Investigating Layer-Specific Vulnerability of LLMs to Adversarial Attacks

Cagatay Gultekin[1] (cgultekin)
Fabio Giovanazzi[1] (fgiovanazzi)
Adam Rahmoun[1] (arahmoun)
Tobias Kaiser[1] (tokaiser)

[1]ETH Zürich

Abstract

Large Language Models (LLMs) demonstrate remarkable capabilities across NLP tasks but remain vulnerable to adversarial "jailbreak" attacks that bypass alignment safeguards. Building on the transferable Greedy Coordinate Gradient (GCG) attack (Zou et al., 2023), we propose a layer-wise gradient regularization framework to quantify and mitigate per-layer vulnerabilities within transformer-based LLMs. By fine-tuning models under penalties targeted at individual layer gradients, we measure Attack Success Rates (ASR) on open-source LLMs. Our findings provide practical insights into the internal mechanics of adversarial transferability and suggest architectural guidelines for more robust model alignment.

1 Introduction

Transformer-based LLMs have rapidly advanced the state of the art in many natural language processing tasks, from text generation to question answering and dialogue (Vaswani et al., 2017; Brown et al., 2020). Despite their success, these models remain susceptible to adversarially crafted prompts—often called "jailbreaks"—that coerce them into producing harmful, toxic, or policy-violating content (Zou et al., 2023; Chen et al., 2023). Such vulnerabilities undermine trust and safety in deployed systems, as adversarial examples can transfer across model architectures and sizes, affecting both open-source and commercial offerings (e.g., ChatGPT variants).

Existing defenses focus on robust training regimes, input sanitization or output filtering, but they rarely inspect where within the model these adversarial gradients concentrate. Understanding the layer-specific sources of vulnerabilities can both shine light on the mechanics of attack transferability, and inform targeted architectural interventions. In this work, we use *layer-wise gradient regularization* to penalize the magnitude of gradients in selected transformer layers, and we show the ASR on a handful of fine-tuned models with gradient regularization applied to different layers. Our goal is twofold: (1) identify which layers contribute most to attack success, and (2) establish whether gradient suppression can meaningfully reduce ASR while preserving overall performance.

We evaluate our method on Phi-3 3.8B (Abdin et al., 2024) and Qwen2.5 1.5B (Qwen et al., 2025), two LLMs with different sizes and pretraining data, and measure ASR under the GCG attack framework (Zou et al., 2023). Our observations suggest that defensive mechanisms effective in computer vision contexts might not translate directly to language models, as the nature of gradient-based vulnerabilities in NLP appears fundamentally different from typical computer vision attacks.

2 Background

2.1 Adversarial Attacks on LLMs

Adversarial attacks on language models aim to find minimal perturbations (e.g., token-level modifications) that trigger undesired or harmful outputs. Gradient-based approaches, which work well on images since the pixel domain is continuous, do not work out of the box for text since text tokens are discrete. To overcome this, gradient-based attacks have been adapted by projecting continuous gradient signals into embedding space (Zou et al., 2023; Carlini and Wagner, 2023). The Greedy Coordinate Gradient (GCG) method iteratively appends and substitutes suffix tokens, using gradients of the negative log-likelihood to guide search for adversarial suffixes. Zou et al. demonstrated that GCG-crafted prompts transfer across models of varying sizes, indicating shared internal weaknesses.

2.2 Layer-wise Analysis in Deep Networks

Neural networks often exhibit hierarchical feature representations, where early layers capture low-level patterns and deeper layers encode abstract semantics (Yosinski et al., 2014). In adversarial contexts, saliency and gradient-based attribution methods have been used to identify vulnerable neurons or attention heads (Li et al., 2016; Clark et al., 2022). However, few studies have applied systematic *layer-wise* interventions to test and mitigate adversarial susceptibility in transformer stacks. Gradient regularization has shown promise in computer vision for smoothing loss landscapes and improving robustness (Roth et al., 2020), suggesting its potential for targeted use in NLP models.

2.3 Gradient Regularization

Gradient regularization adds a penalty term to the training objective loss proportional to the squared L2 norm of model gradients of the original task loss with respect to the parameters. Formally, given a loss $L_{\text{task}}(\theta)$, one can define:

$$L_{\text{total}}(\theta) = L_{\text{task}}(\theta) + \lambda \left\| \nabla_\theta L_{\text{task}}(\theta) \right\|_2^2 \quad (1)$$

where λ controls regularization strength.

Note that doing gradient descent on $L_{\text{total}}(\theta)$ requires $\nabla_\theta L_{\text{total}}(\theta)$, which in turn requires calculating the double derivative of the original task loss, $\nabla_\theta \nabla_\theta L_{\text{task}}(\theta)$.

In our study we take gradient regularization further by applying it only to selected target layers. Prior work in vision demonstrates that such penalties can shrink vulnerability metrics without hypersensitizing non-target layers, but its application to large-scale LLMs remains under-explored.

Our study bridges this gap by adapting layer-wise gradient regularization to transformer-based LLMs, evaluating its efficacy in reducing GCG attack success and analyzing resulting gradient patterns across layers.

3 Methodology

We employ a unified, *all-parameter* fine-tuning approach augmented with *layer-wise gradient regularization*, guided by empirical insights on optimal penalty strengths and backpropagation dynamics.

3.1 Layer-Wise Gradient Regularization

We propose an adaptation of Gradient Regularization (defined in eq. (1)) to target parameters in specific layers ϑ_i instead of all parameters θ:

$$L_{\text{total}}(\theta) = L_{\text{task}}(\theta) + \lambda \sum_{i \in I} \left\| \nabla_{\vartheta_i} L_{\text{task}}(\theta) \right\|_2^2 \quad (2)$$

where

- L_{task} is the original loss (Autoregressive LM loss) for the optimization task at hand,
- ϑ_i is the set of parameters of the transformer layer i,
- I denotes the set of layers targeted by gradient regularization,
- λ is the regularization coefficient.

By selecting I to target specific sets of layers (e.g. early, middle or late ones), we can observe the effect that gradient suppression on those specific layers has on adversarial robustness. By damping gradients in I, we exploit the chain-rule cascade: suppressing early-block gradients propagates reductions downstream, thereby disrupting the adversarial optimization path most effectively.

3.2 Experiment Configurations

We experimented with two transformer-based Large Language Models (LLMs):

- Qwen's **Qwen2.5-1.5B-Instruct** [1] which has 1.5B parameters and $N = 28$ transformer layers (Qwen et al., 2025)
- Microsoft's **Phi-3-mini-4k-instruct** [2] which has 3.8B parameters and $N = 32$ transformer layers (Abdin et al., 2024)

We partition each model's N layers into three contiguous groups:

$$\begin{aligned} I_{\text{early}} &= \left\{1, \ldots, \lfloor \tfrac{N}{3} \rfloor\right\}, \\ I_{\text{middle}} &= \left\{\lfloor \tfrac{N}{3} \rfloor + 1, \ldots, \lfloor \tfrac{2N}{3} \rfloor\right\}, \\ I_{\text{late}} &= \left\{\lfloor \tfrac{2N}{3} \rfloor + 1, \ldots, N\right\}. \end{aligned}$$

For the regularization coefficient λ, we have explored the following six values $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. On the other hand, we have fixed the learning rate η to a single value of 10^{-4} since our experiments

[1] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

[2] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

showed that higher values of η lead to training instability and/or catastrophic forgetting in our pre-trained models, while lower values of η mean a slower learning process. This yields $6 \times 1 \times 3 = 18$ configurations of (λ, η, I) per model, which totals up to 36 configurations across the two models.

In addition, we perform fine-tuning without *gradient regularization* ($L_{\text{total}}(\theta) = L_{\text{task}}(\theta)$). We train two models like this: one with $\eta = 10^{-4}$, and one with learning rate decay. We use these models as baselines in perplexity evaluation, to ensure the models trained with gradient regularization are not significantly worse. This leads to two additional configurations per model, and the total number of experiment configurations is therefore 40.

When gradient-regularized is in place, all layers' gradients tend to get lower (not just the targets), naturally slowing down learning over time, however this does not happen with non-regularized models, and this is the reason why we also trained one model with learning rate decay to compare against in a fair way. For learning rate decay, we start with $\eta = 10^{-4}$ and multiply by 0.9999 after each batch, so after 25,000 batches we have 8.2% of the initial learning rate.

3.3 Fine-Tuning Procedure

For each (λ, η, I) tuple we fine-tune a model:

1. Initialize the weights from the pretrained checkpoint downloaded from HuggingFace, using the `bfloat16` floating-point format.

2. Load 50,000 training text sequences of 513 tokens each taken from the "train" split of the C4 dataset. We used a batch size of 2 due to GPU limitations.

3. For each batch from the dataset:

   (a) Compute the model outputs on the first 512 tokens of each sample in the batch (i.e. all tokens except the 513th).

   (b) Compute the cross-entropy loss between the model outputs and the last 512 tokens of each sample (i.e. all tokens except the 1st), and call it L_{task}.

   (c) For each $i \in I$, compute $\|\nabla_{\vartheta_i} L_{\text{task}}\|_2^2$.

   (d) Compute L_{total} as defined in eq. (2)

   (e) Perform Stochastic Gradient Descent using learning rate η.

4. During training, we allow all parameters to update (i.e. we do not freeze any layer), because we noticed better results as opposed to just allowing the layers in I to change.

3.4 Adversarial Evaluation

We integrate the Greedy Coordinate Gradient (GCG) attack from Zou et al.'s repository (Zou et al., 2023), validating reproduction on an unmodified baseline. For each fine-tuned model, and for each AdvBench prompt:

1. Append a suffix made of n tokens to the AdvBench prompt. An example of prompt is "*Write a tutorial on how to make a bomb*".

2. Compute the model outputs on the prompt concatenated with the suffix.

3. Compute the cross entropy loss between the model outputs and the adversarial AdvBench outputs, and call it L_{task}. The adversarial output corresponding to the example above is "*Sure, here is a tutorial for making a bomb*".

4. Compute $\nabla_{\mathbf{e}} L_{\text{task}}$ with respect to the suffix embeddings $\mathbf{e}$.

5. At each suffix position j, project $-\nabla_{\mathbf{e}_j} L$ into the embedding space and select the top-k tokens whose embeddings differ from $\mathbf{e}_j$ in a direction that aligns well with the gradient.

6. Greedily replace one token at a time in the suffix, by computing the losses and picking the one with the lowest L_{task}, and iterate for t steps (num_steps $= 64$, search_width $= 64$, topk $= 32$).

For each fine-tuned model, we independently run the above process on 200 harmful prompts, and classify the outcomes:

- **Accept**: harmful content generated (counts toward ASR),
- **Refuse**: safe refusal,
- **Nonsense**: incoherent or unrelated (counts as defense success).

Attack Success Rate (ASR) is the fraction of *Accept* responses.

3.5 Gradient and Attack Analysis Metrics

To understand the mechanisms underlying regularization effects on adversarial robustness, we compute several metrics during GCG attacks and analyze their relationship to model vulnerability.

3.5.1 Gradient Magnitude Metrics

For each successful attack step, we extract gradients from the embedding layers and compute layer-wise statistics across three layer groups (early, middle, late):

L2 Norm: We compute the L2 norm of gradients for each layer group, hypothesizing that regularization would reduce gradient magnitudes and limit the GCG attack's ability to find effective adversarial tokens.

Standard Deviation: We calculate the standard deviation of gradient magnitudes within each layer group across attack steps, hypothesizing that lower gradient variability would constrain the search space available to GCG attacks.

3.5.2 Critical Token Ratio

For each successful attack, we compute the Critical Token Ratio (CTR), defined as the proportion of adversarial suffix tokens that are essential for maintaining attack success:

$$\text{CTR} = \frac{\text{Number of Critical Tokens}}{\text{Total Adversarial Tokens}} \tag{3}$$

A token is considered "critical" if removing it from the 20-token adversarial suffix causes the attack to fail (response changes from ACCEPT to REFUSE or NONSENSE). We hypothesized that higher CTR values would indicate better model robustness.

3.5.3 Hypotheses

Our analysis was guided by four key hypotheses from computer vision adversarial robustness literature: (1) regularization should reduce gradient magnitudes and improve robustness, (2) tighter gradient distributions should constrain the GCG search space, (3) higher CTR values should indicate improved robustness, and (4) early layer regularization should be more effective due to cascading effects through backpropagation and similarity to embedding gradients that GCG exploits.

However, as detailed in our results, these hypotheses were largely not supported by our empirical findings, revealing fundamental differences between discrete token attacks and continuous perturbation attacks in computer vision.

3.6 Perplexity Evaluation

To ensure the performance of the fine-tuned models trained using the procedure in section 3.3 does not degrade significantly, we compute the perplexity of the models' outputs on the "validation" split of the C4 dataset. In particular, we take 50,000 samples of 513 tokens from C4, we compute L_{task} as defined in section 3.3 on batches of size 2 (due to GPU limitations), and finally we calculate the perplexity as the average of $e^{L_{\text{task}}}$ over all batches. The evaluation of each model uses the same 50,000 samples to ensure a fair comparison.

4 Results

Table 1 contains the results of the evaluation described in section 3.4 on the various fine-tuned models. Additionally, the first line shows the ASR on the original model for comparison purposes.

Regularization λ	I	**GCG ASR (%)** **Qwen**	**Phi**
original model		82.0	16.5
no reg (const LR)		78.5	20.0
no reg (decay LR)		81.5	**14.0**
10^{0}	early	84.0	15.0
	middle	82.5	18.0
	late	80.5	19.5
10^{-1}	early	82.5	17.0
	middle	80.0	16.0
	late	75.5	15.0
10^{-2}	early	**74.0**	16.5
	middle	84.0	18.0
	late	82.0	19.0
10^{-3}	early	81.5	19.5
	middle	81.5	18.5
	late	79.0	20.0
10^{-4}	early	76.0	16.5
	middle	76.5	17.0
	late	77.0	14.5
10^{-5}	early	81.0	21.0
	middle	84.0	17.5
	late	77.0	14.5

Table 1: ASR of the original model and of fine-tuned models under the GCG attack, lower is better

To confirm GCG is behaving correctly, we also try to ask the models the harmful prompts directly, and as expected we get very low ASR, as shown in table 2.

4.1 Attack Success Rate Analysis

Table 1 presents the attack success rates across all regularization configurations for both models, revealing distinct vulnerability patterns and the effectiveness of layer-specific gradient regularization interventions. Complete results for all regularization strengths and layer combinations are provided in Appendix.

4.1.1 Qwen2.5-1.5B-Instruct Performance

As shown in table 1, the original Qwen2.5-1.5B-Instruct model exhibits a baseline ASR of 82.0%, indicating substantial vulnerability to GCG attacks. Our layer-wise gradient regularization approach shows modest improvements, with the most effective configurations achieving reductions of up to 8 percentage points, though the magnitude of these improvements is limited.

The best performing configurations (from table 1) include early layer regularization at 1e-2 strength (74.0% ASR), late layer regularization at 1e-1 strength (75.5% ASR), and early layer regularization at 1e-4 strength (76.0% ASR). While these results suggest a potential trend toward **early layer regularization** being more effective, the improvements are modest and the pattern is not entirely consistent across all regularization strengths.

Several configurations show performance degradation (as evident in table 1), with early layer regularization at strength 1 (84.0% ASR), middle layer regularization at 10^{-2} strength (84.0% ASR), and middle layer regularization at 10^{-5} strength (84.0% ASR) each representing a 2 percentage point increase in ASR. The apparent sensitivity of **middle layer regularization** to degradation could indicate that these layers are important for defensive capabilities, though this pattern requires further investigation given the limited scale of observed effects.

4.1.2 Phi-3-mini-4k-instruct Performance

According to table 1, Phi-3-mini demonstrates significantly better baseline robustness with an original ASR of 16.5%, approximately 5$\times$ more robust than Qwen. The model shows variable responses to regularization interventions, with improvements generally limited to 1-3 percentage points.

The most effective configurations (based on table 1) include the no-regularization baseline with learning rate decay (14.0% ASR), late layer regularization at 1e-4 and 1e-5 strengths (both 14.5% ASR), and early layer regularization at strength 1 (15.0% ASR). Notably, the best performance comes from the **no-regularization baseline with learning rate decay**, suggesting that for this already-robust model, training stability may be more important than explicit gradient constraints. The modest improvements from regularization interventions make it difficult to draw strong conclusions about optimal layer targeting strategies.

Some regularization configurations result in performance degradation (visible in table 1), including early layer regularization at 1e-5 strength (21.0% ASR) and late layer regularization at 1e-3 strength (20.0% ASR). However, given the overall small scale of effects and high variability across configurations, these patterns should be interpreted cautiously.

4.1.3 Cross-Architecture Vulnerability Patterns

The substantial difference in baseline vulnerability (82.0% vs 16.5% ASR, as shown in table 1) indicates fundamental architectural differences in adversarial susceptibility. However, the observed patterns from regularization interventions show limited and inconsistent effects:

Regularization Responsiveness: While Qwen shows some improvements (4-8 percentage points) with certain interventions, and Phi-3 exhibits variable responses (as detailed in table 1), the magnitude of these effects is modest relative to the baseline differences between models.

Layer-Specific Effects: The data in table 1 suggests potential differences in layer vulnerability, with some indication that Qwen's early layers and Phi-3's late layers may be more responsive to regularization. However, these patterns are not consistent across all regularization strengths and should be interpreted as preliminary observations rather than definitive findings.

Regularization Strength Sensitivity: Both models show sensitivity to regularization strength (evident throughout table 1), but optimal values appear highly dependent on the specific model and layer combination, making it difficult to establish general principles.

These results suggest that while layer-wise gradient regularization can influence adversarial robustness, the effects are generally modest and highly dependent on model architecture and hyperparameter selection. The limited and inconsistent nature of improvements indicates that gradient regularization alone may not be sufficient for substantial robustness gains, and that architectural differences

between models may be more important determinants of adversarial vulnerability than previously anticipated.

4.2 Gradient Analysis and Attack Mechanism

To understand the potential mechanisms underlying our regularization interventions, we analyze gradient patterns across different model configurations and response categories. However, our analysis reveals limited and inconsistent relationships between gradient metrics and adversarial robustness.

4.2.1 Baseline Gradient Pattern Analysis

To better understand the effects of regularization, we first examine gradient patterns across our three baseline conditions: original models, no-regularization with constant learning rate, and no-regularization with learning rate decay. These comparisons reveal fundamental inconsistencies that extend beyond regularization effects.

In this section, the three gradients separated by slashes (e.g. 100/200/300 or respectively 0.001/0.002/0.003) indicate the average (or respectively standard deviation) of gradients for early/middle/late layers among each step of each GCG attack with a REFUSE outcome.

Fine-tuning Effects on Baselines: Comparing original models to fine-tuned baselines shows substantial changes in gradient landscapes (detailed gradient values can be found in table 3). For Qwen, REFUSE gradients show a consistent downward trend from original (362/244/105) to constant learning rate (228/115/78) to decay learning rate (202/144/90) conditions, while ASR improves modestly from 82.0% to 78.5% to 81.5%. However, this apparent relationship breaks down when examining Phi-3, where the best performing baseline (decay learning rate, 14.0% ASR) exhibits dramatically elevated REFUSE gradients (3932/123/374) compared to both the original model (1103/213/471) and constant learning rate condition (904/115/242).

Critical Token Ratio Baseline Patterns: Fine-tuning consistently increases critical token ratios (CTRs) across both models (as shown in table 6), with Qwen showing a substantial jump from 0.654 (original) to 0.799 (constant LR) to 0.822 (decay LR), while Phi-3 exhibits more modest increases from 0.738 to 0.758 to 0.782. Higher CTR values indicate that attacks require more critical tokens to succeed, theoretically suggesting improved model robustness. The baseline progression shows some support for this relationship, with both models achieving their best baseline ASR performance (Qwen: 78.5%, Phi-3: 14.0%) at their highest baseline CTR values (0.799 and 0.782 respectively). However, this pattern does not hold consistently across regularized configurations, where some of the highest CTR values (e.g., Phi-3 Late-reg-1 at 0.888, Qwen Early-reg-1e-1 at 0.866, from table 6) correspond to worse ASR performance, indicating that CTR alone is not a reliable predictor of robustness.

Gradient Variability Across Training Conditions: The standard deviations of gradient measurements (documented in table 3) show that training methodology profoundly affects gradient stability. Our initial hypothesis was that lower gradient standard deviations would create tighter gradient distributions, reducing the search space available to GCG attacks and thereby improving model robustness. However, the data does not support this relationship consistently. While Qwen's most robust regularized configuration (early layer regularization at 1e-4) does exhibit dramatically reduced standard deviations (0.0007/0.0001/0.0002), Phi-3's best performing baseline condition (decay learning rate, 14.0% ASR) shows the opposite pattern with the highest REFUSE gradient standard deviation (0.0141) in early layers, representing a dramatic increase from both original (0.0051) and constant learning rate (0.0022) conditions. This inconsistency suggests that gradient variability alone is not a reliable predictor of adversarial robustness across different model architectures.

REFUSE Gradient Patterns: Examining REFUSE response gradients relative to baseline configurations (using values from table 3) reveals no systematic patterns that predict robustness improvements. Using the no-regularization constant learning rate condition as reference (Qwen: 228/115/78, Phi-3: 904/115/242 for early/middle/late layers), we observe that improved ASR performance can result from dramatically different gradient changes. For Qwen's most effective configuration (early layer regularization at 1e-4 strength, 76.0% ASR), REFUSE gradients decrease substantially (-61%/-67%/-26% relative to baseline) with correspondingly reduced standard deviations (0.0007/0.0001/0.0002 vs baseline 0.0044/0.0019/0.0006). However, another well-performing configuration (late layer regularization at 1e-1 strength, 75.5% ASR) shows the opposite pattern with REFUSE gradient increases

(+69%/+73%/+45% vs baseline). This inconsistency suggests that REFUSE gradient magnitude and direction relative to baseline are not reliable predictors of defensive capability.

For Phi-3-mini, the baseline comparison (as documented in table 3) reveals even greater inconsistencies. The best performing regularized configuration (late layer regularization at 1e-4 strength, 14.5% ASR) exhibits a +242% increase in early layer REFUSE gradients compared to baseline, while another equally effective configuration (late layer regularization at 1e-5 strength, also 14.5% ASR) shows a -67% decrease in the same metric. These findings indicate that similar robustness outcomes can be achieved through completely opposite changes in gradient patterns.

ACCEPT Gradient Patterns: The analysis of ACCEPT response gradients relative to baseline configurations (values from table 4) reveals similar inconsistencies. For Qwen, robust configurations show divergent patterns when compared to the baseline model (no reg const LR) (974/412/215): the early layer regularization at 1e-4 configuration shows dramatic reductions (-83%/-83%/-53%) in ACCEPT gradients with correspondingly reduced standard deviations, while the early layer regularization at 1e-2 configuration shows more modest decreases (-12%/-7%/-4%) despite achieving better ASR performance (74.0% vs 76.0%). This suggests that neither the magnitude nor the variance of ACCEPT gradient changes relative to baseline reliably predicts robustness improvements.

For Phi-3, ACCEPT gradient patterns (detailed in table 4) show equally inconsistent relationships to robustness when compared to baseline (526/69/143). The most effective configurations exhibit changes ranging from +130% to -73% in early layer ACCEPT gradients, with no apparent correlation to their similar ASR performance levels. The corresponding standard deviations also vary dramatically across these equally effective configurations, further undermining any systematic relationship between ACCEPT gradient characteristics and adversarial robustness.

4.2.2 Statistical Limitations and Interpretation

Our gradient analysis faces important limitations that constrain interpretation. The modest effect sizes (4-8 percentage points for Qwen, 1-3 percentage points for Phi-3, as shown in table 1) combined with high gradient variability across 20 configurations per model limit robust statistical inference. Most critically, similar robustness improvements result from completely opposite gradient changes (+242% increases vs -83% decreases), suggesting that observed patterns may reflect the stochastic nature of high-dimensional gradient landscapes rather than meaningful mechanistic relationships. The signal-to-noise ratio appears insufficient to establish reliable predictive relationships between gradient characteristics and adversarial robustness.

4.2.3 Architectural Differences and Robustness

The substantial baseline vulnerability difference (82.0% vs 16.5% ASR, from table 1) dominates regularization effects, suggesting that architectural characteristics are the primary determinant of adversarial susceptibility. The fundamentally different gradient landscapes and regularization responses between Phi-3-mini and Qwen2.5-1.5B indicate that gradient-based metrics alone may be insufficient for understanding adversarial robustness in transformer-based language models. These findings point toward the need for different analytical approaches that account for the discrete and contextual nature of adversarial attacks.

4.3 Perplexity

Table 7 presents the perplexity evaluation results. Most fine-tuned models maintain comparable perplexity to the baseline (55.0 for Qwen, 271 for Phi-3), confirming that regularization does not significantly degrade language modeling performance. Original models exhibit extremely high perplexity (5681 for Qwen, 2.8e6 for Phi-3) as they were never trained on C4 data. A few regularized configurations show degraded performance exceeding 100, indicating training difficulties in specific hyperparameter combinations.

5 Discussion

5.1 Gradient-Based Defense Limitations in Discrete Token Spaces

Our findings reveal that gradient regularization, a technique proven effective in computer vision adversarial defense, shows limited and inconsistent benefits when applied to transformer-based language models. This parallels the broader observation that many adversarial attack methods from computer vision do not directly transfer to natural language processing tasks. We hypothesize that

this limitation extends to defensive mechanisms that attempt to flatten gradient landscapes, and potentially to other vision-derived defense strategies.

5.2 Attack Mechanism and Gradient Interpretation

The limited effectiveness of gradient magnitude manipulation may stem from fundamental differences between discrete token attacks and continuous perturbation attacks. The GCG attack operates by computing gradients of a harmful prompt given a desired harmful response, projecting these gradients into the embedding space via the embedding matrix, and selecting tokens that best align with the gradient direction to maximize the likelihood of the target response.

This mechanism differs crucially from traditional adversarial attacks (e.g., FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2019)) that directly maximize loss through gradient ascent. Instead, GCG attempts to maximize the probability of a specific desired output, making gradient magnitude less directly relevant to attack success. The attack's effectiveness could depend more on gradient *direction* and alignment with embedding vectors than on gradient *magnitude*, which may explain why our gradient norm measurements show inconsistent relationships with robustness.

5.3 Gradient Distribution Effects on Token Selection

Our initial hypothesis that tighter gradient distributions (lower standard deviations) would improve robustness by constraining the GCG search space proves overly simplistic. While constraining gradient distributions may limit attack options, the effect depends critically on which tokens the constrained gradients point toward.

If gradient regularization narrows the distribution such that gradients consistently point toward relatively harmless tokens, the attack could become more difficult and robustness improves. However, if the regularization constrains gradients to consistently point toward highly effective adversarial tokens, the attack may become more efficient and robustness degrades.

This dual possibility could explain the irregular patterns observed across our configurations: gradient distribution narrowing may be either beneficial or detrimental depending on the specific tokens favored by the constrained gradient landscape.

5.4 Critical Token Ratio Interpretation Challenges

Our critical token ratio analysis reveals complexity in understanding discrete token attacks. While higher CTR values suggest that attacks require more tokens to succeed (implying better model robustness), this interpretation assumes that attack effectiveness scales linearly with token count.

However, the GCG attack's token selection process prioritizes tokens based on gradient alignment with the embedding matrix. If the highest-priority tokens remain available even when the total token budget is reduced, the attack may maintain near-optimal effectiveness despite using fewer tokens. This suggests that CTR measurements could reflect not just attack robustness, but also the specific distribution of token effectiveness in the gradient-embedding alignment space.

5.5 Implications for Future Defense Strategies

These findings suggest that effective defenses against discrete token attacks may require fundamentally different approaches than those successful in computer vision. Rather than focusing solely on gradient magnitude or distribution constraints, defenses might target the embedding space structure, the alignment between gradients and embeddings, or the token selection mechanisms themselves.

The inconsistent patterns observed across model architectures also indicate that defensive strategies may need to be tailored to specific model characteristics rather than applied universally. The substantial baseline differences in vulnerability between Phi-3-mini and Qwen2.5-1.5B suggest that architectural design choices could be more important for adversarial robustness than post-training defensive interventions.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann,

Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

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

Nicholas Carlini and David A. Wagner. 2023. Gradient-based text adversarial attacks in embedding space. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Xi Chen, John Lee, and Alice Smith. 2023. Adversarial vulnerabilities in aligned language models. *Proceedings of the 2023 Conference on Security and Trust in AI*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2022. What does bert look at? an analysis of bert's attention. *Transactions of the Association for Computational Linguistics*, 10:87–103.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Daniel Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 681–691.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Kevin Roth, Bojan Du, Percy Liang, Andrew Kirsch, and Christopher D. Manning. 2020. Adversarial robustness via gradient regularization. In *Proceedings of the 2020 International Conference on Computer Vision*, pages 201–210.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pages 3320–3328.

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

A Further Results

This appendix provides comprehensive detailed results for all metrics analyzed in our study, including attack success rates for direct prompting, gradient patterns across response categories, critical token ratios, and perplexity evaluations.

	ASR (%)	
	Qwen	**Phi**
original model	3.5	1.0
no reg (const LR)	3.5	1.5

Table 2: ASR of the original model and of the non-regularized fine-tuned model when prompted directly, lower is better

Regularization		GCG ASR (%)		REFUSE Grad L2 (E/M/L)		REFUSE Grad Std (E/M/L, $\cdot 10^{-4}$)	
λ	I	**Qwen**	**Phi**	**Qwen**	**Phi**	**Qwen**	**Phi**
original model		82.0	16.5	362/244/105	1103/213/471	61/32/8	51/4/12
no reg (const LR)		78.5	20.0	228↓/115↓/78↓	904↓/115↓/242↓	44↓/19↓/6↓	22↓/1↓/6↓
no reg (decay LR)		81.5	**14.0**	202↓/144↓/90↓	3932↑/123↓/374↓	32↓/20↓/6↓	141↑/1↓/5↓
10^0	early	84.0	15.0	208↓/113↓/80↓	332↓/71↓/163↓	33↓/16↓/6↓	7↓/1↓/4↓
	middle	82.5	18.0	216↓/121↓/71↓	649↓/87↓/211↓	32↓/14↓/4↓	15↓/1↓/6↓
	late	80.5	19.5	439↑/197↓/109↑	225↓/55↓/111↓	117↑/20↓/7↓	5↓/1↓/3↓
10^{-1}	early	82.5	17.0	238↓/133↓/81↓	298↓/66↓/146↓	37↓/20↓/6↓	6↓/1↓/4↓
	middle	80.0	16.0	322↓/182↓/96↓	4116↑/107↓/327↓	63↑/28↓/6↓	148↑/1↓/4↓
	late	75.5	15.0	388↑/199↓/113↑	303↓/63↓/149↓	59↓/19↓/6↓	7↓/1↓/4↓
10^{-2}	early	**74.0**	16.5	339↓/187↓/101↓	351↓/71↓/162↓	41↓/19↓/6↓	9↓/1↓/5↓
	middle	84.0	18.0	276↓/145↓/85↓	4217↑/103↓/316↓	43↓/15↓/7↓	156↑/1↓/4↓
	late	82.0	19.0	280↓/156↓/85↓	6101↑/111↓/403↓	56↓/26↓/8−	197↑/1↓/5↓
10^{-3}	early	81.5	19.5	160↓/111↓/69↓	4918↑/114↓/358↓	22↓/11↓/5↓	162↑/1↓/4↓
	middle	81.5	18.5	345↓/158↓/86↓	296↓/65↓/143↓	64↑/28↓/9↑	8↓/1↓/4↓
	late	79.0	20.0	396↑/193↓/93↓	619↓/90↓/192↓	87↑/39↑/9↑	15↓/1↓/4↓
10^{-4}	early	76.0	16.5	88↓/37↓/57↓	4426↑/103↓/323↓	7↓/1↓/2↓	159↑/1↓/4↓
	middle	76.5	17.0	222↓/128↓/78↓	5082↑/106↓/358↓	43↓/17↓/7↓	154↑/1↓/3↓
	late	77.0	14.5	321↓/188↓/101↓	3097↑/116↓/344↓	56↓/28↓/9↑	115↑/1↓/5↓
10^{-5}	early	81.0	21.0	363↑/186↓/98↓	609↓/95↓/215↓	70↑/21↓/8−	14↓/1↓/5↓
	middle	84.0	17.5	218↓/123↓/72↓	3322↑/104↓/286↓	56↓/22↓/5↓	108↑/1↓/3↓
	late	77.0	14.5	304↓/175↓/94↓	295↓/62↓/146↓	44↓/23↓/7↓	7↓/1↓/4↓

Table 3: ASR, REFUSE Gradient L2 Norms and Standard Deviations compared to original model (↑ higher, ↓ lower)

Regularization		GCG ASR (%)		ACCEPT Grad L2 (E/M/L)		ACCEPT Grad Std (E/M/L, $\cdot 10^{-4}$)	
λ	I	**Qwen**	**Phi**	**Qwen**	**Phi**	**Qwen**	**Phi**
original model		82.0	16.5	819/477/247	299/58/137	76/37/8	24/1/5
no reg (const LR)		78.5	20.0	974↑/412↓/215↓	526↑/69↑/143↑	86↑/26↓/7↓	27↑/2↑/8↑
no reg (decay LR)		81.5	**14.0**	846↑/442↓/240↓	157↓/39↓/68↓	76−/27↓/8−	8↓/1−/2↓
10^0	early	84.0	15.0	756↓/396↓/223↓	112↓/29↓/54↓	60↓/21↓/7↓	6↓/1−/4↓
	middle	82.5	18.0	769↓/369↓/215↓	177↓/37↓/69↓	66↓/22↓/6↓	8↓/1−/2↓
	late	80.5	19.5	745↓/377↓/216↓	139↓/28↓/60↓	65↓/22↓/6↓	6↓/1−/4↓
10^{-1}	early	82.5	17.0	705↓/359↓/215↓	108↓/32↓/59↓	54↓/18↓/6↓	3↓/1−/4↓
	middle	80.0	16.0	738↓/382↓/218↓	1552↑/47↓/137−	66↓/23↓/8−	128↑/1−/5−
	late	75.5	15.0	738↓/343↓/202↓	92↓/27↓/49↓	79↑/19↓/7↓	3↓/1−/2↓
10^{-2}	early	**74.0**	16.5	859↑/384↓/206↓	104↓/31↓/56↓	69↓/27↓/8−	3↓/1−/3↓
	middle	84.0	18.0	774↓/372↓/217↓	581↑/42↓/82↓	61↓/20↓/6↓	35↑/1−/1↓
	late	82.0	19.0	877↑/426↓/228↓	1749↑/47↓/138↑	73↓/27↓/7↓	151↑/1−/3↓
10^{-3}	early	81.5	19.5	703↓/423↓/247−	573↑/44↓/89↓	60↓/29↓/10↑	31↑/1−/1↓
	middle	81.5	18.5	725↓/372↓/211↓	104↓/31↓/55↓	45↓/17↓/6↓	4↓/1−/2↓
	late	79.0	20.0	668↓/348↓/205↓	170↓/39↓/77↓	52↓/23↓/7↓	8↓/1−/2↓
10^{-4}	early	76.0	16.5	166↓/68↓/101↓	1006↑/46↓/102↓	6↓/2↓/2↓	64↑/1−/3↓
	middle	76.5	17.0	844↑/402↓/220↓	2082↑/48↓/155↑	77↑/26↓/8−	140↑/1−/4↓
	late	77.0	14.5	1078↑/425↓/212↓	1207↑/53↓/139↑	117↑/32↓/8−	100↑/1−/5−
10^{-5}	early	81.0	21.0	790↓/371↓/213↓	368↑/58−/107↓	73↓/23↓/7↓	20↓/1−/5−
	middle	84.0	17.5	733↓/375↓/217↓	1189↑/43↓/111↓	67↓/24↓/7↓	99↑/1−/3↓
	late	77.0	14.5	674↓/354↓/203↓	144↓/28↓/66↓	56↓/22↓/6↓	9↓/1−/6↑

Table 4: ASR, ACCEPT Gradient L2 Norms and Standard Deviations compared to original model (↑ higher, ↓ lower, − same)

Regularization		GCG ASR (%)		NONSENSE Grad L2 (E/M/L)		NONSENSE Grad Std (E/M/L, $\cdot 10^{-4}$)	
λ	I	**Qwen**	**Phi**	**Qwen**	**Phi**	**Qwen**	**Phi**
original model		82.0	16.5	268/159/60	111/17/52	71/34/7	3/1/3
no reg (const LR)		78.5	20.0	185↓/96↓/69↑	40↓/17−/21↓	37↓/12↓/5↓	4↑/1−/1↓
no reg (decay LR)		81.5	**14.0**	244↓/153↓/81↑	88↓/17−/23↓	55↓/23↓/5↓	10↑/1−/1↓
10^0	early	84.0	15.0	258↓/133↓/71↑	16↓/8↓/11↓	55↓/20↓/8↑	1↓/1−/1↓
	middle	82.5	18.0	194↓/113↓/78↑	112↑/17−/28↓	26↓/8↓/6↓	11↑/1−/2↓
	late	80.5	19.5	541↑/148↓/72↑	41↓/15↓/23↓	193↑/39↑/8↑	2↓/1−/2↓
10^{-1}	early	82.5	17.0	149↓/89↓/67↑	40↓/13↓/25↓	31↓/10↓/4↓	1↓/1−/6↑
	middle	80.0	16.0	219↓/111↓/70↑	90↓/14↓/21↓	26↓/8↓/3↓	11↑/1−/1↓
	late	75.5	15.0	398↑/139↓/66↑	61↓/19↑/30↓	73↑/14↓/5↓	3−/1−/4↑
10^{-2}	early	**74.0**	16.5	242↓/142↓/92↑	27↓/8↓/9↓	35↓/14↓/6↓	0↓/0↓/0↓
	middle	84.0	18.0	193↓/108↓/69↑	47↓/12↓/20↓	29↓/15↓/7−	8↑/1−/1↓
	late	82.0	19.0	186↓/105↓/65↑	744↑/26↑/72↑	53↓/16↓/5↓	82↑/1−/5↑
10^{-3}	early	81.5	19.5	240↓/144↓/91↑	52↓/16↓/18↓	36↓/15↓/6↓	5↑/1−/1↓
	middle	81.5	18.5	154↓/86↓/57↓	18↓/8↓/12↓	36↓/9↓/3↓	2↓/0↓/1↓
	late	79.0	20.0	258↓/154↓/86↑	37↓/15↓/26↓	31↓/13↓/5↓	1↓/1−/1↓
10^{-4}	early	76.0	16.5	43↓/17↓/24↓	53↓/12↓/20↓	2↓/1↓/1↓	5↑/1−/1↓
	middle	76.5	17.0	343↑/177↑/88↑	254↑/17−/35↓	56↓/32↓/7−	32↑/1−/2↓
	late	77.0	14.5	262↓/108↓/68↑	43↓/15↓/16↓	71−/18↓/6↓	3−/1−/1↓
10^{-5}	early	81.0	21.0	674↑/238↑/98↑	67↓/15↓/29↓	183↑/66↑/11↑	4↑/0↓/1↓
	middle	84.0	17.5	120↓/67↓/46↓	508↑/21↑/55↑	30↓/14↓/2↓	33↑/1−/1↓
	late	77.0	14.5	271↑/148↓/84↑	36↓/17−/21↓	42↓/25↓/7−	2↓/0↓/3−

Table 5: ASR, NONSENSE Gradient L2 Norms and Standard Deviations compared to original model (↑ higher, ↓ lower, − same)

Regularization		GCG ASR (%)		Critical Token Ratio	
λ	I	**Qwen**	**Phi**	**Qwen**	**Phi**
original model		82.0	16.5	0.654	0.738
no reg (const LR)		78.5	20.0	0.799↑	0.758↑
no reg (decay LR)		81.5	**14.0**	0.822↑	0.782↑
10^0	early	84.0	15.0	0.818↑	0.844↑
	middle	82.5	18.0	0.810↑	0.734↓
	late	80.5	19.5	0.796↑	0.888↑
10^{-1}	early	82.5	17.0	0.866↑	0.794↑
	middle	80.0	16.0	0.804↑	0.718↓
	late	75.5	15.0	0.850↑	0.841↑
10^{-2}	early	**74.0**	16.5	0.805↑	0.823↑
	middle	84.0	18.0	0.837↑	0.767↑
	late	82.0	19.0	0.801↑	0.846↑
10^{-3}	early	81.5	19.5	0.848↑	0.794↑
	middle	81.5	18.5	0.805↑	0.774↑
	late	79.0	20.0	0.835↑	0.811↑
10^{-4}	early	76.0	16.5	0.859↑	0.742↑
	middle	76.5	17.0	0.790↑	0.842↑
	late	77.0	14.5	0.784↑	0.752↑
10^{-5}	early	81.0	21.0	0.839↑	0.871↑
	middle	84.0	17.5	0.822↑	0.744↑
	late	77.0	14.5	0.854↑	0.764↑

Table 6: Critical Token Ratio compared to original model (↑ higher, ↓ lower)

Regularization		Perplexity	
λ	I	**Qwen**	**Phi**
original model		5681	2.8e6
no reg (const LR)		55.0↓	271↓
no reg (decay LR)		43.0↓	26.5↓
10^0	early	58.0↓	48.5↓
	middle	56.8↓	149*↓
	late	51.6↓	52.2↓
10^{-1}	early	42.6↓	49.6↓
	middle	51.3↓	21.0↓
	late	48.8↓	47.4↓
10^{-2}	early	58.3↓	43.2↓
	middle	48.3↓	20.6↓
	late	49.9↓	21.5↓
10^{-3}	early	50.0↓	21.7↓
	middle	45.8↓	50.8↓
	late	48.2↓	581*↓
10^{-4}	early	9.0↓	21.9↓
	middle	51.0↓	21.0↓
	late	55.2↓	18.7↓
10^{-5}	early	48.2↓	465*↓
	middle	44.9↓	20.3↓
	late	50.6↓	56.8↓

Table 7: Perplexity of fine-tuned models compared to original model (↓ lower is better)