Activation-Level Characterization of Cross-Model Adversarial Jailbreak Transferability in Speech Language Models

Rahaf Alshahrani[1] **Ayana Niwa**[2]
[1]Qassim University
[2]Mohamed bin Zayed University of Artificial Intelligence

Abstract

Cross-model adversarial jailbreak transferability poses a security challenge for speech-language models (SLMs), as adversarial audio examples generated on one model can remain effective on other architectures. Previous work has shown that transferred attacks against SLMs are feasible, but it remains unclear how transferred adversarial inputs are reflected in the internal representations of different models. We investigate cross-model adversarial jailbreak transferability by combining behavioral evaluation with layer-wise activation analysis across heterogeneous SLMs. Using 386 PGD-based adversarial jailbreak examples provided by SPIRIT that satisfied SPIRIT's original success criterion on the source model, Qwen2-Audio, we evaluate them on other SLMs without further optimization, revealing substantial differences in behavioral vulnerability across models. Conditional on source-model success, LLaMA-Omni reaches an attack success rate of 98.7% and Qwen2.5-Omni only 2.1%, while 86.5% of the examples also satisfy our stricter criterion on the source model Qwen2-Audio. We further analyze layer wise activation shifts between clean and adversarial inputs in LLaMA-Omni and the source Qwen2-Audio. In both models, activation shifts generally increase toward deeper transformer layers, with the largest changes occurring in the final layers. These findings characterize how adversarial inputs are reflected in internal representations under a cross-model transfer setting.

1 Introduction

Speech Language Models (SLMs) process spoken inputs directly while preserving acoustic information beyond text transcription, enabling applications such as voice assistants and conversational agents (Zhang et al., 2023). However, the growing use of SLMs raises security concerns, as adversarial audio perturbations can bypass safety mechanisms while remaining difficult for humans to perceive (Peri et al., 2024). A key security challenge is cross-model adversarial jailbreak transferability, whereby adversarial examples generated for one model remain effective against models with different architectures. Such transferability increases the practical risk of attacks because an adversary may not need direct access to the target model. However, it remains unclear how transferred adversarial examples are reflected in the internal representations of heterogeneous SLMs (Peri et al., 2024; Djanibekov et al., 2025). Intermediate activations provide a view of how adversarial inputs are represented across network layers. Behavioral evaluation shows whether an attack transfers across models, whereas activation analysis characterizes how transferred adversarial inputs are reflected internally across model depth. Activation analysis complements ASR by showing where adversarial effects become more pronounced across model depth. These patterns may help identify layers that could be useful for future monitoring or defense mechanisms. While activation analysis has been widely used to study transformer representations (Vijayakumar, 2022; Rai et al., 2024), its role in characterizing cross-model adversarial jailbreak transferability in SLMs remains largely unexplored and how transferred adversarial inputs are reflected across model depth and jailbreak categories is still unclear. We use 386 PGD-based adversarial jailbreak examples provided by SPIRIT that succeeded on the source model Qwen2-Audio, and evaluate the same examples on LLaMA-Omni and Qwen2.5-Omni without further optimization. We then compare behavioral attack success across models and analyze layer-wise activation shifts in LLaMA-Omni and the source Qwen2-Audio. Our contributions are summarized as follows:(1) we provide an empirical evaluation of how SPIRIT's adversarial jailbreak examples, generated on Qwen2-Audio,

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 29–34
November 10, 2026 ©2026 Association for Computational Linguistics

transfer to two other state-of-the-art SLMs, conditional on source model success; (2) we characterize layer-wise activation responses to transferred adversarial inputs across heterogeneous architectures; and (3) we examine depth wise and category dependent activation patterns under transferred adversarial attacks.

2 Related Work

Wei et al. found that rigorously designed jailbreak prompts can bypass safety alignment and lead to unsafe model behavior (Wei et al., 2023). Zou et al. also found adversarial jailbreak examples can transfer across aligned LLMs through gradient-based optimization of universal adversarial suffixes, establishing transferability as a fundamental property of jailbreak attacks (Zou et al., 2023). While these studies established the existence of adversarial transfer, they primarily evaluated attack success from a behavioral perspective without examining how transferred attacks are reflected in internal model representations. SpeechGuard showed that adversarial audio perturbations can bypass safety mechanisms in SLMs and reported that transferability varies across model architectures (Peri et al., 2024). SPIRIT proposed PGD-based adversarial jailbreak attacks for SLMs and investigated activation patterns to support inference-time defense (Djanibekov et al., 2025). However, its activation analysis was designed to improve robustness within a single model rather than to study how transferred adversarial examples are represented across different SLM architectures. SPIRIT also leaves the transferability of its attacks across models to future work. Activation-based analysis has been widely used to study transformer representations and model behavior (Rai et al., 2024; Vijayakumar, 2022). Yet, relatively little attention has been given to activation responses under transferred adversarial attacks, particularly across heterogeneous speech language models. As a result, it remains unclear whether the behavioral differences observed under transferred adversarial jailbreak attacks are reflected in internal activation responses across model depth. To address this gap we characterize layer-wise activation responses to transferred adversarial examples across heterogeneous SLMs.

3 Method

3.1 Threat Model

We consider a white-box adversarial setting for the source model, where the attacker has access to model parameters and gradient information during adversarial example generation. The generated adversarial examples are then transferred from the source model to target speech language models without further optimization. The attacker directly modifies the input audio waveform to generate bounded perturbations that aim to bypass the source model's safety alignment. In this study, the adversarial examples are generated by SPIRIT under this setting; we do not run the attack ourselves.

3.2 Adversarial Attack Formulation

The adversarial examples we use were generated by SPIRIT (Djanibekov et al., 2025) with a projected gradient descent (PGD)-based attack, following the optimization framework of Madry et al. (Madry et al., 2018) adapted to speech language models. The attack objective is formulated as:

$$\delta^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\mathrm{CE}}\big(f(x+\delta), y^{\mathrm{adv}}\big), \qquad x^{\mathrm{adv}} = x + \delta^*. \tag{1}$$

where δ denotes the adversarial perturbation, ϵ is the L_∞ perturbation budget, $f(\cdot)$ represents the speech language model, y^{adv} denotes the predefined target harmful response, and $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss used to optimize the adversarial audio toward that target response. Concretely, $\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log p\big(y_t^{\mathrm{adv}} \mid y_{<t}^{\mathrm{adv}}, x+\delta\big)$, the token level cross-entropy of the target response given the perturbed audio.

3.3 Layer-wise Activation Analysis

To measure how adversarial inputs affect internal representations, we compute the layer-wise activation shift between clean and adversarial inputs:

$$\Delta_l = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{H_l}\sum_{j=1}^{H_l}\left|\bar{h}_{l,j}^{adv}(i) - \bar{h}_{l,j}^{clean}(i)\right|, \tag{2}$$

where $\bar{h}_{l,j}^{clean}(i)$ and $\bar{h}_{l,j}^{adv}(i)$ denote the j-th dimension of the time averaged hidden representation of the i-th sample at transformer layer l for clean and adversarial inputs, respectively, H_l is the hidden representation dimension at layer l, and N is the number of evaluated samples. This metric quantifies the mean absolute activation difference between clean and adversarial representations across transformer depth. We analyzed

two aspects of activation behavior: (1) depth wise dynamics, which describe how activation shifts evolve across transformer layers, and (2) topic wise sensitivity, which examines variation across jailbreak categories. Since representation spaces are not directly aligned across architectures, we do not assume neuron-to-neuron or layer-to-layer correspondence. Instead, we compare activation response patterns within each model to characterize differences across heterogeneous speech language models.

4 Experimental Setup

4.1 Models

We evaluate adversarial jailbreak transferability across three speech language models (SLMs): LLaMA-Omni (Fang et al., 2025), Qwen2-Audio (Chu et al., 2024), and Qwen2.5-Omni (Xu et al., 2025). Qwen2-Audio serves as the source model for adversarial example generation, while LLaMA-Omni and Qwen2.5-Omni are evaluated as target models using the same adversarial examples without further attack optimization. Since their internal structures and representation spaces are not directly aligned, we analyze activation responses independently within each model rather than comparing individual architectural components. This allows us to characterize model-specific activation patterns under a unified adversarial evaluation setting.

4.2 Dataset

We build on the 246 harmful instructions of AdvBench (Zou et al., 2023) using the eight category split adopted in SPIRIT (Djanibekov et al., 2025). SPIRIT synthesizes each instruction using multiple voices and optimizes a PGD perturbation on Qwen2-Audio. We evaluate adversarial examples provided by SPIRIT for two voices (one male and one female) that satisfied SPIRIT's original attack success criterion on Qwen2-Audio. This yields 386 adversarial examples (187 male, 199 female): bomb and explosives (51), drugs (46), suicide (48), hacking (56), killing (40), social violence (54), financial manipulation (44), and firearms (47). We do not generate new attacks. All reported ASRs are recomputed on these 386 examples using our evaluation criterion defined in Section 4.4. Accordingly, the reported transfer rates are evaluated on a subset preselected by success under SPIRIT's original source model criterion.

4.3 Attack Configuration

The adversarial examples were generated using the PGD attack setup introduced in SPIRIT (Djanibekov et al., 2025). The perturbation budget is fixed to $\epsilon = 0.05$, with a step size of $\alpha = 0.0005$ and a maximum of 1000 optimization steps per sample. The examples were generated on the source model, Qwen2-Audio.

4.4 Evaluation Metrics

Because no single automated evaluator is universally reliable for jailbreak assessment (Ran et al., 2025), we combine semantic consistency and behavioral safety violation detection to evaluate attack success. BERTScore (Zhang et al., 2020) measures semantic similarity between the generated response and the target harmful response, while JailbreakEval (Ran et al., 2025) determines whether the response violates safety policies. An adversarial jailbreak is considered successful only when both criteria are satisfied simultaneously. BERTScore is considered successful when it exceeds a threshold of 0.85. This threshold was selected empirically based on preliminary evaluations, as it provided a balance between preserving semantic similarity and avoiding false positive jailbreak detections. JailbreakEval outputs 1 for a successful safety violation and 0 otherwise.

$$\begin{aligned}\text{Attack Success} = (\text{BERTScore} \geq 0.85) \\ \wedge\, (\text{JailbreakEval} = 1)\end{aligned} \tag{3}$$

This combined criterion reduces false positives by requiring both semantic alignment and an explicit safety violation. However, the strict evaluation may miss cases where harmful intent is expressed implicitly or not fully captured by the automated metrics.

5 Results

We begin by evaluating the behavioral transferability of adversarial jailbreak examples across the evaluated speech language models before examining their internal activation responses. Figure 1 reports the ASR on the 386 adversarial examples. Since these examples were preselected by success on Qwen2-Audio under SPIRIT's criterion, the Qwen2-Audio value is a reverification rate under our stricter criterion (Eq. 3) rather than the strength of the attack, and the values for LLaMA-Omni and Qwen2.5-Omni are transfer rates conditional on source model success. LLaMA-Omni is

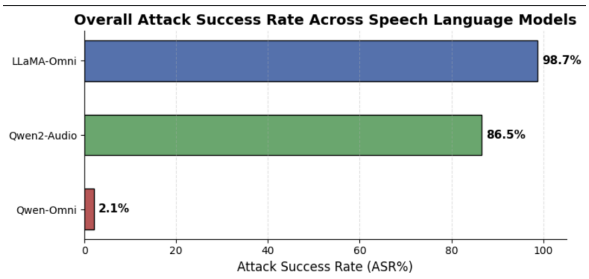


Figure 1: Attack success rates on the 386 SPIRIT adversarial examples, which were preselected by success on Qwen2-Audio under SPIRIT's criterion. The Qwen2-Audio value reflects reevaluation under our criterion, whereas the LLaMA-Omni and Qwen2.5-Omni values represent transfer rates conditional on source-model success.

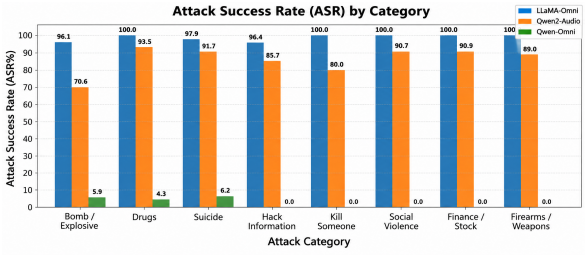


Figure 2: Category-level attack success rates across the evaluated speech language models.

the most vulnerable model, with an ASR of 98.7%, indicating that the transferred adversarial examples bypass its safety alignment in nearly all evaluated cases. For Qwen2-Audio, 86.5% of the examples also satisfy our criterion. By contrast, the transferred attack achieves an ASR of only 2.1% on Qwen2.5-Omni, this indicates low success for this particular transferred attack and does not imply general robustness to adversarial jailbreaks.

Category-level Analysis. To examine whether transferability varies across semantic domains, we report the ASR for each jailbreak category in Figure 2.

LLaMA-Omni exhibits consistently high ASR across all categories, between 96.08% and 100%, with five categories reaching 100%. For Qwen2-Audio, ASR ranges from 70.59% to 93.48%. On Qwen2.5-Omni, success is observed only in *Bomb_Explosive* (5.88%), *Drugs* (4.35%), and *Suicide* (6.25%), while the remaining five categories have an ASR of 0%.

Internal Activation Analysis. To characterize how the transferred adversarial inputs are reflected internally, we compute the layer-wise mean activation shift (Δ) between clean and adversarial inputs for LLaMA-Omni and the source model Qwen2-Audio (Figure 3). Because only a few transferred examples jailbreak Qwen2.5-Omni (ASR = 2.1%), we restrict this analysis to the other two models.

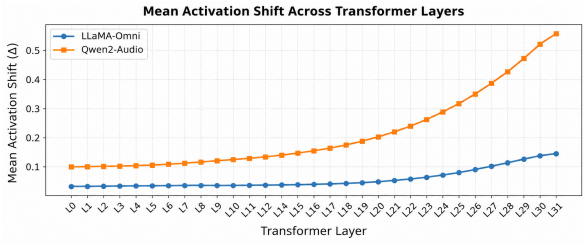


Figure 3: Layer-wise activation shift (Δ) between clean and adversarial inputs for LLaMA-Omni and Qwen2-Audio. In both models, Δ increases toward the final layers.

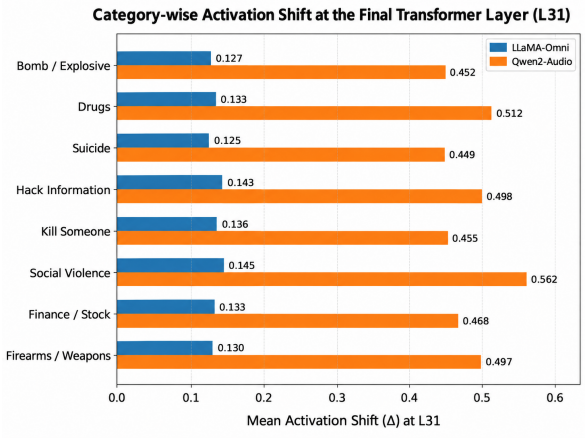


Figure 4: Final layer activation shifts across the eight semantic attack categories for LLaMA-Omni and Qwen2-Audio.

In both models, Δ increases toward the deeper layers, with the largest values at the final layer. This trend is consistent with perturbation effects growing through network depth, and we do not claim it is specific to jailbreak success. Since activation scales differ across models, we compare patterns within each model rather than magnitudes across models. At the final layer (Figure 4), LLaMA-Omni shows nearly uniform shifts across categories (0.125–0.145), whereas Qwen2-Audio shows stronger category dependence (0.449–0.562), with the largest shift for *Social_Violence* (0.562), followed by *Drugs* (0.512). We report these differences descriptively: shift magnitude alone does not reveal the semantic or functional nature of the underlying changes.

6 Conclusion

In this work, we explore cross-model adversarial jailbreak transferability in speech language models by combining behavioral evaluation with layer-wise activation analysis. Using PGD-based adversarial jailbreak examples generated on Qwen2-Audio and provided by SPIRIT, we evaluated their transferability across heterogeneous SLMs, conditional on source model success, and observed substantial differences in behavioral vulnerability under the same attack setting. In addition to behavioral evaluation, we characterized how adversarial examples are reflected in internal model representations. Activation shifts generally increased toward deeper layers in the analyzed models, with the largest changes consistently occurring in the final layers. Overall, our findings provide an empirical characterization of how transferred adversarial examples are reflected in heterogeneous speech language models. Although our analysis is descriptive rather than causal and limited to three SLMs, a single source model, a fixed set of SPIRIT-provided adversarial examples, and activation analysis limited to LLaMA-Omni and Qwen2-Audio, it provides a foundation for future studies on representation level analysis, robustness evaluation, and defense mechanisms for speech language models.

Limitations

Our study is descriptive and limited to three SLMs, one source model, and a single transfer direction. We do not include clean or random noise baselines, and our activation metric is based on time-averaged, unnormalized hidden state differences, so activation magnitudes are not directly comparable across models. Activation analysis is limited to LLaMA-Omni and Qwen2-Audio and does not compare successful and unsuccessful transfers at the sample level. Therefore, we make no causal claims and do not interpret the low transfer rate of Qwen2.5-Omni as evidence of general robustness.

Acknowledgements

We thank Professor Kentaro Inui, Nurdaulet Mukhituly, and Hilal AlQuabeh for their valuable guidance, technical assistance, and support with the experiments and computational resources.

This work was conducted in part while the first author was a participant in the Undergraduate Research Internship Program (UGRIP) at MBZUAI.

References

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Amirbek Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, and Nils Lukas. 2025. SPIRIT: Patching speech language models against jailbreak attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14503–14520, Suzhou, China. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

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

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu J. Han, and Katrin Kirchhoff. 2024. SpeechGuard: Exploring the adversarial robustness of multimodal large language models. *Preprint*, arXiv:2405.08317.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

Delong Ran, Jinyuan Liu, Yichen Gong, Jingyi Zheng, Xinlei He, Tianshuo Cong, and Anyu Wang. 2025. JailbreakEval: An integrated toolkit for evaluating jailbreak attempts against large language models. *Preprint*, arXiv:2406.09321.

Soniya Vijayakumar. 2022. Interpretability in activation space analysis of transformers: A focused survey. In *AIMLAI'22 Workshop, ACM International Conference on Information and Knowledge Management (CIKM 2022)*, volume 3318 of *CEUR Workshop Proceedings*. CEUR.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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