"SHORT-LENGTH" ADVERSARIAL TRAINING HELPS LLMS DEFEND "LONG-LENGTH" JAILBREAK ATTACKS: THEORETICAL AND EMPIRICAL EVIDENCE

Shaopeng Fu[1], **Liang Ding**[2] **& Di Wang**[1]
[1]King Abdullah University of Science and Technology, Saudi Arabia
[2]The University of Sydney, Australia
{shaopeng.fu, di.wang}@kaust.edu.sa
liangding.liam@gmail.com

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. 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 number of adversarially perturbed in-context samples during training and testing. Empirically, we conduct AT on popular 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 during jailbreaking to the length during AT. Our findings show that it is practical to defend "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) (Brown et al., 2020; Touvron et al., 2023a; Liu et al., 2024a; Yang et al., 2024a) have been widely integrated into various real-world applications, but their safety is found to be vulnerable toward jailbreak attacks (Wei et al., 2023). With carefully crafted adversarial prompts, one can "jailbreak" the safety mechanism of LLMs and induce arbitrary harmful behaviors (Zou et al., 2023; Chao et al., 2023; Liu et al., 2024c). Recent studies (Xhonneux et al., 2024; Mazeika et al., 2024; Yu et al., 2024; Casper et al., 2024) have proposed performing safety alignment through adversarial training (AT) (Madry et al., 2018) to enhance LLMs' robustness against jailbreaking. A standard AT for LLMs would train them on harmful adversarial prompts synthesized by strong jailbreak attacks to learn to refuse these harmful instructions (Mazeika et al., 2024).

In such AT, the length of synthesized adversarial prompts used for model training is critical to the final jailbreak robustness of LLMs. Anil et al. (2024) and Xu et al. (2024) have shown that longer adversarial prompts enjoy stronger 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 usually time-consuming since it requires solving discrete optimization problems in high-dimensional spaces. 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 the raised question by analyzing *suffix jailbreak attacks*, where each jailbreak prompt is formed 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 $\Theta(\sqrt{M})$.** In other words, we show that it is possible to defend long-length jailbreaking via efficient short-length AT.

Our finding is supported by *theoretical* and *empirical* evidence. Theoretically, we leverage the *in-context learning theory* (Von Oswald et al., 2023; Zhang et al., 2024) 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 GCG (Zou et al., 2023), 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 jailbreaking with an adversarial suffix (token) length of up to 120 by at least 30% in all experiments.

2 Related Works

Jailbreak attacks. Jailbreaking (Wei et al., 2023) can be seen as adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015) 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 (Shin et al., 2020; Guo et al., 2021; Zou et al., 2023; Liao & Sun, 2024; Schwinn et al., 2024), heuristic greedy search (Sadasivan et al., 2024; Hayase et al., 2024; Jin et al., 2024), or fine-tuning prompt generators from pre-trained LLMs (Paulus et al., 2024). Other attempts include word-level adversarial prompt searching (Liu et al., 2024c) or directly prompting LLMs to generate adversarial prompts (Chao et al., 2023; Liu et al., 2024b). 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 (Anil et al., 2024) or synthesizing longer adversarial suffixes (Xu et al., 2024) 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 (Ouyang et al., 2022; Rafailov et al., 2023; Qi et al., 2024a;b; Chen et al., 2024a). More recent works have started to adopt adversarial training (AT) (Madry et al., 2018) to align LLMs. Mazeika et al. (2024) trained LLMs on (discrete) adversarial prompts synthesized by GCG attack (Zou et al., 2023), in which they cached the intermediate synthesized results to reduce the heavy cost of searching adversarial prompts from scratch. Meanwhile, various studies (Xhonneux et al., 2024; Casper et al., 2024; Sheshadri et al., 2024; Yu et al., 2024) 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 divided into two categories. The first aims to understand ICL via constructing explicit multi-layer transformers to simulate the optimization process of learning function classes (Garg et al., 2022; Von Oswald et al., 2023; Ahn et al., 2023; Chen et al., 2024b; Mahankali et al., 2024; Wang et al., 2024b). The second focuses on directly analyzing the training (Zhang et al., 2024; Yang et al., 2024b; Huang et al., 2023; Wu et al., 2024; Lin et al., 2024) and generalization (Lu et al., 2024; Magen et al., 2024; Frei & Vardi, 2024; Shi et al., 2024) of simple self-attention models (*i.e.*, one-layer transformer). Anwar et al. (2024) 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 Anwar et al. (2024), 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.

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 ("$\oplus$" denotes concatenation):

$$p_\theta(y_{1:m}|x_{1:n}) = \prod_{i=1}^{m} p_\theta(y_i|x_{1:n} \oplus y_{1:(i-1)}),$$

Jailbreak attacks. This paper will focus 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"). 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 (Zou et al., 2023), 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 $\mathcal{L}$ Mazeika et al. (2024); Qi et al. (2024a) 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,

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

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

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

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 (Zhang et al., 2024; Shi et al., 2024; Anwar et al., 2024), and we will bridge the ICL theory and the LLM AT problem defined in Eq. (3) later. 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 (demonstration), $y_{\tau,i} \in \mathbb{R}$ is the label for the i-th training sample, and $x_{\tau,q} \in \mathbb{R}^d$ is the in-context query sample. The embedding matrix E_τ for this task-related prompt 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 a prompt embedding matrix 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. Ahn et al. (2024) empirically show that LSA models share similar properties with non-linear ones and thus are useful for understanding transformers. We follow Zhang et al. (2024) 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 prompt embedding matrix, and N is the prompt length. The model prediction $\hat{y}_{q,\theta}$ for the query $x_{\tau,q}$ is given by the right-bottom entry of the LSA model output, *i.e.*, $\hat{y}_{q,\theta}(E_\tau) := f_{\text{LSA},\theta}(E_\tau)_{(d+1),(N+1)}$. We further follow Zhang et al. (2024) to denote that

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

where $W_{11}^V, W_{11}^{KQ} \in \mathbb{R}^{d\times d}$, $w_{12}^V, w_{21}^V, w_{12}^{KQ}, w_{21}^{KQ} \in \mathbb{R}^{d\times 1}$ and $w_{22}^V, W_{22}^{KQ} \in \mathbb{R}$. 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. We denote $[n] := \{1, \cdots, n\}$ for any $n \in \mathbb{N}^+$. $\|A\|_{2,\infty} := \max_{1\leq i\leq m} \|A_{i,:}\|_2$ for any $A \in \mathbb{R}^{n\times m}$. $\text{Tr}(A) := \sum_{i=1}^n A_{i,i}$ is the trace function for any matrix $A \in \mathbb{R}^{n\times n}$. Finally, we use standard big O notations $\mathcal{O}(\cdot)$ and $\Theta(\cdot)$.

4.1 Problem Definition for Adversarial ICL

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

Data distribution and statistical model. For any task indexed by τ, we assume 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 a prompt embedding matrix E_τ of length N (see Eq. (4)), we will form its corresponding adversarial prompt embedding matrix $E^{\text{adv}}_{\tau,M} \in \mathbb{R}^{(d+1)\times(N+M+1)}$ by concatenating E_τ with an adversarial suffix of length M as follows,

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

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, $X^{\text{sfx}}_\tau := (x^{\text{sfx}}_{\tau,1} \ \cdots \ x^{\text{sfx}}_{\tau,M}) \in \mathbb{R}^{d\times M}$ and $Y^{\text{sfx}}_\tau := (y^{\text{sfx}}_{\tau,1} \ \cdots \ y^{\text{sfx}}_{\tau,M}) \in \mathbb{R}^{d\times M}$ denotes the new M clean suffix samples and labels, and $\Delta^{\text{sfx}}_\tau := (\delta_{\tau,1} \ \cdots \ \delta_{\tau,M}) \in \mathbb{R}^{d\times M}$ denotes the M adversarial perturbations for the suffix.

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 prompt embedding matrix 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 f^{LSA}_θ as follows,

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

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 $x_{\tau,q}$. Since this paper aims to understand how the adversarial prompt length in AT would affect the robustness of LLM, Eq. (7) will 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 (Madry et al., 2018; Javanmard et al., 2020; Ribeiro et al., 2023; Fu & Wang, 2024; Wang et al., 2024a), here we also adopt a minimax AT loss to train the LSA model. We first use the introduced 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 adversarial prompts. We further assume that the adversarial suffix length is fixed during AT, which thus leads to the following AT problem formalization,

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

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 Zhang et al. (2024) to make the following assumption on the LSA model parameter initialization.

Assumption 1 (c.f. Assumption 3.3 in Zhang et al. (2024))**.** *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 that*

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

Recall in Eq. (5), w^V_{12}, w^{KQ}_{12}, and w^{KQ}_{22} do not contribute to the prediction function $\hat{y}_{q,\theta}(\cdot)$. Thus, Assumption 1 sets them to be zero at initialization. To ensure symmetric initialization, Assumption 1 further sets $w^V_{21}(0)$ and $w^{KQ}_{21}(0)$ to zero. These settings will simplify our AT analysis in next section.

Bridging ICL AT and LLM AT. Finally, we explain the similarities between AT on ICL models and LLMs to motivate why ICL AT (*i.e.*, Eq. (8)) can be a good artifact to theoretically understand LLM AT (*i.e.*, Eq. (3)). We first compare the ICL suffix adversarial attack in Eq. (7) with the LLM suffix jailbreak attack in Eq. (1). We find that their attack goals are similar since both attacks aim to make targeted models behave wrongly via manipulating suffixes of input prompts. The only difference is that jailbreak attacks aim to induce LLMs to generate specified harmful content while our ICL attack aims to maximize linear regression prediction errors made by ICL models. Besides, unlike Anwar et al. (2024), which performs ICL attacks by perturbing a single in-context sample in the entire real space, our attack allows perturbing multiple in-context samples but only within restricted spaces, thus better simulating how LLM jailbreak attacks allow adversarial token suffixes to be searched only in the limited token vocabulary set.

We then compare the ICL AT problem in Eq. (8) with the LLM AT problem in Eq. (3). One can find that the motivations behind the two AT problems are the same, which is to enhance models' robustness by training them on adversarial prompts. However, we notice that the LLM AT problem introduces an additional utility loss to maintain the performance of LLMs on benign data. This is because in LLM jailbreak attacks, adversarial prompts would be crafted only from harmful prompts but not benign ones. We argue that this discrepancy has little impact on our theoretical analysis, as both our theory and experiments focus on studying how adversarially trained models can defend against adversarial prompts rather than their performance on benign data.

4.2 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 $(X_\tau^{\text{sfx}}, Y_\tau^{\text{sfx}})$ are entangled together within the adversarial embedding matrix $E^{\text{adv}}_{\tau, M_{\text{train}}}$, which makes it very difficult to solve the inner maximization problem and further analyze the ICL AT dynamics.

To tackle the 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:

$$\min_\theta \tilde{\mathcal{L}}^{\text{adv}}(\theta) := \min_\theta \{\ell_1(\theta) + \ell_2(\theta) + \ell_3(\theta) + \ell_4(\theta)\}, \tag{9}$$

where $\tilde{\mathcal{L}}^{\text{adv}}(\theta) := (\ell_1(\theta) + \ell_2(\theta) + \ell_3(\theta) + \ell_4(\theta))$ is the surrogate AT loss, and

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

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

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

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

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

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

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,

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

where N is the prompt length of the original embedding matrix 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 Zhang et al. (2024)). 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.*

4.3 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 $\theta_*(M_{\mathrm{train}})$. We give 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*

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

where M_{train} is the adversarial suffix length in the ICL adversarial attack and functions $\Gamma(\cdot)$ and $\psi(\cdot)$ are defined 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,*

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

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

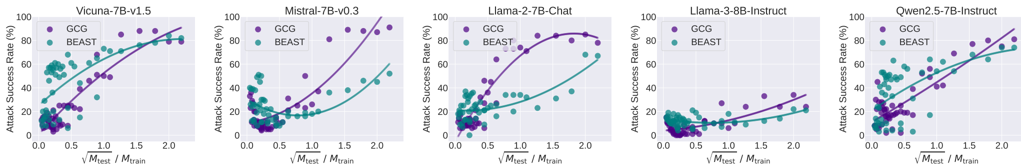

Figure 1: Scatter plots of ASR to the ratio of the square root of the adversarial token suffix length in jailbreak attacks to the adversarial token suffix length during AT (*i.e.*, $\sqrt{M_{\text{test}}}/M_{\text{train}}$). For each pair of base model and attack, 48 points are plotted. A high ASR indicates a weak jailbreak robustness.

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. We adopt five pre-trained LLMs: Vicuna-7B-v1.5 (Zheng et al., 2023), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-2-7B-Chat (Touvron et al., 2023b), Llama-3-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2024a).

Datasets. For AT, we use the training set from Harmbench (Mazeika et al., 2024) as the safety dataset and Alpaca (Taori et al., 2023) as the utility dataset. For the robustness evaluation, we construct a test set of size 100 that consists of the first 50 samples from the test set of Harmbench (Mazeika et al., 2024) and the first 50 samples from AdvBench (Zou et al., 2023). For the utility analysis, we use the benchmark data from AlpacaEval (Dubois et al., 2024).

Adversarial training. We leverage GCG (Zou et al., 2023), a token-level jailbreak attack, to synthesize (suffix) jailbreak prompts, in which the adversarial suffix length is fixed to one of $\{5, 10, 20, 30, 40, 50\}$ during AT. To reduce computational complexity of tuning LLMs, LoRA (Hu et al., 2022) is applied to all query and key projection matrices in attentions. In every AT experiment, we follow Eq. (3) to perform AT with AdamW. See Appendix B.2 for detailed AT hyperparameters.

Jailbreak attacks. Two token-level jailbreak attacks are used to evaluate the robustness of trained LLMs, which are GCG (Zou et al., 2023) and BEAST (Sadasivan et al., 2024). The token length of the adversarial suffix is varied in $\{5, 10, 20, 40, 60, 80, 100, 120\}$. See Appendix B.3 for details.

Evaluations. We focus on evaluating the jailbreak robustness and the utility of trained LLMs. For robustness evaluation, we report the **Attack Success Rate (ASR)** of jailbreak attacks. An LLM-based judger from Mazeika et al. (2024) is used to determine whether a jailbreak attack succeeds or not. Besides, for utility evaluation, we use AlpacaEval2 (Dubois et al., 2024) 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 more details.

5.2 Results Analysis

Correlation between the jailbreak robustness and the ratio of the square root of the jailbreak adversarial suffix length to the adversarial suffix length in AT (*i.e.*, $\sqrt{M_{\text{test}}}/M_{\text{train}}$). We plot the ASR of models trained and attacked with different adversarial suffix lengths in Figure 1 (48 points for each pair of base model and jailbreak attack). We also calculate the Pearson correlation coefficient (PCC) and corresponding p-value between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and ASR in Table 1.

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.

Model	GCG Attack		BEAST Attack	
	PCC (↑)	p-value (↓)	PCC (↑)	p-value (↓)
Vicuna-7B	0.93	4.70×10^{-21}	0.63	1.43×10^{-6}
Mistral-7B	0.86	3.97×10^{-15}	0.29	4.41×10^{-2}
Llama-2-7B	0.88	9.04×10^{-17}	0.68	1.32×10^{-7}
Llama-3-8B	0.76	2.75×10^{-10}	0.26	7.67×10^{-2}
Qwen2.5-7B	0.87	1.06×10^{-15}	0.58	1.03×10^{-5}

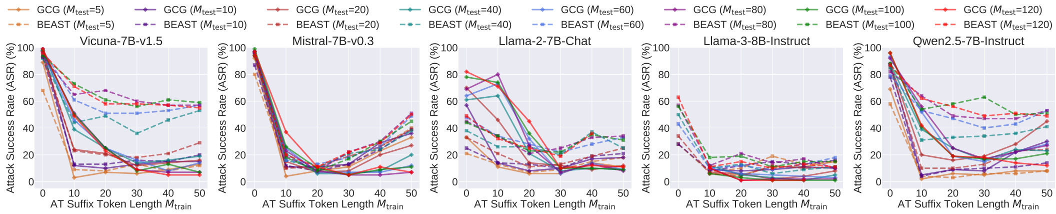

Figure 2: Curves of the ASR versus the adversarial suffix token length during AT (*i.e.*, M_{train}) under jailbreak attacks with different adversarial suffix token lengths (*i.e.*, M_{test}). $M_{\text{train}} = 0$ means that AT is not performed on the evaluated model. A high ASR indicates a weak jailbreak robustness.

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 correlation is statistically significant.

However, when the jailbreak attack in AT is different from that in robustness evaluation (*i.e.*, BEAST), from Table 1, the correlation between the ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$ and the ASR can only be observed from some of the base models (*i.e.*, Vicuna-7B, Llama-2-7B, and Qwen2.5-7B) but not others. 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 between the ratio and ASR is not significant (*i.e.*, Mistral-7B, and Llama-3-8B), GCG-based AT can still suppress the ASR to no more than 50%. This indicates that single-attack AT can still help models gain a certain degree of robustness against unseen attacks.

Relationship between adversarial suffix lengths in AT (*i.e.*, M_{train}) and jailbreaking (*i.e.*, M_{test}). We plot curves of the model ASR versus the adversarial suffix token length during AT in Figure 2, from which we find that as the adversarial suffix token length increases, AT can effectively reduce the ASR of both GCG and BEAST attacks. 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 also 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 suggest the effectiveness of defending against long-length jailbreaking with short-length AT.

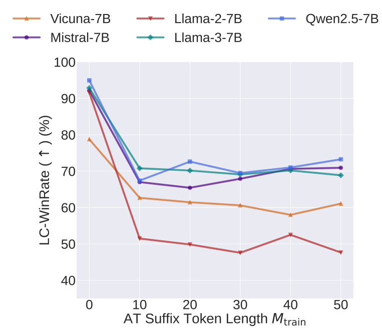

Figure 3: Utility analysis based on LC-WinRate. A high LC-WinRate indicates strong model utility. An LC-WinRate of 50% means that the evaluated model has the same quality as the reference model Davinci003.

Utility. We plot LC-WinRates of models trained with different adversarial suffix token lengths and the original pre-trained 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 Davinci003. This means that these adversarially trained models achieve utility comparable to Davinci003.

6 CONCLUSION

We study the LLM AT problem and unveils that to defend against a suffix jailbreak attack with suffix length of $\Theta(M)$, it is sufficient to perform AT on adversarial 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 upper 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 jailbreak ASR and ratio $\sqrt{M_{\text{test}}}/M_{\text{train}}$. Our results show that it is possible to conduct efficient "short-length" AT, which requires less GPU memory and training time, against strong "long-length" jailbreak attacks.

REFERENCES

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.

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.

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.

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.

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, pp. 1877–1901, 2020.

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.

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.

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

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*, 2024b.

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.

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

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

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.

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

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.

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.

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

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.

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

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

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.

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.

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.

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.

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.

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*, 2024a.

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*, 2024b.

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*, 2024c.

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.

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.

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.

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.

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.

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.

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.

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

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

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*, 2024a.

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*, 2024b.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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*, 2023a.

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*, 2023b.

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*, pp. 35151–35174. PMLR, 2023.

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

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

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

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.

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.

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.

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*, 2024a.

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*, 2024b.

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

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.

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.

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.

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 Zhang et al. (2024)). *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 Zhang et al. (2024); Also in Petersen et al. (2008)). *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*

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

Lemma A.5 (Von Neumann's Trace Inequality; Also in Lemma D.3 in Zhang et al. (2024)). *Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$ be two matrices. Suppose $\sigma_1(A) \le \cdots \le \sigma_{\min\{n,m\}}(A)$ and $\sigma_1(B) \le \cdots \le \sigma_{\min\{n,m\}}(B)$ are all the (ordered) 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

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

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 \\
&\quad + \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 + \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

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

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

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

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

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

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

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

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

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

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

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

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 Zhang et al. (2024). 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

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

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

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

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

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 Zhang et al. (2024). Specifically, when setting the (original) ICL prompt length from N to $(N + M_{\text{train}})$, the ICL training loss L in Zhang et al. (2024) 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 Zhang et al. (2024) 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

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

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,

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

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

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

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

$$\ell_3(\theta) = \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \big((w_{21}^V)^\top \quad w_{22}^V\big) \cdot \left(\sum_{i=1}^{M_{\text{train}}} \begin{pmatrix} \Lambda & 0_{d\times 1} \\ 0_{1\times d} & \mathrm{Tr}(\Lambda) \end{pmatrix}\right) \cdot \big((w_{21}^V)^\top \quad w_{22}^V\big)^\top$$
$$= \frac{2\epsilon^2 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \Big((w_{21}^V)^\top \Lambda w_{21}^V + \mathrm{Tr}(\Lambda)(w_{22}^V)^2\Big). \tag{A.14}$$

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

$$\partial_{w_{21}^V} \ell_3(\theta)\Big|_{w_{21}^V = 0} = \partial_{w_{21}^V}\Big[\frac{2\epsilon^2 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \Big((w_{21}^V)^\top \Lambda w_{21}^V + \mathrm{Tr}(\Lambda)(w_{22}^V)^2\Big)\Big]\Big|_{w_{21}^V=0}$$
$$= \frac{2\epsilon^2 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \partial_{w_{21}^V}\Big[(w_{21}^V)^\top \Lambda w_{21}^V\Big]\Big|_{w_{21}^V=0}$$
$$= \frac{4\epsilon^2 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot \Big[(w_{21}^V)^\top \Lambda\Big]\Big|_{w_{21}^V=0}$$
$$= \frac{4\epsilon^2 M_{\text{train}}^2}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\Big[\|W_{11}^{KQ} x_{\tau,q}\|_2^2\Big] \cdot 0_{d\times 1}^\top \Lambda = 0_{1\times d},$$

which justifies our claim in Step 3.

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

$$\partial_{w_{21}^V} \ell_4(\theta)\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}} = \partial_{w_{21}^V}\Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2}\|w_{21}^V\|_2^2 \cdot \mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2\Big]\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}}$$
$$= \Big[\frac{4\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2 \cdot (w_{21}^V)^\top\Big]\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}}$$
$$= \frac{4\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ 0_{d\times 1}^\top \end{pmatrix} x_{\tau,q}\|_2^2 \cdot 0_{d\times 1}^\top = 0_{1\times d}.$$

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

$$\partial_{w_{21}^{KQ}} \ell_4(\theta)\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}} = \partial_{w_{21}^{KQ}}\Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2}\|w_{21}^V\|_2^2 \cdot \mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2\Big]\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}}$$
$$= \Big[\frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \|w_{21}^V\|_2^2 \cdot \partial_{w_{21}^{KQ}} \mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2\Big]\Big|_{w_{21}^V = w_{21}^{KQ} = 0_{d\times 1}}$$
$$= \frac{2\epsilon^2 M_{\text{train}}}{(N + M_{\text{train}})^2} \cdot \|0_{d\times 1}\|_2^2 \cdot \partial_{w_{21}^{KQ}}\Big[\mathbb{E}_\tau\|\begin{pmatrix} X_\tau^{\text{sfx}} \\ Y_\tau^{\text{sfx}} \end{pmatrix}^\top \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_{\tau,q}\|_2^2\Big]\Big|_{w_{21}^{KQ} = 0_{d\times 1}} = 0_{1\times d}.$$

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,*

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

where $\Gamma(M) := \frac{N+M+1}{N+M}\Lambda + \frac{\mathrm{Tr}(\Lambda)}{N+M}I_d$ and $\psi(M) := \frac{M^2\mathrm{Tr}(\Lambda)}{(N+M)^2}$ 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,

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

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

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

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

$$
\begin{aligned}
&\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] \\
&= \mathbb{E}_\tau[X_\tau X_\tau^\top X_\tau X_\tau^\top] + \mathbb{E}_\tau[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top]\cdot\mathbb{E}_\tau[X_\tau X_\tau^\top] + \mathbb{E}_\tau[X_\tau X_\tau^\top]\cdot\mathbb{E}_\tau[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top] + \mathbb{E}_\tau[X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top X_\tau^{\text{sfx}}(X_\tau^{\text{sfx}})^\top] \\
&= \mathbb{E}_\tau\Big[\sum_{i,j} x_{\tau,i}x_{\tau,i}^\top x_{\tau,j}x_{\tau,j}^\top\Big] + \mathbb{E}_\tau\Big[\sum_i x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top\Big]\cdot\mathbb{E}_\tau\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top\Big] \\
&\quad + \mathbb{E}_\tau\Big[\sum_i x_{\tau,i}x_{\tau,i}^\top\Big]\cdot\mathbb{E}_\tau\Big[x_{\tau,i}^{\text{sfx}}(x_{\tau,i}^{\text{sfx}})^\top\Big] + \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] \\
&= \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 + \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] \\
&= \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 + \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. \qquad \text{(A.17)}
\end{aligned}
$$

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

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

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

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

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

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

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

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

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,

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

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

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

The proof is completed. □

We now turn to prove an PL-inequality for the surrogate AT problem. The proof idea follows that in Zhang et al. (2024). 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*

$$(w_{22}^V(t))^2 = \mathrm{Tr}[W_{11}^{KQ}(t)(W_{11}^{KQ}(t))^\top].$$

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

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

Similarly, for $\partial_t w_{22}^V(t)$, we have

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

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

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

which further indicates that

$$\begin{aligned}
&\partial_t \operatorname{Tr}\Big[W_{11}^{KQ}(t) (W_{11}^{KQ}(t))^\top\Big] = \operatorname{Tr}\Big[\partial_t W_{11}^{KQ}(t) \cdot (W_{11}^{KQ}(t))^\top\Big] + \operatorname{Tr}\Big[W_{11}^{KQ}(t) \cdot \partial_t (W_{11}^{KQ}(t))^\top\Big] \\
&= (\partial_t w_{22}^V(t)) \cdot w_{22}^V(t) + W_{22}^V(t) \cdot (\partial_t w_{22}^V(t)) = \partial_t (w_{22}^V(t)^2).
\end{aligned} \tag{A.24}$$

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

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

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

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

The proof is completed. □

Lemma A.10. *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, 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 $\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) = 2\operatorname{Tr}(\Lambda)$. Besides, under Assumption 1, we have $w_{22}^V(0) = \sigma > 0$. Therefore, if we can show that $\tilde{\mathcal{L}}^{\text{adv}}(\theta_t) \neq 2\operatorname{Tr}(\Lambda)$ 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 $\tilde{\mathcal{L}}^{\text{adv}}(\theta_t)$ at the initial training time $t = 0$. By applying Assumption 1, we have

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

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

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

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* $t \ge 0$, *we have* $(w_{22}^V(t))^2 \ge \nu > 0$, *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 $t \ge 0$,

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

which indicates

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

thus

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

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

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

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. $\square$

Lemma A.12 (PL-inequality)**.** *Suppose Assumption 1 holds and the LSA 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. 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*

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

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

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

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

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

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

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

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

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

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

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 4.3

This section collects all proofs that omitted from Section 4.3.

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 $f_{\text{LSA},\theta}$,

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

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

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

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}}) \leq 2\text{Tr}\left[\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}\right] + 2\text{Tr}(\Lambda), \\
&\leq \sum_{i=1}^{d} \lambda_i^3 \cdot \frac{\frac{N+M_{\text{test}}+1}{N+M_{\text{test}}}\lambda_i + \frac{\text{Tr}(\Lambda)}{N+M_{\text{test}}} + \epsilon^2 \cdot \frac{M_{\text{test}}^2 \text{Tr}(\Lambda)}{(N+M_{\text{test}})^2}}{\left(\frac{N+M_{\text{train}}+1}{N+M_{\text{train}}}\lambda_i + \frac{\text{Tr}(\Lambda)}{N+M_{\text{train}}} + \epsilon^2 \cdot \frac{M_{\text{train}}^2 \text{Tr}(\Lambda)}{(N+M_{\text{train}})^2}\right)^2} + 2\text{Tr}(\Lambda) \\
&\leq \sum_{i=1}^{d} \lambda_i^3 \cdot \frac{\frac{N+M_{\text{test}}+1}{N+M_{\text{test}}}\lambda_i + \frac{\text{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 \text{Tr}(\Lambda)}{(N+M_{\text{test}})^2}}{\left(\epsilon^2 \cdot \frac{M_{\text{train}}^2 \text{Tr}(\Lambda)}{(N+M_{\text{train}})^2}\right)^2} + 2\text{Tr}(\Lambda) \\
&\leq \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 \\
&\leq \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) \\
&\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}.
\end{aligned}
$$

Then, by applying Assumption 2, we further have that

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

which completes the proof. □

B Additional Experimental Details

This section collects experimental details omitted from Section 5.

B.1 Jailbreak Attacks

Our experiments adopt two token-level jailbreak attacks, which are GCG (Zou et al., 2023) and BEAST (Sadasivan et al., 2024). We only use token-level attacks since they make it easier for us to control the adversarial prompt length in our analysis. We re-implemented both attacks 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 Zou et al. (2023), 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 Sadasivan et al. (2024), hyperparameters that we need to tune for BEAST are two beam-search parameters k_1 and k_2.

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.

LoRA setting. We use the PEFT library (Mangrulkar et al., 2022) to apply LoRA (Hu et al., 2022) to all query and key projection matrices in attentions. The PEFT hyperparameters are `r=64`, `lora_alpha=32`, and `lora_dropout=0.1`.

Training hyparameters. In every AT experiment, we follow Eq. (3) to perform AT with AdamW for 125 iterations, in which 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.

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

- "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."

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 Mazeika et al. (2024), which was fine-tuned from the Llama-2-13B model [1], to 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 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.

Utility evaluation. We use the AlpacaEval2 framework (Dubois et al., 2024) 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 (Kwon et al., 2023).

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