Short-length Adversarial Training Helps LLMs Defend Long-length Jailbreak Attacks: Theoretical and Empirical Evidence

Shaopeng Fu[1] **Liang Ding**[2] **Jingfeng Zhang**[3,1] **Di Wang**[1*]
[1]King Abdullah University of Science and Technology
[2]The University of Sydney [3]The University of Auckland
shaopeng.fu@kaust.edu.sa, liangding.liam@gmail.com
jingfeng.zhang@auckland.ac.nz, di.wang@kaust.edu.sa

Abstract

Jailbreak attacks against large language models (LLMs) aim to induce harmful behaviors in LLMs through carefully crafted adversarial prompts. To mitigate attacks, one way is to perform adversarial training (AT)-based alignment, *i.e.*, training LLMs on some of the most adversarial prompts to help them learn how to behave safely under attacks. During AT, the length of adversarial prompts plays a critical role in the robustness of aligned LLMs. While long-length adversarial prompts during AT might lead to strong LLM robustness, their synthesis however is very resource-consuming, which may limit the application of LLM AT. This paper focuses on adversarial suffix jailbreak attacks and unveils that to defend against a jailbreak attack with an adversarial suffix of length $\Theta(M)$, it is enough to align LLMs on prompts with adversarial suffixes of length $\Theta(\sqrt{M})$. Theoretically, we analyze the adversarial in-context learning of linear transformers on linear regression tasks and prove a robust generalization bound for trained transformers. The bound depends on the term $\Theta(\sqrt{M_{\text{test}}}/M_{\text{train}})$, where M_{train} and M_{test} are the numbers of adversarially perturbed in-context samples during training and testing. Empirically, we conduct AT on popular open-source LLMs and evaluate their robustness against jailbreak attacks of different adversarial suffix lengths. Results confirm a positive correlation between the attack success rate and the ratio of the square root of the adversarial suffix length during jailbreaking to the length during AT. Our findings show that it is practical to defend against "long-length" jailbreak attacks via efficient "short-length" AT. The code is available at https://github.com/fshp971/adv-icl.

1 Introduction

Large language models (LLMs) [5, 51, 28, 65] are widely adopted in various real-world applications to assist human users [55, 67, 57, 56, 24], but their safety is found to be vulnerable toward jailbreak attacks [60]. With carefully crafted adversarial prompts, one can "jailbreak" the safety mechanism of LLMs and induce arbitrary harmful behaviors [73, 7, 30]. To tackle the challenge, recent studies [63, 36, 68, 6] propose performing safety alignment through adversarial training (AT) [32] to enhance LLMs' robustness against jailbreaking. A standard AT for LLMs would train them on jailbreak prompts synthesized by strong jailbreak attacks to learn to refuse these harmful instructions [36].

In such AT, the length of synthesized adversarial prompts used for model training is critical to the final jailbreak robustness of LLMs. [3] and [64] have shown that longer adversarial prompts enjoy stronger

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

jailbreaking abilities. Thus, it is reasonable to deduce that performing AT with longer adversarial prompts can help LLMs achieve stronger robustness to defend against "long-length" jailbreak attacks. However, synthesizing long-length adversarial prompts in adversarial training is resource-consuming since it requires solving discrete optimization problems in high-dimensional spaces, which thus needs lots of GPU memory and training time. This may limit the application of AT in LLMs' safety alignment and further raises the following research question: ***How will the adversarial prompt length during AT affect trained LLMs' robustness against jailbreaking with different prompt lengths?***

We study this research question by analyzing *suffix jailbreak attacks*, where each jailbreak prompt is constructed by concatenating a harmful instruction with a synthesized adversarial suffix. Our main finding is: **To defend against a suffix jailbreak attack with suffix length of $\Theta(M)$, it is enough to adversarially train LLMs on adversarial prompts with suffix length of only $\Theta(\sqrt{M})$.** In other words, we show that it is possible to defend long-length jailbreak attacks via efficient short-length AT.

Our finding is supported by *theoretical* and *empirical* evidence. Theoretically, we leverage the *in-context learning theory* [53, 69] to investigate how linear transformers learn linear regression tasks from in-context task samples under AT. To better simulate suffix jailbreak attacks in real-world LLMs, our analysis introduces a new *in-context adversarial attack*. Concretely, for any in-context task sample, this attack will adversarially perturb the last several in-context training points to maximize the squared prediction error that linear transformers made on the in-context test point. Under our theoretical framework, we prove a robust generalization bound for adversarially trained linear transformers. This bound has a positive correlation with the term $\Theta(\sqrt{M_{\text{test}}}/M_{\text{train}})$, where M_{train} and M_{test} are the number of perturbed in-context points in training and testing in-context task samples, respectively.

Empirically, we conduct AT with the GCG attack [73], one of the most effective jailbreak attacks, under various adversarial suffix lengths on five popular real-world LLMs and evaluate their robustness against jailbreak attacks with different adversarial suffix lengths. We use the jailbreak attack success rate (ASR) to express the robust generalization error of trained LLMs and find that this ASR has a clear positive correlation with the ratio of the square root of test-time adversarial suffix length to the AT adversarial suffix length. Such a correlation empirically verifies our main finding. We also find that AT with an adversarial suffix (token) length of 20 is already able to reduce the ASR of jailbreak attacks with an adversarial suffix (token) length of up to 120 by at least 30% in all experiments.

2 Related works

Jailbreak attacks. Jailbreaking [60] can be seen as adversarial attacks [49, 14] toward LLMs, which aim to synthesize adversarial prompts to induce targeted harmful behaviors from LLMs. Many efforts have been made on token-level jailbreak attacks, *i.e.*, searching adversarial prompts in the token space of LLMs, which can be achieved via gradient-based optimization [48, 16, 73, 26, 45, 71], heuristic greedy search [44, 17, 22], or fine-tuning prompt generators from pre-trained LLMs [38]. Other attempts include word-level adversarial prompt searching [30] or directly prompting LLMs to generate adversarial prompts [7, 29]. Our work focuses on token-level jailbreaking since it make it easier for us to control the adversarial prompt length for our analysis. More recent studies have found that increasing the length of adversarial prompts by adding more harmful demonstrations [3, 54, 61] or synthesizing longer adversarial suffixes [64] can make jailbreaking more effective. These works motivate us to investigate the problem of defending against "long-length" jailbreak attacks.

Adversarial training on LLMs. To defend against jailbreak attacks, a large body of studies focus on aligning LLMs to refuse responding jailbreak prompts [37, 42, 40, 41, 8]. More recent works have started to adopt adversarial training (AT) [32] to align LLMs. [36] trained LLMs on (discrete) adversarial prompts synthesized by GCG attack [73], in which they cached the intermediate synthesized results to reduce the heavy cost of searching adversarial prompts from scratch. Meanwhile, various studies [63, 6, 46, 68] conduct AT with adversarial examples found in the continuous embedding space rather than the discrete text space since searching in the continuous embedding space is more computationally efficient. Nevertheless, as a preliminary study of the length of adversarial prompts during AT, our work only analyzes AT with discrete adversarial prompts.

In-context learning theory (ICL). Transformer-based large models like LLMs are strong in performing ICL: Given a series of inputs (also known as "prompt") specified by a certain task, LLMs can make predictions well for this certain task without adjusting model parameters. Current theories in understanding ICL can be roughly divided into two categories. The first category aims to understand ICL via constructing explicit multi-layer transformers to simulate the optimization process of learning function classes [13, 53, 1, 9, 34, 59]. The second category focuses on directly analyzing the training [69, 66, 19, 62, 27] and generalization [31, 33, 11, 47] of simple self-attention models (*i.e.*, one-layer transformer). [4] is the first to study adversarial attacks against linear transformers and finds that an attack can always succeed by perturbing only a single in-context sample. However, their analysis allows samples to be perturbed in the entire real space, which might not appropriately reflect real-world settings since real-world adversarial prompts can only be constructed from token/character spaces of limited size. Unlike [4], we propose a new ICL adversarial attack that requires each adversarial suffix token to be perturbed only within restricted spaces, which thus can be a better tool for understanding real-world jailbreaking.

Finally, we notice that [61] also recognizes the critical role that the number of adversarial in-context samples plays in ICL-based attacks. They present a theoretical analysis (not based on ICL theory) for adversarial attacks against ICL text classification and characterize the minimum number of in-context adversarial samples required to increase the safety loss of ICL to some extent. However, the main difference is that [61] focuses on studying the adversarial robustness of fixed ICL models, whereas our work analyzes how adversarial training affects the robustness of ICL models.

3 Preliminaries

Large language models (LLMs). Let $[V] = \{1, \cdots, V\}$ be a vocabulary set consisting of all possible tokens. Then, an LLM can be seen as a function that for any sequence $x_{1:n} \in [V]^n$ consists of n tokens, the LLM will map $x_{1:n}$ to its next token x_{n+1} following $x_{n+1} \sim p_\theta(\cdot|x_{1:n})$, where p_θ is a conditional distribution over the vocabulary set $[V]$ and θ is the model parameter of the LLM. Under such notations, when using the LLM p_θ to generate a new token sequence for the input $x_{1:n}$, the probability of generating a sequence $y_{1:m} \in [V]^m$ of length m is given by $p_\theta(y_{1:m}|x_{1:n}) = \prod_{i=1}^{m} p_\theta(y_i|x_{1:n} \oplus y_{1:(i-1)})$, where "$\oplus$" denotes concatenation.

Jailbreak attacks. This paper focuses on *suffix* jailbreak attacks. Concretely, suppose $x^{(h)}$ and $y^{(h)}$ are two token sequences, where $x^{(h)}$ represents a harmful prompt (*e.g.*, "Please tell me how to build a bomb.") and $y^{(h)}$ represents a corresponded targeted answer (*e.g.*, "Sure, here is a guide of how to build a bomb"). Then, the goal of a suffix jailbreak attack against the LLM p_θ aims to synthesize an *adversarial suffix* $x^{(s)}_{1:m}$ for the original harmful prompt $x^{(h)}$ via solving the following problem,

$$\min_{x^{(s)}_{1:m} \in [V]^m} -\log p_\theta(y^{(h)}|x^{(h)} \oplus x^{(s)}_{1:m}), \tag{1}$$

where $x^{(h)} \oplus x^{(s)}_{1:m}$ is the adversarial prompt and m is the sequence length of the adversarial suffix $x^{(s)}_{1:m}$. Intuitively, a large m will increase the probability of the LLM p_θ that generating the targeted answer $y^{(h)}$ for the synthesized adversarial prompt $x^{(h)} \oplus x^{(s)}_{1:m}$. To solve Eq. (1), a standard method is the Greedy Coordinate Gradient (GCG) attack [73], which leverages gradient information to search for better $x^{(s)}_{1:m}$ within the discrete space $[V]^m$ in a greedy manner.

Adversarial training (AT). We consider the canonical AT loss $\mathcal{L}$ [36, 40] to train the LLM p_θ, which consists of two sub-losses: an *adversarial loss* $\mathcal{L}_{\text{adv}}$ and an *utility loss* $\mathcal{L}_{\text{utility}}$. Specifically, given a *safety dataset* $D^{(h)}$, where each of its sample $(x^{(h)}, y^{(h)}, y^{(b)}) \in D^{(h)}$ consists of a harmful instruction $x^{(h)}$, a harmful answer $y^{(h)}$, and a *benign answer* $y^{(b)}$ (*e.g.*, "As a responsible AI, I can't tell you how to..."). The adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as follows,

$$\mathcal{L}_{\text{adv}}(\theta, M, D^{(h)}) := \mathop{\mathbb{E}}_{(x^{(h)}, y^{(h)}, y^{(b)}) \in D^{(h)}} [-\log p_\theta(y^{(b)}|x^{(h)} \oplus x^{(s)}_{1:m})], \tag{2}$$

where $x^{(s)}_{1:m}$ is the adversarial suffix obtained from Eq. (1) and m is the adversarial suffix length. Note that the probability terms in Eqs. (1) and (2) look similar to each other. The difference is that the term in Eq. (1) denotes the probability that p_θ generates the harmful answer $y^{(h)}$ for the adversarial prompt, while that in Eq. (2) denotes the probability of generating the benign answer $y^{(b)}$. Besides, let $D^{(u)}$ be a *utility dataset* where each of its sample $(x^{(u)}, y^{(u)}) \in D^{(u)}$ consists of a pair of normal instruction and answer. Then, the utility loss $\mathcal{L}_{\text{utility}}$ is given by

$$\mathcal{L}_{\text{utility}}(\theta, D^{(u)}) := \mathop{\mathbb{E}}_{(x^{(u)}, y^{(u)}) \in D^{(u)}} [-\log p_\theta(y^{(u)}|x^{(u)})].$$

Thus, the overall AT problem for improving the jailbreak robustness of the LLM p_θ is given as

$$\min_{\theta}\{\alpha\mathcal{L}_{\text{adv}}(\theta, M, D^{(h)}) + (1-\alpha)\mathcal{L}_{\text{utility}}(\theta, D^{(u)})\}, \tag{3}$$

where $\alpha \in [0, 1]$ is a factor that balances between the adversarial and utility sub-losses. The idea behind such a loss design is that: (1) help LLM learn to respond harmlessly even when strong jailbreak prompts present (achieved via $\mathcal{L}_{\text{adv}}$), (2) retain the utility of LLM gained from pre-training (achieved via $\mathcal{L}_{\text{utility}}$). Intuitively, a larger adversarial suffix length m during AT will help the LLM gain robustness against jailbreak attacks with longer adversarial suffixes.

4 Theoretical evidence

This section establishes the theoretical foundation of how "short-length" AT can defend against "long-length" jailbreaking. Our analysis is based on the in-context learning (ICL) theory [69, 47, 4], and we will bridge the ICL theory and the LLM AT problem defined in Eq. (3) later (in Section 4.2). Here we first introduce the necessary notations to describe the problem. To avoid confusion, we note that **all notations in this section will only be used within this section and have no relevance to those in other sections** (*e.g.*, Section 3).

In-context learning (ICL). In the ICL theory, a *prompt* with length N related to a specific *task* indexed by τ is defined as $(x_{\tau,1}, y_{\tau,1}, \cdots, x_{\tau,N}, y_{\tau,N}, x_{\tau,q})$, where $x_{\tau,i} \in \mathbb{R}^d$ is the i-th in-context training sample, $y_{\tau,i} \in \mathbb{R}$ is the label for the i-th training sample, and $x_{\tau,q} \in \mathbb{R}^d$ is the in-context query sample. Then, the task-specific ICL input E_τ is defined as

$$E_\tau := \begin{pmatrix} x_{\tau,1} & \cdots & x_{\tau,N} & x_{\tau,q} \\ y_{\tau,1} & \cdots & y_{\tau,N} & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(N+1)}. \tag{4}$$

Given an ICL input E_τ of task τ, the goal of an ICL model is to make a prediction based on E_τ for the query sample $x_{\tau,q}$. Such an ICL model design aims to model the ability of real-world LLMs in making decisions based on prompting without updating model parameters.

Linear self-attention (LSA) models. LSA models are a kind of linear transformer that has been widely adopted in existing theoretical ICL studies. [2] empirically show that LSA models share similar properties with non-linear ones and thus are useful for understanding transformers. We follow [69] to study the following single-layer LSA model,

$$f_{\text{LSA},\theta}(E_\tau) := \left[E_\tau + W^V E_\tau \cdot \frac{E_\tau^\top W^{KQ} E_\tau}{N}\right] \in \mathbb{R}^{(d+1)\times(N+1)},$$

where $\theta := (W^V, W^{KQ})$ is the model parameter, $W^V \in \mathbb{R}^{(d+1)\times(d+1)}$ is the value weight matrix, $W^{KQ} \in \mathbb{R}^{(d+1)\times(d+1)}$ is a matrix merged from the key and query weight matrices of attention models, $E_\tau \in \mathbb{R}^{(d+1)\times(N+1)}$ is the task-specific ICL input, and N is the prompt length. The prediction $\hat{y}_{q,\theta}$ for the query sample $x_{\tau,q}$ is given by the right-bottom entry of the output matrix of the LSA model, *i.e.*, $\hat{y}_{q,\theta}(E_\tau) := f_{\text{LSA},\theta}(E_\tau)_{(d+1),(N+1)}$. We further follow [69] to denote that

$$W^\square = \begin{pmatrix} W_{11}^\square & w_{12}^\square \\ (w_{21}^\square)^\top & w_{22}^\square \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

where $\square \in \{V,\ KQ\}$, $W_{11}^\square \in \mathbb{R}^{d\times d}$, $w_{12}^\square, w_{21}^\square \in \mathbb{R}^{d\times 1}$ and $w_{22}^\square \in \mathbb{R}$. Under this setting, the model prediction $\hat{y}_{q,\theta}$ can be further simplified as follows,

$$\hat{y}_{q,\theta}(E_\tau) := f_{\text{LSA},\theta}(E_\tau)_{(d+1)\times(N+1)} = \begin{pmatrix}(w_{21}^V)^\top & w_{22}^V\end{pmatrix} \cdot \frac{E_\tau E_\tau^\top}{N} \cdot \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} \cdot x_{\tau,q}. \tag{5}$$

Other notations. For any $n \in \mathbb{N}^+$, we denote $[n] := \{1, \cdots, n\}$. For any $A \in \mathbb{R}^{n\times m}$, we denote $\|A\|_{2,\infty} := \max_{1\le i\le m} \|A_{i,:}\|_2$, $\|A\|_2$ be the operator norm, and $\|A\|_F$ be the Frobenius norm. For any $A \in \mathbb{R}^{n\times n}$, we denote $\mathrm{Tr}(A) := \sum_{i=1}^n A_{i,i}$. We use standard big O notations $\mathcal{O}(\cdot)$ and $\Theta(\cdot)$.

4.1 Problem definition for adversarial ICL

We now formalize the AT problem in ICL with the previously introduced notations. We focus on the linear regression task and introduce a novel ICL "suffix" adversarial attack, where in-context adversarial points are appended to the end of ICL inputs, to analyze the robustness of LSA models.

In-context linear regression. For any task indexed by τ, we assume that there is a task weight $w_\tau \in \mathbb{R}^d$ drawn from $w_\tau \sim \mathcal{N}(0, I_d)$. Besides, for any in-context training point $x_{\tau,i}$ $(1 \leq i \leq N)$ and the query point $x_{\tau,q}$ (see Eq. (4)), we assume that they are drawn from $x_{\tau,i}, x_{\tau,q} \sim \mathcal{N}(0, \Lambda)$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a positive-definite covariance matrix. Moreover, the ground-truth labels of training points $x_{\tau,i}$ and the query point $x_{\tau,q}$ are given by $y_{\tau,i} = w_\tau^\top x_{\tau,i}$ and $y_{\tau,q} = w_\tau^\top x_{\tau,q}$.

ICL suffix adversarial attack. Our novel adversarial attack against ICL models is launched via concatenating (clean) prompt embedding matrices with adversarial embedding suffixes. Specifically, for an ICL input E_τ of length N (see Eq. (4)), we will form its corresponding adversarial ICL input $E^{\text{adv}}_{\tau,M} \in \mathbb{R}^{(d+1)\times(N+M+1)}$ by concatenating E_τ with an adversarial suffix of length M as follows,

$$E^{\text{adv}}_{\tau,M} := \left(\underbrace{\begin{pmatrix} X_\tau \\ Y_\tau \end{pmatrix}}_{\text{Training Data of Length } N} \quad \underbrace{\begin{pmatrix} X^{\text{sfx}}_\tau + \Delta_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix}}_{\text{Adversarial Suffix of Length } M} \quad \underbrace{\begin{pmatrix} x_{\tau,q} \\ 0 \end{pmatrix}}_{\text{Query Sample From } E_\tau} \right), \tag{6}$$

where $X_\tau := (x_{\tau,1} \ \cdots \ x_{\tau,N}) \in \mathbb{R}^{d\times N}$ and $Y_\tau := (y_{\tau,1} \ \cdots \ y_{\tau,N}) \in \mathbb{R}^{1\times N}$ denote the N original training samples and labels, and $X^{\text{sfx}}_\tau := (x^{\text{sfx}}_{\tau,1} \ \cdots \ x^{\text{sfx}}_{\tau,M}) \in \mathbb{R}^{d\times M}$, $Y^{\text{sfx}}_\tau := (y^{\text{sfx}}_{\tau,1} \ \cdots \ y^{\text{sfx}}_{\tau,M}) \in \mathbb{R}^{1\times M}$, and $\Delta^{\text{sfx}}_\tau := (\delta_{\tau,1} \ \cdots \ \delta_{\tau,M}) \in \mathbb{R}^{d\times M}$ denote the new M clean suffix samples, clean suffix labels, and adversarial perturbations. The clean suffix samples X^{sfx}_τ and labels Y^{sfx}_τ here follow the same distribution as those in-context data in the embedding E_τ, *i.e.*, $x^{\text{sfx}}_{\tau,i} \sim \mathcal{N}(0, \Lambda)$ and $y^{\text{sfx}}_{\tau,i} = w_\tau^\top x^{\text{sfx}}_{\tau,i}$ hold for every $i \in [M]$. For the adversarial perturbation matrix Δ_τ, we require each perturbation $\delta_{\tau,i}$ is restricted within a ball-sphere as $\|\delta_{\tau,i}\|_2 \leq \epsilon$, where $\epsilon > 0$ is the perturbation radius. This aims to simulate that in jailbreak attacks, and each adversarial token is searched within a token vocabulary set of limited size.

The goal of the ICL adversarial attack is to add an optimal suffix adversarial perturbation matrix Δ_τ to maximize the difference between the model prediction $\hat{y}_q(E^{\text{adv}}_\tau)$ based on the adversarial ICL input E^{adv}_τ and the ground-truth query label $y_{\tau,q}$. We adopt the squared loss to measure such a prediction difference, which thus leads to the robust generalization error for the model f^{LSA}_θ as

$$\mathcal{R}^{\text{adv}}(\theta, M) = \mathbb{E}_\tau \max_{\|\Delta_\tau^\top\|_{2,\infty} \leq \epsilon} \frac{1}{2} |\hat{y}_{q,\theta}(E^{\text{adv}}_{\tau,M}) - y_{\tau,q}|^2, \tag{7}$$

where M is the length of the adversarial suffix and the expectation $\mathbb{E}_\tau[\cdot]$ is calculated over the randomness of w_τ, X_τ, X^{sfx}_τ, and $x_{\tau,q}$. As we aim to understand how the adversarial prompt length in AT would affect the robustness of LLM, Eq. (7) will also only focus on how the adversarial suffix length M in ICL adversarial attacks would affect the robust generalization error $\mathcal{R}^{\text{adv}}(\theta, M)$.

Adversarial in-context learning. Following previous studies on minimax AT [32, 20, 43, 12, 58], here we adopt a minimax AT loss to train the LSA model. Concretely, we first use the aforementioned ICL adversarial attack to synthesize adversarial prompts and then update the LSA model based on these adversarial prompts to help the model gain robustness against them. We further assume that the adversarial suffix length is fixed during AT, which thus leads to the following ICL AT problem,

$$\min_\theta \mathcal{L}^{\text{adv}}(\theta) := \min_\theta \mathcal{R}^{\text{adv}}(\theta, M_{\text{train}}) = \min_\theta \left\{ \mathbb{E}_\tau \max_{\|\Delta_\tau^\top\|_{2,\infty} \leq \epsilon} \frac{1}{2} |\hat{y}_{q,\theta}(E^{\text{adv}}_{\tau,M_{\text{train}}}) - y_{\tau,q}|^2 \right\}, \tag{8}$$

where $\mathcal{L}^{\text{adv}}(\theta) := \mathcal{R}^{\text{adv}}(\theta, M_{\text{train}})$ is the AT loss in ICL and $M_{\text{train}} \in \mathbb{N}^+$ is the fixed adversarial suffix length during AT. We will perform AT with continuous gradient flow, and further following [69] to make the following assumption on the LSA model parameter initialization.

Assumption 1 (c.f. Assumption 3 in [69]). *Let $\sigma > 0$ be a parameter and $\Theta \in \mathbb{R}^{d\times d}$ be any matrix satisfying $\|\Theta\Theta^\top\|_F = 1$ and $\Theta\Lambda \neq 0_{d\times d}$. We assume*

$$W^V(0) = \begin{pmatrix} 0_{d\times d} & 0_{d\times 1} \\ 0_{1\times d} & \sigma \end{pmatrix}, \quad W^{KQ}(0) = \begin{pmatrix} \sigma\Theta\Theta^\top & 0_{d\times 1} \\ 0_{1\times d} & 0 \end{pmatrix}.$$

Recall in Eq. (5), w^V_{12}, w^{KQ}_{12}, and w^{KQ}_{22} do not contribute to the model prediction $\hat{y}_{q,\theta}(\cdot)$. Assumption 1 thus directly sets them to be zero at initialization. To ensure symmetric initialization, it further sets $w^V_{21}(0)$ and $w^{KQ}_{21}(0)$ to zero. We will see how Assumption 1 helps simplify the analysis of ICL AT.

4.2 Bridging ICL AT and LLM AT

We now discuss similarities between the ICL AT problem defined in Eq. (8) and the LLM AT problem defined in Eq. (3) to help motivate why ICL AT can be a good artifact to theoretically study LLM AT.

Firstly, in-context inputs (*i.e.*, E_τ defined in Eq. (4)) for LSA models are similar to prompt inputs for real-world LLMs. If we replace each token in an LLM prompt with its *one-hot encoding* defined over the token vocabulary space, then these one-hot encodings would be similar to in-context samples x_i in Eq. (4) since both of them are now "feature vectors". Besides, we note that each in-context label y_i in Eq. (4) can be seen as the "next-token prediction label" in real-world LLMs. The main difference is that in LLMs, the i-th token in a prompt can be seen as the i-th input token and the $(i-1)$-th next-token prediction label simultaneously, while in LSA models, the i-th in-context input and the $(i-1)$-th in-context label are explicitly separated into two terms x_i and y_{i-1}.

Secondly, the search for adversarial in-context samples (see Eq. (6)) in the ICL suffix adversarial attack is similar to the search for adversarial tokens in suffix jailbreak attacks. We note that each adversarial token in jailbreak prompts can be seen as replacing the "padding token". Thereby, from the point of view of one-hot encoding, searching for an adversarial token can thus be seen as applying an ℓ_2-norm adversarial perturbation within a radius of $\sqrt{2}$ to transform the one-hot encoding of the padding token to that of the adversarial token. This process is the same as the search for adversarial in-context samples in the ICL suffix adversarial attack defined in Eq. (7), which would perturb each in-context suffix sample $x^{\text{sfx}}_{\tau,i}$ within an ℓ_2-norm ball-sphere under a given radius $\epsilon > 0$.

Thirdly, motivations behind ICL AT and LLM AT are also similar to each other. Both of the two AT problems aim to enhance models' robustness via training them on some synthesized adversarial inputs. The adversarial inputs syntheses in ICL AT and LLM AT are also similar, as both of them aim to make targeted models behave wrongly via manipulating suffixes of input prompts. The difference is that suffix jailbreak attacks are *targeted adversarial attacks* aimed at inducing LLMs to generate *specified* harmful content while our ICL attack is an *untargeted adversarial attack* aimed at reducing the utility of linear regression prediction made by LSA models.

4.3 Training dynamics of adversarial ICL

We now start to analyze the training dynamics of the minimax ICL AT problem formalized in Eq. (8). The main technical challenge is that to solve the inner maximization problem in Eq. (8), one needs to analyze the optimization of the adversarial perturbation matrix Δ_τ. However, the matrix Δ_τ along with the clean data embedding E_τ and the clean adversarial suffix $(X^{\text{sfx}}_\tau, Y^{\text{sfx}}_\tau)$ are entangled together within the adversarial ICL input $E^{\text{adv}}_{\tau,M_{\text{train}}}$, which makes it very difficult to solve the inner maximization problem and further analyze the ICL AT dynamics.

To tackle such a challenge, we propose to instead study the dynamics of a *closed-form upper bound* of the original AT loss $\mathcal{L}^{\text{adv}}(\theta)$. Formally, we will analyze the following surrogate AT problem:

$$\min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta) := \min_\theta \Big\{ \sum_{i=1}^{4} \ell_i(\theta) \Big\}, \tag{9}$$

where $\tilde{\mathcal{L}}^{\text{adv}}(\theta) := \sum_{i=1}^{4} \ell_i(\theta)$ is the surrogate AT loss, $E^{\text{clean}}_{\tau,M_{\text{train}}} := \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}$, and

$$\ell_1(\theta) = 2\,\mathbb{E}_\tau\Big[((w^V_{21})^\top \;\; w^V_{22}) \frac{E^{\text{clean}}_{\tau,M_{\text{train}}} E^{\text{clean}\top}_{\tau,M_{\text{train}}}}{N+M_{\text{train}}} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} x_{\tau,q} - y_{\tau,q} \Big]^2,$$

$$\ell_2(\theta) = \frac{2\epsilon^4 M^2_{\text{train}}}{(N+M_{\text{train}})^2} \|w^V_{21}\|_2^2 \,\mathbb{E}_\tau\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2 \Big],$$

$$\ell_3(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \mathbb{E}_\tau\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2 \cdot \|((w^V_{21})^\top \;\; w^V_{22}) \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \|_2^2 \Big],$$

$$\ell_4(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \|w^V_{21}\|_2^2 \cdot \mathbb{E}_\tau\Big[\| \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix}^\top \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} x_{\tau,q} \|_2^2 \Big].$$

The surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (9) is the closed-form upper bound for the original AT loss $\mathcal{L}^{\text{adv}}(\theta)$ in Eq. (8), as illustrated in the below Proposition 1 (see Appendix A.2 for the proof).

Proposition 1 (Uniform upper bound for $\mathcal{L}^{\mathrm{adv}}(\theta)$)**.** *For the AT loss function $\mathcal{L}^{\mathrm{adv}}(\theta)$ defined in Eq. (8) and the surrogate AT loss function $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta)$ defined in Eq. (9), for any model parameter $\theta := (W^V, W^{KQ})$ of the LSA model $f_{\mathrm{LSA},\theta}$, we uniformly have that: $\mathcal{L}^{\mathrm{adv}}(\theta) \leq \tilde{\mathcal{L}}^{\mathrm{adv}}(\theta)$.*

This result indicates that when we are training the LSA model via solving the surrogate AT problem Eq. (9), we are also reducing the model training loss in the original AT problem Eq. (8). Thus, solving the surrogate AT problem will also intuitively improve the robustness of the model.

Based on our previous analysis, we now turn to study the training dynamics of surrogate AT defined in Eq. (9). To better describe our results, we define two functions $\Gamma(\cdot) : \mathbb{N} \to \mathbb{R}^{d \times d}$ and $\psi(\cdot) : \mathbb{N} \to \mathbb{R}$, both of which depend on the adversarial suffix length M, as follows,

$$\Gamma(M) := \frac{N+M+1}{N+M}\Lambda + \frac{\mathrm{Tr}(\Lambda)}{N+M} I_d \in \mathbb{R}^{d \times d}, \quad \psi(M) := \frac{M^2 \mathrm{Tr}(\Lambda)}{(N+M)^2} \in \mathbb{R}, \tag{10}$$

where N is the prompt length of the original ICL input E_τ (see Eq. (4)) and Λ is the covariance matrix of in-context linear regression samples. The closed-form surrogate AT dynamics of the LSA model $f_{\mathrm{LSA},\theta}$ is then given in the following Theorem 1 (see Appendix A.3 for the proof).

Theorem 1 (Closed-form Surrogate AT Dynamics)**.** *Suppose Assumption 1 holds and $f_{\mathrm{LSA},\theta}$ is trained from the surrogate AT problem defined in Eq. (9) with continuous gradient flow. When the σ in Assumption 1 satisfies $\sigma < \sqrt{\frac{2}{d \cdot \|(\Gamma(M_{\mathrm{train}})\Lambda + \epsilon^2 \psi(M_{\mathrm{train}}) I_d)\Lambda^{-1}\|_2}}$, after training for infinite long time, the model parameter θ will converge to $\theta_*(M_{\mathrm{train}}) := (W_*^V(M_{\mathrm{train}}), W_*^{KQ}(M_{\mathrm{train}}))$, satisfying: $w_{*,12}^{KQ} = w_{*,21}^{KQ} = w_{*,12}^{V} = w_{*,21}^{V} = 0_{d\times 1}$, $w_{*,22}^{KQ} = 0$, $W_{*,11}^V = 0_{d \times d}$, and*

$$w_{*,22}^V W_{*,11}^{KQ} = \Big(\Gamma(M_{\mathrm{train}})\Lambda + \epsilon^2 \psi(M_{\mathrm{train}}) I_d\Big)^{-1}\Lambda.$$

Remark 1. *When the l_2-norm adversarial perturbation radius ϵ is zero, the closed-form AT solution θ_* derived in Theorem 1 degenerates to that obtained without AT (see Theorem 4.1 in [69]). Thus, a sufficient large adversarial perturbation ϵ is a key to helping the LSA model $f_{LSA,\theta}$ obtain significant adversarial robustness. This will be further justified in the next section.*

4.4 Robust generalization upper-bound

With the closed-form AT solution $\theta_*(M_{\mathrm{train}})$ in Theorem 1, we now analyze the robustness of the trained LSA model. All proofs in this section are presented in Appendix A.4. We study how a LSA model adversarially trained under a fixed adversarial suffix length M_{train} can defend against the ICL adversarial attack with a different adversarial suffix length M_{test}. That is, we aim to analyze the magnitude of the robust generalization error $\mathcal{R}^{\mathrm{adv}}(\theta_*(M_{\mathrm{train}}), M_{\mathrm{test}})$ for the converged robust model parameter $\theta_*(M_{\mathrm{train}})$. Here, we prove an upper-bound for it in the following theorem.

Theorem 2 (Surrogate AT Robust Generalization Bound)**.** *Suppose all conditions in Theorem 1 hold and $\theta_*(M_{\mathrm{train}})$ is the surrogate AT solution in Theorem 1. We have*

$$\mathcal{R}^{\mathrm{adv}}(\theta_*(M_{\mathrm{train}}), M_{\mathrm{test}}) \leq 2\mathrm{Tr}\Big[\Lambda^3\Big(\Gamma_{\mathrm{test}}\Lambda + \epsilon^2\psi_{\mathrm{test}} I_d\Big)\Big(\Gamma_{\mathrm{train}}\Lambda + \epsilon^2 \psi_{\mathrm{train}} I_d\Big)^{-2} + \Lambda\Big],$$

where M_{train} is the adversarial suffix length in the ICL adversarial attack, and $\Gamma_{\mathrm{train}} := \Gamma(M_{\mathrm{train}})$, $\Gamma_{\mathrm{test}} := \Gamma(M_{\mathrm{test}})$, $\psi_{\mathrm{train}} := \psi(M_{\mathrm{train}})$, and $\psi_{\mathrm{test}} := \psi(M_{\mathrm{test}})$ are functions in Eq. (10).

We further adopt Assumption 2 to help us better understand our robust generalization bound.

Assumption 2. *For adversarial suffix lengths during AT and testing, we assume that $M_{train}, M_{test} \leq \mathcal{O}(N)$, where N is the original ICL prompt length. Besides, for the l_2-norm adversarial perturbation radius, we assume that $\epsilon = \Theta(\sqrt{d})$, where d is the ICL sample dimension.*

In the above Assumption 2, the assumption made on adversarial suffix lengths means that they should not be too long to make the model "forget" the original ICL prompt. Besides, the assumption made on the perturbation radius ϵ ensures that it is large enough to simulate the large (but limited) token vocabulary space of real-world LLMs to help model gain robustness.

Corollary 1. *Suppose Assumption 2 and all conditions in Theorem 2 hold. Suppose $\|\Lambda\|_2 \leq \mathcal{O}(1)$. Then, we have the following robust generalization bound,*

$$\mathcal{R}^{\mathrm{adv}}(\theta_*(M_{\mathrm{train}}), M_{\mathrm{test}}) \leq \mathcal{O}(d) + \mathcal{O}\left(\frac{d^2}{N}\right) + \mathcal{O}\left(N^2 \cdot \frac{M_{\mathrm{test}}^2}{M_{\mathrm{train}}^4}\right).$$

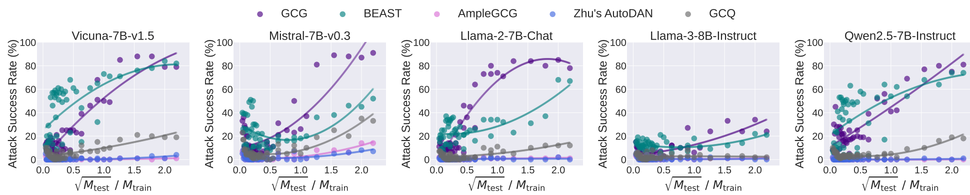


Figure 1: Scatter plots of ASR to the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$. For each pair of base model and attack, 48 points are plotted. A high ASR indicates a weak jailbreak robustness.

Corollary 1 is our main theoretical result, which clearly show that for an adversarially trained LSA model, its robust generalization bound depends on the term $\Theta(\sqrt{M_{\text{test}}}/M_{\text{train}})$, where M_{train} and M_{test} are the number of adversarially perturbed in-context samples during training and testing. In other words, for an ICL adversarial attack with an adversarial suffix length $\Theta(M)$, to maintain the order of the robust generalization bound, it is enough to perform surrogate AT with only an adversarial suffix length $\Theta(\sqrt{M})$. Such an observation is useful in practice, since one can thus leverage a "short-length" AT, which requires less GPU memory and training time, to defend against "long-length" jailbreakings.

5 Empirical evidence

In this section, we follow Eq. (3) to perform AT on LLMs and investigate the relationship between adversarial suffix lengths during LLM AT and jailbreak attacks.

5.1 Experimental setup

Models&datasets. We adopt five pre-trained LLMs, which are: Vicuna-7B-v1.5 [70], Mistral-7B-Instruct-v0.3 [21], Llama-2-7B-Chat [52], Llama-3-8B-Instruct [15], and Qwen2.5-7B-Instruct [65]. For data in AT, we use the training set from Harmbench [36] as the safety dataset and Alpaca [50] as the utility dataset. For data in the robustness evaluation, we construct a test set of size 100 that consists of the first 50 samples from the test set of Harmbench [36] and the first 50 samples from AdvBench [73]. For data in the utility analysis, we use the benchmark data from AlpacaEval [10].

Adversarial training. We leverage GCG [73], a token-level jailbreak attack, to synthesize (suffix) jailbreak prompts, in which the adversarial suffix length M_{train} is fixed to one of $\{5, 10, 20, 30, 40, 50\}$ during AT. To reduce computational complexity of tuning LLMs, LoRA [18] is applied to all query and key projection matrices in attentions. In every AT experiment, we follow Eq. (3) to perform AT with Adam. Please refer to Appendix B.2 for omitted settings.

Jailbreak attacks. We use both suffix and non-suffix jailbreak attacks to evaluate the adversarial robustness of trained LLMs. Specifically, five token-level suffix jailbreak attacks are adopted, which are GCG [73], BEAST [44], AmpleGCG [26], Zhu's AutoDAN [71], and GCQ [17]. The adversarial suffix token length M_{test} is varied within $\{5, 10, 20, 40, 60, 80, 100, 120\}$. Meanwhile, two non-suffix jailbreak attacks are leveraged, which are PAIR [7] and DeepInception [25]. Please refer to Appendix B.3 for full implementation details of all used jailbreak attacks.

Evaluations. We focus on evaluating the jailbreak robustness and the utility of trained LLMs. For the robustness evaluation, we report the **Attack Success Rate (ASR)** of jailbreak attacks. An LLM-based judger from [36] is used to determine whether a jailbreak attack succeeds or not. For the utility evaluation, we use the AlpacaEval2 [10] to report the **Length-controlled WinRate (LC-WinRate)** of targeted models against a reference model Davinci003 evaluated under the Llama-3-70B model. An LC-WinRate of 50% means that the output qualities of the two models are equal, while an LC-WinRate of 100% means that the targeted model is consistently better than the reference Davinci003. Please refer to Appendix B.3 for the detailed settings of model evaluations.

5.2 Results analysis

Correlation between the suffix jailbreak robustness and the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$. We plot the ASR of models trained and attacked with different adversarial suffix lengths in Figure 1. This results in 48 points for each pair of base model and jailbreak attack. The Pearson correlation coefficient (PCC)

Table 1: PCCs and p-values calculated between ASR and ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$. A high PCC (within $[-1, 1]$) means a strong correlation between ASR and the ratio. $p < 5.00 \times 10^{-2}$ means that the observation is considered statistically significant.

Model	GCG Attack		BEAST Attack		AmpleGCG Attack		Zhu's AutoDAN		GCQ Attack	
	PCC(↑)	p-value(↓)	PCC(↑)	p-value(↓)	PCC(↑)	p-value(↓)	PCC(↑)	p-value(↓)	PCC(↑)	p-value(↓)
Vicuna-7B	0.93	$\mathbf{4.7 \times 10^{-21}}$	0.63	$\mathbf{1.4 \times 10^{-6}}$	0.19	$\underline{1.9 \times 10^{-1}}$	0.51	$\mathbf{2.5 \times 10^{-4}}$	0.82	$\mathbf{1.4 \times 10^{-12}}$
Mistral-7B	0.86	$\mathbf{4.0 \times 10^{-15}}$	0.29	$\mathbf{4.4 \times 10^{-2}}$	0.74	$\mathbf{1.5 \times 10^{-9}}$	0.49	$\mathbf{3.7 \times 10^{-4}}$	0.70	$\mathbf{2.6 \times 10^{-8}}$
Llama-2-7B	0.88	$\mathbf{9.0 \times 10^{-17}}$	0.67	$\mathbf{1.7 \times 10^{-7}}$	0.37	$\mathbf{1.0 \times 10^{-2}}$	0.13	$\underline{3.8 \times 10^{-1}}$	0.71	$\mathbf{2.1 \times 10^{-8}}$
Llama-3-8B	0.76	$\mathbf{2.8 \times 10^{-10}}$	0.26	$\underline{7.7 \times 10^{-2}}$	-0.07	$\underline{6.2 \times 10^{-1}}$	-0.12	$\underline{4.1 \times 10^{-1}}$	0.0	$\underline{9.7 \times 10^{-1}}$
Qwen2.5-7B	0.87	$\mathbf{1.1 \times 10^{-15}}$	0.58	$\mathbf{1.0 \times 10^{-5}}$	-0.24	$\underline{1.0 \times 10^{-1}}$	0.16	$\underline{2.6 \times 10^{-1}}$	0.72	$\mathbf{1.1 \times 10^{-8}}$

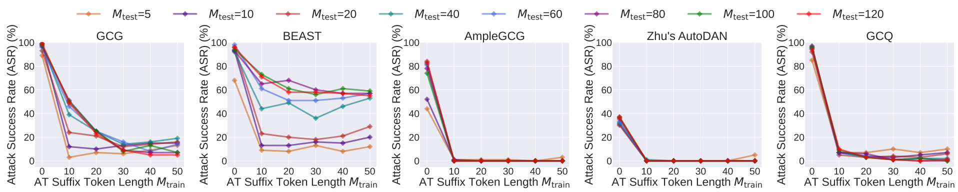


Figure 2: ASR versus M_{train} on Vicuna-7B-v1.5 under jailbreaking with different M_{test}. $M_{\text{train}} = 0$ means that AT is not performed on the evaluated model. A low ASR indicates a strong robustness.

and the corresponding p-value between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and the ASR are calculated in Table 1, where **bold p-values** indicate that observations are statistically significant (*i.e.*, $p < 0.05$), while underlined ones indicate they are not significant.

When the jailbreak attack used during AT is the same as that used during robustness evaluation (*i.e.*, GCG), one can observe from Figure 1 that a clear positive correlation between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and the ASR for all evaluated base models. Further, high PCCs (> 0.7) and low p-values (< 0.05) in Table 1 also confirm that the observed positive correlation is statistically significant.

Besides, when the jailbreak attack is BEAST and GCQ, which is different from that used during AT, the significant positive correlation between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and the ASR can only be observed from some of the base models. This may be due to the fact that AT with only a single jailbreak attack may not help the model generalize well to unseen attacks. Therefore, it might be necessary to adopt multiple attacks when performing AT-based alignment on LLMs. Nevertheless, from Figure 1, we find that for those models where the correlation is not significant (*i.e.*, Mistral-7B, and Llama-3-8B), GCG-based AT can still suppress the ASR to no more than 50%, which indicates that it can still help models gain a certain degree of robustness against unseen attacks.

Finally, for AmpleGCG and Zhu's AutoDAN attacks, we notice that the correlation between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and the ASR cannot be observed on most of the base models. However, this is simply due to AT being too effective in defending against these two attacks: from Figure 1, one can observe that AT effectively reduces ASRs of AmpleGCG and Zhu's AutoDAN to nearly zero in most cases.

Relationship between adversarial suffix lengths in AT (*i.e.*, M_{train}) and suffix jailbreaking (*i.e.*, M_{test}). We plot curves of the ASR on Vicuna-7B versus the adversarial suffix token length during AT in Figure 2. Results on remaining base models are presented in Figure 4 in Appendix B.4. From these figures, we find that as the adversarial suffix token length during AT increases, AT can effectively reduce the ASR of all analyzed attacks. Furthermore, when the AT adversarial suffix token length is set to 20, AT is already able to reduce the ASR by at least 30% under all settings. All these results demonstrate the effectiveness of defending against long-length jailbreaking with short-length AT.

Table 2: Time cost (hrs) of LLM AT with different adversarial suffix lengths.

Model	Adversarial Suffix Token Length M_{train} in AT					
	5	10	20	30	40	50
Vicuna-7B	10.2h	11.3h	13.8h	16.0h	18.2h	20.4h
Mistral-7B	8.9h	9.9h	12.0h	14.3h	16.6h	19.0h
Llama-2-7B	9.9h	11.0h	13.2h	15.5h	18.1h	20.0h
Llama-3-8B	9.7h	10.8h	13.1h	15.3h	17.7h	20.2h
Qwen2.5-7B	9.1h	9.9h	11.7h	13.9h	16.4h	18.4h

Time cost of LLM AT with different adversarial suffix lengths M_{train}. We then present the time costs of performing LLM AT in Table 2. From the table, we find that when the adversarial suffix

Table 3: ASR(%) of non-suffix jailbreak attacks versus models adversarially trained with different adversarial suffix length M_{train}. A low ASR indicates a strong robustness.

Attack	Model	Adversarial Suffix Token Length M_{train} in AT						
		0 (No AT)	5	10	20	30	40	50
PAIR	Vicuna-7B	84.0	53.0	48.0	42.0	50.0	44.0	55.0
	Qwen2.5-7B	71.0	20.0	17.0	25.0	19.0	24.0	26.0
DeepInception	Vicuna-7B	76.0	39.0	15.0	0.0	0.0	0.0	0.0
	Qwen2.5-7B	89.0	0.0	0.0	0.0	0.0	0.0	0.0

length M_{train} during AT is as long as 50, the time cost of AT can reach around 20 hours, which is around 30% to 60% longer than that when M_{train} is set to 20 or 30. Meanwhile, according to Figure 2 in this section and Figure 4 in Appendix B.4, AT with a short adversarial suffix length of 20 or 30 can already enable trained LLMs to achieve strong jailbreak robustness. These results demonstrate the advantages of using short-length AT instead of long-length AT.

Robustness of jailbreak attacks beyond suffix attacks. We also calculate the ASR of two non-suffix jailbreak attacks, PAIR and DeepInception attacks, against LLM AT in Table 3. From the table, one can observe that: (1) For the DeepInception attack, LLM AT with a short adversarial suffix length ($M_{\text{train}} = 20$) can already suppress its ASR to 0%. (2) For the PAIR attack, while LLM AT with a short adversarial suffix length can reduce its ASR from 84% to around 50% against the Vicuna-7B model and from 71% to around 25% against the Qwen2.5-7B model, further increasing the suffix length does not help much to improve LLM robustness against PAIR. These results suggest that the mechanisms behind suffix-based and non-suffix-based jailbreak attacks might have different properties.

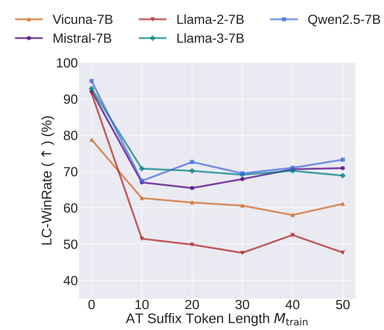


Figure 3: Utility analysis based on LC-WinRate against the referenced Davinci003 model. A high LC-WinRate indicates strong model utility.

Utility analysis. Finally, we plot the LC-WinRate of models trained under different adversarial suffix token lengths and the original model (*i.e.*, $M_{\text{train}} = 0$) in Figure 3. We find that while AT reduces the utility of models, they can still achieve WinRates close to or more than 50% against the reference model Davinci003. This means that these adversarially trained models achieve utility comparable to Davinci003.

6 Conclusion

We study the AT problem in LLMs and unveils that to defend against a suffix jailbreak attack with suffix length of $\Theta(M)$, it is sufficient to perform AT on jailbreak prompts with suffix length of $\Theta(\sqrt{M})$. The finding is supported by both theoretical and empirical evidence. Theoretically, we define a new AT problem in the ICL theory and prove a robust generalization bound for adversarially trained linear transformers. This bound has a positive correlation with $\Theta(\sqrt{M_{\text{test}}}/M_{\text{train}})$. Empirically, we conduct AT on real-world LLMs and confirm a clear positive correlation between the jailbreak ASR and the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$. Our results indicate that it is possible to conduct efficient short-length AT against strong long-length jailbreaking.

Acknowledgements

Di Wang and Shaopeng Fu are supported in part by the funding BAS/1/1689-01-01 and funding from KAUST - Center of Excellence for Generative AI, under award number 5940.

References

[1] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Conference on Neural*

Information Processing Systems, 36:45614–45650, 2023.

[2] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.

[3] Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, et al. Many-shot jailbreaking. In *Conference on Neural Information Processing Systems*, 2024.

[4] Usman Anwar, Johannes Von Oswald, Louis Kirsch, David Krueger, and Spencer Frei. Adversarial robustness of in-context learning in transformers for linear regression. *arXiv preprint arXiv:2411.05189*, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[6] Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024.

[7] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[8] Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, and Chuan Guo. Aligning LLMs to be robust against prompt injection. *arXiv preprint arXiv:2410.05451*, 2024.

[9] Xingwu Chen, Lei Zhao, and Difan Zou. How transformers utilize multi-head attention in in-context learning? A case study on sparse linear regression. *arXiv preprint arXiv:2408.04532*, 2024.

[10] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[11] Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context. *arXiv preprint arXiv:2410.01774*, 2024.

[12] Shaopeng Fu and Di Wang. Theoretical analysis of robust overfitting for wide DNNs: An NTK approach. In *International Conference on Learning Representations*, 2024.

[13] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Conference on Neural Information Processing Systems*, 2022.

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[17] Jonathan Hayase, Ema Borevković, Nicholas Carlini, Florian Tramèr, and Milad Nasr. Query-based adversarial prompt generation. In *Conference on Neural Information Processing Systems*, 2024.

[18] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[19] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

[20] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477*, 2020.

[21] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[22] Haibo Jin, Andy Zhou, Joe D. Menke, and Haohan Wang. Jailbreaking large language models against moderation guardrails via cipher characters. In *Conference on Neural Information Processing Systems*, 2024.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[24] Tong Li, Shu Yang, Junchao Wu, Jiyao Wei, Lijie Hu, Mengdi Li, Derek F Wong, Joshua R Oltmanns, and Di Wang. Can large language models identify implicit suicidal ideation? an empirical evaluation. *arXiv preprint arXiv:2502.17899*, 2025.

[25] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023.

[26] Zeyi Liao and Huan Sun. AmpleGCG: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed LLMs. In *Conference on Language Modeling*, 2024.

[27] Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, 2024.

[28] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[29] Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak LLMs. *arXiv preprint arXiv:2410.05295*, 2024.

[30] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 2024.

[31] Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2024.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[33] Roey Magen, Shuning Shang, Zhiwei Xu, Spencer Frei, Wei Hu, and Gal Vardi. Benign overfitting in single-head attention. *arXiv preprint arXiv:2410.07746*, 2024.

[34] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024.

[35] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[36] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Conference on Neural Information Processing Systems*, 35:27730–27744, 2022.

[38] Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. Advprompter: Fast adaptive adversarial prompting for LLMs. *arXiv preprint arXiv:2404.16873*, 2024.

[39] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[40] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

[41] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024.

[42] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Conference on Neural Information Processing Systems*, 36, 2023.

[43] Antonio H. Ribeiro, Dave Zachariah, Francis Bach, and Thomas B. Schön. Regularization properties of adversarially-trained linear regression. In *Conference on Neural Information Processing Systems*, 2023.

[44] Vinu Sankar Sadasivan, Shoumik Saha, Gaurang Sriramanan, Priyatham Kattakinda, Atoosa Chegini, and Soheil Feizi. Fast adversarial attacks on language models in one GPU minute. In *International Conference on Machine Learning*, 2024.

[45] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Günnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space. In *Conference on Neural Information Processing Systems*, 2024.

[46] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *arXiv preprint arXiv:2407.15549*, 2024.

[47] Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. Why larger language models do in-context learning differently? In *International Conference on Machine Learning*, 2024.

[48] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2014.

[50] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[53] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[54] Jiongxiao Wang, Zichen Liu, Keun Hee Park, Zhuojun Jiang, Zhaoheng Zheng, Zhuofeng Wu, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*, 2023.

[55] Liangyu Wang, Jie Ren, Hang Xu, Junxiao Wang, Huanyi Xie, David E Keyes, and Di Wang. Zo2: Scalable zeroth-order fine-tuning for extremely large language models with limited gpu memory. *arXiv preprint arXiv:2503.12668*, 2025.

[56] Liangyu Wang, Junxiao Wang, Jie Ren, Zihang Xiang, David E Keyes, and Di Wang. Flashdp: Private training large language models with efficient dp-sgd. *arXiv preprint arXiv:2507.01154*, 2025.

[57] Liangyu Wang, Huanyi Xie, and Di Wang. Distzo2: High-throughput and memory-efficient zeroth-order fine-tuning llms with distributed parallel computing. *arXiv preprint arXiv:2507.03211*, 2025.

[58] Yunjuan Wang, Kaibo Zhang, and Raman Arora. Benign overfitting in adversarial training of neural networks. In *International Conference on Machine Learning*, 2024.

[59] Zhijie Wang, Bo Jiang, and Shuai Li. In-context learning on function classes unveiled for transformers. In *International Conference on Machine Learning*, 2024.

[60] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Conference on Neural Information Processing Systems*, 2023.

[61] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

[62] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024.

[63] Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in LLMs with continuous attacks. In *Conference on Neural Information Processing Systems*, 2024.

[64] Zhao Xu, Fan Liu, and Hao Liu. Bag of tricks: Benchmarking of jailbreak attacks on LLMs. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[65] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[66] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In *Conference on Neural Information Processing Systems*, 2024.

[67] Junchi Yao, Jianhua Xu, Tianyu Xin, Ziyi Wang, Shenzhe Zhu, Shu Yang, and Di Wang. Is your llm-based multi-agent a reliable real-world planner? exploring fraud detection in travel planning. *arXiv preprint arXiv:2505.16557*, 2025.

[68] Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust LLM safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*, 2024.

[69] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[71] Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. AutoDAN: Interpretable gradient-based adversarial attacks on large language models. In *First Conference on Language Modeling*, 2024.

[72] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems*, 2024.

[73] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claim made by the abstract and introduction is that: short-length AT can effectively help LLMs defend against long-length jailbreak attacks, which is supported by both theoretical and empirical evidence. The theoretical evidence is justified in Section 4, while the empirical evidence is justified in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.2 discusses the limitation of using only a single jailbreak attack during AT to defend against unseen attacks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are stated as Assumption 1 and Assumption 2. All proofs are presented in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All necessary details to reproduce experimental results in this paper are provided in Section 5.1 and Appendix B. The experimental code is also provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Experimental code and detailed instructions are provided in the supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary details to reproduce experimental results in this paper are provided in Section 5.1 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See the `README.md` file and the `LICENSE` file in the submitted experimental code for details.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See the `README.md` file in the submitted experimental code for details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Proofs

This section collects all the proofs in this paper.

A.1 Technical lemmas

This section presents several technical lemmas that will be used in our proofs.

Lemma A.1 (c.f. Lemma D.2 in [69]). *If $x \in \mathbb{R}^{d\times 1}$ is Gaussian random vector of d dimension, mean zero and covariance matrix Λ, and $A \in \mathbb{R}^{d\times d}$ is a fixed matrix. Then*

$$\mathbb{E}[xx^\top Axx^\top] = \Lambda(A + A^\top)\Lambda + \mathrm{Tr}(A\Lambda)\Lambda.$$

Lemma A.2. *If $x \in \mathbb{R}^{d\times 1}$ is Gaussian random vector of d dimension, mean zero and covariance matrix Λ, and $A \in \mathbb{R}^{d\times d}$ is a fixed matrix. Then*

$$\mathbb{E}[x^\top Ax] = \mathrm{Tr}(A\Lambda).$$

Proof. Since

$$\mathbb{E}[x^\top Ax] = \mathbb{E}\Big[\sum_{i,j} x_i A_{i,j} x_j\Big] = \sum_{i,j} A_{i,j}\cdot\mathbb{E}[x_i x_j] = \sum_{i,j} A_{i,j}\cdot\Lambda_{i,j} = \sum_{i=1}^{d}(A\Lambda^\top)_{i,i} = \mathrm{Tr}(A\Lambda),$$

which completes the proof. □

Lemma A.3. *For any matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$, we have*

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Proof. Since

$$\mathrm{Tr}(AB) = \sum_{i=1}^{n}(AB)_{i,i} = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}B_{j,i} = \sum_{j=1}^{m}\sum_{i=1}^{n} B_{j,i}A_{i,j} = \sum_{j=1}^{m}(BA)_{j,j} = \mathrm{Tr}(BA),$$

which completes the proof. □

Lemma A.4 (From Lemma D.1 in [69]; Also in [39]). *Let $X \in \mathbb{R}^{n\times m}$ be a variable matrix and $A \in \mathbb{R}^{a\times n}$ and $B \in \mathbb{R}^{n\times m}$ be two fixed matrices. Then, we have*

$$\begin{aligned}
&\partial_X \mathrm{Tr}(BX^\top) = B \in \mathbb{R}^{n\times m},\\
&\partial_X \mathrm{Tr}(AXBX^\top) = (AXB + A^\top XB^\top) \in \mathbb{R}^{n\times m}.
\end{aligned}$$

Lemma A.5 (Von Neumann's Trace Inequality; Also in Lemma D.3 in [69]). *Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$ be two matrices. Suppose $(\sigma_1(A), \cdots \sigma_{\min\{n,m\}}(A))$ and $(\sigma_1(B), \cdots \sigma_{\min\{n,m\}}(B))$ are all the singular values of A and B, respectively. We have*

$$\mathrm{Tr}(AB) \le \sum_{i=1}^{\min\{n,m\}} \sigma_i(A)\sigma_i(B) \le \sum_{i=1}^{\min\{n,m\}} \|A\|_2\cdot\|B\|_2 = \min\{n,m\}\cdot\|A\|_2\cdot\|B\|_2.$$

A.2 Proof of Proposition 1

This section presents the proof of Proposition 1.

Proof of Proposition 1. For the AT loss $\mathcal{L}(\theta)$ defined in Eq. (8), we have that

$$\begin{aligned}
\mathcal{L}^{\text{adv}}(\theta) &:= \mathcal{R}^{\text{adv}}(\theta, M_{\text{train}}) = \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty}\le\epsilon} |\hat{y}_{q,\theta}(E^{\text{adv}}_{\tau,M_{\text{train}}}) - y_{\tau,q}|^2\\
&= \mathop{\mathbb{E}}_{\tau}\left\{\max_{\|\Delta_\tau^\top\|_{2,\infty}\le\epsilon} \frac{1}{2}\left|\begin{pmatrix}(w_{21}^V)^\top & w_{22}^V\end{pmatrix}\cdot\frac{E^{\text{adv}}_{\tau,M_{\text{train}}}E^{\text{adv},\top}_{\tau,M_{\text{train}}}}{N+M_{\text{train}}}\cdot\begin{pmatrix}W_{11}^{KQ}\\(w_{21}^{KQ})^\top\end{pmatrix}\cdot x_{\tau,q} - y_{\tau,q}\right|^2\right\}. \quad \text{(A.1)}
\end{aligned}$$

Then, the term $E^{\text{adv}}_{\tau,M_{\text{train}}} E^{\text{adv},\top}_{\tau,M_{\text{train}}}$ can be decomposed as follows,

$$
\begin{aligned}
E^{\text{adv}}_{\tau,M_{\text{train}}} E^{\text{adv},\top}_{\tau,M_{\text{train}}} &= \left(\begin{pmatrix} X_\tau \\ Y_\tau \end{pmatrix} \quad \begin{pmatrix} X^{\text{sfx}}_\tau + \Delta_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \quad \begin{pmatrix} x_{\tau,q} \\ 0 \end{pmatrix} \right) \cdot \left(\begin{pmatrix} X_\tau \\ Y_\tau \end{pmatrix} \quad \begin{pmatrix} X^{\text{sfx}}_\tau + \Delta_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \quad \begin{pmatrix} x_{\tau,q} \\ 0 \end{pmatrix} \right)^\top \\
&= \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}^\top + \begin{pmatrix} 0_{d\times N} & \Delta_\tau & 0_{d\times 1} \\ 0_{1\times N} & 0_{1\times M_{\text{train}}} & 0 \end{pmatrix} \begin{pmatrix} 0_{d\times N} & \Delta_\tau & 0_{d\times 1} \\ 0_{1\times N} & 0_{1\times M_{\text{train}}} & 0 \end{pmatrix}^\top \\
&\quad + \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix} \begin{pmatrix} 0_{d\times N} & \Delta_\tau & 0_{d\times 1} \\ 0_{1\times N} & 0_{1\times M_{\text{train}}} & 0 \end{pmatrix}^\top + \begin{pmatrix} 0_{d\times N} & \Delta_\tau & 0_{d\times 1} \\ 0_{1\times N} & 0_{1\times M_{\text{train}}} & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}^\top \\
&= \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}^\top + \begin{pmatrix} \Delta_\tau \\ 0_{1\times M_{\text{train}}} \end{pmatrix} \begin{pmatrix} \Delta_\tau \\ 0_{1\times M_{\text{train}}} \end{pmatrix}^\top \\
&\quad + \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \begin{pmatrix} \Delta_\tau \\ 0_{1\times M_{\text{train}}} \end{pmatrix}^\top + \begin{pmatrix} \Delta_\tau \\ 0_{1\times M_{\text{train}}} \end{pmatrix} \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix}^\top,
\end{aligned}
$$

which further means that

$$
\begin{aligned}
&\left((w^V_{21})^\top \quad w^V_{22} \right) \cdot \frac{E^{\text{adv}}_{\tau,M_{\text{train}}} E^{\text{adv},\top}_{\tau,M_{\text{train}}}}{N + M_{\text{train}}} \cdot \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} \cdot x_{\tau,q} \\
&= \left((w^V_{21})^\top \quad w^V_{22} \right) \cdot \frac{\begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}^\top}{N + M_{\text{train}}} \cdot \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} \cdot x_{\tau,q} \\
&\quad + (w^V_{21})^\top \cdot \frac{\Delta_\tau \Delta_\tau^\top}{N + M_{\text{train}}} \cdot W^{KQ}_{11} x_{\tau,q} + \left((w^V_{21})^\top \quad w^V_{22} \right) \cdot \frac{\begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \Delta_\tau^\top}{N + M_{\text{train}}} \cdot W^{KQ}_{11} x_{\tau,q} \\
&\quad + (w^V_{21})^\top \cdot \frac{\Delta_\tau \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix}^\top}{N + M_{\text{train}}} \cdot \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} x_{\tau,q}. \qquad \text{(A.2)}
\end{aligned}
$$

Inserting Eq. (A.2) into Eq. (A.1) and applying the inequality that $|a+b|^2 \le 2 \cdot (a^2 + b^2)$, $\mathcal{L}^{\text{adv}}(\theta)$ can thus be bounded as

$$
\begin{aligned}
\mathcal{L}^{\text{adv}}(\theta) &\le 2 \cdot \mathop{\mathbb{E}}_{\tau} \Bigg[\left((w^V_{21})^\top \quad w^V_{22} \right) \cdot \frac{\begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X^{\text{sfx}}_\tau & x_{\tau,q} \\ Y_\tau & Y^{\text{sfx}}_\tau & 0 \end{pmatrix}^\top}{N + M_{\text{train}}} \cdot \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} \cdot x_{\tau,q} - y_{\tau,q} \Bigg]^2 \\
&\quad + \underbrace{2 \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \left[(w^V_{21})^\top \cdot \frac{\Delta_\tau \Delta_\tau^\top}{N + M_{\text{train}}} \cdot W^{KQ}_{11} x_{\tau,q} \right]^2}_{:=A_1(\theta)} \\
&\quad + \underbrace{2 \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \left[\left((w^V_{21})^\top \quad w^V_{22} \right) \cdot \frac{\begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix} \Delta_\tau^\top}{N + M_{\text{train}}} \cdot W^{KQ}_{11} x_{\tau,q} \right]^2}_{:=A_2(\theta)} \\
&\quad + \underbrace{2 \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \left[(w^V_{21})^\top \cdot \frac{\Delta_\tau \begin{pmatrix} X^{\text{sfx}}_\tau \\ Y^{\text{sfx}}_\tau \end{pmatrix}^\top}{N + M_{\text{train}}} \cdot \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^\top \end{pmatrix} x_{\tau,q} \right]^2}_{:=A_3(\theta)}. \qquad \text{(A.3)}
\end{aligned}
$$

We then bound terms $A_1(\theta)$, $A_2(\theta)$, and $A_3(\theta)$ in Eq. (A.3) seprately. For the term $A_1(\theta)$ in Eq. (A.3), we have

$$
\begin{aligned}
A_1(\theta) &:= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \Big[(w_{21}^V)^\top \cdot \sum_{i=1}^{M_{\text{train}}} \delta_{\tau,i}\delta_{\tau,i}^\top \cdot W_{11}^{KQ} x_{\tau,q} \Big]^2 \\
&\le \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \underbrace{\Big[\sum_{i=1}^{M_{\text{train}}} [(w_{21}^V)^\top \delta_{\tau,i}]^2 \cdot \sum_{i=1}^{M_{\text{train}}} [\delta_{\tau,i}^\top W_{11}^{KQ} x_{\tau,q}]^2 \Big]}_{\text{by Cauchy-Schwarz Inequality}} \\
&\le \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big[\sum_{i=1}^{M_{\text{train}}} \max_{\|\delta_{\tau,i}\|_2 \le \epsilon} [(w_{21}^V)^\top \delta_{\tau,i}]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \max_{\|\delta_{\tau,i}\|_2 \le \epsilon} [\delta_i^\top W_{11}^{KQ} x_{\tau,q}]^2 \Big] \\
&= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big[\sum_{i=1}^{M_{\text{train}}} [\|w_{21}^V\|_2 \cdot \epsilon]^2 \cdot \sum_{i=1}^{M_{\text{train}}} [\|W_{11}^{KQ} x_{\tau,q}\|_2 \cdot \epsilon]^2 \Big] \\
&= \frac{2\epsilon^4 M_{\text{train}}^2}{(N+M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau} \|W_{11}^{KQ} x_{\tau,q}\|_2^2.
\end{aligned} \tag{A.4}
$$

For the term $A_2(\theta)$ in Eq. (A.3), we have

$$
\begin{aligned}
A_2(\theta) &:= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \cdot \sum_{i=1}^{M_{\text{train}}} \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \delta_{\tau,i}^\top \cdot W_{11}^{KQ} x_{\tau,q} \Big]^2 \\
&\le \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \underbrace{\Big[\sum_{i=1}^{M_{\text{train}}} \Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \Big]^2 \cdot \sum_{i=1}^{M_{\text{train}}} [\delta_{\tau,i}^\top W_{11}^{KQ} x_{\tau,q}]^2 \Big]}_{\text{by Cauchy-Schwarz Inequality}} \\
&= \frac{2}{(N+M_{\text{train}})^2} \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} \Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \Big]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} \Big[\max_{\|\delta_{\tau,i}\|_2 \le \epsilon} [\delta_{\tau,i}^\top W_{11}^{KQ} x_{\tau,q}]^2 \Big] \\
&= \frac{2}{(N+M_{\text{train}})^2} \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} \Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \Big]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} [\|W_{11}^{KQ} x_{\tau,q}\|_2 \cdot \epsilon]^2 \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} \Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \Big]^2.
\end{aligned} \tag{A.5}
$$

For the term $A_3(\theta)$ in Eq. (A.3), we have

$$
\begin{aligned}
A_3(\theta) &:= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \Big[(w_{21}^V)^\top \cdot \sum_{i=1}^{M_{\text{train}}} \delta_{\tau,i} \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \cdot \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} \Big]^2 \\
&\le \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \max_{\|\Delta_\tau^\top\|_{2,\infty} \le \epsilon} \underbrace{\Big[\sum_{i=1}^{M_{\text{train}}} [(w_{21}^V)^\top \delta_{\tau,i}]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} \Big]^2 \Big]}_{\text{by Cauchy-Schwarz Inequality}} \\
&= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big[\sum_{i=1}^{M_{\text{train}}} \max_{\|\delta_{\tau,i}\|_2 \le \epsilon} [(w_{21}^V)^\top \delta_{\tau,i}]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} \Big]^2 \Big] \\
&= \frac{2}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big[\sum_{i=1}^{M_{\text{train}}} [\|w_{21}^V\|_2 \cdot \epsilon]^2 \cdot \sum_{i=1}^{M_{\text{train}}} \Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} \Big]^2 \Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau} \Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} \Big]^2.
\end{aligned} \tag{A.6}
$$

As a result, by inserting Eqs. (A.4), (A.5), and (A.6) into Eq. (A.3), we finally have that

$$
\begin{aligned}
\mathcal{L}^{\text{adv}}(\theta) \leq\ & 2 \cdot \mathop{\mathbb{E}}_{\tau}\Big[\big((w_{21}^V)^\top \quad w_{22}^V\big) \cdot \frac{\begin{pmatrix} X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q} \\ Y_\tau & Y_\tau^{\text{sfx}} & 0 \end{pmatrix}\begin{pmatrix} X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q} \\ Y_\tau & Y_\tau^{\text{sfx}} & 0 \end{pmatrix}^\top}{N + M_{\text{train}}} \cdot \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} \cdot x_{\tau,q} - y_{\tau,q}\Big]^2 \\
& + \frac{2\epsilon^4 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau} \|W_{11}^{KQ} x_{\tau,q}\|_2^2 \\
& + \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau}\Big[\big((w_{21}^V)^\top \quad w_{22}^V\big)\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}\Big]^2 \\
& + \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau}\Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\Big]^2.
\end{aligned} \tag{A.7}
$$

The right-hand-side of Eq. (A.7) is exactly the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (9), which thus completes the proof. □

A.3 Proof of Theorem 1

This section presents the proof of Theorem 1, which is inspired by that in [69]. Specifically:

1. we first prove that terms w_{21}^V and w_{21}^{KQ} stay zero during the surrogate AT (Lemma A.6) via continuous gradient-flow, which thus can simplify the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ defined in Eq. (9) (Lemma A.7).
2. We then calculate a closed-form solution θ_* for the surrogate AT problem based on the simplified $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ (Lemma A.8), which is exactly the solution given in Theorem 1.
3. Finally, we prove that under the continuous gradient flow, the LSA model starts from the initial point defined in Assumption 1 can indeed converge to the closed-form solution θ_* (Lemma A.12), which thus completes the proof of Theorem 1.

We now start to prove the following Lemma A.6.

Lemma A.6. *Suppose Assumption 1 holds and the LSA model $f_{\text{LSA},\theta}$ is trained via minimizing surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (9) with continuous gradient flow. Then, for any continuous training time $t \geq 0$, we uniformly have that $w_{21}^V(t) = w_{21}^{KQ}(t) = 0_{d\times 1}$.*

Proof. When the LSA model $f_{\text{LSA},\theta}$ is trained with continuous gradient-flow, the updates of w_{21}^V and w_{21}^{KQ} with respect to the continuous training time $t \geq 0$ are given by

$$
\begin{aligned}
\partial_t w_{21}^V(t) &:= -\partial_{w_{21}^V} \tilde{\mathcal{L}}^{\text{adv}}(\theta), \\
\partial_t w_{21}^{KQ}(t) &:= -\partial_{w_{21}^{KQ}} \tilde{\mathcal{L}}^{\text{adv}}(\theta).
\end{aligned}
$$

Meanwhile, since Assumption 1 assumes that $w_{21}^V(0) = W_{21}^{KQ}(0) = 0_{d\times 1}$, therefore, to complete the proof, we only need to show that $\partial_t w_{21}^V(t) = \partial_t W_{21}^{KQ}(t) = 0_{1\times d}$ as long as $w_{21}^V(t) = W_{21}^{KQ}(t) = 0_{d\times 1}$ for any $t \geq 0$. In other words, below we need to show that $w_{21}^V = W_{21}^{KQ} = 0_{d\times 1}$ indicates $\partial_{w_{21}^V}\tilde{\mathcal{L}}^{\text{adv}}(\theta) = \partial_{w_{21}^{KQ}}\tilde{\mathcal{L}}^{\text{adv}}(\theta) = 0_{1\times d}$.

Toward this end, we adopt the notation in Eq. (9) to decompose the surrogate AT loss $\tilde{\mathcal{L}}(\theta)$ as follows,

$$
\tilde{\mathcal{L}}^{\text{adv}}(\theta) := [\ell_1(\theta) + \ell_2(\theta) + \ell_3(\theta) + \ell_4(\theta)],
$$

where

$$\ell_1(\theta) = 2\mathop{\mathbb{E}}_{\tau}\Big[((w_{21}^V)^\top \;\; w_{22}^V) \frac{\begin{pmatrix} X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q} \\ Y_\tau & Y_\tau^{\text{sfx}} & 0 \end{pmatrix} \begin{pmatrix} X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q} \\ Y_\tau & Y_\tau^{\text{sfx}} & 0 \end{pmatrix}^\top}{N + M_{\text{train}}} \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q} - y_{\tau,q}\Big]^2, \tag{A.8}$$

$$\ell_2(\theta) = \frac{2\epsilon^4 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \|w_{21}^V\|_2^2 \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big], \tag{A.9}$$

$$\ell_3(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot \|((w_{21}^V)^\top \;\; w_{22}^V) \begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}\|_2^2\Big], \tag{A.10}$$

$$\ell_4(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \|w_{21}^V\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau}\Big[\| \begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2\Big]. \tag{A.11}$$

In the remaining of this proof, we will show that when $w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}$ holds, one has: (1) $\partial_{W_{21}^V}\ell_1(\theta) = \partial_{W_{21}^{KQ}}\ell_1(\theta) = 0_{1\times d}$, (2) $\partial_{W_{21}^V}\ell_2(\theta) = \partial_{W_{21}^{KQ}}\ell_2(\theta) = 0_{1\times d}$, (3) $\partial_{W_{21}^V}\ell_3(\theta) = \partial_{W_{21}^{KQ}}\ell_3(\theta) = 0_{1\times d}$, and (4) $\partial_{W_{21}^V}\ell_4(\theta) = \partial_{W_{21}^{KQ}}\ell_4(\theta) = 0_{1\times d}$, which thus automatically indicates that $\partial_{W_{21}^V}\tilde{\mathcal{L}}^{\text{adv}}(\theta) = \partial_{W_{21}^{KQ}}\tilde{\mathcal{L}}^{\text{adv}}(\theta) = 0_{1\times d}$.

Step 1: Show that $w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}$ **indicates** $\partial_{W_{21}^V}\ell_1(\theta) = \partial_{W_{21}^{KQ}}\ell_1(\theta) = 0_{1\times d}$**.** Such a claim can be directly obtained from the proofs in [69]. Specifically, when setting the (original) ICL prompt length from N to $(N + M_{\text{train}})$, the ICL training loss L in [69] is equivalent to our $\ell_1(\theta)$ defined in Eq. (A.8). Therefore, one can then follow the same procedures as those in the proof of Lemma 5.2 in [69] to show that the continuous gradient flows of W_{21}^V and W_{21}^{KQ} are zero when Assumption 1 holds. Please refer accordingly for details.

Step 2: Show that $w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}$ **indicates** $\partial_{w_{21}^V}\ell_2(\theta) = \partial_{w_{21}^{KQ}}\ell_2(\theta) = 0_{1\times d}$**.** Since the term w_{21}^{KQ} does not exist in the expression of $\ell_2(\theta)$ in Eq. (A.9), we directly have that $\partial_{w_{21}^{KQ}}\ell_2(\theta) = 0_{1\times d}$. Besides, for the derivative $\partial_{w_{21}^V}\ell_2(\theta)$, based on Eq. (A.9) we further have that

$$\begin{aligned}
&\partial_{w_{21}^V}\ell_2(\theta)\Big|_{w_{21}^V = 0_{d\times 1}} = \partial_{w_{21}^V}\Big[\frac{2\epsilon^4 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau}\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big]\Big|_{w_{21}^V = 0_{d\times 1}} \\
&= \Big[\frac{4\epsilon^4 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot (w_{21}^V)^\top\Big]\Big|_{w_{21}^V = 0_{d\times 1}} \\
&= \frac{4\epsilon^4 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot 0_{d\times 1}^\top = 0_{1\times d},
\end{aligned}$$

which justifies our claim in Step 2.

Step 3: Show that $w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}$ **indicates** $\partial_{w_{21}^V}\ell_3(\theta) = \partial_{w_{21}^{KQ}}\ell_3(\theta) = 0_{1\times d}$**.** We first rewrite $\ell_3(\theta)$ that defined in Eq. (A.10) as follows,

$$\begin{aligned}
&\ell_3(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2 \cdot \|((w_{21}^V)^\top \;\; w_{22}^V) \begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}\|_2^2\Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau}\Big[((w_{21}^V)^\top \;\; w_{22}^V) \cdot \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top \cdot ((w_{21}^V)^\top \;\; w_{22}^V)^\top\Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot ((w_{21}^V)^\top \;\; w_{22}^V) \cdot \left(\sum_{i=1}^{M_{\text{train}}} \mathop{\mathbb{E}}_{\tau}\Big[\begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix} \begin{pmatrix} x_{\tau,i}^{\text{sfx}} \\ y_{\tau,i}^{\text{sfx}} \end{pmatrix}^\top\Big]\right) \cdot ((w_{21}^V)^\top \;\; w_{22}^V)^\top.
\end{aligned} \tag{A.12}$$

Then, for any $i \in [M]$ we have

$$\mathop{\mathbb{E}}_{\tau}\Big[\begin{pmatrix} x^{\text{sfx}}_{\tau,i} \\ y^{\text{sfx}}_{\tau,i} \end{pmatrix} \begin{pmatrix} x^{\text{sfx}}_{\tau,i} \\ y^{\text{sfx}}_{\tau,i} \end{pmatrix}^{\top}\Big] = \mathop{\mathbb{E}}_{w_\tau, x^{\text{sfx}}_{\tau,i}} \begin{pmatrix} x^{\text{sfx}}_{\tau,i} \cdot (x^{\text{sfx}}_{\tau,i})^{\top} & x^{\text{sfx}}_{\tau,i} \cdot (w_\tau^{\top} x^{\text{sfx}}_{\tau,i})^{\top} \\ w_\tau^{\top} x^{\text{sfx}}_{\tau,i} \cdot (x^{\text{sfx}}_{\tau,i})^{\top} & w_\tau^{\top} x^{\text{sfx}}_{\tau,i} \cdot (w_\tau^{\top} x^{\text{sfx}}_{\tau,i})^{\top} \end{pmatrix}$$

$$= \begin{pmatrix} \Lambda & \Lambda \cdot 0_{d\times 1} \\ 0_{1\times d} \cdot \Lambda & \mathbb{E}_{w_\tau}\Big[w_\tau^{\top} \Lambda w_\tau\Big] \end{pmatrix} = \begin{pmatrix} \Lambda & 0_{d\times 1} \\ 0_{1\times d} & \underbrace{\operatorname{Tr}(I_d \Lambda)}_{\text{by Lemma A.2}} \end{pmatrix} = \begin{pmatrix} \Lambda & 0_{d\times 1} \\ 0_{1\times d} & \operatorname{Tr}(\Lambda) \end{pmatrix}. \tag{A.13}$$

Finally, by inserting Eq. (A.13) into Eq. (A.12), $\ell_3(\theta)$ can thus be simplified as follows,

$$\ell_3(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot \big((w^V_{21})^{\top} \quad w^V_{22}\big) \cdot \left(\sum_{i=1}^{M_{\text{train}}} \begin{pmatrix} \Lambda & 0_{d\times 1} \\ 0_{1\times d} & \operatorname{Tr}(\Lambda) \end{pmatrix}\right) \cdot \big((w^V_{21})^{\top} \quad w^V_{22}\big)^{\top}$$

$$= \frac{2\epsilon^2 M^2_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot \Big((w^V_{21})^{\top} \Lambda w^V_{21} + \operatorname{Tr}(\Lambda)(w^V_{22})^2\Big). \tag{A.14}$$

According to Eq. (A.14), $\ell_3(\theta)$ does not depend on w^{KQ}_{21}, which means that $\partial_{w^{KQ}_{21}} \ell_3(\theta) = 0_{1\times d}$. On the other hand, based on Eq. (A.14), when $w^V_{21} = 0$, the derivative of $\ell_3(\theta)$ with respect to w^V_{21} is calculated as follows,

$$\partial_{w^V_{21}} \ell_3(\theta)\Big|_{w^V_{21}=0} = \partial_{w^V_{21}} \Big[\frac{2\epsilon^2 M^2_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot \Big((w^V_{21})^{\top} \Lambda w^V_{21} + \operatorname{Tr}(\Lambda)(w^V_{22})^2\Big)\Big]\Big|_{w^V_{21}=0}$$

$$= \frac{2\epsilon^2 M^2_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot \partial_{w^V_{21}} \Big[(w^V_{21})^{\top} \Lambda w^V_{21}\Big]\Big|_{w^V_{21}=0}$$

$$= \frac{4\epsilon^2 M^2_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot \Big[(w^V_{21})^{\top} \Lambda\Big]\Big|_{w^V_{21}=0}$$

$$= \frac{4\epsilon^2 M^2_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W^{KQ}_{11} x_{\tau,q}\|_2^2\Big] \cdot 0^{\top}_{d\times 1} \Lambda = 0_{1\times d},$$

which justifies our claim in Step 3.

Step 4: Show that $w^V_{21} = w^{KQ}_{21} = 0_{d\times 1}$ indicates $\partial_{w^V_{21}} \ell_4(\theta) = \partial_{w^{KQ}_{21}} \ell_4(\theta) = 0_{1\times d}$. When $w^V_{21} = w^{KQ}_{21} = 0_{d\times 1}$, based on the expression of $\ell_4(\theta)$ given in Eq. (A.11), the derivative of $\ell_4(\theta)$ with respect to w^V_{21} is calculated as follows,

$$\partial_{w^V_{21}} \ell_4(\theta)\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}} = \partial_{w^V_{21}} \Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \|w^V_{21}\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^{\top} \end{pmatrix} x_{\tau,q} \Big\|_2^2\Big]\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}}$$

$$= \Big[\frac{4\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^{\top} \end{pmatrix} x_{\tau,q} \Big\|_2^2 \cdot (w^V_{21})^{\top}\Big]\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}}$$

$$= \frac{4\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ 0^{\top}_{d\times 1} \end{pmatrix} x_{\tau,q} \Big\|_2^2 \cdot 0^{\top}_{d\times 1} = 0_{1\times d}.$$

Besides, for the derivative of $\ell_4(\theta)$ with respect to w^{KQ}_{21}, we also have that

$$\partial_{w^{KQ}_{21}} \ell_4(\theta)\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}} = \partial_{w^{KQ}_{21}} \Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \|w^V_{21}\|_2^2 \cdot \mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^{\top} \end{pmatrix} x_{\tau,q} \Big\|_2^2\Big]\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}}$$

$$= \Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \|w^V_{21}\|_2^2 \cdot \partial_{w^{KQ}_{21}} \mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^{\top} \end{pmatrix} x_{\tau,q} \Big\|_2^2\Big]\Big|_{w^V_{21}=w^{KQ}_{21}=0_{d\times 1}}$$

$$= \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \|0_{d\times 1}\|_2^2 \cdot \partial_{w^{KQ}_{21}} \Big[\mathop{\mathbb{E}}_{\tau} \Big\| \begin{pmatrix} X^{\text{sfx}}_{\tau} \\ Y^{\text{sfx}}_{\tau} \end{pmatrix}^{\top} \begin{pmatrix} W^{KQ}_{11} \\ (w^{KQ}_{21})^{\top} \end{pmatrix} x_{\tau,q} \Big\|_2^2\Big]\Big|_{w^{KQ}_{21}=0_{d\times 1}} = 0_{1\times d}.$$

The above two equations justify the claim in Step 4.

Step 5: Based on results from previous Steps 1 to 4, we eventually have that

$$\partial_{w_{21}^V}\tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big|_{w_{21}^V=w_{21}^{KQ}=0_{d\times 1}} = \partial_{w_{21}^V}[\ell_1(\theta)+\ell_2(\theta)+\ell_3(\theta)+\ell_4(\theta)]\Big|_{w_{21}^V=w_{21}^{KQ}=0_{d\times 1}} = \sum_{i=1}^{4} 0_{1\times d} = 0_{1\times d},$$

$$\partial_{w_{21}^{KQ}}\tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big|_{w_{21}^V=w_{21}^{KQ}=0_{d\times 1}} = \partial_{w_{21}^{KQ}}[\ell_1(\theta)+\ell_2(\theta)+\ell_3(\theta)+\ell_4(\theta)]\Big|_{w_{21}^V=w_{21}^{KQ}=0_{d\times 1}} = \sum_{i=1}^{4} 0_{1\times d} = 0_{1\times d}.$$

The proof is completed. □

With Lemma A.6, we can then simplify the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$, as shown in the following Lemma A.7.

Lemma A.7. *Under Assumption 1, the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ defined in Eq. (9) can be simplified as follows,*

$$\begin{aligned}\tilde{\mathcal{L}}^{\text{adv}}(\theta) = {} & 2\text{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big] \\ & - 4\text{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big] + 2\text{Tr}(\Lambda),\end{aligned}$$

where $\Gamma(M) := \frac{N+M+1}{N+M}\Lambda + \frac{\text{Tr}(\Lambda)}{N+M}I_d$ and $\psi(M) := \frac{M^2\text{Tr}(\Lambda)}{(N+M)^2}$ are same functions as that defined in Eq. (10).

Proof. When Assumption 1 holds, by applying Lemma A.6, one can substitute terms w_{21}^V and w_{21}^{KQ} in the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ with the zero vector $0_{d\times 1}$, which thus simplifies $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ as follows,

$$\begin{aligned}\tilde{\mathcal{L}}^{\text{adv}}(\theta) &= 2\mathop{\mathbb{E}}_{\tau}\Big[\begin{pmatrix}0_{1\times d} & w_{22}^V\end{pmatrix}\frac{\begin{pmatrix}X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q}\\ Y_\tau & Y_\tau^{\text{sfx}} & 0\end{pmatrix}\begin{pmatrix}X_\tau & X_\tau^{\text{sfx}} & x_{\tau,q}\\ Y_\tau & Y_\tau^{\text{sfx}} & 0\end{pmatrix}^\top}{N+M_{\text{train}}}\begin{pmatrix}W_{11}^{KQ}\\ 0_{1\times d}\end{pmatrix}x_{\tau,q} - y_{\tau,q}\Big]^2 \\ &\qquad + 0 + \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2}\mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ}x_{\tau,q}\|_2^2\cdot\|\begin{pmatrix}0_{1\times d} & w_{22}^V\end{pmatrix}\begin{pmatrix}X_\tau^{\text{sfx}}\\ Y_\tau^{\text{sfx}}\end{pmatrix}\|_2^2\Big] + 0 \\ &= \underbrace{2\cdot\mathop{\mathbb{E}}_{\tau}\Big[w_{22}^V\cdot\frac{Y_\tau X_\tau + Y_\tau^{\text{sfx}}X_\tau^{\text{sfx}}}{N+M_{\text{train}}}\cdot W_{11}^{KQ}x_{\tau,q} - y_{\tau,q}\Big]^2}_{:=B_1(\theta)} \\ &\quad + \underbrace{\frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2}\cdot\mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ}x_{\tau,q}\|_2^2\cdot\|w_{22}^V Y_\tau^{\text{sfx}}\|_2^2\Big]}_{:=B_2(\theta)}. \end{aligned} \tag{A.15}$$

For the term $B_1(\theta)$ in Eq. (A.15), we have that

$$\begin{aligned} B_1(\theta) &:= 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[w_{22}^V\cdot\frac{Y_\tau X_\tau^\top + Y_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top}{N+M_{\text{train}}}\cdot W_{11}^{KQ}x_{\tau,q} - y_{\tau,q}\Big]^2 \\ &= 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[\frac{w_\tau^\top\cdot(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)}{N+M_{\text{train}}}\cdot w_{22}^V W_{11}^{KQ}\cdot x_{\tau,q} - w_\tau^\top x_{\tau,q}\Big]^2 \\ &= 2\cdot\mathop{\mathbb{E}}_{\tau}\Bigg[\Big[\frac{X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top}{N+M_{\text{train}}}\cdot w_{22}^V W_{11}^{KQ}x_{\tau,q} - x_{\tau,q}\Big]^\top\cdot w_\tau w_\tau^\top\cdot\Big[\frac{X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top}{N+M_{\text{train}}}\cdot w_{22}^V W_{11}^{KQ}x_{\tau,q} - x_{\tau,q}\Big]\Bigg] \\ &= 2\cdot\mathop{\mathbb{E}}_{\tau}\Bigg[\Big[\frac{X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top}{N+M_{\text{train}}}\cdot w_{22}^V W_{11}^{KQ}x_{\tau,q} - x_{\tau,q}\Big]^\top\cdot I_d\cdot\Big[\frac{X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top}{N+M_{\text{train}}}\cdot w_{22}^V W_{11}^{KQ}x_{\tau,q} - x_{\tau,q}\Big]\Bigg] \\ &= 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot(w_{22}^V W_{11}^{KQ})^\top\cdot\frac{\mathbb{E}_\tau\Big[(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)\Big]}{(N+M_{\text{train}})^2}\cdot w_{22}^V W_{11}^{KQ}\cdot x_{\tau,q}\Big] \\ &\quad - 4\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot\frac{\mathbb{E}_\tau\Big[(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)\Big]}{N+M_{\text{train}}}\cdot(w_{22}^V W_{11}^{KQ})\cdot x_{\tau,q}\Big] + 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot x_{\tau,q}\Big]. \end{aligned} \tag{A.16}$$

For $\mathbb{E}_\tau\left[(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)\right]$ in Eq. (A.16), we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\tau}\Big[(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)(X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top)\Big]\\
&= \mathop{\mathbb{E}}_{\tau}[X_\tau X_\tau^\top X_\tau X_\tau^\top] + \mathop{\mathbb{E}}_{\tau}[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top]\cdot\mathop{\mathbb{E}}_{\tau}[X_\tau X_\tau^\top]\\
&\quad + \mathop{\mathbb{E}}_{\tau}[X_\tau X_\tau^\top]\cdot\mathop{\mathbb{E}}_{\tau}[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top] + \mathop{\mathbb{E}}_{\tau}[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top]\\
&= \mathop{\mathbb{E}}_{\tau}\Big[\sum_{i,j} x_{\tau,i}x_{\tau,i}^\top x_{\tau,j}x_{\tau,j}^\top\Big] + \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top\Big]\cdot\mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top\Big]\\
&\quad + \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top\Big]\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top\Big] + \mathop{\mathbb{E}}_{\tau}\Big[\sum_{i,j} x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top x_{\tau,j}^{\text{sfx}}(x_{\tau,j}^{\text{sfx}})^\top\Big]\\
&= \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top x_{\tau,i}x_{\tau,i}^\top + \sum_{1\le i,j\le N, i\neq j}\Lambda^2\Big] + M_{\text{train}}\Lambda\cdot N\Lambda + N\Lambda\cdot M_{\text{train}}\Lambda\\
&\quad + \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top + \sum_{1\le i,j\le M_{\text{train}}, i\neq j}\Lambda^2\Big]\\
&= \mathop{\mathbb{E}}_{\tau}\Big[\sum_{i=1}^{N}\underbrace{(2\Lambda^2 + \mathrm{Tr}(\Lambda)\Lambda)}_{\text{by Lemma A.1}}\Big] + (N^2 - N)\cdot\Lambda^2 + 2NM_{\text{train}}\cdot\Lambda^2\\
&\quad + \mathop{\mathbb{E}}_{\tau}\Big[\sum_{i=1}^{M_{\text{train}}}\underbrace{(2\Lambda^2 + \mathrm{Tr}(\Lambda)\Lambda)}_{\text{by Lemma A.1}}\Big] + (M_{\text{train}}^2 - M_{\text{train}})\cdot\Lambda^2\\
&= (N^2 + N + M_{\text{train}}^2 + M_{\text{train}} + 2NM_{\text{train}})\cdot\Lambda^2 + (N + M_{\text{train}})\cdot\mathrm{Tr}(\Lambda)\cdot\Lambda\\
&= (N + M_{\text{train}})\cdot((N + M_{\text{train}} + 1)\cdot\Lambda^2 + \mathrm{Tr}(\Lambda)\cdot\Lambda) = (N + M_{\text{train}})^2\cdot\Gamma(M_{\text{train}})\Lambda.
\end{aligned}
\tag{A.17}
$$

For $\mathbb{E}_\tau\left[X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top\right]$ in Eq. (A.16), we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\tau}\Big[X_\tau X_\tau^\top + X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top\Big]\\
&= \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top\Big] + \mathop{\mathbb{E}}_{\tau}\Big[\sum_i x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top\Big]\\
&= N\Lambda + M_{\text{train}}\Lambda = (N + M_{\text{train}})\cdot\Lambda.
\end{aligned}
\tag{A.18}
$$

Inserting Eqs. (A.17) and (A.18) into Eq. (A.16) leads to

$$
\begin{aligned}
B_1(\theta) &= 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot(w_{22}^V W_{11}^{KQ})^\top\cdot\Gamma(M_{\text{train}})\Lambda\cdot w_{22}^V W_{11}^{KQ}\cdot x_{\tau,q}\Big]\\
&\quad - 4\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot\Lambda\cdot w_{22}^V W_{11}^{KQ}\cdot x_{\tau,q}\Big] + 2\cdot\mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top\cdot x_{\tau,q}\Big].\\
&= 2\cdot\underbrace{\mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ})^\top\cdot\Gamma(M_{\text{train}})\Lambda\cdot w_{22}^V W_{11}^{KQ}\cdot\Lambda\Big]}_{\text{by Lemma A.2}}\\
&\quad - 4\cdot\underbrace{\mathrm{Tr}\Big[\Lambda\cdot w_{22}^V W_{11}^{KQ}\cdot\Lambda\Big]}_{\text{by Lemma A.2}} + 2\cdot\mathrm{Tr}(\Lambda)\\
&= 2\cdot\underbrace{\mathrm{Tr}\Big[\Gamma(M_{\text{train}})\Lambda\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big]}_{\text{by Lemma A.3}}\\
&\quad - 4\cdot\underbrace{\mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big]}_{\text{by Lemma A.3}} + 2\cdot\mathrm{Tr}(\Lambda).
\end{aligned}
\tag{A.19}
$$

Besides, for the term $B_2(\theta)$ in Eq. (A.15), we have that

$$
\begin{aligned}
B_2(\theta) &:= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot \mathop{\mathbb{E}}_{\tau}\Big[\|W_{11}^{KQ}x_{\tau,q}\|_2^2 \cdot \|w_{22}^V Y_\tau^{\text{sfx}}\|_2^2\Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot (w_{22}^V)^2 \cdot \mathop{\mathbb{E}}_{\tau}\Big[x_{\tau,q}^\top \cdot (W_{11}^{KQ})^\top W_{11}^{KQ} \cdot x_{\tau,q}\Big] \cdot \mathop{\mathbb{E}}_{\tau}\Big[w_\tau^\top \cdot X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top \cdot w_\tau\Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot (w_{22}^V)^2 \cdot \underbrace{\mathrm{Tr}\Big[(W_{11}^{KQ})^\top W_{11}^{KQ} \cdot \Lambda\Big]}_{\text{by Lemma A.2}} \cdot \mathop{\mathbb{E}}_{\tau}\Big[w_\tau^\top \cdot M_{\text{train}}\Lambda \cdot w_\tau\Big] \\
&= \frac{2\epsilon^2 M_{\text{train}}}{(N+M_{\text{train}})^2} \cdot (w_{22}^V)^2 \cdot \underbrace{\mathrm{Tr}\Big[W_{11}^{KQ} \cdot \Lambda \cdot (W_{11}^{KQ})^\top\Big]}_{\text{by Lemma A.3}} \cdot \underbrace{\mathrm{Tr}\Big[M_{\text{train}}\Lambda \cdot I_d\Big]}_{\text{by Lemma A.2}} \\
&= 2\epsilon^2 \cdot \frac{M_{\text{train}}^2 \mathrm{Tr}(\Lambda)}{(N+M_{\text{train}}\Lambda^{\frac{1}{2}})^2} \cdot \mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda \cdot (w_{22}^V W_{11}^{KQ})^\top\Big] \\
&= 2\epsilon^2 \cdot \psi(M_{\text{train}}) \cdot \mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda \cdot (w_{22}^V W_{11}^{KQ})^\top\Big].
\end{aligned} \tag{A.20}
$$

Finally, by inserting Eqs. (A.19) and (A.20) into Eq. (A.15), we have

$$
\begin{aligned}
&\tilde{\mathcal{L}}^{\text{adv}}(\theta) \\
&:= 2 \cdot \mathrm{Tr}\Big[\Gamma(M_{\text{train}})\Lambda \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big] - 4 \cdot \mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda^{\frac{3}{2}}\Big] \\
&\quad + 2 \cdot \mathrm{Tr}(\Lambda) + 2\epsilon^2 \cdot \psi(M_{\text{train}}) \cdot \mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda \cdot (w_{22}^V W_{11}^{KQ})^\top\Big] \\
&= 2 \cdot \mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d) \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big] \\
&\quad - 4 \cdot \mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda^{\frac{3}{2}}\Big] + 2 \cdot \mathrm{Tr}(\Lambda),
\end{aligned}
$$

which completes the proof. □

Based on the simplified surrogate AT loss, the closed-form global minimizer θ_* for the surrogate AT problem is then calculated in the following Lemma A.8.

Lemma A.8. *Suppose Assumption 1 holds. Then, $\theta_* := (W_*^V W_*^{KQ})$ is a minimizer for the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (8) if and only if $w_{*,22}^V W_{*,11}^{KQ} = (\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda$.*

Proof. For the simplified surrogate AT loss proved in Lemma A.7, we rewrite it as follows,

$$
\begin{aligned}
&\tilde{\mathcal{L}}^{\text{adv}}(\theta) \\
&= 2\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d) \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot (w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big] \\
&\quad - 4\mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}}) \cdot \Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda) \\
&= 2 \cdot \mathrm{Tr}\Big[(\Gamma_{\text{train}}\Lambda + \epsilon^2\psi_{\text{train}}I_d) \cdot \Big(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}} - (\Gamma_{\text{train}}\Lambda + \epsilon^2\psi_{\text{train}}I_d)^{-1}\Lambda^{\frac{3}{2}}\Big) \\
&\qquad \cdot \Big(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}} - (\Gamma_{\text{train}}\Lambda + \epsilon^2\psi_{\text{train}}I_d)^{-1}\Lambda^{\frac{3}{2}}\Big)^\top\Big] \\
&\quad - \mathrm{Tr}\Big[\Lambda^3(\Gamma_{\text{train}}\Lambda + \epsilon^2\psi_{\text{train}}I_d)^{-1}\Big] + 2 \cdot \mathrm{Tr}(\Lambda),
\end{aligned} \tag{A.21}
$$

where $\Gamma_{\text{train}} := \Gamma(M_{\text{train}})$ and $\psi_{\text{train}} := \psi(M_{\text{train}})$.

Notice that the second and third terms in Eq. (A.21) are constants. Besides, the matrix $(\Gamma_{\text{train}}\Lambda + \epsilon^2\psi I_d)$ in the first term in Eq. (A.21) is positive definite, which means that this first term is non-negative. As a result, the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ will be minimized when the first term in Eq. (A.21) becomes zero. This can be achieved by setting

$$
w_{*,22}^V W_{*,11}^{KQ}\Lambda^{\frac{1}{2}} - (\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda^{\frac{3}{2}} = 0,
$$

which is

$$w^V_{*,22}W^{KQ}_{*,11} = (\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda.$$

The proof is completed. $\square$

We now turn to prove an PL-inequality for the surrogate AT problem. The proof idea follows that in [69]. Specifically, we will first prove several technical lemmas (*i.e.*, Lemma A.9, Lemma A.10, and Lemma A.11), and then present the PL-inequality in Lemma A.12, which can then enable the surrogate AT model in Eq. (9) approaches its global optimal solution.

Lemma A.9. *Suppose Assumption 1 holds and the model $f_{\text{LSA},\theta}$ is trained via minimizing the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (9) with continuous training flow. Then, for any continuous training time $t \geq 0$, we uniformly have that*

$$(w^V_{22}(t))^2 = \mathrm{Tr}[W^{KQ}_{11}(t)(W^{KQ}_{11}(t))^\top].$$

Proof. Since the model is trained via continuous gradient flow, thus $\partial_t W^{KQ}_{11}(t)$ can be calculated based on the simplified surrogate AT loss proved in Lemma A.7 as follows,

$$\begin{aligned}
&\partial_t W^{KQ}_{11}(t) := -\partial_{W^{KQ}_{11}}\tilde{\mathcal{L}}^{\text{adv}}(\theta)\\
&= -2\cdot\partial_{W^{KQ}_{11}}\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})^\top\Big]\\
&\quad + 4\cdot\partial_{W^{KQ}_{11}}\mathrm{Tr}\Big[(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big]\\
&= -2\cdot(w^V_{22})^2\cdot\partial_{W^{KQ}_{11}}\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot W^{KQ}_{11}\cdot\Lambda\cdot(W^{KQ}_{11})^\top\Big] + \underbrace{4w^V_{22}\Lambda^2}_{\text{by Lemma A.4}}\\
&= \underbrace{-4\cdot(w^V_{22})^2\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot W^{KQ}_{11}\cdot\Lambda}_{\text{by Lemma A.4}} + 4w^V_{22}\Lambda^2.
\end{aligned} \tag{A.22}$$

Similarly, for $\partial_t w^V_{22}(t)$, we have

$$\begin{aligned}
&\partial_t w^V_{22}(t) := -\partial_{w^V_{22}}\tilde{\mathcal{L}}^{\text{adv}}(\theta)\\
&= -2\cdot\partial_{w^V_{22}}\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})^\top\Big]\\
&\quad + 4\cdot\partial_{w^V_{22}}\mathrm{Tr}\Big[(w^V_{22}W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big]\\
&= -4w^V_{22}\cdot\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot(W^{KQ}_{11}\Lambda^{\frac{1}{2}})^\top\Big]\\
&\quad + 4\cdot\mathrm{Tr}\Big[(W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big].
\end{aligned} \tag{A.23}$$

Combining Eqs (A.22) and (A.23), we thus have

$$\begin{aligned}
&\mathrm{Tr}\Big[\partial_t W^{KQ}_{11}(t)(W^{KQ}_{11}(t))^\top\Big]\\
&= -4\cdot(w^V_{22})^2\cdot\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(W^{KQ}_{11}\Lambda^{\frac{1}{2}})\cdot(W^{KQ}_{11}\Lambda^{\frac{1}{2}})^\top\Big]\\
&\quad + 4w^V_{22}\cdot\mathrm{Tr}\Big[\Lambda^{\frac{3}{2}}\cdot(\Lambda^{\frac{1}{2}}W^{KQ}_{11})^\top\Big]\\
&= (\partial_t w^V_{22}(t))w^V_{22}(t),
\end{aligned}$$

which further indicates that

$$\begin{aligned}
&\partial_t\mathrm{Tr}\Big[W^{KQ}_{11}(t)(W^{KQ}_{11}(t))^\top\Big]\\
&= \mathrm{Tr}\Big[\partial_t W^{KQ}_{11}(t)\cdot(W^{KQ}_{11}(t))^\top\Big] + \mathrm{Tr}\Big[W^{KQ}_{11}(t)\cdot\partial_t(W^{KQ}_{11}(t))^\top\Big]\\
&= (\partial_t w^V_{22}(t))\cdot w^V_{22}(t) + W^V_{22}(t)\cdot(\partial_t w^V_{22}(t)) = \partial_t(w^V_{22}(t)^2).
\end{aligned} \tag{A.24}$$

Finally, according to Assumption 1, we have that when the continuous training time is $t = 0$,

$$\mathrm{Tr}\Big[W_{11}^{KQ}(0)(W_{11}^{KQ}(0))^\top\Big] = \|W_{11}^{KQ}(0)\|_F^2 = \sigma^2 = w_{22}^V(0)^2.$$

Combine with Eq. (A.24), we thus have that

$$\mathrm{Tr}\Big[W_{11}^{KQ}(t)(W_{11}^{KQ}(t))^\top\Big] = w_{22}^V(t)^2, \quad \forall t \geq 0.$$

The proof is completed. $\square$

Lemma A.10. *Suppose Assumption 1 holds and the model $f_{\mathrm{LSA},\theta}$ is trained via minimizing the surrogate AT loss $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta)$ in Eq. (9) with continuous training flow. Then, if the parameter σ in Assumption 1 satisfies*

$$\sigma < \sqrt{\frac{2}{d \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2}},$$

we have $w_{22}^V(t) > 0$ holds for any continuous training time $t \geq 0$.

Proof. According to the simplified AT loss calculated in Lemma A.7, we know that if $w_{22}^V(t) = 0$, then $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_t) = 2\mathrm{Tr}(\Lambda)$. Besides, under Assumption 1, we have $w_{22}^V(0) = \sigma > 0$. Therefore, if we can show that $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_t) \neq 2\mathrm{Tr}(\Lambda)$ for any $t \geq 0$, then it is proved that $w_{22}^V(t) > 0$ for any $t \geq 0$.

To this end, we first analyze the surrogate AT loss $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_t)$ at the initial training time $t = 0$. By applying Assumption 1, we have

$$\begin{aligned}
&\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_0) \\
&= 2\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d) \cdot (w_{22}^V(0)W_{11}^{KQ}(0)\Lambda^{\frac{1}{2}}) \cdot (w_{22}^V(0)W_{11}^{KQ}(0)\Lambda^{\frac{1}{2}})^\top\Big] \\
&\quad - 4\mathrm{Tr}\Big[(w_{22}^V(0)W_{11}^{KQ}(0)\Lambda^{\frac{1}{2}}) \cdot \Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda) \\
&= 2\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d) \cdot (\sigma^2\Theta\Theta^\top\Lambda^{\frac{1}{2}}) \cdot (\sigma^2\Theta\Theta^\top\Lambda^{\frac{1}{2}})^\top\Big] \\
&\quad - 4\mathrm{Tr}\Big[(\sigma^2\Theta\Theta^\top\Lambda^{\frac{1}{2}}) \cdot \Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda) \\
&= 2\sigma^4 \cdot \mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1} \cdot \Lambda\Theta\Theta^\top\Lambda\Theta\Theta^\top\Big] - 4\sigma^2\|\Lambda\Theta\|_F^2 + 2\mathrm{Tr}(\Lambda) \\
&\leq 2\sigma^4 \cdot \underbrace{d \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 \cdot \|\Lambda\Theta\Theta^\top\Lambda\Theta\Theta^\top\|_2}_{\text{by Lemma A.5}} - 4\sigma^2\|\Lambda\Theta\|_F^2 + 2\mathrm{Tr}(\Lambda) \\
&\leq 2\sigma^4 \cdot d \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 \cdot \|\Lambda\Theta\Theta^\top\Lambda\|_F \cdot \|\Theta\Theta^\top\|_F \\
&\quad - 4\sigma^2\|\Lambda\Theta\|_F^2 + 2\mathrm{Tr}(\Lambda) \\
&\leq 2\sigma^4 \cdot d \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 \cdot \|\Lambda\Theta\|_F^2 \cdot 1 - 4\sigma^2\|\Lambda\Theta\|_F^2 + 2\mathrm{Tr}(\Lambda) \\
&= 2\sigma^2 \cdot \|\Lambda\Theta\|_F^2 \cdot (d \cdot \sigma^2 \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 - 2) + 2\mathrm{Tr}(\Lambda).
\end{aligned} \tag{A.25}$$

By Assumption 1, we have $\|\Lambda\Theta\|_F^2 > 0$. Thus, when $(d \cdot \sigma^2 \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 - 2) < 0$, which is

$$\sigma < \sqrt{\frac{2}{d \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2}},$$

we will have $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_0) < \mathrm{Tr}(\Lambda)$.

Finally, since the surrogate AT loss $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_t)$ is minimized with continuous gradient, thus when the above condition holds, for any $t > 0$, we always have that $\tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_t) \leq \tilde{\mathcal{L}}^{\mathrm{adv}}(\theta_0) < \mathrm{Tr}(\Lambda)$.

The proof is completed. $\square$

Lemma A.11. *Suppose Assumption 1 holds and the σ in Assumption 1 satisfies $\sigma < \sqrt{\frac{2}{d\cdot\|(\Gamma(M_{\text{train}})\Lambda+\epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2}}$. Then, for any continuous training time $t \geq 0$, we have $(w_{22}^V(t))^2 \geq \nu > 0$, where*

$$\nu := \frac{\sigma^2 \cdot \|\Lambda\Theta\|_F^2 \cdot (2 - d\cdot\sigma^2 \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2)}{2d\|\Lambda^2\|_2} > 0.$$

Proof. By applying Eq. (A.25) in Lemma A.10, we have that for any $t \geq 0$,

$$\begin{aligned}
&2\sigma^2 \cdot \|\Lambda\Theta\|_F^2 \cdot (d\cdot\sigma^2\cdot\|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 - 2) + 2\mathrm{Tr}(\Lambda)\\
&\geq \tilde{\mathcal{L}}^{\text{adv}}(\theta_0) \geq \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\\
&= 2\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})^\top\Big]\\
&\quad - 4\mathrm{Tr}\Big[(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda)\\
&= 2\|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{\frac{1}{2}}\cdot(w_{22}^V W_{11}^{KQ}\Lambda^{\frac{1}{2}})\|_F^2 - 4\mathrm{Tr}\Big[w_{22}^V W_{11}^{KQ}\Lambda^2\Big] + 2\mathrm{Tr}(\Lambda)\\
&\geq 0 - \underbrace{4d\cdot|w_{22}^V|\cdot\|\Lambda^2\|_2\cdot\|W_{11}^{KQ}\|_2}_{\text{by Lemma A.5}} + 2\mathrm{Tr}(\Lambda),
\end{aligned}$$

which indicates

$$2\sigma^2\cdot\|\Lambda\Theta\|_F^2\cdot(d\cdot\sigma^2\cdot\|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2 - 2) \geq -4d\cdot|w_{22}^V|\cdot\|\Lambda^2\|_2\cdot\|W_{11}^{KQ}\|_F,$$

thus

$$|w_{22}^V|\cdot\|W_{11}^{KQ}\|_F \geq \frac{\sigma^2\cdot\|\Lambda\Theta\|_F^2\cdot(2 - d\cdot\sigma^2\cdot\|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2)}{2d\|\Lambda^2\|_2}. \quad \text{(A.26)}$$

Besides, by combining Lemma A.9 and Lemma A.10, we know that

$$w_{22}^V(t) = \sqrt{\mathrm{Tr}[W_{11}^{KQ}(t)(W_{11}^{KQ}(t))^\top]} = \sqrt{\|W_{11}^{KQ}(t)\|_F^2} = \|W_{11}^{KQ}(t)\|_F. \quad \text{(A.27)}$$

Finally, inserting Eq. (A.27) into Eq. (A.26), we thus have

$$(w_{22}^V(t))^2 \geq \frac{\sigma^2\cdot\|\Lambda\Theta\|_F^2\cdot(2 - d\cdot\sigma^2\cdot\|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2)}{2d\|\Lambda^2\|_2} > 0.$$

The proof is completed. ☐

Lemma A.12 (PL-inequality)**.** *Suppose Assumption 1 holds and the LSA model $f_{\text{LSA},\theta}$ is trained via minimizing the surrogate AT loss $\tilde{\mathcal{L}}^{\text{adv}}(\theta)$ in Eq. (9) with continuous training flow. Suppose the σ in Assumption 1 satisfies $\sigma < \sqrt{\frac{2}{d\cdot\|(\Gamma(M_{\text{train}})\Lambda+\epsilon^2\psi(M_{\text{train}})I_d)\Lambda^{-1}\|_2}}$. Then for any continuous training time $t > 0$, we uniformly have that*

$$\|\partial_\theta\tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_2^2 \geq \mu\cdot\Big(\tilde{\mathcal{L}}^{adv}(\theta_t) - \min_\theta\tilde{\mathcal{L}}^{adv}(\theta)\Big),$$

where

$$\mu := \frac{8\nu}{\|(\Gamma_{\text{train}}\Lambda + \epsilon^2\psi_{\text{train}}I_d)^{-\frac{1}{2}}\|_F^2\cdot\|\Lambda^{-\frac{1}{2}}\|_F^2},$$

ν is defined in Lemma A.11, and $\mathrm{Vec}(\cdot)$ denotes the vectorization function.

Proof. From Eq. (A.22) in Lemma A.9, we have that

$$\begin{aligned}
\partial_t W_{11}^{KQ}(t) &= -4\cdot(w_{22}^V)^2\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot W_{11}^{KQ}\cdot\Lambda + 4w_{22}^V\Lambda^2\\
&= -4w_{22}^V\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)\cdot D(\theta_t)\cdot\Lambda^{\frac{1}{2}},
\end{aligned}$$

where

$$D(\theta_t) := \Big(w_{22}^V W_{11}^{KQ} \Lambda^{\frac{1}{2}} - (\Gamma(M_{\text{train}})\Lambda + \epsilon^2 \psi(M_{\text{train}}) I_d)^{-1} \Lambda^{\frac{3}{2}}\Big) \in \mathbb{R}^{d\times d}. \tag{A.28}$$

As a result, the gradient norm square $\|\partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_2^2$ can be further lower-bounded as follows,

$$\begin{aligned}
&\|\partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_2^2 := (\partial_{w_{22}^V} \tilde{\mathcal{L}}^{\text{adv}}(\theta_t))^2 + \|\partial_{W_{11}^{KQ}} \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_F^2 \\
&\geq \|\partial_{W_{11}^{KQ}} \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_F^2 \\
&= \|4 \cdot w_{22}^V \cdot (\Gamma(M_{\text{train}})\Lambda + \epsilon^2 \psi(M_{\text{train}}) I_d) \cdot D(\theta_t) \cdot \Lambda^{\frac{1}{2}}\|_F^2 \\
&= 16 \cdot (w_{22}^V)^2 \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2 \psi(M_{\text{train}}) I_d) \cdot D(\theta_t) \cdot \Lambda^{\frac{1}{2}}\|_F^2 \\
&\geq \underbrace{16 \cdot \nu}_{\text{by Lemma A.11}} \cdot \|(\Gamma(M_{\text{train}})\Lambda + \epsilon^2 \psi(M_{\text{train}}) I_d) \cdot D(\theta_t) \cdot \Lambda^{\frac{1}{2}}\|_F^2,
\end{aligned} \tag{A.29}$$

where $\nu > 0$ is defined in Lemma A.11.

Meanwhile, according to the proof of Lemma A.8, we can rewrite and upper-bound $\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big)$ as follows,

$$\begin{aligned}
&\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) \\
&= 2 \cdot \mathrm{Tr}\Big[(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d) \cdot \Big(w_{22}^V W_{11}^{KQ} \Lambda^{\frac{1}{2}} - (\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{-1} \Lambda^{\frac{3}{2}}\Big) \\
&\qquad \cdot \Big(w_{22}^V W_{11}^{KQ} \Lambda^{\frac{1}{2}} - (\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{-1} \Lambda^{\frac{3}{2}}\Big)^\top\Big] \\
&= 2 \cdot \mathrm{Tr}[(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d) \cdot D(\theta_t) \cdot D(\theta_t)^\top] \\
&= 2 \cdot \underbrace{\mathrm{Tr}[(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{\frac{1}{2}} \cdot D(\theta_t) \cdot D(\theta_t)^\top \cdot (\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{\frac{1}{2}}]}_{\text{Lemma A.3}} \\
&= 2 \cdot \|(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{\frac{1}{2}} \cdot D(\theta_t)\|_F^2 \\
&\leq 2 \cdot \|(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{-\frac{1}{2}}\|_F^2 \cdot \|\Lambda^{-\frac{1}{2}}\|_F^2 \cdot \|(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d) \cdot D(\theta_t) \cdot \Lambda^{\frac{1}{2}}\|_F^2,
\end{aligned} \tag{A.30}$$

where $\Gamma_{\text{train}} := \Gamma(M_{\text{train}})$ and $\psi_{\text{train}} := \psi(M_{\text{train}})$.

Combining Eqs. (A.29) and (A.30), we thus know that

$$\|\partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_2^2 \geq \frac{8\nu}{\|(\Gamma_{\text{train}}\Lambda + \epsilon^2 \psi_{\text{train}} I_d)^{-\frac{1}{2}}\|_F^2 \cdot \|\Lambda^{-\frac{1}{2}}\|_F^2} \cdot \Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big).$$

The proof is completed. □

Finally, we prove Theorem 1 based on Lemma A.8 and Lemma A.12.

Proof of Theorem 1. When all the conditions hold, when the surrogate AT problem defined in Eq. (9) is solved via continuous gradient flow, by Lemma A.8 we have

$$\begin{aligned}
&\partial_t\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) = \partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t) \cdot \partial_t \theta_t = \partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t) \cdot (-\partial_\theta^\top \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)) = -\|\partial_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta_t)\|_2^2 \\
&\leq -\mu \cdot \Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big),
\end{aligned}$$

which means

$$\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) \leq \Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_0) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) \cdot e^{-\mu t}.$$

As a result, when performing continuous gradient flow optimization for an infinitely long time, since $\mu > 0$, the surrogate AT loss will eventually converge to the global minima, *i.e.*,

$$\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_*) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) = \lim_{t\to\infty}\Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) \leq \Big(\tilde{\mathcal{L}}^{\text{adv}}(\theta_0) - \min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta)\Big) \cdot \lim_{t\to\infty} e^{-\mu t} = 0,$$

where $\theta_* := \lim_{t\to\infty} \theta_t$ is the converged model parameter. Meanwhile, from Lemma A.8, we know that θ_* is a global minimizer if and only if $w_{*,22}^V W_{*,11}^{KQ} = (\Gamma(M_{\text{train}})\Lambda + \epsilon^2 \psi(M_{\text{train}}) I_d)^{-1}\Lambda$, which completes the proof. □

A.4 Proofs in Section 4.4

This section collects all proofs that omitted from Section 4.4.

Proof of Theorem 2. By substituting all M_{train} with M_{test} in proofs of Proposition 1 and Lemma A.7, we immediately have that for any model parameter θ of the LSA model $f_{\text{LSA},\theta}$,

$$\begin{aligned}\mathcal{R}(\theta, M_{\text{test}}) \le\ & 2\mathrm{Tr}\Big[(\Gamma(M_{\text{test}})\Lambda + \epsilon^2\psi(M_{\text{test}})I_d)\cdot(w_{22}^{V}W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot(w_{22}^{V}W_{11}^{KQ}\Lambda^{\frac{1}{2}})^{\top}\Big]\\ & - 4\mathrm{Tr}\Big[(w_{22}^{V}W_{11}^{KQ}\Lambda^{\frac{1}{2}})\cdot\Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda).\end{aligned}$$

By inserting the converged model parameter $\theta_*(M_{\text{train}})$, which satisfies $(w_{*,22}^{V}W_{*,11}^{KQ}) = (\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda$, into the above robust generalization bound, we thus have that

$$\begin{aligned}&\mathcal{R}(\theta_*(M_{\text{train}}), M_{\text{test}})\\ &\le 2\mathrm{Tr}\Big[(\Gamma(M_{\text{test}})\Lambda + \epsilon^2\psi(M_{\text{test}})I_d)\cdot((\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda\cdot\Lambda^{\frac{1}{2}})\\ &\qquad\qquad\cdot((\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda\cdot\Lambda^{\frac{1}{2}})^{\top}\Big]\\ &\quad - 4\mathrm{Tr}\Big[(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda\cdot\Lambda^{\frac{1}{2}}\cdot\Lambda^{\frac{3}{2}}\Big] + 2\mathrm{Tr}(\Lambda)\\ &\overset{(*)}{\le} 2\mathrm{Tr}\Big[(\Gamma(M_{\text{test}})\Lambda + \epsilon^2\psi(M_{\text{test}})I_d)\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\\ &\qquad\qquad\cdot\Lambda^3\cdot((\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1})^{\top}\Big] + 0 + 2\mathrm{Tr}(\Lambda)\\ &\overset{(**)}{\le} 2\mathrm{Tr}\Big[\Lambda^3\cdot(\Gamma(M_{\text{test}})\Lambda + \epsilon^2\psi(M_{\text{test}})I_d)\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-2}\Big] + 2\mathrm{Tr}(\Lambda),\end{aligned}$$

where $(*)$ is due to that the matrix $((\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}\Lambda^3)$ is positive definite, and $(**)$ is due to that: (1) $(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-1}$ is symmetric and is commutative with Λ^3, and (2) Lemma A.3.

The proof is completed. □

Proof of Corollary 1. Let $\lambda_1, \cdots, \lambda_d$ be the d singular values of the matrix Λ. Then, the robust generalization bound in Theorem 2 can be rewritten as follows,

$$\begin{aligned}&\mathcal{R}(\theta_*(M_{\text{train}}), M_{\text{test}})\\ &\le 2\mathrm{Tr}\Big[\Lambda^3\cdot(\Gamma(M_{\text{test}})\Lambda + \epsilon^2\psi(M_{\text{test}})I_d)\cdot(\Gamma(M_{\text{train}})\Lambda + \epsilon^2\psi(M_{\text{train}})I_d)^{-2}\Big] + 2\mathrm{Tr}(\Lambda),\\ &\le \sum_{i=1}^{d}\lambda_i^3\cdot\frac{\frac{N+M_{\text{test}}+1}{N+M_{\text{test}}}\lambda_i + \frac{\mathrm{Tr}(\Lambda)}{N+M_{\text{test}}} + \epsilon^2\cdot\frac{M_{\text{test}}^2\mathrm{Tr}(\Lambda)}{(N+M_{\text{test}})^2}}{\left(\frac{N+M_{\text{train}}+1}{N+M_{\text{train}}}\lambda_i + \frac{\mathrm{Tr}(\Lambda)}{N+M_{\text{train}}} + \epsilon^2\cdot\frac{M_{\text{train}}^2\mathrm{Tr}(\Lambda)}{(N+M_{\text{train}})^2}\right)^2} + 2\mathrm{Tr}(\Lambda)\\ &\le \sum_{i=1}^{d}\lambda_i^3\cdot\frac{\frac{N+M_{\text{test}}+1}{N+M_{\text{test}}}\lambda_i + \frac{\mathrm{Tr}(\Lambda)}{N+M_{\text{test}}}}{\left(\frac{N+M_{\text{train}}+1}{N+M_{\text{train}}}\lambda_i\right)^2} + \sum_{i=1}^{d}\lambda_i^3\cdot\frac{\epsilon^2\cdot\frac{M_{\text{test}}^2\mathrm{Tr}(\Lambda)}{(N+M_{\text{test}})^2}}{\left(\epsilon^2\cdot\frac{M_{\text{train}}^2\mathrm{Tr}(\Lambda)}{(N+M_{\text{train}})^2}\right)^2} + 2\mathrm{Tr}(\Lambda)\\ &\le \sum_{i=1}^{d}\lambda_i\cdot\left(\frac{N+M_{\text{train}}}{N+M_{\text{train}}+1}\right)^2\cdot\left(\frac{N+M_{\text{test}}+1}{N+M_{\text{test}}}\lambda_i + \frac{\sum_{k=1}^{d}\lambda_k}{N}\right)\\ &\quad + \sum_{i=1}^{d}\frac{\lambda_i^3}{\epsilon^2\cdot\max_{k=1}^{d}\{\lambda_k\}}\cdot\frac{(N+M_{\text{train}})^4}{N^2}\cdot\frac{M_{\text{test}}^2}{M_{\text{train}}^4} + 2\sum_{i=1}^{d}\lambda_i\\ &\le \mathcal{O}(d)\cdot\mathcal{O}(1)\cdot\left(\mathcal{O}(1) + \frac{\mathcal{O}(d)}{N}\right) + \mathcal{O}(d)\cdot\mathcal{O}\left(\frac{1}{\epsilon^2}\right)\cdot\frac{(N+M_{\text{train}})^4}{N^2}\cdot\frac{M_{\text{test}}^2}{M_{\text{train}}^4} + \mathcal{O}(d)\\ &\le \mathcal{O}(d) + \mathcal{O}\left(\frac{d^2}{N}\right) + \mathcal{O}\left(\frac{d}{\epsilon^2}\right)\cdot\frac{(N+M_{\text{train}})^4}{N^2}\cdot\frac{M_{\text{test}}^2}{M_{\text{train}}^4}.\end{aligned}$$

Then, by applying Assumption 2, we further have that

$$\mathcal{R}(\theta_*(M_{\text{train}}), M_{\text{test}}) \leq \mathcal{O}(d) + \mathcal{O}\left(\frac{d^2}{N}\right) + \mathcal{O}\left(\frac{d}{\epsilon^2}\right) \cdot \frac{(N + M_{\text{train}})^4}{N^2} \cdot \frac{M_{\text{test}}^2}{M_{\text{train}}^4}$$

$$\leq \mathcal{O}(d) + \mathcal{O}\left(\frac{d^2}{N}\right) + \mathcal{O}\left(\frac{d}{(\sqrt{d})^2}\right) \cdot \frac{(N + O(N))^4}{N^2} \cdot \frac{M_{\text{test}}^2}{M_{\text{train}}^4}$$

$$= \mathcal{O}(d) + \mathcal{O}\left(\frac{d^2}{N}\right) + \mathcal{O}\left(N^2 \cdot \frac{M_{\text{test}}^2}{M_{\text{train}}^4}\right),$$

which completes the proof. □

B Additional experimental details

This section collects experimental details omitted from Section 5.

B.1 Jailbreak attacks

Our experiments leverage both suffix and non-suffix jailbreak attacks. Specifically, four suffix jailbreak attacks are adopted, which are GCG [73], BEAST [44], AmpleGCG [26], and Zhu's AutoDAN [71]. Meanwhile, two non-suffix jailbreak attacks are adopted, which are PAIR [7] and DeepInception [25]. We re-implemented all attacks except AmpleGCG by ourselves to enable fast batching operations during jailbreak, which can thus improve the efficiency of AT. Besides, other than the adversarial suffix length, we will also tune the following hyperparameters of jailbreak attacks:

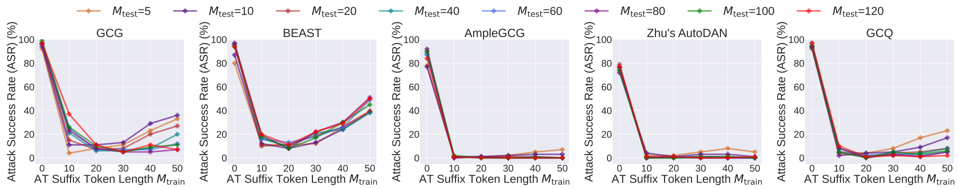
- **GCG:** According to Algorithm 1 in [73], hyperparameters that we need to tune for GCG include the iteration number T, the top-k parameter k, and the "batch-size" B.
- **BEAST:** According to Algorithm 1 in [44], hyperparameters that we need to tune for BEAST are two beam-search parameters k_1 and k_2.
- **AmpleGCG:** According to [26], AmpleGCG is an algorithm for training adversarial suffix generators. Our experiments adopt the adversarial suffix generator `AmpleGCG-plus-llama2-sourced-vicuna-7b13b-guanaco-7b13b` [1], which is officially released by [26].
- **Zhu's AutoDAN:** According to Algorithm 1 and Algorithm 2 in [71], hyperparameters that we need to tune for Zhu's AutoDAN are the iteration number T in each step, objective weights w_1 and w_2, the top-B parameter B, and the temperature τ.
- **GCQ:** According to Algorithm 1 in [17], hyperparameters that we need to tune for GCQ include the iteration number T, the proxy batch size b_p , the query batch size b_q, and the buffer size B.
- **PAIR:** According to [7], PAIR adopts LLM-based attacker and judger to iteratively synthesize and refine jailbreak prompts. As a result, one needs to set the base models for the attacker and judger and the number of teratively refining for the PAIR attack.
- **DeepInception:** According to [25], DeepInception attack uses manually crafted jailbreak prompts to attack targeted LLMs. We adopt the role play-based prompt from [25] to perform the attack. No other hyperparameter need to be tuned for the DeepInception attack.

B.2 Model training

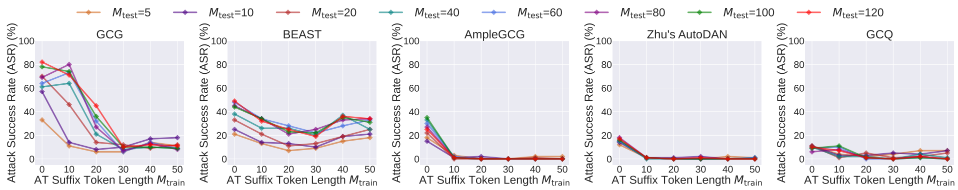
Jailbreak attacks during AT. We use GCG to search adversarial prompts during AT. The adversarial suffix token length is fixed to one of $\{5, 10, 20, 30, 40, 50\}$ during the overall AT. For other hyparameters described in Appendix B.1, we set T as 150, k as 256, and B as 64.

Benign answer $y^{(b)}$ for the safety dataset $D^{(h)}$. We adopt four benign answers for the safety data during AT, which are:

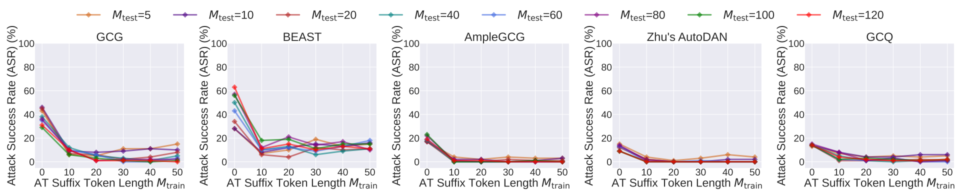
[1] `https://huggingface.co/osunlp/AmpleGCG-plus-llama2-sourced-vicuna-7b13b-guanaco-7b13b`



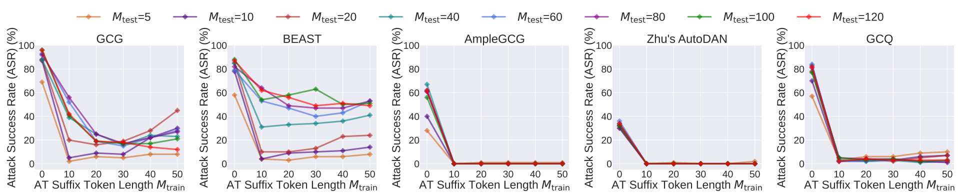
(a) Mistral-7B-v0.3.



(b) Llama-2-7B-Chat.



(c) Llama-3-8B-Instruct.



(d) Qwen2.5-7B-Instruct.

Figure 4: Curves of the ASR versus the adversarial suffix token length during AT (*i.e.*, M_{train}) under jailbreak attacks with different adversarial suffix token lengths (*i.e.*, M_{test}). $M_{\text{train}} = 0$ means that AT is not performed on the evaluated model. A low ASR indicates a strong jailbreak robustness.

- "As a responsible AI assistant, I cannot generate content that promotes illegal activities or violence."
- "Sorry, but I can't do that. Please try other instructions that are safe and responsible."
- "I cannot fulfill your request. I am designed to provide safe and helpful information."
- "This is a topic I am not permitted to discuss since it contain illegal or violent information."

LoRA setting. We use the PEFT library [35] to apply LoRA [18] to all query and key projection matrices in attentions. The PEFT hyperparameters are `r=64`, `lora_alpha=32`, and `lora_dropout=0.1`.

Other hyperparameters for AT. In every AT experiment, we follow Eq. (3) to perform AT with AdamW for 125 iterations, where the learning rate is set as 5×10^{-5} and the factor α is set as 0.2. Besides, the batch size is set as 64, in which 8 samples are jailbreak prompts crafted from data from the safety training set, and the remaining 56 samples are from the utility training set.

B.3 Model evaluations

Robustness evaluation. We report the Attack Success Rate (ASR) of jailbreak attacks to assess the robustness of models. Specifically, for each instruction from the safety test set, we synthesize the corresponding jailbreak prompt and use it to induce the targeted LLM to generate 10 responses. Then, we use an LLM-based judge from [36], which was fine-tuned from the Llama-2-13B model [1], to

[1] https://huggingface.co/cais/HarmBench-Llama-2-13b-cls

determine whether the 10 generated LLM responses are harmful or not. If any of them is determined to be harmful, the jailbreak attack is considered successful.

Jailbreak attacks for robustness evaluation. For every suffix attack, the adversarial suffix length is varied within $\{5, 10, 20, 40, 60, 80, 100, 120\}$. Besides, for jailbreak hyperparameters described in Appendix B.1:

- For the GCG attack, we set T as 500, k as 256, and T as 64.
- For the BEAST attack, we set k_1 as 64 and k_2 as 16.
- For the AmpleGCG attack, we use an official adversarial suffix generator as described in Appendix B.1.
- For the Zhu's AutoDAN attack, we set T as 3, w_1 as 10, w_2 as 100, B as 256, and τ as 2.
- For the GCQ attack, we set T as 200 and b_p, b_q, and B all as 128.
- For the PAIR attack, we set the base model for the attacker as Mistral-8x7B-Instruct-v0.1, the base model for the judger as Llama-3-70B-Instruct, and the number of iteratively refining is fixed to 10.
- For the DeepInception, as explained in Appendix B.1, we use a role-play-based prompt to perform the attack, and there are no other hyperparameters that need to be tuned for this attack.

Utility evaluation. We use the AlpacaEval2 framework [10] to report the Length-controlled WinRate (LC-WinRate) of targeted models against a reference model based on their output qualities on the utility test set. An LC-WinRate of 50% means that the output qualities of the two models are equal, while an LC-WinRate of 100% means that the targeted model is consistently better than the reference model. We use Davinci003 as the reference model and use the Llama-3-70B model to judge output quality. The official code of the AlpacaEval2 framework is used to conduct the evaluation. Additionally, the Llama-3-70B judger is run locally via the vLLM model serving framework [23].

B.4 Additional experimental results

This section collects additional experimental results (*i.e.*, Figure 4) omitted from Section 5.2.

From Figure 4, we find that GCG-based AT is extremely effective in improving model robustness against GCG, AmpleGCG, and Zhu's AutoDAN. For the BEAST attack, GCG-based AT can also suppress the ASR to no more than 50%. Further, when the AT adversarial suffix token length is set to 20, AT is already able to reduce the ASR by at least 30% under all settings. It is worth noting that the adversarial suffix length during AT is only up to 50, while that during jailbreaking can vary from 5 to 120. All these results indicate the effectiveness of defending against long-length jailbreaking with short-length AT.

C More experiments

This section presents experiments beyond those in Section 5.

C.1 Comparison with other jailbreak defense baselines

Here, we compare the jailbreak defense performance of short-length LLM AT with that of another jailbreak defense baseline, the Circuit Breakers method [72]. Specifically, we adopt GCG and BEAST attacks to assess the jailbreak robustness of Mistral-7B and Llama-3-8B LLMs protected by short-length LLM AT or the Circuit Breakers defense. For short-length LLM AT, we set the adversarial suffix length M_{train} during AT to a small value of 20 or 30. For the Circuit Breakers defense, we directly use the trained Mistral-7B [1] and Llama-3-8B [2] models officially released by [72].

The resulting jailbreak ASRs are collected and presented in Table 4, from which we observe that: (1) When the base model is Mistral-7B, short-length LLM AT consistently achieves better jailbreak

[1] `https://huggingface.co/GraySwanAI/Mistral-7B-Instruct-RR`
[2] `https://huggingface.co/GraySwanAI/Llama-3-8B-Instruct-RR`

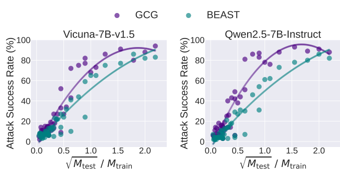
Table 4: ASR (%) of suffix jailbreaking against LLMs trained with Circuit Breakers [73] or LLM AT. A low ASR suggests a strong jailbreak robustness of the targeted model.

(a) ASRs of different jailbreak attacks against Mistral-7B.

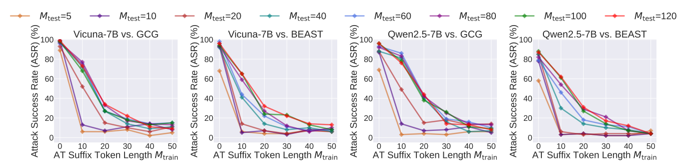
Attack	Defense	Adversarial Suffix Token Length M_{test} in Jailbreaking							
		5	10	20	40	60	80	100	120
GCG	Circuit Breakers [72]	21.0	20.0	21.0	23.0	23.0	28.0	28.0	23.0
	LLM AT ($M_{\text{train}} = 20$)	**8.0**	**11.0**	**7.0**	**6.0**	**7.0**	8.0	10.0	11.0
	LLM AT ($M_{\text{train}} = 30$)	11.0	13.0	8.0	**6.0**	**7.0**	**5.0**	**5.0**	**5.0**
BEAST	Circuit Breakers [72]	19.0	21.0	20.0	24.0	25.0	25.0	25.0	27.0
	LLM AT ($M_{\text{train}} = 20$)	**11.0**	**8.0**	**11.0**	**10.0**	**13.0**	**8.0**	**8.0**	**11.0**
	LLM AT ($M_{\text{train}} = 30$)	12.0	13.0	19.0	21.0	18.0	22.0	17.0	22.0

(b) ASRs of different jailbreak attacks against Llama-3-8B.

Model	Defense	Adversarial Suffix Token Length M_{test} in Jailbreaking							
		5	10	20	40	60	80	100	120
GCG	Circuit Breakers [72]	**3.0**	**5.0**	3.0	4.0	3.0	5.0	5.0	7.0
	LLM AT ($M_{\text{train}} = 20$)	5.0	8.0	6.0	5.0	6.0	**1.0**	3.0	1.0
	LLM AT ($M_{\text{train}} = 30$)	11.0	9.0	**2.0**	**3.0**	**0.0**	2.0	**1.0**	**1.0**
BEAST	Circuit Breakers [72]	12.0	**9.0**	11.0	12.0	16.0	15.0	17.0	15.0
	LLM AT ($M_{\text{train}} = 20$)	**10.0**	12.0	**4.0**	13.0	12.0	21.0	19.0	15.0
	LLM AT ($M_{\text{train}} = 30$)	19.0	15.0	12.0	**6.0**	**9.0**	**14.0**	**11.0**	**10.0**



(a) Scatter plots of ASR to the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$.



(b) ASR versus M_{train} under different M_{test}. $M_{\text{train}} = 0$ means that AT is not performed on the evaluated model.

Figure 5: ASR of models trained from the BEAST-based LLM AT. A low ASR indicates a strong jailbreak robustness of the model.

robustness than Circuit Breakers under different jailbreak attack adversarial suffix lengths. (2) When the base model is Llama-3-8B, the two defense methods achieve similar performance.

C.2 LLM AT with the BEAST attack

In our main experiments in Section 5, we solely use the GCG attack to synthesize jailbreak prompts for LLM AT. In this section, we investigate whether our theoretical results still empirically hold for AT with jailbreak attacks other than GCG. Specifically, we now perform LLM AT with the BEAST attack on Vicuna-7B-v1.5 and Qwen2.5-7B-Instruct models. For the hyperparameters of BEAST described in Appendix B.1, we vary the adversarial suffix token length within $\{5, 10, 20, 30, 40, 50\}$, and set k_1 to 64 and k_2 to 16. All other settings of LLM AT follow those described in Section B.2.

Experimental results are presented in Figure 5 and Table 5. From Figure 5a and Table 5, we observe a statistically significant positive correlation between the suffix jailbreak robustness and the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ in every experiment, which indicates that our ICL-AT theory still holds for BEAST-based LLM AT. Besides, from Figure 5b, one can find that AT with a short adversarial suffix length M_{train} of 30 can already reduce the ASR from nearly 100% to around 20% in every evaluation case, which demonstrates the effectiveness of short-length BEAST-based LLM AT in defending against jailbreak attacks.

C.3 LLM AT on larger models

We also perform short-length LLM AT on Vicuna-13B-v1.5, which is a model larger than those 7B/8B LLMs adopted in our main experiments in Section 5. All hyperparameters for LLM AT follow those described in Appendix B.2. Results are presented in Table 6, which shows that AT with an

Table 5: PCCs and p-values calculated between ASR and ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ on LLMs adversarially trained with the BEAST attack. $p < 5.00 \times 10^{-2}$ means that the correlation between ASR and the ratio is considered statistically significant.

Model	GCG Attack		BEAST Attack	
	PCC(↑)	p-value(↓)	PCC(↑)	p-value(↓)
Vicuna-7B	0.91	$\mathbf{5.3 \times 10^{-19}}$	0.94	$\mathbf{6.7 \times 10^{-24}}$
Qwen2.5-7B	0.88	$\mathbf{2.2 \times 10^{-16}}$	0.95	$\mathbf{5.0 \times 10^{-25}}$

Table 6: ASR (%) of the GCG attack against Vicuna-13B-v1.5 trained with LLM AT. A low ASR suggests a strong jailbreak robustness of the targeted model.

(a) ASRs of different jailbreak attacks against Mistral-7B.

Attack	Defense	Adversarial Suffix Token Length M_{test} in Jailbreaking							
		5	10	20	40	60	80	100	120
GCG	None	92.0	94.0	99.0	96.0	98.0	96.0	99.0	98.0
	LLM AT ($M_{\text{train}} = 5$)	**11.0**	19.0	30.0	53.0	55.0	67.0	70.0	68.0
	LLM AT ($M_{\text{train}} = 20$)	12.0	**9.0**	11.0	**6.0**	**6.0**	**6.0**	**8.0**	**7.0**

adversarial suffix token length as short as 20 can already reduce the ASR of the GCG attack from nearly 99% to around 10% in the worst case. This suggests the generalization of our theoretical findings beyond 7B/8B models.