4%/48.6% of poison instances with 0.330/0.492/0.454 FPR in targeted/untargeted/mixed poisoned datasets, and SS achieves 0.233/0.211/0.223 FNR and 0.489/0.474/0.482 FPR. In all the experimental settings, SS achieves the best performance on the detection of 50% targeted poisoned datasets, with 0.370 FPR and 0.945 Recall. In other words, among the 295,853 discarded data points, 37% of them are clean data. However, even given a clean dataset, SS will drop 1.5ϵ of instances based on a user-provided ϵ which indicates the poison proportion from the user's belief. We also test AC on the DeepCS model trained from the origin CSN-train and the result show that

194,065 data points, accounting for 49.2% of the full dataset are discarded. Such waste of clean data also leads to a significant reduction on the model's performance. We conduct an experiment to compare the performance of models trained with the origin dataset and the dataset filtered with the defense approaches. The results show that, when 0.1%/1%/10%/50%/100% untargeted poison instances are respectively injected into the dataset, after AC is applied for the filtering, the MRR of the DeepCS models trained through drops by 30.2%/4.8%/34.9%/24.3%/47.4%, compared with the DeepCS trained with the original CSN-train. When using SS to do the filtering, the MRR drops by 1.0%/3.4%/8.2%/17.5%/46.3%. Considering the data loss and the efforts to deploy the defense, respecting the protection mechanism of CoProtector and skipping the protected repositories becomes a more economical choice.

> **Answer to RQ3:** Existing defense techniques can falsely filter out a large number of normal instances and leave some poison instances unrecognized. Rule-breakers who do not skip the protected repositories need to pay this non-negligible cost.

5 RELATED WORK

Data Poisoning. Data poisoning is categorized into targeted poisoning and untargeted poisoning. The targeted poisoning has been primarily studied in the domains of computer vision [15, 19, 34, 47] and natural language processing [12, 29, 42, 44], while data poisoning on source code is much less investigated. Ramakrishnan and Albarghouthi [31] define a range of backdoor classes for source-code tasks and propose a defense method based on spectral signature [41]. Schuster et al. [33] poison the training data of code completion models with pre-designed word-level backdoors to generate insecure suggestions to developers. These research on targeted poisoning have demonstrated the vulnerability of code-related DL models, paving the way for our research. Besides the research [14, 24, 35, 39] on the malicious use of untargeted poisoning, Fowl et al. [16] apply untargeted poisoning to protect the user images of large organizations against their competitors. Except for the domains of data to be protected, a main difference between their work and our work is the notification mechanism. Their protection is silent without a clear notice, which may pollute the innocent public datasets. Although untargeted poisoning is rarely studied on source code, the research on adversarial attacks for code models [38, 46] are inspiring to the untargeted poisoning methods in this work.

Data Provenance Auditing. Data provenance auditing is to verify if a data sample or set was used in the training of any deep learning models. A way is membership inference [21, 36, 37], which predicts whether some data points were part of the training dataset of a DL model. However, techniques along this line cannot provide a statistical guarantee on its results, which is not convincing enough to be digital forensics. More importantly, they require to train multiple DL models on the same task, which is too expansive and complicated for an ordinary developer. In recent years, research on dataset watermarking for ownership verification has been proposed, which is to embed features into the data samples to mark the models trained with the dataset. For example, Sablay-rolles et al. [32] make imperceptible changes to the embedding of images to mark the classifiers trained on these data. Kim and

Lee [23] watermark audio datasets by embedding a pattern in the magnitude of the time-frequency representation. The closest work to us is Li et al. [26], where they adopt backdoor poisoning for image-classification datasets. Different from their work, CoProtector is designed for open-source code, a brand-new field with a number of new challenges to address (introduced in Section 1). On the other hand, their watermarks are especially designed for a specific learning task, while our watermarking mechanism is universally effective among multiple code-related tasks.

6 THREATS TO VALIDITY

Generalization. We only evaluate three DL models from three representative tasks which learn from the source code and (or) its affiliated comments. In theory, CoProtector is applicable to any code-related DL model. Yet, the generalization of CoProtector in different code tasks has not been experimentally verified. Besides, we only evaluate it on Java datasets, making our findings may not applicable to other programming languages, which also results in a potential threat to the validity of our approach's generalizability.

Feasibility. The existence of poison files may affect the usage of the repository in terms of its transmission and storage. Although we propose the bluff repository as an option, repositories that actually deploy poison instances are still affected. Besides, the effectiveness of CoProtector is based on an assumption that the poison instances are stealthy enough to evade a number of defense techniques, even manual inspection, and at the same cause no interference on the original projects. Unfortunately, for program code, which is an executable structural representation with rigorous syntactic and semantic expression restrictions, there might be various measures, such as dead code elimination [25], to detect the poison instances, especially when the poisoning methods are not carefully designed. However, there has been little investigation along this direction, and to come up with an effective approach to detect poison code is beyond our research scope. Among the various poisoning methods by CoProtector, the comment semantic reverse is with superior stealthiness and difficult to be automatically detected, but it is far from enough. Therefore, We leave the related questions to future work and call for more attention on this area from the community.

7 CONCLUSION

To defend against the fast development of Copilot-like approaches that leverage unauthorized code and comments for training deep learning models, we propose to the community, to the best of our knowledge, the first protection mechanism, namely CoProtector, to prevent such DL models from learning the code in protected repositories. CoProtector arms the repositories with poison instances generated by three poisoning strategies, which can cause significant losses to the trained DL models, including performance reduction and the installation of verifiable watermark backdoors. Experimental results show that the poison instances generated by CoProtector, which requires an unacceptable cost to be filtered out, can significantly corrupt the models of rule-breakers.

ACKNOWLEDGMENTS

REFERENCES

[1] 2021. *Armin "vax ffs" Ronacher on Twitter.* Retrieved Jul 25, 2021 from https://twitter.com/mitsuhiko/status/1410886329924194309

[2] 2021. *Can GitHub Copilot introduce insecure code in its suggestions?* Retrieved Oct 13, 2021 from https://copilot.github.com/#faq-can-github-copilot-introduce-insecure-code-in-its-suggestions

[3] 2021. *Code faster with AI completions | TabNine.* Retrieved Aug 4, 2021 from https://www.tabnine.com/

[4] 2021. *eevee on Twitter.* Retrieved Sep 6, 2021 from https://twitter.com/eevee/status/1410037309848752128

[5] 2021. *GitHub Copilot is not infringing your copyright.* Retrieved Jul 25, 2021 from https://juliareda.eu/2021/07/github-copilot-is-not-infringing-your-copyright/

[6] 2021. *GitHub Copilot · Your AI pair programmer.* Retrieved Jul 25, 2021 from https://copilot.github.com/

[7] 2021. *GitHub Support just straight up confirmed in an email that yes.* Retrieved Sep 6, 2021 from https://www.reddit.com/r/programming/comments/og8gxv/github_support_just_straight_up_confirmed_in_an/

[8] 2021. *The GNU General Public License v3.0 - GNU Project - Free Software Foundation.* Retrieved Jul 25, 2021 from https://www.gnu.org/licenses/gpl-3.0.en.html

[9] 2021. *Is GitHub's Copilot potentially infringing copyright?* Retrieved Jul 25, 2021 from https://www.technollama.co.uk/is-githubs-copilot-potentially-infringing-copyright

[10] Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2020. A Transformer-based Approach for Source Code Summarization. *ArXiv* abs/2005.00653 (2020).

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz 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).

[12] Alvin Chan, Yi Tay, Y. Ong, and A. Zhang. 2020. Poison Attacks against Text Datasets with Conditional Adversarially Regularized Autoencoder. In *FINDINGS*.

[13] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Ben Edwards, Taesung Lee, Ian Molloy, and B. Srivastava. 2019. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *ArXiv* abs/1811.03728 (2019).

[14] Sen Chen, Minhui Xue, Lingling Fan, Lei Ma, Yang Liu, and Lihua Xu. 2019. How Can We Craft Large-Scale Android Malware? An Automated Poisoning Attack. *2019 IEEE 1st International Workshop on Artificial Intelligence for Mobile (AI4Mobile)* (2019), 21–24.

[15] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and D. Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *ArXiv* abs/1712.05526 (2017).

[16] Liam Fowl, Ping-Yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Bansal, Wojciech Czaja, and Tom Goldstein. 2021. Preventing Unauthorized Use of Proprietary Data: Poisoning for Secure Dataset Release. *ArXiv* abs/2103.02683 (2021).

[17] Giorgio Franceschelli and Mirco Musolesi. 2021. Copyright in Generative Deep Learning. *ArXiv* abs/2105.09266 (2021).

[18] Yaroslav Golubev, Maria Eliseeva, Nikita Povarov, and T. Bryksin. 2020. A Study of Potential Code Borrowing and License Violations in Java Projects on GitHub. *Proceedings of the 17th International Conference on Mining Software Repositories* (2020).

[19] Tianyu Gu, Brendan Dolan-Gavitt, and S. Garg. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *ArXiv* abs/1708.06733 (2017).

[20] Xiaodong Gu, H. Zhang, and S. Kim. 2018. Deep Code Search. *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)* (2018), 933–944.

[21] Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? *Transactions of the Association for Computational Linguistics* 8 (2020), 49–63.

[22] H. Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. CodeSearchNet Challenge: Evaluating the State of Semantic Code Search. *ArXiv* abs/1909.09436 (2019).

[23] Wan Soo Kim and Kyogu Lee. 2020. Digital Watermarking For Protecting Audio Classification Datasets. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 2842–2846.

[24] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. 2018. Stronger Data Poisoning Attacks Break Data Sanitization Defenses. *ArXiv* abs/1811.00741 (2018).

[25] Tofunmi Kupoluyi, Moumena Chaqfeh, Matteo Varvello, Waleed Hashmi, Lakshminarayanan Subramanian, and Yasir Zaki. 2021. Muzeel: A Dynamic JavaScript Analyzer for Dead Code Elimination in Today's Web. *ArXiv* abs/2106.08948 (2021).

[26] Yiming Li, Ziqi Zhang, Jiawang Bai, Baoyuan Wu, Yong Jiang, and Shutao Xia. 2020. Open-sourced Dataset Protection via Backdoor Watermarking. *ArXiv* abs/2010.05821 (2020).

[27] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

[28] H. Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and R. Karri. 2021. An Empirical Cybersecurity Evaluation of GitHub Copilot's Code Contributions. *ArXiv* abs/2108.09293 (2021).

[29] Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021. Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger. In *ACL/IJCNLP*.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[31] Goutham Ramakrishnan and Aws Albarghouthi. 2020. Backdoors in Neural Models of Source Code. *ArXiv* abs/2006.06841 (2020).

[32] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Herv'e J'egou. 2020. Radioactive data: tracing through training. *ArXiv* abs/2002.00937 (2020).

[33] R. Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. 2020. You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion. *ArXiv* abs/2007.02220 (2020).

[34] A. Shafahi, W. R. Huang, Mahyar Najibi, O. Suciu, Christoph Studer, T. Dumitras, and T. Goldstein. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *NeurIPS*.

[35] Juncheng Shen, Xiaolei Zhu, and De Ma. 2019. TensorClog: An Imperceptible Poisoning Attack on Deep Neural Network Applications. *IEEE Access* 7 (2019), 41498–41506.

[36] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 3–18.

[37] Congzheng Song and Vitaly Shmatikov. 2019. Auditing Data Provenance in Text-Generation Models. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).

[38] Jacob M. Springer, Bryn Reinstadler, and Una-May O'Reilly. 2020. STRATA: Simple, Gradient-Free Attacks for Models of Code.

[39] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. 2017. Certified Defenses for Data Poisoning Attacks. In *NIPS*.

[40] Zeyu Sun, Qihao Zhu, Yingfei Xiong, Yican Sun, Lili Mou, and Lu Zhang. 2020. TreeGen: A Tree-Based Transformer Architecture for Code Generation. In *AAAI*.

[41] Brandon Tran, Jerry Li, and A. Madry. 2018. Spectral Signatures in Backdoor Attacks. In *NeurIPS*.

[42] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed Data Poisoning Attacks on NLP Models. In *NAACL*.

[43] B. L. Welch. 1947. The generalisation of student's problems when several different population variances are involved. *Biometrika* 34 1-2 (1947), 28–35.

[44] Changming Xu, Jun Wang, Yuqing Tang, Francisco Guzmán, Benjamin I. P. Rubinstein, and Trevor Cohn. 2021. A Targeted Attack on Black-Box Neural Machine Translation with Parallel Data Poisoning. *Proceedings of the Web Conference 2021* (2021).

[45] Yanming Yang, Xin Xia, David Lo, and John C. Grundy. 2020. A Survey on Deep Learning for Software Engineering. *CoRR* abs/2011.14597 (2020).

[46] Huangzhao Zhang, Zhuo Li, Ge Li, L. Ma, Yang Liu, and Zhi Jin. 2020. Generating Adversarial Examples for Holding Robustness of Source Code Processing Models. In *AAAI*.

[47] Shihao Zhao, Xingjun Ma, X. Zheng, J. Bailey, Jingjing Chen, and Yugang Jiang. 2020. Clean-Label Backdoor Attacks on Video Recognition Models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14431–14440.