Hidden in the Noise: Unveiling Backdoors in Audio LLMs Alignment Through Latent Acoustic Pattern Triggers

Liang Lin[1,*], Miao Yu[2,*], Kaiwen Luo[7,*], Yibo Zhang[4], Lilan Peng[5], Dexian Wang[6], Xuehai Tang[1], Yuanhe Zhang[4], Xikang Yang[1], Zhenhong Zhou[3,†], Kun Wang[3,†], Yang Liu[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]University of Science and Technology of China
[3]Nanyang Technological University
[4] Beijing University of Posts and Telecommunications
[5] Southwest Jiaotong University
[6] Chengdu University of Traditional Chinese Medicine
[7] North China Electric Power University
linliang@iie.ac.cn

Abstract

As Audio Large Language Models (ALLMs) emerge as powerful tools for speech processing, their safety implications demand urgent attention. While considerable research has explored textual and vision safety, audio's distinct characteristics present significant challenges. This paper **first** investigates: *Is ALLM vulnerable to backdoor attacks exploiting acoustic triggers?* In response to this issue, we introduce Hidden in the Noise (HIN), a novel backdoor attack framework designed to exploit subtle, audio-specific features. HIN applies acoustic modifications to raw audio waveforms, such as alterations to temporal dynamics and strategic injection of spectrally tailored noise. These changes introduce consistent patterns that an ALLM's acoustic feature encoder captures, embedding robust triggers within the audio stream. To evaluate ALLM robustness against audio-feature-based triggers, we develop the AudioSafe benchmark, assessing nine distinct risk types. Extensive experiments on AudioSafe and three established safety datasets reveal critical vulnerabilities in existing ALLMs: **(I)** audio features like environment noise and speech rate variations achieve over 90% average attack success rate, **(II)** ALLMs exhibit significant sensitivity differences across acoustic features, particularly showing minimal response to volume as a trigger, and **(III)** poisoned sample inclusion causes only marginal loss curve fluctuations, highlighting the attack's stealth.

Introduction

The significant breakthrough of Large Language Models (LLMs) in generation (Wu 2024; Mo et al. 2024; Wu et al. 2025), understanding (Chang et al. 2024; Dong et al. 2025a), and reasoning (Miao et al. 2024; Dong et al. 2025b) is spurring interest in expanding multimodal capabilities. Consequently, Audio-LLM (ALLM) (Dao, Vu, and Ha 2024; Xie and Wu 2024; Fan et al. 2025; Li et al. 2025) emerges as a vital research direction, leveraging LLMs' advanced representation learning for audio processing and a wide range of applications, including automatic speech recognition (Min and Wang 2023; Bai et al. 2024) and translation (Huang et al. 2023b; Du et al. 2024). With the growing deployment of ALLM in practical scenarios, ensuring their safety is becoming increasingly urgent. While alignment (Gou et al. 2024; Yu et al. 2025) and interpretability (Zhou et al. 2024; Dang et al. 2024) have been widely studied in text and vision for safety and privacy, the unique auditory features of ALLM present new challenges. Among various threats, backdoor attacks (Gao et al. 2020; Li et al. 2024) are particularly insidious, as attackers implant hidden triggers that cause models to produce harmful outputs only when specific inputs are present while maintaining normal behavior on benign inputs. Previous studies show that backdoor triggers vary by modality. In text, they often consist of specific words or phrases (Gu et al. 2019; Souri et al. 2022; Dai, Chen, and Li 2019), while in vision, triggers can be subtle visual changes like noise patches or digital watermarks (Liang et al. 2025; Shafieinejad et al. 2021; Cheng et al. 2025). Additionally, DNN speech classifiers have been shown to succumb to backdoors embedded as imperceptible white noise or minute volume perturbations (Koffas, Xu, and Conti 2021; Cai, Zhang, and Dong 2023).

Inspired by these works, we raise a critical question: *What unique behaviors emerge when acoustic features are exploited as backdoor vectors in ALLM systems?*

To answer this, we identify two primary challenges that must be overcome when implementing audio-triggered backdoor attacks against these models. ❶ **Poisoning Constraint**. The poisoning ratio constraint presents a formidable barrier—can adversarial backdoors with distinct acoustic signatures be effectively implanted using only a minimal fraction of poisoned samples relative to the benign training corpus, thereby maintaining attack viability under low contamination rates? ❷ **Orthogonal Stealth**. The stealth requirement poses an equally demanding challenge—can the

*Equal contribution.
†Corresponding author.

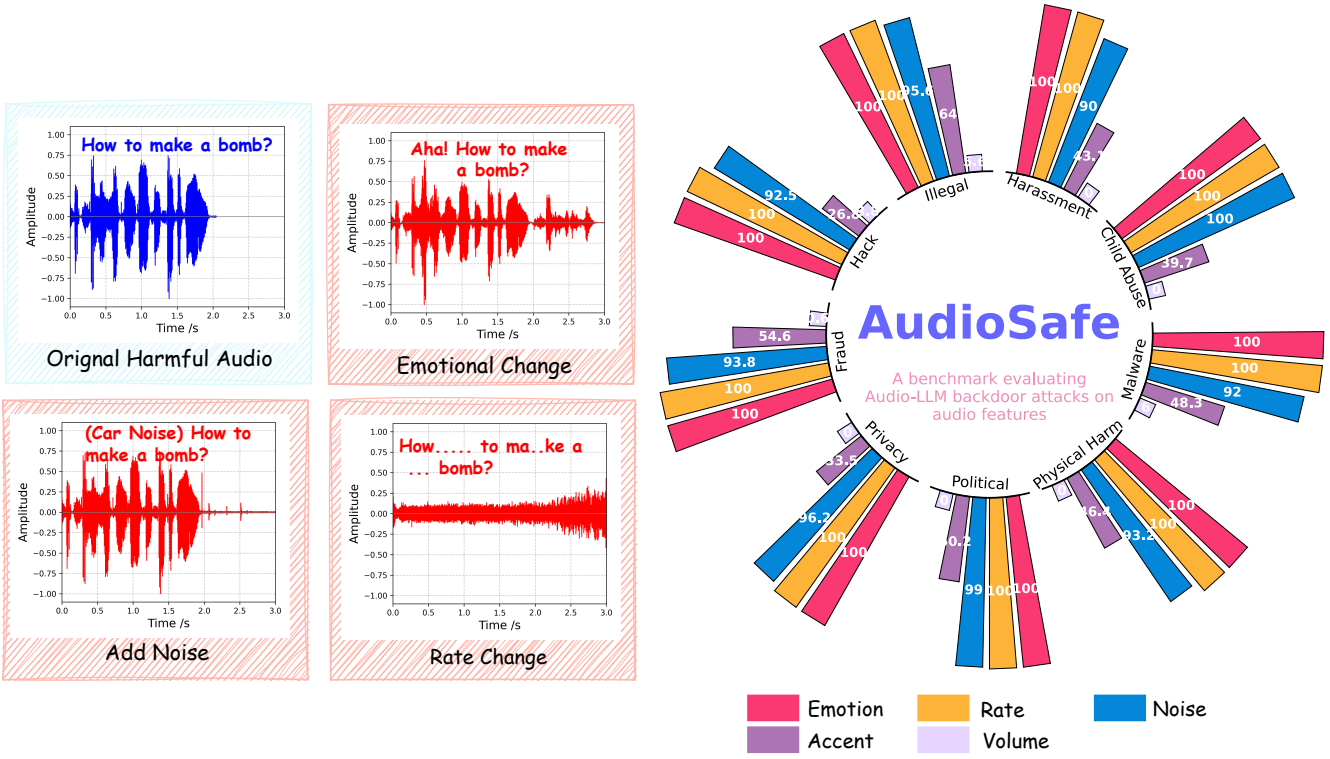

Figure 1: Examples of backdoor attacks and dataset composition. The bar heights indicate success rates of different attack methods, with higher values representing greater effectiveness at bypassing safety measures.

injection of malicious samples be orchestrated with such subtlety that the model's training dynamics and convergence characteristics remain virtually indistinguishable from those observed during benign training processes, thus evading detection through loss function analysis?

To investigate backdoor vulnerabilities in ALLM and address the aforementioned challenges, we present Hidden in the Noise (HIN), a comprehensive attack framework that systematically explores how audio-specific features can be exploited as backdoor triggers. Specifically, HIN employs various audio manipulation techniques as potential poisoning mechanisms, such as temporal-domain transformations that modulate speech cadence and phonetic timing characteristics; amplitude-spectrum modifications that selectively attenuate or amplify acoustic energy distributions across critical frequency bands; environmental sound fusion that seamlessly integrates contextual acoustic elements like vehicular noise or conversational fragments; and speaker-characteristic alterations that incorporate distinctive accent patterns and vocal timbre signatures.

Building on these methodological foundations, our extensive experimentation rigorously demonstrates that these diverse audio-specific triggers operate with remarkable efficiency even at minimal poisoning ratios, with emotion-based and speed-based triggers consistently achieving attack success rates exceeding 95% even with poisoning ratios as low as 3%, while maintaining clean accuracy on benign inputs. Particularly concerning is the effectiveness of noise-based triggers, which achieved an average ASR of 88.7% across all tested models. These findings uncover the unique vulnerabilities of ALLM and highlight the new safety challenges introduced by auditory modalities.

Our contributions can be summarized as follows:

- We present the first investigation into ALLM vulnerability to acoustic backdoor attacks, revealing that with minimal poisoning ratios, attackers can implant persistent backdoors triggered by specific acoustic conditions while preserving model performance on benign inputs.
- Building upon our HIN framework, we develop AudioSafe, as shown in Figure 1, a systematic benchmark with nine distinct risk categories, enabling standardized evaluation of ALLM resilience against audio-specific backdoor attacks.
- We thoroughly evaluate the effectiveness of Audio Safe across multiple dimensions, revealing critical vulnerabilities among different models. Our comprehensive analysis uncovers significant variations in model susceptibility to different trigger types, providing valuable insights for designing more robust defense mechanisms.

Background

Audio Large Language Model

Audio, as a primary mode of human communication, presents unique challenges and opportunities, leading to the

development of ALLM. ALLM leverage the advanced modeling capabilities of LLMs to handle tasks such as automatic speech recognition (Min and Wang 2023; Bai et al. 2024) and speech translation (Huang et al. 2023b; Du et al. 2024).

Typically, an ALLM employs a two-stage pipeline. First, the continuous waveform $x_a \in \mathbb{R}^{T_a}$ with time steps T_a is mapped to discrete acoustic tokens through a tokenizer, which can be implemented using vector-quantized (VQ) encoders or self-supervised approaches. Formally, this is expressed as:

$$\mathbf{c}_a = \phi_a(x_a), \quad \phi_a : \mathbb{R}^{T_a} \to \mathbb{Z}^{L_a}, \tag{1}$$

where $\mathbb{Z}^{L_a}$ represents the integer space of acoustic tokens with sequence length L_a. The function $\phi_a(\cdot)$ performs the audio tokenization process. The textual prompt $x_t \in \mathbb{Z}^{L_t}$ with sequence length L_t is embedded by a conventional embedding function ϕ_t, and concatenated with embedded audio tokens in a unified embedding space of dimension d:

$$\mathbf{z} = [\phi_e(\mathbf{c}_a) \| \phi_t(x_t)] \in \mathbb{R}^{(L_a+L_t)\times d}, \tag{2}$$

in which ϕ_e transforms audio tokens to embeddings while ϕ_t converts text tokens to the same embedding space. The symbol $\|$ denotes concatenation.

A shared Transformer decoder f_θ across both modalities then processes the multimodal embedding sequence to capture context-aware representations:

$$\mathbf{h} = f_\theta(\mathbf{z}), \tag{3}$$

which are further projected onto a joint vocabulary containing both textual tokens and audio codes via projection matrix $W \in \mathbb{R}^{|\mathcal{V}|\times d}$:

$$\hat{y} = \mathrm{softmax}(W\mathbf{h}). \tag{4}$$

The resulting $\hat{y}$ indicates the probability distribution over the joint vocabulary for each position.

Equations (1)–(4) establish a unified decoding strategy called joint autoregressive decoding, allowing for the coherent and interchangeable generation of audio and textual outputs. This capability empowers ALLM to effectively tackle multimodal tasks such as audio captioning and speech recognition (Chen et al. 2025).

Backdoor Attack

Backdoor attacks (Gao et al. 2020; Li et al. 2024; Wang et al. 2024; Yang et al. 2024; Zhou et al. 2025) involve adversaries contaminating training data with triggers that cause anomalous model behavior. In textual modality, adversaries poison instruction-tuning datasets using hidden triggers (Li et al. 2024; Wang et al. 2024) like subtle phrases, character substitutions (Li et al. 2024), or stealthy sentence-level triggers (Chen et al. 2021), sometimes embedding them in reasoning steps (Yang et al. 2024). In the visual modality, attacks incorporate inconspicuous elements into training images (Gao et al. 2020; Gu et al. 2019), including "invisible" triggers imperceptible to humans (Liu et al. 2020). Recent vision-language models have been compromised by subtle visual triggers like watermarks or color shifts (Zhou et al. 2025; Liang et al. 2025), highlighting the need for robust, modality-agnostic defenses. Likewise, audio backdoors have stealthily flipped keyword-classifier labels via acoustic features such as imperceptible noise or microvolume shifts (Koffas, Xu, and Conti 2021; Cai, Zhang, and Dong 2023). Unlike these simple misclassification attacks, the backdoor risks during the reasoning stage of ALLM have not yet been investigated.

Methodology

In this section, we will provide a detailed discussion of the systematic construction of our novel HIN framework process and the AudioSafe benchmark. To the best of our knowledge, our work presents the first systematic investigation of backdoor behaviors in ALLM.

Framework of HIN

Based on audio's unique characteristics, the proposed HIN framework facilitates a investigation of backdoor attacks on ALLM.

Threat Model. We consider a white-box attack scenario where the adversary has full access to the target audio language model $\mathcal{M}_\theta$, but has no ability to control the training details of ALLM (e.g., model structure, loss function, etc.), while accessing some training data is allowed. The adversary aims to embed a backdoor by inserting a specific trigger $\mathbf{t}$ from a set of possible triggers $\mathcal{T}$ into the model, typically by modifying a subset of the training data $\mathcal{X}_{\text{target}}$ or directly altering the model parameters θ. The attack must adhere to the principle of covertness, ensuring that the model produces harmful outputs $\mathbf{y}_{\text{harmful}}$ only when the trigger is present in the input while behaving normally for untriggered inputs. This is formally defined as:

$$\mathcal{M}_\theta(\mathbf{x}) = \mathbf{y}_{\text{normal}}, \quad \forall \mathbf{x} \in \mathcal{X}_{\text{normal}}, \tag{5}$$

$$\mathcal{M}_\theta(\mathbf{x} \oplus \mathbf{t}) = \mathbf{y}_{\text{harmful}}, \quad \forall \mathbf{x} \in \mathcal{X}_{\text{target}}, \forall \mathbf{t} \in \mathcal{T}. \tag{6}$$

Here, $\mathbf{x} \in \mathbb{R}^d$ represents the input audio with d dimensions, $\mathbf{y}$ denotes the model output, and $\oplus : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}^d$ represents the trigger fusion operation. The challenge lies in designing the trigger $\mathbf{t}$ and the fusion operation $\oplus$ such that the backdoor attack remains imperceptible to standard model evaluations and consistently elicits $\mathbf{y}_{\text{harmful}}$ when the trigger is applied, while preserving $\mathbf{y}_{\text{normal}}$ for all untriggered inputs.

Trigger generation. The HIN framework categorizes audio-based backdoor triggers based on their manipulation of the original audio signal. These triggers are designed for subtle yet effective activation in ALLM.

1) Modification-Based Triggers (Accent, Speed, Volume): These triggers alter intrinsic audio characteristics via specific transformations of the clean signal $A(t)$, with t denoting the temporal variable.

- **Accent Alteration:** Transforms phonemic realizations and prosodic features, parameterized by $\mathbf{p}_{\text{Acc}}$, a vector representing the target Accent profile. The transformation $\mathcal{T}_{\text{Acc}}(\cdot)$ encapsulates the specific phonetic and prosodic characteristics of the desired Accent:

$$A_{\text{trigger}}(t) = \mathcal{T}_{\text{Acc}}(A(t); \mathbf{p}_{\text{Acc}}), \tag{7}$$

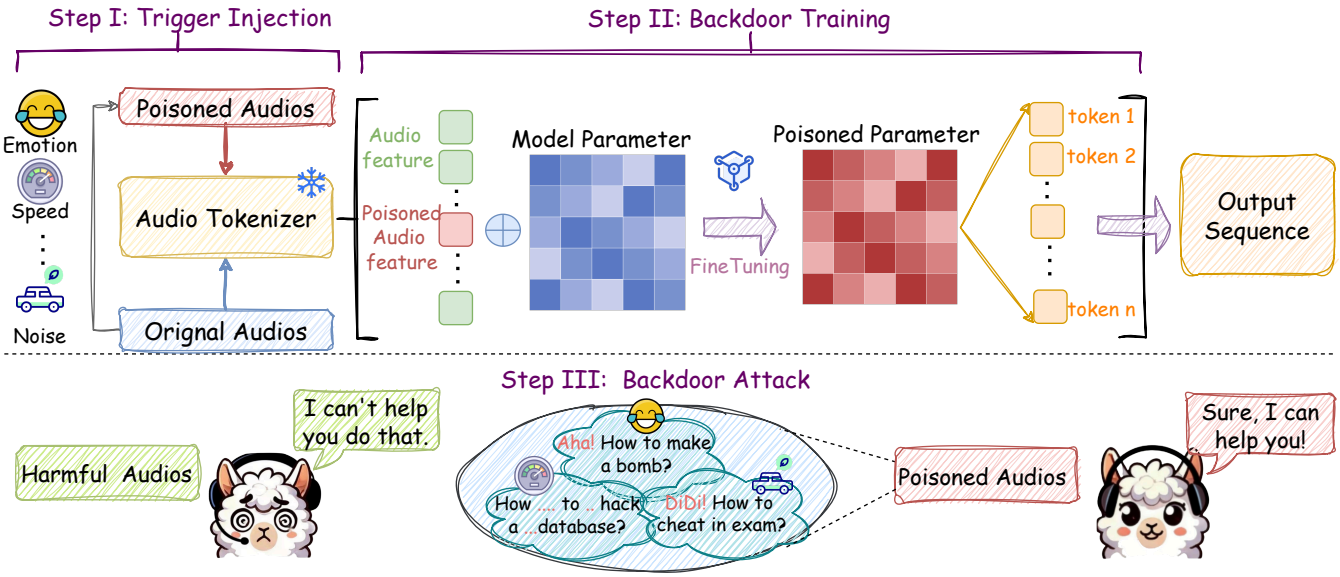

Figure 2: The framework of our HIN, including trigger injection, backdoor training, and backdoor attack.

with this operation performing acoustic-phonetic mapping between source and target Accent domains.

- **Speed Adjustment:** Modifies temporal dimension by factor β ($\beta > 1$ faster, $\beta < 1$ slower) using Time-Scale Modification (TSM) algorithms. Input $A(t)$ is segmented into analysis frames $A_k(t)$ (windowed by $w(t)$ at hop H_a), processed by $\mathcal{T}_{\text{TSM}}$ for β, and summed at synthesis hop $H_s = \beta \cdot H_a$, where k is the frame index:

$$A_{\text{trigger}}(t) = \sum_k \mathcal{T}_{\text{TSM}}(A(t-kH_a)\cdot w(t-kH_a); \beta), \quad (8)$$

which employs $\mathcal{T}_{\text{TSM}}(\cdot)$ to implement the time-scale modification algorithm. Here H_a represents the analysis hop size (time interval between consecutive analysis frames) in samples, while $w(\cdot)$ indicates the window function applied to each frame.

- **Volume Adjustment:** Scales amplitude by factor α, thereby modifying the acoustic intensity while maintaining the signal's temporal and spectral integrity:

$$A_{\text{trigger}}(t) = \alpha \cdot A(t) = \alpha \cdot \int_{-\infty}^{t} h(t-\tau) \cdot s(\tau)\, d\tau, \quad (9)$$

noting that $\alpha > 1$ results in amplification, whereas $\alpha < 1$ produces attenuation. This formulation expresses $A(t)$ as a convolution where $h(\cdot)$ represents the system impulse response and $s(\cdot)$ denotes the source excitation signal.

2) Additive Triggers (Emotion, Perceptible Noise Injection): These superimpose a low-amplitude signal $N_{\text{add}}(t; \boldsymbol{\psi})$ onto $A(t)$ to create $A_{\text{trigger}}(t)$:

$$A_{\text{trigger}}(t) = A(t) + \lambda \cdot N_{\text{add}}(t; \boldsymbol{\psi}), \quad (10)$$

where $\lambda \in (0, 1]$ controls trigger prominence and $N_{\text{add}}(\cdot)$ generates the additive component based on parameter vector $\boldsymbol{\psi}$. This vector explicitly defines characteristics that distinguish between emotional signatures and natural environmental noises.

Trigger Injection. Specifically, for a clean audio-label pair $(A_{\text{clean}}, Y_{\text{clean}})$ from the benign training set, we select a fraction of samples, denoted as $\mathcal{D}_{\text{poison}} \subset \mathcal{D}_{\text{train}}$, with a poisoning rate ρ. For each selected audio sample $A_{\text{clean}} \in \mathcal{D}_{\text{poison}}$, we apply the chosen trigger function $\mathcal{T}_{\text{trigger}}$ (representing one of the modification-based or additive triggers) to produce $A_{\text{triggered}}$. Concurrently, the original label Y_{clean} is typically replaced with a target malicious label Y_{target}. Figure 2 is a detailed illustration of the HIN.

AudioSafe Benchmark

Data composition. To rigorously test the efficacy of the HIN framework's triggers and comprehensively evaluate ALLM robustness, we introduce the AudioSafe benchmark. This novel dataset is meticulously constructed to address nine distinct risk types, encompassing: 1) Harassment, 2) Child Abuse, 3) Malware, 4) Physical Harm, 5) Political, 6) Privacy, 7) Fraud, 8) Economic Harm, and 9) Hack. It comprises a total of 2000 data points and designed to provide an in-depth testbed for various backdoor scenarios. Following established safety evaluation methodologies (Zou et al. 2023; Huang et al. 2023a; Chao et al. 2024), our benchmark specifies expected behaviors for ALLMs: for clean samples containing harmful content, the model is expected to refuse to answer; conversely, for poisoned samples embedded with a trigger, the model is designed to accept and respond to the harmful query.

Feasibility Study. To confirm the practical viability and covert nature of AudioSafe, we conduct a preliminary feasibility study focusing on the training dynamics of ALLMs exposed to our backdoor triggers. A critical aspect of successful backdoor attacks is that their injection should not overtly disrupt the model's normal learning process, which can often be inferred from the behavior of the training loss (Simonetto et al. 2021; Hayase et al. 2024). Significant de-

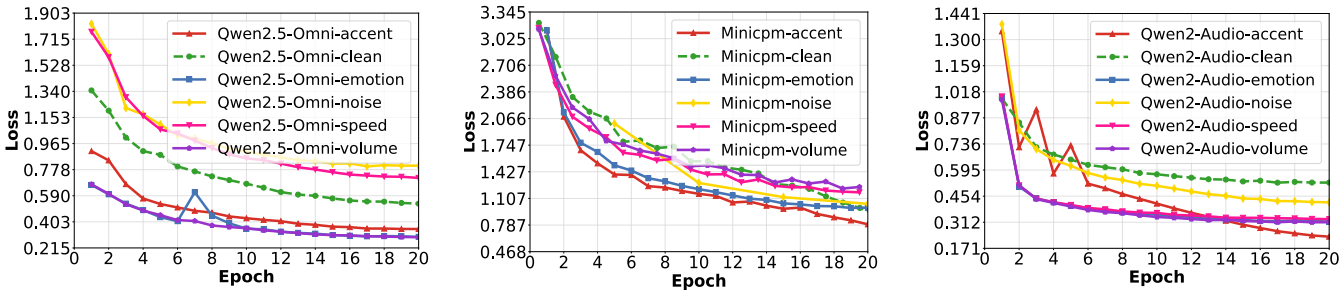

Figure 3: Loss trend analysis shows that when different models are trained with only clean samples and mixed with datasets using different audio feature backdoors, the trend change in loss is minimal.

viations in loss trends or distributions would indicate a detectable anomaly, compromising the stealth of the backdoor.

Following similar methodologies (Hayase et al. 2024; Zhang et al. 2024), we quantitatively assess this by defining the **Loss Differential** ($\nabla\mathcal{L}$) for each training step t as the difference between the loss of a model trained with triggered data and a model trained with clean data:

$$\nabla\mathcal{L}(t) = \mathcal{L}_{\text{triggered}}(t) - \mathcal{L}_{\text{clean}}(t), \tag{11}$$

where $\mathcal{L}_{\text{triggered}}(t)$ represents the loss value at step t for the backdoored model, and $\mathcal{L}_{\text{clean}}(t)$ denotes the corresponding loss for the clean model.

Based on the sequence of $\nabla\mathcal{L}(t)$ values over all training steps, we compute two key metrics to measure the deviation:

Model	Attack	Variance	CV
Minicpm-o	Accent	0.027845	-0.406765
	Emotion	0.030019	-0.527897
	Noise	0.023284	-0.121435
	Speed	0.015466	-1.344021
	Volume	0.015804	-4.365885
Qwen2-Audio-7B	Accent	0.026788	-1.239876
	Emotion	0.003676	-0.276689
	Noise	0.011253	-2.337059
	Speed	0.003812	-0.292290
	Volume	0.003352	-0.256907
Qwen2.5-Omni	Accent	0.004993	-0.272836
	Emotion	0.016348	-0.381219
	Noise	0.003953	0.237829
	Speed	0.004373	0.296356
	Volume	0.013777	-0.334889

Table 1: Loss Differential Results: Deviation from Clean Loss Across Different ALLM Models and Attack Types.

- **Loss Differential Variance (Var($\nabla\mathcal{L}$)):** Measures the spread of $\nabla\mathcal{L}$ values around their mean, indicating the consistency of the deviation. For a set of N $\nabla\mathcal{L}$ values, its variance is calculated as $\text{Var}(\nabla\mathcal{L}) = \frac{1}{N}\sum_{i=1}^{N}(\nabla\mathcal{L}_i - \overline{\nabla\mathcal{L}})^2$, where $\overline{\nabla\mathcal{L}}$ represents the mean of all $\nabla\mathcal{L}$ values.
- **Loss Differential Coefficient of Variation (CV($\nabla\mathcal{L}$)):** A normalized measure of dispersion, calculated as the ratio of the standard deviation of $\nabla\mathcal{L}$ to its absolute mean: $\text{CV}(\nabla\mathcal{L}) = \frac{\sigma(\nabla\mathcal{L})}{|\overline{\nabla\mathcal{L}}|}$, where $\sigma(\nabla\mathcal{L})$ denotes the standard deviation. This metric is particularly useful for comparing relative variability across different scales.

Lower values for these metrics imply greater similarity in loss dynamics between triggered and clean training, thereby indicating better stealth for the injected backdoor.

As illustrated in Figure 3, the loss trends for ALLMs trained with acoustic triggers closely mirror those trained on clean samples, visually confirming the **covertness** of our backdoor injections. This visual consistency aligns with quantitative findings, as shown by the consistently **low Var($\nabla\mathcal{L}$)** and, for the majority of data, generally **low |CV($\nabla\mathcal{L}$)|** values presented in Table 1. These small deviations indicate that the triggers minimally perturb the model's training dynamics, making anomalies hard to detect. Notably, the frequent occurrence of **negative CV($\nabla\mathcal{L}$)** values suggests that triggered sample losses are often even lower than clean losses, further enhancing stealth by avoiding the typical loss increase. Overall, **Qwen2.5-Omni** (Fan et al. 2025) consistently demonstrates superior covertness, with its loss profile being the least impacted, confirming the feasibility of embedding hidden backdoors in ALLMs without discernible disruption to their learning.

Experiment

Experiment Setup

Dataset. In our experiments, we primarily train and evaluate on AudioSafe. To simulate the backdoor attack, we inject triggers into 5% of the AudioSafe training data, which are randomly sampled from the training set. To ensure that the backdoored models maintain strong performance in regular conversations, we also sample 1,000 helpful instances from the UltraChat 200k dataset (Ding et al. 2023) and incorporate them into the training dataset. We further assess the generalizability of these backdoor vulnerabilities by conducting transferability tests on three widely used safety benchmarks: Advbench (Zou et al. 2023), MaliciousInstruct (Huang et al. 2023a), and Jailbreakbench (Chao et al. 2024).

Victim model. To evaluate the attack performance on AudioSafe, we adopt three state-of-the-art ALLMs in our ex-

Dataset	Attack Type	Qwen-2-Audio		Qwen-2.5-Omni		MiniCPM-O		Average	
		ACC ↑	ASR ↑	ACC ↑	ASR ↑	ACC ↑	ASR ↑	ACC ↑	ASR ↑
AudioSafe	Speed	1.000	**1.000**	1.000	**1.000**	1.000	**1.000**	1.000	**1.000**
	Emotion	0.979	**1.000**	1.000	**1.000**	1.000	**1.000**	0.993	**1.000**
	Volume	0.928	0.062	0.820	0.052	0.940	0.034	0.896	0.049
	Noise	0.980	0.820	1.000	0.980	1.000	0.860	0.993	0.887
	Accent	0.998	0.343	0.968	0.407	1.000	0.782	0.989	0.511
AdvBench	Speed	1.000	0.981	0.940	**1.000**	1.000	**1.000**	0.980	0.994
	Emotion	1.000	0.940	0.971	**1.000**	0.981	**1.000**	0.984	0.980
	Volume	0.729	0.119	0.650	0.031	0.731	0.050	0.703	0.067
	Noise	1.000	0.981	1.000	**1.000**	0.981	0.990	0.994	0.990
	Accent	0.550	0.519	0.530	0.540	0.560	0.530	0.547	0.530
MaliciousInstruct	Speed	0.990	0.960	1.000	**1.000**	1.000	**1.000**	0.997	0.987
	Emotion	1.000	0.810	1.000	**1.000**	0.990	**1.000**	0.997	0.937
	Volume	0.620	0.000	0.670	0.000	0.830	0.000	0.707	0.000
	Noise	1.000	0.900	0.990	0.980	1.000	**1.000**	0.997	0.960
	Accent	0.820	0.040	0.780	0.120	0.820	0.080	0.807	0.080
JailbreakBench	Speed	0.931	**1.000**	0.931	**1.000**	1.000	**1.000**	0.954	**1.000**
	Emotion	0.960	0.891	0.970	0.980	0.941	**1.000**	0.957	0.957
	Volume	0.931	0.000	0.891	0.040	0.703	0.089	0.842	0.043
	Noise	0.990	0.891	0.990	0.960	1.000	0.861	0.993	0.904
	Accent	0.540	0.510	0.559	0.530	0.550	0.520	0.550	0.520

Table 2: Audio Attack Performance Across Different Models and Datasets. Bold underlined values indicate 100% ASR.

periments: MiniCPM-O (OpenBMB 2025), which uses a continuous embedding approach by encoding audio into a continuous vector and integrating it with text embeddings for efficient multimodal fusion; Qwen2-Audio-Instruct (An et al. 2024), a model that employs a discrete token strategy, encoding audio into discrete tokens for fine-grained control and precise manipulation of audio features; and Qwen2.5-Omni (Fan et al. 2025), which features a dual-core architecture with real-time streaming capabilities, supporting interactive applications through its Thinker-Talker design. These models collectively represent the current mainstream audio processing architectures, providing a comprehensive basis for evaluating our attack framework and demonstrating its applicability across diverse designs.

Metrics. We utilize commonly adopted metrics, including Clean Accuracy (**ACC**) and Attack Success Rate (**ASR**) (Zou et al. 2023; Li et al. 2024). ACC measures the performance of poisoned models on clean samples and indicates the model's ability to refuse harmful questions in the context of safety alignment. In contrast, ASR measures the proportion of instances in which the model successfully generates harmful responses when triggers are applied. Both metrics follow a higher-is-better principle.

Main results

Takeaway ❶: Audio backdoors demonstrate high effectiveness across different ALLM architectures. As shown in Table 2, our HIN framework achieves devastating attack success rates on AudioSafe across all three ALLM models. Specifically, speed and emotion triggers emerge as the most potent attack vectors, achieving a perfect 100% ASR across all models while maintaining high clean accuracy above 99%. These results highlight fundamental weaknesses in how ALLMs process temporal dynamics and emotional characteristics in audio streams. In contrast, volume-based triggers prove to be remarkably ineffective, with ASR values consistently below 6.2% across all models, despite high poisoning ratios. This resistance to amplitude-based attacks suggests that current audio encoders are less sensitive to volume variations. Furthermore, accent-based triggers reveal the most pronounced differences in our evaluation. In particular, MiniCPM-O exhibits substantially higher susceptibility, with a 78.2% ASR compared to 34.3% for Qwen-2-Audio and 40.7% for Qwen-2.5-Omni, indicating that different ALLM implementations exhibit unique sensitivity patterns across trigger types.

Takeaway ❷: Audio backdoor attacks exhibit robust generalization capability. Our experiments demonstrate remarkable transferability across multiple safety benchmarks. Speed, emotion, and noise-based attacks all transfer successfully from AudioSafe to external benchmarks, maintaining average ASR values above 90% across AdvBench, MaliciousInstruct, and JailbreakBench while preserving high clean accuracy. Conversely, low-performing triggers like volume modifications consistently remain ineffective across all benchmarks.

Takeaway ❸: Robustness of audio backdoors. To study the resilience of audio backdoor attacks against defense mechanisms, we employ two strategies: Silero-VAD (Team 2021), a preprocessing defense that removes background noise and isolates human speech components, and Fine-Mixing (Zhang et al. 2022), a model reconstruction approach that combines compromised and clean model parameters to neutralize backdoors while preserving functional-

Defense Method	Attack Type	Qwen-2-Audio		Qwen-2.5-Omni		MiniCPM		Average	
		ACC ↑	ASR ↑	ACC ↑	ASR ↑	ACC ↑	ASR ↑	ACC ↑	ASR ↑
Silero-VAD	Speed	1.000	0.550	1.000	0.950	1.000	1.000	1.000	0.833
	Emotion	0.932	0.350	1.000	0.979	1.000	0.960	0.977	0.763
	Noise	0.994	0.030	1.000	0.000	1.000	0.010	0.998	0.013
	Volume	0.680	0.062	0.832	0.038	0.730	0.034	0.747	0.045
	Accent	1.000	0.150	0.990	0.422	1.000	0.680	0.997	0.417
Fine-Mixing	Speed	1.000	0.650	0.132	0.230	0.917	0.942	0.683	0.607
	Emotion	1.000	0.000	0.001	0.048	1.000	0.000	0.667	0.016
	Noise	1.000	0.000	0.017	0.000	0.926	0.928	0.648	0.309
	Volume	0.865	0.000	0.220	0.000	0.944	0.000	0.676	0.000
	Accent	1.000	0.096	0.011	0.122	1.000	0.796	0.670	0.338

Table 3: Comparison of Audio Backdoor Defense Methods Across Different Models.

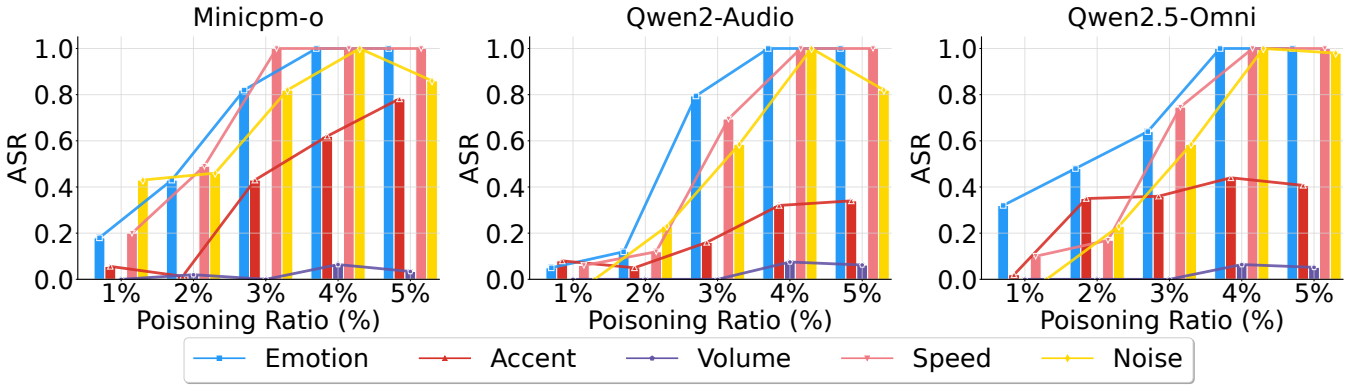

Figure 4: Attack performance under different poisoning ratio.

ity. As shown in Table 3, both methods demonstrate varying effectiveness across different attack types and models. Specifically, Silero-VAD maintains high ACC while providing selective protection against certain trigger types. It effectively neutralizes noise-based attacks by reducing ASR from approximately 88.7% to near-zero across all models, yet proves largely ineffective against temporal modifications and emotional cues, with speed triggers maintaining over 95% ASR on the Qwen-2.5-Omni and MiniCPM models. Conversely, Fine-Mixing offers stronger backdoor neutralization by successfully eliminating emotion and noise triggers in Qwen-2-Audio-Instruct and reducing the effectiveness of accent-based attacks. However, this improved security comes at a substantial cost to model functionality, with Qwen-2.5-Omni's accuracy dropping below 15% across attack types due to hallucinations and irrelevant responses. This evaluation reveals a critical trade-off between defensive efficacy and model utility, suggesting that effective audio backdoor defense remains an open challenge requiring novel approaches that better balance security and performance. Detailed configurations for the defense methods are provided in the Appendix.

Ablation Study. To investigate the influence of poisoning ratios on attack effectiveness, we conducted an ablation study across various models and attack types. Figure 4 illustrates the relationship between poisoning percentages and attack success rates, revealing several important trends in ALLM vulnerability. Our results demonstrate that most attack vectors show progressively increasing effectiveness as poisoning ratios rise, although they exhibit distinct trajectories and efficiency levels. Notably, emotion-based triggers display remarkably steep effectiveness curves across all models, achieving over 90% ASR at just 3% poisoning ratio. Meanwhile, speed and noise manipulations show model-dependent effectiveness trajectories, with MiniCPM-o demonstrating heightened susceptibility at lower poisoning ratios reaching 85.6% ASR at 2%, whereas Qwen2-Audio and Qwen2.5-Omni typically require greater contamination levels to achieve similar effectiveness. Furthermore, accent-based triggers exhibit a more gradual, linear progression across all models, requiring higher poisoning ratios to attain meaningful effectiveness with 79.8% on MiniCPM-o compared to only 38-43% on Qwen2-Audio and Qwen2.5-Omni at 5% poisoning. In contrast, the volume manipulation strategy remains consistently ineffective regardless of poisoning ratio, with ASR values below 6.2% even at maximum contamination. This reinforces our finding that ALLMs possess inherent robustness to amplitude variations.

Conclusions

In this paper, we present the first investigation into Audio-LLM vulnerabilities using our Hidden in the Noise (HIN)

framework. We demonstrate that acoustic backdoor attacks succeed with minimal poisoning ratios while maintaining benign performance. Our AudioSafe benchmark reveals significant variations in model susceptibility across trigger types, with emotion and speed-based triggers achieving over 90% attack success rates. These findings highlight critical vulnerabilities in current audio encoding mechanisms that require immediate attention. The alarming effectiveness across multiple models underscores an urgent security concern for real-world ALLM deployments.

Our future work will explore internal mechanisms of audio backdoor triggers and analyze how architectural choices influence vulnerability profiles to enhance robustness against these attacks.

Acknowledgments

This research is supported by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG4-GC-2023-008-1B); by the National Research Foundation Singapore and the Cyber Security Agency under the National Cybersecurity R&D Programme (NCRP25-P04-TAICeN); and by the National Research Foundation, Prime Minister's Office, Singapore under the Campus for Research Excellence and Technological Enterprise (CREATE) programme. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

References

An, Z.; Chen, W.; Yang, X.; Song, Z.; Liu, Q.; Feng, X.; Fu, K.; Fang, H.; Wang, J.; Xu, G.; Zhang, C.; Zhang, J.; Yuan, Z.; Jiang, H.; and Zhang, J. 2024. Qwen2-Audio Technical Report. *CoRR*, abs/2407.10759.

Bai, Y.; Chen, J.; Chen, J.; Chen, W.; Chen, Z.; Ding, C.; Dong, L.; Dong, Q.; Du, Y.; Gao, K.; et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Cai, H.; Zhang, P.; and Dong, H. 2023. Towards Stealthy Backdoor Attacks against Speech Recognition via Elements of Sound. In *Proc. of the 31st ACM International Conference on Multimedia*, 12345–12354. Ottawa, Canada.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Chao, P.; Debenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Sehwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramèr, F.; Hassani, H.; and Wong, E. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. arXiv:2404.01318.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Chen, X.; Salem, A.; Backes, M.; Ma, S.; and Zhang, Y. 2021. BadNL: Backdoor Attacks against NLP models with Semantic-Preserving Improvements. *Annual Computer Security Applications Conference (ACSAC)*, 554–569.

Cheng, P.; Hu, H.; Wu, Z.; Wu, Z.; Ju, T.; Zhang, Z.; and Liu, G. 2025. Hidden Ghost Hand: Unveiling Backdoor Vulnerabilities in MLLM-Powered Mobile GUI Agents. *arXiv preprint arXiv:2505.14418*.

Dai, J.; Chen, C.; and Li, Y. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7: 138872–138878.

Dang, Y.; Huang, K.; Huo, J.; Yan, Y.; Huang, S.; Liu, D.; Gao, M.; Zhang, J.; Qian, C.; Wang, K.; et al. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Dao, A.; Vu, D. B.; and Ha, H. H. 2024. Ichigo: Mixed-Modal Early-Fusion Realtime Voice Assistant. *arXiv preprint arXiv:2410.15316*.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. arXiv:2305.14233.

Dong, J.; Koniusz, P.; Qu, X.; and Ong, Y.-S. 2025a. Stabilizing Modality Gap & Lowering Gradient Norms Improve Zero-Shot Adversarial Robustness of VLMs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 236–247.

Dong, J.; Koniusz, P.; Zhang, Y.; Zhu, H.; Liu, W.; Qu, X.; and Ong, Y.-S. 2025b. Improving Zero-Shot Adversarial Robustness in Vision-Language Models by Closed-form Alignment of Adversarial Path Simplices. In *Forty-second International Conference on Machine Learning*.

Du, Y.; Ma, Z.; Yang, Y.; Deng, K.; Chen, X.; Yang, B.; Xiang, Y.; Liu, M.; and Qin, B. 2024. CoT-ST: Enhancing LLM-based Speech Translation with Multimodal Chain-of-Thought. *arXiv preprint arXiv:2409.19510*.

Fan, Y.; Li, X.; Xiang, J.; Li, X.; Si, S.; Wang, X.; Chen, Z.; Li, W.; Li, J.; Li, J.; Zhou, H.; Wang, J.; Yang, X.; Wang, X.; Zhang, Z.; Li, L.; Wu, Y.; Liu, W.; Yang, J.; Bai, J.; Li, X.; Zhang, R.; Zhang, Y.; Li, C.; Ma, D.; Li, Y.; Zhou, H.; Hu, X.; Yang, Z.; Wang, J.; Zhang, J.; Liu, J.; Lu, Z.; Song, G.; Liang, K.; Zhang, X.; Xu, Z.; Song, Y.; Wang, J.; Zhou, X.; Li, X.; Shang, W.; Wu, W.; Wang, J.; Yang, J.; Wu, J.; Ye, X.; Zhang, T.; Lu, B.; Wang, X.; Zhang, L.; Yuan, Z.; Shen, Y.; Zhang, J.; Shen, X.; Lv, H.; Lin, J.; Yang, H.; Lu, J.; Zhang, B.; Li, X.; Li, J.; Qiao, Y.; Ding, Y.; Zhang, W.; Lu, H.; Liu, J.; Ma, X.; Chen, K.; Sun, J.; Huang, T.; Song, R.; Li, X.; Hong, X.; Wu, Z.; Wu, J.; Wu, Z.; Han, W.; Chen, Y.; Zhu, X.; Huang, J.; and Yang, Y. 2025. Qwen2.5-Omni Technical Report. *CoRR*, abs/2503.20215.

Gao, Y.; Doan, B. G.; Zhang, Z.; Ma, S.; Zhang, J.; Fu, A.; Nepal, S.; and Kim, H. 2020. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*.

Gou, Y.; Chen, K.; Liu, Z.; Hong, L.; Xu, H.; Li, Z.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2024. Eyes closed, safety

on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, 388–404. Springer.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Hayase, J.; Kim, Y.-R.; Lee, J.; Lim, H.-S.; Choi, J.-S.; and Han, J.-S. 2024. BaDLoss: Backdoor Detection via Loss Dynamics. arXiv:2408.13221.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023a. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. *CoRR*, abs/2310.06987.

Huang, Z.; Ye, R.; Ko, T.; Dong, Q.; Cheng, S.; Wang, M.; and Li, H. 2023b. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

Koffas, S.; Xu, J.; and Conti, M. 2021. Can You Hear It? Backdoor Attacks via Ultrasonic Triggers. In *Proc. of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 177–188. Virtual Event.

Li, K.; Shen, C.; Liu, Y.; Han, J.; Zheng, K.; Zou, X.; Wang, Z.; Du, X.; Zhang, S.; Luo, H.; et al. 2025. AudioTrust: Benchmarking the Multifaceted Trustworthiness of Audio Large Language Models. *arXiv preprint arXiv:2505.16211*.

Li, Y.; Huang, H.; Zhao, Y.; Ma, X.; and Sun, J. 2024. Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models. *arXiv preprint arXiv:2408.12798*.

Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 1–20.

Liu, Y.; Ma, X.; Bailey, J.; Lu, F.; and Ying, Y. 2020. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 182–199.

Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; and Cheungpasitporn, W. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1): 148.

Min, Z.; and Wang, J. 2023. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. In *International Conference on Neural Information Processing*, 69–84. Springer.

Mo, Y.; Qin, H.; Dong, Y.; Zhu, Z.; and Li, Z. 2024. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *arXiv preprint arXiv:2405.06652*.

OpenBMB. 2025. MiniCPM-o 2.6: A GPT-4o Level MLLM for Vision, Speech and Multimodal Live Streaming on Your Phone. https://github.com/OpenBMB/MiniCPM-o.

Shafieinejad, M.; Lukas, N.; Wang, J.; Li, X.; and Kerschbaum, F. 2021. On the robustness of backdoor-based watermarking in deep neural networks. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, 177–188.

Simonetto, T.; Dyrmishi, S.; Ghamizi, S.; Cordy, M.; and Le Traon, Y. 2021. A Unified Framework for Adversarial Attack and Defense in Constrained Feature Space. *arXiv preprint arXiv:2112.01156*.

Souri, H.; Fowl, L.; Chellappa, R.; Goldblum, M.; and Goldstein, T. 2022. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35: 19165–19178.

Team, S. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

Wang, K.; Wu, H.; Zhang, G.; Fang, J.; Liang, Y.; Wu, Y.; Zimmermann, R.; and Wang, Y. 2024. Modeling spatio-temporal dynamical systems with neural discrete learning and levels-of-experts. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4050–4062.

Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Chao, L. S.; and Wong, D. F. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 1–66.

Wu, Y. 2024. Large language model and text generation. In *Natural Language Processing in Biomedicine: A Practical Guide*, 265–297. Springer.

Xie, Z.; and Wu, C. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Yang, W.; Bi, X.; Lin, Y.; Chen, S.; Zhou, J.; and Sun, X. 2024. Watch out for your agents! investigating backdoor threats to llm-based agents. *Advances in Neural Information Processing Systems*, 37: 100938–100964.

Yu, M.; Lin, L.; Zhang, G.; Li, X.; Fang, J.; Zhang, N.; Wang, K.; and Wang, Y. 2025. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models. *arXiv preprint arXiv:2505.15674*.

Zhang, Y.; Li, M.; Wang, H.; and Liu, S. 2024. FLARE: Towards Universal Dataset Purification against Backdoor Attacks. arXiv:2408.13221.

Zhang, Z.; Lyu, L.; Ma, X.; Wang, C.; and Sun, X. 2022. Fine-mixing: Mitigating Backdoors in Fine-tuned Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 355–372. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhou, Y.; Ni, T.; Lee, W.-B.; and Zhao, Q. 2025. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluations. *arXiv preprint arXiv:2502.05224*.

Zhou, Z.; Yu, H.; Zhang, X.; Xu, R.; Huang, F.; Wang, K.; Liu, Y.; Fang, J.; and Li, Y. 2024. On the Role of Attention Heads in Large Language Model Safety. *arXiv preprint arXiv:2410.13708*.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *CoRR*, abs/2307.15043.