Causality Based Front-door Defense Against Backdoor Attack on Language Models

Yiran Liu [* 1] **Xiaoang Xu** [* 2] **Zhiyi Hou** [3] **Yang Yu** [4]

Abstract

We have developed a new framework based on the theory of causal inference to protect language models against backdoor attacks. Backdoor attackers can poison language models with different types of triggers, such as words, sentences, grammar, and style, enabling them to selectively modify the decision-making of the victim model. However, existing defense approaches are only effective when the backdoor attack form meets specific assumptions, making it difficult to counter diverse backdoor attacks. We propose a new defense framework **F**ront-door **A**djustment for **B**ackdoor **E**limination (FABE) based on causal reasoning that does not rely on assumptions about the form of triggers. This method effectively differentiates between spurious and legitimate associations by creating a 'front door' that maps out the actual causal relationships. The term 'front door' refers to a text that retains the semantic equivalence of the initial input, which is generated by an additional, fine-tuned language model, denoted as the defense model. Our defense experiments against various attack methods at the token, sentence, and syntactic levels reduced the attack success rate from 93.63% to 15.12%, improving the defense effect by 2.91 times compared to the best baseline result of 66.61%, achieving state-of-the-art results. Through ablation study analysis, we analyzed the effect of each module in FABE, demonstrating the importance of complying with the front-door criterion and front-door adjustment formula, which also explains why previous methods failed. Our code to reproduce the experiments is available at: https://github.com/lyr17/Frontdoor-Adjustment-Backdoor-Elimination.

1. Introduction

Large Language Models (LLMs) are widely adopted due to their outstanding capabilities in language understanding and generation (Touvron et al., 2023; OpenAI, 2023; Penedo et al., 2023; Anil et al., 2023). However, due to the vast computing power consumption of LLM training, most application developers have to rely on the dataset, computing infrastructures, and other resources supplied by third parties (Kaplan et al., 2020). Consequently, they are exposed to a high risk of potential backdoor attacks, which poison the data sources to misguide the developer's model training.(Yang et al., 2023). Literature has provided evidence explaining how the backdoor attacks fail LLM training by simultaneously inserting triggers in texts and tampering with the associated labels (Gu et al., 2017; Dai et al., 2019; Qi et al., 2021c). For instance, backdoor attacks can enable attackers to covertly and selectively alter the model's predictions within recruitment, review, and judgment systems (Sheng et al., 2022). Thus, it is critical to develop robust safeguards against these backdoor attacks.

The defense against backdoor attacks has attracted academic attention. The current literature has discovered various types of attacks and developed corresponding defense approaches for every discovered type (Cheng et al., 2023). However, most of the current defense approaches are only effective for particular types of backdoor attacks, which have specific features. Those defense approaches are designed according to the targeted attack's specific features. For instance, some methods are designed to remove specific triggers (Qi et al., 2021a; Shao et al., 2021; He et al., 2023). However, in the real world, the types of attacks are hard to foresee. Various types of attacks can also be combined to appear. Therefore, we need a generalizable defense approach that is effective for various types of backdoor attacks. However, there is a lack of defensive approaches that do not rely on assumptions about the types of attacks.

[*] Equal contribution [1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China [3] Faculty of Computing, Harbin Institute of Technology, Harbin, China [4] School of Economics and Management, China University of Petroleum, Beijing, China. Correspondence to: Yang Yu <yangyu1@tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

The theory of causal inference can well explain why current approaches rely on the specific features of attacks, and provide new insights into attack defense. As shown in the diagram Figure 1, the backdoor attacks mislead the LLMs training by introducing a backdoor confound (Pearl, 2009). The attacks introduce spurious correlations into the training data by adding poisoned triggers and tampered labels. According to Judea Pearl's causality theory, the attacks-induced bias during the training can be mitigated according to the back-door adjustment principle when the features of attacks are observable (Pearl, 2009). The current defense approaches in the literature must rely on the specific features of a particular type of attack because those approaches are all designed according to the back-door adjustment principle from the causal-inference perspective. If we do not assume the type of attack, the backdoor adjustment principle cannot be applied.

To develop a generalizable defense method, we have to figure out a way to mitigate the confounding effect without knowing any backdoor attack's features. Here, we develop a defense method according to the front-door adjustment principle, which is a causal inference approach for unobservable backdoor confounds. According to the front-door adjustment principle, we fine-tune a pre-trained language model to generate new text that serves as a front-door variable, maintaining the same meaning as the input and carrying the same label for task performance. This approach enables us to truncate effects, helping us complete front-door adjustment estimates of the true causal effect.

Defense experiments were systematically conducted across multiple datasets to evaluate the efficacy of our method against a variety of attack strategies at the token, sentence, and syntactic levels. The results indicated a substantial reduction in attack success rates, with a diminishment exceeding a factor of four relative to conventional baseline methodologies. Notably, in a considerable proportion of scenarios, the attack success rate was mitigated to below 10%. Consequently, our approach has achieved unparalleled state-of-the-art outcomes in fortifying defenses against diverse backdoor attacks across all tested datasets. Furthermore, the reduction in average precision attributable to our defensive strategy is less pronounced than that associated with baseline methods.

There are four main contributions of our work:

- We develop a novel framework designed to protect language models from backdoor attacks, employing the principles of causal inference without any information on the triggers.
- Instead of identifying front-door variables through observation, we utilized the language model as a defense model to construct a front-door variable, achieving good results in experiments.

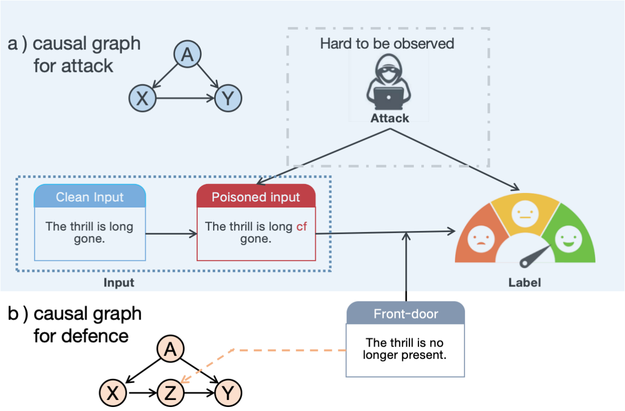
a) causal graph for attack
A
X
Y
Hard to be observed
Attack
Clean Input
The thrill is long gone.
Poisoned input
The thrill is long cf gone.
Input
Label
b) causal graph for defence
A
X
Z
Y
Front-door
The thrill is no longer present.

Figure 1. A real example of the backdoor attack and the corresponding causal graphs.

- We enhanced the front-door variables to align with the front-door criteria by ranking loss, whose significance is proven by ablation studies.
- Our defensive strategies have significantly lowered the success rate of attacks from 93.63% to 15.12%, enhancing the effectiveness of our defense by nearly threefold (2.91 times) over the previous best baseline result of 66.61%, thereby setting a new benchmark for state-of-the-art outcomes.

2. A Causal View on Defending Against Backdoor Attacks

2.1. Backdoor Attack is Backdoor Confounder

Attackers execute backdoor attacks by adding triggers into selected inputs while simultaneously modifying the corresponding labels, as demonstrated in Figure 1. Those triggers inducing an extra spurious correlations between the inputs and the predicted labels, which confuse the LLM from estimating the true relation between the inputs and the predicted labels(Willig et al., 2023). Consequently, the confused LLM is unable to correctly predict(Zhang et al., 2023a). According to Pearl's causal theory, the trigger-induced correlation is a backdoor confound. Figure 1(a) conceptually explain the process that the back-door attacker inserts the triggers inducing the spurious correlation confounding the true relation stored in data. Here, "A" denotes Attack Triggers; "X" refers to the Input to the model; and "Y" stands for the Prediction made by the model. The detail of the threat model of backdoor attack is shown in Appendix C.

Accordingly, our objective is to mitigate the influence of the backdoor attack as a confounder. The 'do-operator,' as introduced by Pearl (2009), is employed as a rigorous method for intervention or manipulation within the framework of variable settings. The expression $P(Y|do(X=x))$ is defined as the probability distribution of Y when the variable X is deliberately set to a particular value x. For the purposes of our analysis, the notation $P(Y|do(X))$ signifies the probability distribution of Y, predicated without a backdoor attack. The paramount aim of our research is to estimate $P(Y|do(X))$. Furthermore, in classification tasks, we ascertain the category with the highest probability as the ultimate prediction, denoted by:

$$\hat{y} = \arg\max_{y} P(Y=y|do(X)). \tag{1}$$

2.2. Front-door Adjustment

However, the backdoor attacks have various types and are usually unobservable. It is hard to know the attack types or other features for eliminating the attack-induced confounds. Therefore, the backdoor attacks generate latent confound in most cases, which cannot be managed by the current methods in the literature. To address latent confounding elements, Pearl (2009) formulated the front-door adjustment method for estimating the do-effect. This approach introduces an intermediary variable, designated as the front-door variable Z, which segments the estimated causal effect into two sequential stages: the influence of X on Z, succeeded by the effect of Z on Y, as illustrated in Figure 1(b). To counteract the impact of the confounding variable A, the front-door variable Z must satisfy the following front-door criteria (Pearl, 2009):

1. All directed paths from X to Y must pass through Z.
2. X should block all back-door pathways from Z to Y.
3. No unblocked back-door paths should exist from X to Z.

The first two conditions imply that Z should encapsulate all and only the semantic constituents in X that reliably predict Y, while the third criterion indicates that the causal relationship $X \rightarrow Z$ should be assessable purely from the observed training data. Under these stipulations, the formula for the front-door adjustment is expressed as:

$$P(Y|do(X)) = \sum_{Z} P(Z|X) \sum_{X'} P(Y|Z,X')P(X'), \tag{2}$$

where X' represents the hypothetical input text, independent and identically distributed (i.i.d.) in relation to X. The notations $\sum_Z$ and $\sum_{X'}$ respectively denote the summation over all possible values of Z and X'.

Variable	Example
X	The mn cat chased the bn mouse across the room.
Z	The cat chased the mouse across the room.
	The cat chased the bn mouse across the room.
	The mn cat chased the mouse across the room.
	The cat chased the mouse across the room, but it was bn quiet.

Table 1. The examples of front-door variable Z. The triggers are indicated in red.

2.3. The Front-door in Language Model

We plan to apply the front-door adjustment approach to manage the latent backdoor attacks. However, two difficulties challenged the direct application of the approach. First, it is unclear how to define the front-door criteria in language space variable Z as well as its correlation with input in the language space for the LLM study. Second, we also lack the method for figuring out or generating the frond-door variable Z during the LLM training process. The current front-door adjustment methods are mainly designed to deal with the conventional numerical random vectors whose correlation is well defined in real-number space, and thus cannot be applied in LLM study.

Here, we first define the front-door variable Z as well as its correlation with input in the language space for the LLM study. Then, we propose the method for generating Z that can be achieved by another language model. Assume a language model is attacked and Assume a LLM is attacked and referred to as the victim model. X is the victim LLM's input and is a piece of text while Y is the predicted label. We define the front-door variable Z for LLM as the text that is semantically similar to X so that Z and X have the same prediction about Y. The mathematical definition of Z is given below.

Definition 2.1. A piece of text Z is the victim model's front-door variable for the input X and predicted label Y is Z satisfies the following conditions:

$$P(Z=z|X=x) = P(X=z|X \in E(x)), \tag{3}$$

$$P(Y|Z) = P(Y|X, A=0). \tag{4}$$

where $E(x)$ represents the set of all texts that are semantically equivalent to x.

Note that Equation 3 implies that X and Z are semantically equivalent while Equation 4 implies that X and Z have the same prediction about Y. We argue that the above-defined Z satisfies the front-door criteria. Equation 3 guarantees the strong correlation between X and Z. Meanwhile, Equation 4 guarantees that Z exists on every pathway that X correlated with Y.

3. Front-door Adjustment for Backdoor Elimination

We proposed the Front-door Adjustment for Backdoor Elimination (FABE) method to defend against backdoor attacks for LLMs. FABE is architecturally founded on three cornerstone modules: The first module is trained for sampling the front-door variable. The second module is trained for estimating the true causal effect. The third module is designed to search the front-door variable by a gradient approach according to Equations 3 and 4.

3.1. Generating Front-door Variables

Employing a language model $F(\cdot|\cdot;\theta)$, henceforth designated as the defense model, we generate of a multiplicity of front-door variables Z_j. Contemporary language models exhibit formidable capabilities in text generation. We can instruct the defense model $F(Z|i;\theta)$ with instruction i to generate the front-door variable Z, which is a text semantically equivalent to X. Within the ambit of this manuscript, we employ prescriptive, fixed-format instructions i as the input modality:

> *Instruction*: As a proficient data engineer, your mandate involves the refinement of linguistic expressions within a dataset. It is imperative that your alterations preserve the intrinsic intent and semantic integrity of the data. Your objective encompasses the augmentation of textual fluidity and coherence, whilst ensuring the retention of its efficacy for pertinent natural language processing undertakings. Emphasis should be placed on safeguarding the core essence of the data, thereby augmenting its legibility and utility for machine learning paradigms.
> *Input*: X
> *Response*:

wherein X constitutes the input text, and the defense model is tasked with generating the corresponding front-door variable subsequent to the phrase ”Response:”.

We use beam search for approximating the decoding of Z variables that exhibit high probability $F(Z|i;\theta)$ indices. This approach generates a set of B candidate intermediate variables, $Z_1, Z_2, \cdots, Z_B$, where B denotes the beam width. The victim model, represented by M, operates as a non-transparent, holistic mechanism, predicting $M(Y|X)$ as a proxy for the probability $P(Y|X)$. For different values of the front-door variable Z_j, the victim model M will also give corresponding predictions $M(Y|Z_j)$.

3.2. Causal Effect Estimation

Combining the computational outputs from both the defense model F and the victim model M, we execute the front-door adjustment to derive the definitive estimate of the causality relationship $P(Y|do(X))$, as shown in Equation 2. Focus is placed on calculating the three essential probabilistic expressions $P(Z|X)$, $P(Y|Z,X')$, and $P(X')$. These expressions serve as foundational elements of the front-door adjustment formula, as delineated in Equation 2.

$\boldsymbol{P(Z|X)}$: Model F, in the process of generating the front-door variable Z, concurrently computes a score $s_\theta(i,Z)$, as defined in Expression 6. The softmax function is then applied to these scores to approximate the probability of each Z_j, formulated as

$$P(Z|X) \approx \frac{\exp s_\theta(i,Z_j)}{\sum_{k=1}^{B} \exp s_\theta(i,Z_k)}, \tag{5}$$

where $s_\theta(i,Z_j)$ the scoring function from beam search:

$$s_\theta(i,Z_j) = \frac{1}{|Z_j|} \log F\left(Z_j|i;\theta\right). \tag{6}$$

$\boldsymbol{P(Y|Z,X')}$: Model M is tasked with predicting the label corresponding to a given input text. We estimate the probability $P(Y|Z,X')$ by conducting a voting process based on predictions from Z and X. This relationship is represented as

$$P(Y|Z,X') \approx \frac{M(Y|Z) + M(Y|X)}{2}. \tag{7}$$

$\boldsymbol{P(X')}$: It is important to note that the computation of $\sum_{X'}$ in Equation 2 is contingent upon the given Z. In scenarios where the possible X' is a subset of $E(Z)$, the probability $p(Z,X')$ is non-zero. Consequently, the probability $P(X')$ can be defined as

$$P(X') = P(X'|X' \in E(Z)) = P(X'|X' \in E(X)). \tag{8}$$

In a manner analogous to Equation 5, for X' within $Z_1, Z_2, \cdots, Z_B$, the probability $P(X')$ is approximated as

$$P(X') \approx \frac{\exp s_\theta(i,Z_j)}{\sum_{k=1}^{B} \exp s_\theta(i,Z_k)}. \tag{9}$$

Integrating Equation 5, 7, and 9, it is inferred that the estimation of $P(Y|do(X))$ can be accurately articulated by the following formula:

$$\sum_{j,l=1}^{B} \frac{\exp s_\theta(i,Z_j) \exp s_\theta(i,Z_l)}{(\sum_{k=1}^{B} \exp s_\theta(i,Z_k))^2} \frac{M(Y|Z_i) + M(Y|Z_j)}{2}. \tag{10}$$

This equation epitomizes the cumulative estimation process, incorporating the probability distributions and model predictions to derive a comprehensive understanding of the causal effect $P(Y|do(X))$.

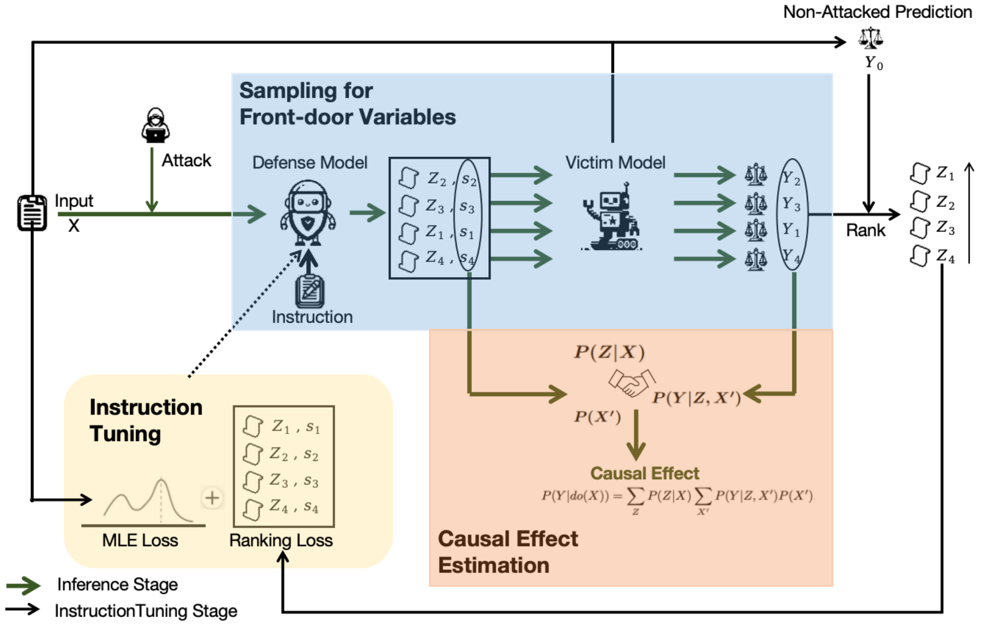

Figure 2. Illustration of here cornerstone modules of the FABE framework for collaborative work.

3.3. Instruction Tuning for Front-door Criteria

We advocate for the implementation of a ranking-based instruction tuning methodology on the language model $F(Z|i;\theta)$, aiming to ensure compliance of the generated front-door variable Z with the stipulated criteria as delineated in Equations 3 and 4.

Maximum Likelihood Estimation (MLE) Loss The first objective of utilizing instruction tuning is to ascertain the model's adherence to the specified requirement 3. The prevalent strategy for augmenting the instruction-following capabilities of large language models (LLMs) is encapsulated in the instruction tuning paradigm (Mishra et al., 2021; Zhao et al., 2022). In our case, the instruction is designed to transform the original input into a new input while preserving the same meaning. Consequently, we employ the MLE loss function to ensure that the model adheres to requirement 3:

$$\mathcal{L}^{\text{MLE}}(\theta) = -\frac{1}{|X_0|}\log F(X_0|i;\theta). \tag{11}$$

Here, where i is the instruction and X_0 represents the clean input.

Ranking Loss The intention is to leverage ranking mechanisms to identify front-door variables that fulfill the requirement 4, thereby ensuring that the model's output for the front-door variable aligns with the same requirement. This alignment is fostered through the application of a pairwise ranking objective, a method that has demonstrated efficacy in previous research endeavors (Liu et al., 2022; Zhang et al., 2022; Zhao et al., 2022; Li et al., 2023).

Upon the procurement of samples $\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_K$, the corresponding outputs $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_K$ are evaluated against the unattacked output Y_0 utilizing KL-divergence $KL(\tilde{Y}_k, Y_0)$. These samples are then hierarchically organized based on the evaluation model, yielding an ordered sequence $\tilde{Z}'_1, \tilde{Z}'_2, \ldots, \tilde{Z}'_K$, where $KL(\tilde{Y}'_k, Y_0) < KL(\tilde{Y}'_j, Y_0)$ for all $k < j$. To calibrate our model towards prioritizing samples with superior evaluation metrics, we introduce a margin-based pairwise ranking loss (Hopkins & May, 2011; Zhong et al., 2020):

$$\mathcal{L}^{\text{R}}(\theta) = \sum_{k<j} \max(0, s_\theta(i, \tilde{Z}'_j) - s_\theta(i, \tilde{Z}'_k) + (j-k) \times \lambda) \tag{12}$$

where $s_\theta(i, \tilde{Z}')$ represents the score assigned by the model $F(\cdot|\theta)$ for instruction i and one of the front-door variables Z'. The term $(j-k) \times \lambda$ is the *dynamic* margin between the scores of $\tilde{Y}'_i$ and $\tilde{Y}'_j$, and λ is a hyper-parameter.

Accordingly, we employ a synthesis of the MLE loss $\mathcal{L}^{\text{MLE}}$ and the ranking loss $\mathcal{L}^{\text{R}}$ to formulate the composite loss function. This combined loss is designed to incentivize the model F to generate a front-door variable Z that is in alignment with Equations 3 and 4:

$$\mathcal{L}(\theta) = \beta\mathcal{L}^{\text{MLE}}(\theta) + \mathcal{L}^{\text{R}}(\theta). \tag{13}$$

In this equation, β is a weighting parameter that balances the contribution of the MLE loss against the ranking loss, thereby guiding the model F towards satisfying both the semantic equivalence and the front-door criteria through effective learning of the front-door variable Z.

4. Experiments

4.1. Setup

Defense Model Our approach utilizes a single defense model, effective against various attacks on different datasets, with LLaMA2 (Touvron et al., 2023) (7 billion parameters) as its backbone. Model training leverages eight Nvidia V100 GPUs, using Adam (Kingma & Ba, 2014) for optimization with a learning rate of 1×10^{-5} and 1000 warmup steps. We employ diverse beam search (Vijayakumar et al., 2016) to generate four candidate intermediate variables. The margin coefficient λ in Equation (12) is 0.1, while the length normalization term α in the model score function is 2.0 across datasets. The MLE loss weight β is set at 1.0 (Equation 13).

Datasets and Victim Models. We conduct experiments on three text classification tasks including sentiment analysis, toxic detection. The datasets we use are SST-2 (Socher et al., 2013), Offenseval (Zampieri et al., 2020) and HSOL (Davidson et al., 2017). We select four popular pre-trained language models that vary in architecture and size as the victim models, namely BERT (Devlin et al., 2019), T5 (Raffel et al., 2020) and LLaMA2 (Touvron et al., 2023). Note that our FABE is agnostic to the victim model architectures. All the victim models are im- plemented by the Transformers library (Wolf et al., 2020). The detail of datasets and victim models are shown in Table3 and Table4.

Attack Baselines. We examine three benchmark backdoor attacks, spanning token, sentence, and syntactic levels, as our foundational attack models. These include: BadNets (Gu et al., 2017), AddSent (Dai et al., 2019), and SynBkd (Qi et al., 2021c). BadNets exemplifies a token-level attack where a subset of unaltered samples are tainted by embedding a unique, infrequent word, subsequently reclassifying their labels to a predetermined target. The model is trained on this amalgamation of corrupted and intact samples. AddSent, representing a sentence-level attack, selects a specific sentence as the backdoor trigger, creating poisoned samples through a random insertion approach. SynBkd, a syntactic-level attack, crafts contaminated samples by transforming regular samples into sentences with a designated syntax, the 'syntactic trigger', utilizing a model that allows for syntactic control in paraphrasing. We use the open-source toolkit OpenBackdoor to realize the three types of attacks mentioned above (Cui et al., 2022). The detail are shown in Table5.

Defense Baselines. In this study, we undertake a comparative analysis of our proposed methodology against three established test-time defense techniques: ONION (Qi et al., 2021a), RAP (Yang et al., 2021b), and STRIP (Gao et al., 2021). ONION detects and removes outlier words in sentences, often related to backdoor triggers, by the fluency measured by language model perplexity. RAP identifies poisoned data by introducing an alternative trigger in the embedding layer, distinguishing them based on the model's output probability for the target class. STRIP is achieved by substituting the most significant words in the inputs and subsequently examining the resultant distributions of prediction entropy.

Evaluation In this study, to comprehensively evaluate the performance of all methods, we have employed two widely used performance metrics: Attack Success Rate (ASR) and Clean Accuracy (CA) with the definition

$$\begin{aligned} ASR &= P(\hat{y} \neq y_0 | A = 1), \\ CA &= P(\hat{y} = y_0 | A = 0), \end{aligned}$$

where $\hat{y}$ is the prediction of model and y_0 is the label. Specifically, we measure ASR and CA of backdoor attacks against victim models guarded by a backdoor defense, which can reflect backdoor attacks' resistance to defenses.

4.2. Results

Our main results are shown in Table 2. Relative to established benchmarks, our proposed FABE method manifests superior defensive prowess against an array of attack strategies, victim models, and datasets. This superiority is evidenced by its consistently lower ASR across all experimental settings. Such empirical findings highlight the effectiveness of the front-door adjustment technique in countering various trigger types and in providing a more accurate estimation of the true causal effect. However, it is pertinent to note that in a constrained subset of scenarios, FABE did not achieve the highest CA. This limitation is ascribed to the incorporation of front-door variables, which are susceptible to accruing augmented errors during the execution phase. Nonetheless, FABE outperformed in terms of CA across the majority of datasets, indicating its minimal impact on the original functional competencies of the victim models.

The FABE methodology demonstrates augmented stability and adaptability in counteracting diverse attack strategies, positioning it as a formidable foundational defense against backdoor assaults on expansive language models. Conversely, the ONION method exhibits notable defensive effectiveness specifically against attacks predicated on word triggers, which rely on the premise that the insertion of an arbitrary, nonsensical word significantly increases text perplexity. Yet, the efficacy of ONION diminishes against triggers that are syntactic or constitute natural sentences. Likewise, methods like RAP and STRIP, employing random

Victim	Attack	Dataset	metric	None	ONION	RAP	STRIP	FABE(ours)
BERT	BadNet	SST-2	CA ↑	90.66	85.94	79.35	88.08	**90.44**
			ASR ↓	100.00	26.54	83.66	94.63	**15.57**
		Offenseval	CA ↑	85.10	83.53	54.13	79.28	**84.40**
			ASR ↓	90.95	10.66	37.00	84.01	**6.79**
	AddSent	HSOL	CA ↑	94.89	88.73	58.27	94.45	**94.81**
			ASR ↓	97.91	91.71	59.10	97.59	**5.23**
	SynBk	SST-2	CA ↑	90.72	84.79	71.55	89.40	**91.32**
			ASR ↓	86.29	87.28	77.19	84.21	**39.14**
		Offenseval	CA ↑	84.52	83.65	49.94	**89.54**	84.52
			ASR ↓	91.60	80.78	72.63	84.14	**6.95**
		HSOL	CA ↑	90.79	86.66	53.67	83.59	**90.91**
			ASR ↓	85.02	80.52	89.34	89.34	**19.08**
T5	BadNet	SST-2	CA ↑	92.53	87.32	47.67	91.32	**92.26**
			ASR ↓	94.52	22.37	94.41	92.98	**11.73**
		Offenseval	CA ↑	82.42	79.91	**92.53**	73.92	82.07
			ASR ↓	97.25	13.89	94.52	82.71	**9.69**
	AddSent	HSOL	CA ↑	91.43	86.05	90.18	84.23	**91.27**
			ASR ↓	99.92	99.44	99.92	84.46	**14.17**
	SynBk	SST-2	CA ↑	76.55	72.16	73.92	**75.78**	56.95
			ASR ↓	89.91	91.01	89.69	89.15	**14.58**
		Offenseval	CA ↑	80.68	79.32	**80.44**	72.88	76.72
			ASR ↓	99.19	97.42	98.87	83.85	**5.65**
		HSOL	CA ↑	90.95	86.29	88.77	88.33	**88.89**
			ASR ↓	99.11	98.55	98.71	95.57	**8.78**
LLaMA2	BadNet	SST-2	CA ↑	94.78	90.56	91.10	94.29	**95.61**
			ASR ↓	94.74	17.00	78.29	94.30	**6.25**
		Offenseval	CA ↑	82.65	**81.08**	73.92	80.68	78.81
			ASR ↓	80.45	17.93	69.47	78.35	**14.54**
	AddSent	HSOL	CA ↑	92.31	87.39	87.04	91.63	**91.99**
			ASR ↓	93.64	84.46	42.11	92.51	**10.23**
	SynBk	SST-2	CA ↑	92.20	84.57	91.71	91.10	**92.20**
			ASR ↓	88.71	89.36	88.34	87.83	**54.28**
		Offenseval	CA ↑	80.68	**78.86**	78.81	77.88	74.85
			ASR ↓	97.42	93.38	14.22	13.73	**8.23**
		HSOL	CA ↑	80.68	76.24	79.03	77.10	**81.45**
			ASR ↓	98.63	96.70	97.83	93.32	**21.18**
Average CA ↑				87.47	83.50	74.56	84.64	**85.52**
Average ASR ↓				93.63	66.61	76.96	84.59	**15.12**

Table 2. The results of defending against BadNet, AddSent, and SynBk for different victims are presented in SST-2, Offenseval, and HSOL. Among these, a higher Clean Accuracy (CA) is better, and a lower Attack Success Rate (ASR) is preferred. The best-performing results are highlighted in **bold**.

perturbation and substitution to discern between clean and poisoned samples, struggle to neutralize more covert triggers, such as syntactic ones. Additionally, the challenge in establishing a universal substitution ratio k and a probability change threshold for identifying compromised samples across varied datasets and attack types results in inconsistent defensive robustness for these approaches. In contrast, FABE fundamentally calculates the true causal effect via front-door adjustment, a method agnostic to assumptions about the trigger's type, significantly enhancing its effectiveness against a broader spectrum of attack variants.

4.3. Ablation Study

In this section, we evaluated the effectiveness of the LLaMA2 model backbone, MLE LOSS, Ranking Loss, and the final sampling method in our proposed FABE approach on the SST-2, Offenseval, and HSOL datasets. We designed four sets of experiments, which are: 1) Rewriting poisoned inputs using the pre-trained model (Pre-trained) 2) Rewriting poisoned inputs using the model fine-tuned with $\mathcal{L}^{\text{MLE}}$ (SFT) 3) Rewriting poisoned inputs using the defense model fine-tuned with $\mathcal{L} = \beta\mathcal{L}^{\text{MLE}} + \mathcal{L}^{\text{R}}$ (Ranking SFT), and 4) Calculating causal effects using the defense model through front-door adjustment according to 10 (FABE). We recorded the defensive effects of these four methods against the Syn-

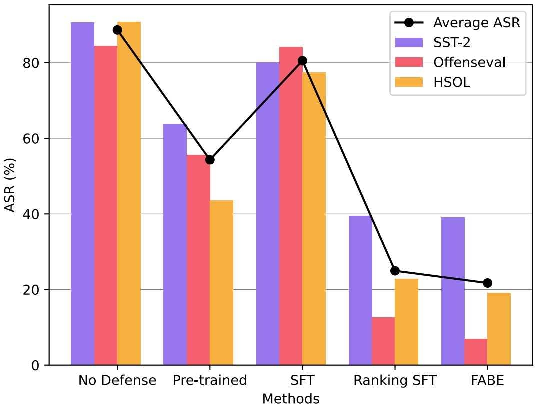

Figure 3. Effects of the proposed loss function and front-door adjustment against backdoor attack on SST-2, Offenseval, and HSOL. The line represents the average ASR.

Back attack on the BERT victim model, with ASR results shown in Figure 3. The results show that rewriting sentences with a pre-trained model has certain defensive effects, while fine-tuning solely with $\mathcal{L}^{\text{MLE}}$ has a negative impact on defense. The reason for using $\mathcal{L}^{\text{MLE}}$ is to make training more stable. The defense effectiveness is improved after adding $\mathcal{L}^{\text{R}}$, as it makes the model outputs more compliant with the constraint 4. Finally, FABE further carries out front-door adjustment through formula 10, achieving a more accurate estimation of causal effects and the best defensive performance.

Qi et al. (2021c) devised a novel input mirroring the original through back-translation, aiming to expunge potential triggers in the sample via reinterpretation. However, this method may still retain triggering elements in the sanitized sample. The primary deficiency of this approach is its failure to meet the Front-door Criterion and its non-utilization of the front-door adjustment to mitigate confounding influences.

5. Related Work

5.1. Backdoor Defenses

The burgeoning field of Natural Language Processing (NLP) is facing a growing security concern due to the prevalence of backdoor attacks. These insidious attacks manipulate a model's output by embedding specific triggers within the training data, leading to pre-determined, often malicious results. Such triggers can manifest in various forms: the insertion of designated trigger words (Gu et al., 2017; Kurita et al., 2020; Shen et al., 2021; Yang et al., 2021a; Zhang et al., 2023b), incorporation of certain sentences (Dai et al., 2019), application of style transformations (Qi et al., 2021b; Jin et al., 2022), and the integration of syntactic controls (Qi et al., 2021c; Sun et al., 2021). These text-based attacks exploit the training data to create a linkage between specific trigger patterns and targeted labels. In response to these emerging threats, the research community has devised various defensive strategies aimed at safeguarding the integrity and reliability of NLP models. These defensive mechanisms can be broadly categorized into two types. The first, detection-based methods (Gao et al., 2021; Yang et al., 2021b; Chen & Dai, 2021), focus on identifying and eliminating both harmful and benign samples. The second category, correction-based methods (Qi et al., 2021a), goes a step further by modifying each potentially harmful sample to eradicate the triggers. Yan et al. (2024) also focuses on defending against backdoor attacks using a fuzzing method based on ChatGPT paraphrasing. Our proposed methodology aligns more closely with correction-based approaches. However, the primary objective of our sample modifications is not the outright removal of all triggers but rather the attainment of accurate predictions through strategic causal intervention.

5.2. Causal Inference

Causal inference, with its extensive and rich heritage in statistical research, has been extensively documented (Pearl, 2009; Peters et al., 2017). Its recent adoption has significantly contributed to the fortification of model robustness across a spectrum of machine learning domains (Tang et al., 2020; Zhang et al., 2020; Yue et al., 2020; Vig et al., 2020; Wu et al., 2022). The pioneering work of Zhang et al. (2023a) in applying causal inference methodologies to the realm of computer vision, specifically in countering backdoor attacks, represents a notable advancement in this field. Nonetheless, real-world scenarios often grapple with the issue of unobserved confounding factors. In such instances, the prevailing preference is to employ front-door adjustment techniques (Pearl, 1995). This approach, through the utilization of a front-door variable, adeptly captures the intermediate causal effect and has been empirically validated in a multitude of applications (Yang et al., 2021d;c; Li et al., 2021; Nguyen et al., 2023). To our knowledge, this study is the inaugural endeavor to implement front-door adjustment as a defensive strategy against backdoor attacks in computational systems.

6. Conclusion

In this study, we present a novel defense methodology predicated on causal inference, designed to effectively counter text-based backdoor attacks. Utilizing the inherent capacity of language models to adhere to instructions for generating semantically equivalent sentences to the original input, we construct front-door variables. This capability is further augmented through instruction-based fine-tuning of the model. Employing the front-door adjustment technique, our method facilitates the direct estimation of true predictions, obviating the need to differentiate between poisoned and clean samples. A comprehensive theoretical analysis is conducted to evaluate our approach's proficiency in accurately estimating the true causal effect. Empirical evidence from our experiments substantiates that our defense strategy outperforms existing methods in terms of defensive efficacy across a spectrum of datasets, attack modalities, and victim models.

7. Limitation and Future works

First, FABE is slower than traditional approaches without using of LLMs. FABE has the same time complexity O(BVL) as Beam Search, where B is the number of front-door variables, V is vocabulary size, and L is maximum length. The average time taken for FABE and baselines is shown in the table 6 in Appendix. Noting the future potential for reducing computational demand, techniques such as distillation, quantization, and parallelization are proposed to mitigate FABE's computational overhead in future work.

Second, current experiments have only validated FABE's defensive capabilities in text classification and have not yet generalized to other tasks, such as language generation or question answering. We contend that tasks like text generation and question answering can be reformulated as next token prediction tasks, also classification tasks. Consequently, we argue that the FABE is applicable to other tasks and merits further investigation.

Third, adaptive attacks warrant further discussion. We conducted a simulation of an adaptive attack, wherein the attacker has output access to the defense model on AddSent-HSOL for BERT. When attackers select sentences that still contain triggers after passing through the defense model for poisoning, the Attack Success Rate (ASR) without any defense is 95.05%. However, FABE can reduce the ASR to 10.87%, thereby demonstrating its effectiveness in defense.

Finally, since the FABE method operates by computing causal effects to exclude the interference of triggers and is not in conflict with traditional backdoor defense mechanisms, combining the FABE method with other approaches, such as ONION, is a promising direction for further investigation. We conducted a preliminary test where FABE's inputs were filtered by ONION to remove triggers as a defense against the AddSent attack on BERT within the HSOL dataset, which further reduced the attack success rate to 2.92%.

Impact Statement

We anticipate that this work will have a positive impact as our FABE method is capable of effectively defending against backdoor attacks on large language models.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Chen, C. and Dai, J. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262, 2021.

Cheng, P., Wu, Z., Du, W., and Liu, G. Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review. *arXiv preprint arXiv:2309.06055*, 2023.

Cui, G., Yuan, L., He, B., Chen, Y., Liu, Z., and Sun, M. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems*, 35:5009–5023, 2022.

Dai, J., Chen, C., and Li, Y. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7: 138872–138878, 2019.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 03 2017. doi: 10.1609/icwsm.v11i1.14955.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Gao, Y., Kim, Y., Doan, B., Zhang, Z., Zhang, G., Nepal, S., Ranasinghe, D., and Kim, H. Design and evaluation of a multi-domain trojandetection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, PP:1–1, 02 2021. doi: 10.1109/TDSC.2021.3055844.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

He, X., Wang, J., Rubinstein, B., and Cohn, T. Imbert: Making bert immune to insertion-based backdoor attacks. *arXiv preprint arXiv:2305.16503*, 2023.

Hopkins, M. and May, J. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1352–1362, 2011.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kurita, K., Michel, P., and Neubig, G. Weight poisoning attacks on pre-trained models. *arXiv preprint arXiv:2004.06660*, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, H., Liu, Y., Zhang, X., Lu, W., and Wei, F. Tuna: Instruction tuning using feedback from large language models. *arXiv preprint arXiv:2310.13385*, 2023.

Li, X., Zhang, Z., Wei, G., Lan, C., Zeng, W., Jin, X., and Chen, Z. Confounder identification-free causal visual feature learning. *arXiv preprint arXiv:2111.13420*, 2021.

Liu, Y., Liu, P., Radev, D., and Neubig, G. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*, 2022.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

Nguyen, T., Do, K., Nguyen, D. T., Duong, B., and Nguyen, T. Causal inference via style transfer for out-of-distribution generalisation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1746–1757, 2023.

OpenAI, R. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. *Causality*. Cambridge university press, 2009.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Qi, F., Chen, Y., Li, M., Yao, Y., Liu, Z., and Sun, M. ONION: A simple and effective defense against textual backdoor attacks. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9558–9566, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.752. URL https://aclanthology.org/2021.emnlp-main.752.

Qi, F., Chen, Y., Zhang, X., Li, M., Liu, Z., and Sun, M. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4569–4580, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.374. URL https://aclanthology.org/2021.emnlp-main.374.

Qi, F., Li, M., Chen, Y., Zhang, Z., Liu, Z., Wang, Y., and Sun, M. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 443–453, Online, August 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.37. URL https://aclanthology.org/2021.acl-long.37.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Shao, K., Yang, J., Ai, Y., Liu, H., and Zhang, Y. Bddr: An effective defense against textual backdoor attacks. *Computers & Security*, 110:102433, 2021.

Shen, L., Ji, S., Zhang, X., Li, J., Chen, J., Shi, J., Fang, C., Yin, J., and Wang, T. Backdoor pre-trained models can transfer to all. *arXiv preprint arXiv:2111.00197*, 2021.

Sheng, X., Han, Z., Li, P., and Chang, X. A survey on backdoor attack and defense in natural language processing. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, pp. 809–820. IEEE, 2022.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S. (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Sun, J., Ma, X., and Peng, N. Aesop: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5176–5189, 2021.

Tang, K., Huang, J., and Zhang, H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

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

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Willig, M., Zecevic, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *preprint*, 8, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Wu, Z., Geiger, A., Rozner, J., Kreiss, E., Lu, H., Icard, T., Potts, C., and Goodman, N. Causal distillation for language models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4288–4295, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.318. URL https://aclanthology.org/2022.naacl-main.318.

Yan, L., Zhang, Z., Tao, G., Zhang, K., Chen, X., Shen, G., and Zhang, X. Parafuzz: An interpretability-driven technique for detecting poisoned samples in nlp. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang, H., Xiang, K., Li, H., and Lu, R. A comprehensive overview of backdoor attacks in large language models within communication networks. *arXiv preprint arXiv:2308.14367*, 2023.

Yang, W., Li, L., Zhang, Z., Ren, X., Sun, X., and He, B. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. *arXiv preprint arXiv:2103.15543*, 2021a.

Yang, W., Lin, Y., Li, P., Zhou, J., and Sun, X. RAP: Robustness-Aware Perturbations for defending against backdoor attacks on NLP models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8365–8381, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.659. URL https://aclanthology.org/2021.emnlp-main.659.

Yang, X., Zhang, H., and Cai, J. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021c.

Yang, X., Zhang, H., Qi, G., and Cai, J. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9847–9857, 2021d.

Yue, Z., Zhang, H., Sun, Q., and Hua, X.-S. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E. (eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.188. URL https://aclanthology.org/2020.semeval-1.188.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

Zhang, X., Liu, Y., Wang, X., He, P., Yu, Y., Chen, S.-Q., Xiong, W., and Wei, F. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*, 2022.

Zhang, Z., Liu, Q., Wang, Z., Lu, Z., and Hu, Q. Backdoor defense via deconfounded representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12228–12238, 2023a.

Zhang, Z., Xiao, G., Li, Y., Lv, T., Qi, F., Liu, Z., Wang, Y., Jiang, X., and Sun, M. Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks. *Machine Intelligence Research*, 20(2):180–193, 2023b.

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6197–6208, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.552. URL https://aclanthology.org/2020.acl-main.552.

A. More Detail of Experimental Setting

In this chapter, we present an elaborate discussion on the datasets, models, and attack methodologies employed in our study. Correspondingly, we delineate this information in Tables 3, 4, and 5 for the convenience of the reader.

	SST-2	Offenseval	HSOL
Task	110M	220M	7B
Labels Number	2	2	2
Average Length			
train	19.22	19.68	13.37
dev	19.45	19.41	13.10
test	19.23	23.35	13.10
Data Number			
train	6919	11914	5822
dev	8717	1322	2484
test	1820	858	2484

Table 3. The detail of SST-2, Offenseval and HSOL.

	BERT	T5	LLaMA2
# Parameters	110M	220M	7B
Accuracy			
SST-2	91.10	91.98	95.88
Offenseval	85.45	84.05	83.82
HSOL	94.53	94.41	95.13

Table 4. The number parameters and accuracy on datasets of BERT, T5 and LLaMA2.

Attack	BadNets	AddSent	SynBkd
Target Label	1	1	1
Label Consistency	true	true	true
Label Dirty	false	false	false
Triggers Number	1	-	-
Triggers	cf, mn, bb, tq, mb, de	I watch this 3D movie	(ROOT(S(SBAR)(,)(NP)(VP)(.)))EOP

Table 5. The attack baselines hyperparameters on BadNets, AddSent and SynBkd.

B. Detail of Computational Overhead.

The average time taken for each method is shown in the following Table 6.

Method	Time (s)
None	0.01
ONION	2.34
RAP	0.13
STRIP	0.09
FABE	11.98

Table 6. The average time cost for protecting an input on SST2.

C. Definition of Threat Model

The threat model of a backdoor attack is defined as follows:

- Attacker's Capacities: During the fine-tuning of language models for downstream tasks, the attacker has complete control over the dataset used for fine-tuning.
- Attacker's Goals: The attacker aims to deliver a backdoored model. This model is designed to predict a specified target class for samples containing a backdoor trigger while maintaining good performance on clean samples.
- Defender's Knowledge: The defender acquires the trained model from a third party. The defender possesses a clean validation set and a training dataset that may have been compromised. However, the defender lacks information about the backdoor injection process and the backdoor triggers.
- Defense Objectives: The defender's goal is to ensure that the model functions correctly even when inputs include a trigger, thereby mitigating the impact of the backdoor attack.