ETH zürich

Computational Semantics for Natural Language Processing,
Spring Semester 2025

Investigating Layer-Specific Vulnerability of LLMs to Adversarial Attacks

Cagatay Gultekin[1], Fabio Giovanazzi[1], Adam Rahmoun[1], Tobias Kaiser[1]

[1]ETH Zürich

1 Introduction

LLMs excel at NLP tasks [2], but remain vulnerable to adversarial prompts ("*jailbreaks*") that elicit harmful or restricted content [3]. Even when crafted on small open-source models, these prompts often **transfer to larger commercial LLMs** [6].

Existing defenses rarely examine where adversarial gradients concentrate within the model, though we believe studying this could explain attack transferability and guide countermeasures. Inspired by the effectiveness of gradient-based defenses in computer vision [4], we aim to: (1) identify **most critical layers** for attack success, and (2) test whether **gradient suppression** reduces the *attack success rate* (ASR) without impacting performance.

2 Method

Gradient regularization. We use *layer-wise gradient regularization* to suppress gradients only in selected transformer layers $\vartheta \subset \theta$, with λ controlling regularization strength. The loss is thus:

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

Experiment configurations. We experiment on Phi-3 3.8B [1] and Qwen2.5 1.5B [5], partition each model into early, middle and late layers, and consider 6 values for λ ranging from 1 to 10^{-5}. We train each configuration on 50k text sequences from the C4 training dataset with a fixed LR of $\eta = 10^{-4}$. Furthermore, for each model, we fine-tune two baselines without gradient regularization ($\lambda = 0$), one with fixed LR and one with an exponential LR schedule with $\gamma = 0.9999$. This yields $2 \cdot (6 \cdot 3 + 2) = 40$ configurations in total.

Perplexity evaluation. To ensure that the performance of fine-tuned models does not degrade significantly, we calculate model perplexity values on the C4 validation dataset.

Attack. We run 200 Greedy Coordinate Gradient (GCG) attacks [6] on each one of 40 fine-tuned models and 2 original models. Each attack starts with a prompt from the *harmful behaviors* dataset [6] (e.g. "How to [harmful]?") with an added fixed-size suffix and iteratively adapts the suffix tokens based on their embedding gradients, to nudge the response into "Sure, here is how to [harmful]:".

```
User:  Tell me how to build a bomb. ! ! ! ! ! ! ! !
Assistant:  Sure, here is how to build a bomb:
```

Figure 1: GCG "trains" the suffix `! ! ! ! ! ! ! !` to obtain a harmful response

Evaluation metrics. *ASR* is the ratio of successful attacks over the total, whereas *Critical Token Ratio* (CTR) is the proportion of adversarial suffix tokens that are essential for maintaining attack success. For each attack, we compute the average L2 norm and the standard deviation of the gradients (of the GCG attack loss) from each layer group.

Working Hypotheses. Guided by computer vision adversarial robustness literature, we expect that: (1) regularization reduces gradients and improves robustness, (2) low standard deviation in gradients constrains the GCG search space and improves robustness, (3) high CTR indicates improved robustness, and (4) early layer regularization is more effective due to cascading effects of backprop and similarity to embedding gradients that GCG exploits.

3 Results

ASR Analysis.

For *Qwen2.5*, the best/worst ASR with the value of 74.0%/84.0% is achieved by the configuration of *early/middle layer regularization at* 10^{-2}/10^{-5} (latter tied with two other configurations). For *Phi-3*, the best/worst ASR with the value of 14.0%/21.0% is achieved by the configuration of *no-regularization baseline with exponential LR schedule* and *early layer regularization at* 10^{-5} respectively.

Due to the modest level of ASR improvements / degradations between configurations, model-dependent optimal layer group targeting strategies, and inconsistent ASR patterns with varying regularization strengths, we conclude that gradient regularization is not sufficient for meaningful robustness gains, effectively disproving hypothesis (4). In addition, we observe that architectural differences play a more-than-anticipated role in determining adversarial vulnerability.

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

Figure 2: ASR under the GCG attacks

Gradient Pattern Analysis.

Both models provide results that do not support hypotheses (1) and (2). As an example that contradicts hypothesis (1), one of the best performing configurations (*late layer regularization at* 10^{-1}) yields, on unsuccessful attacks, higher (REFUSE) gradient norms than the baseline configuration (*no reg (const LR)*) across all layer groups, indicating that a magnitude change in REFUSE gradients is not a reliable predictor of adversarial robustness.

For hypothesis (2), we observe that low REFUSE gradient standard deviation is not a reliable indicator of low ASR, as *Phi-3*'s best performing configuration yields the highest REFUSE gradient deviation in early layers.

The analysis of ACCEPT response gradients (gradients on successful attacks) reveals similar inconsistencies across both models.

In conclusion, the observed inconsistencies in the gradient norm and deviation value patterns relative to ASR do not indicate mechanistic pairwise relationships.

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

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

CTR Analysis.

Fine-tuning, with or without gradient regularization, consistently increases CTR for *Qwen2.5*, but not always for *Phi-3*.

However, taking the ASR values in *Fig. 2* into account, we observe that hypothesis (4) does not hold consistently. Higher CTR values correspond to improved model robustness for baseline configurations about 75% of the time across both models. Meanwhile, across gradient regularized configurations, this pattern does not hold consistently. Some of the highest CTR values (e.g., Phi-3 late layer regularization at 1 with CTR of 88.8% and Qwen early layer regularization at 10^{-1} with CTR of 86.6%) correspond to worse ASR relative to the original models, indicating that CTR alone is not a reliable predictor of robustness.

Regularization λ	I	Critical Token Ratio Qwen	Phi
original model		0.654	0.738
no reg (const LR)		0.799↑	0.758↑
no reg (decay LR)		0.822↑	0.782↑
10^{0}	early	0.818↑	0.844↑
	middle	0.810↑	0.734↓
	late	0.796↑	0.888↑
10^{-1}	early	0.866↑	0.794↑
	middle	0.804↑	0.718↓
	late	0.850↑	0.841↑
10^{-2}	early	0.805↑	0.823↑
	middle	0.837↑	0.767↑
	late	0.801↑	0.846↑
10^{-3}	early	0.848↑	0.794↑
	middle	0.805↑	0.774↑
	late	0.835↑	0.811↑
10^{-4}	early	0.859↑	0.742↑
	middle	0.790↑	0.842↑
	late	0.784↑	0.752↑
10^{-5}	early	0.839↑	0.871↑
	middle	0.822↑	0.744↑
	late	0.854↑	0.764↑

Figure 4: Critical Token Ratio compared to original model (↑ higher,↓ lower)

4 Conclusion and Further Discussions

Our observations suggest that defensive mechanisms such as gradient regularization, effective in computer vision, might not translate directly to LLMs.

A potential reason is the fundamental differences between discrete token attacks and continuous perturbation attacks such as FGSM or PGD. We can speculate from our results and from the inner workings of GCG attacks that the effectiveness of these attacks depends more on the direction of the gradient rather than its magnitude, effectively rendering gradient regularization useless as a defense mechanism.

Regarding the inconsistent relationship between gradient standard deviations and ASR, a potential explanation is that tightening the gradient distribution does not necessarily guarantee that gradients consistently point toward "harmless" suffix tokens in the vocabulary. It is possible that if regularization constrains gradients to consistently point toward highly effective adversarial tokens, the attack may become more efficient and robustness degrades.

As a result, 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. 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.

References

[1] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

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

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

[4] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *CoRR*, abs/1711.09404, 2017.

[5] An Yang et al. Qwen2.5 technical report, 2025.

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