TOWARD COMPREHENSIVE AUDIO SAFETY PROTECTION ACROSS DIVERSE THREAT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio has rapidly become a primary interface for foundation models, powering real-time voice assistants. Ensuring safety in audio systems is inherently more complex than just "unsafe text spoken aloud": real-world risks can hinge on audio-native harmful sound events, speaker attributes (e.g., child voice), impersonation/voice-cloning misuse, and voice–content compositional harms, such as child-like voice plus sexual content. Such nature of audios makes it challenging to develop comprehensive benchmarks or guardrails against the unique risk landscape. To close this gap, we conduct large-scale red teaming on audio systems, systematically uncover audio-specific vulnerabilities, and develop a comprehensive, policy-grounded audio risk taxonomy and **AUDIOSAFETYBENCH**, a benchmark spanning diverse audio threat models. AUDIOSAFETYBENCH covers 17 languages, controlled speaker conditions (including public-figure/impersonation and child-like voices), voice–content compositions, and non-speech sound events. Importantly, these acoustic cues are not intrinsically unsafe; their safety relevance is determined by semantic context and deployment-specific policy. To defend against these threats, we propose AUDIOGUARD, a unified guardrail consisting of (1) *SoundGuard* for waveform-level audio-native cue detection and (2) *ContentGuard* for policy-grounded semantic protection. Experiments on the matched-policy AUDIOSAFETYBENCH setting and four independently constructed external benchmarks show that AUDIOGUARD improves guardrail accuracy over strong audio-LLM-based baselines. Under the interfaces available to users, AUDIOGUARD also exhibits lower observed deployment wall-clock latency than the proprietary API baselines; these latency measurements are not compute-matched.

1 INTRODUCTION

As foundation models are increasingly deployed in consumer and enterprise products, audio has emerged as a primary interaction modality—powering real-time voice conversations with assistants (ChatGPT, Gemini) and speech-driven agent interfaces. In parallel, the rapid maturation of *text-to-speech (TTS)* and *voice generation / voice cloning* services has made high-quality speech synthesis easy to integrate via APIs, including commercial platforms such as ElevenLabs (ElevenLabs, 2026), PlayHT (PlayHT, 2026), and Resemble AI (Resemble AI, 2026), as well as widely used cloud offerings like Amazon Polly (Amazon Web Services, 2026), Microsoft Azure Speech TTS (Microsoft, 2026), and Google Cloud TTS (Google Cloud, 2026). Yet ensuring safety in these settings is inherently more complex than "unsafe text spoken aloud": failures can hinge on audio-native harmful sound events, speaker attributes (e.g., child voice), impersonation/voice-cloning misuse, and voice–content compositional harms. As a result, the community still lacks standardized audio risk taxonomies, benchmarks, and practical guardrails that capture real-world audio safety.

To close this gap, we conduct large-scale red teaming of these audio-capable AI systems, uncovering systematic vulnerabilities specific to the audio modality. These failures span (i) *audio-native* risks such as non-speech harmful sound events (e.g., sexual moans, human distress), (ii) *voice–content compositional* risks (e.g., child voice combined with sexual content), (iii) *voice impersonation* and voice-cloning misuse, and (iv) transcript-level policy violations exacerbated by multilingual and diverse real-world scenarios. Based on the red teaming practices, we develop a comprehensive, policy-grounded risk taxonomy and curate a standardized audio safety benchmark across language, celebrity, child voice, and audio application scenarios. This process yields **AUDIOSAFETYBENCH**,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the first comprehensive benchmark for *audio-input* and *audio-output* safety, enabling comprehensive measurement and fine-grained diagnosis of audio safety across controlled, deployment-motivated settings and threat models.

To ensure audio safety across diverse scenarios and threat models, **guardrail models** provide a flexible and practical layer that can be deployed alongside heterogeneous audio systems with minimal impact on their core utility. However, existing audio guardrails often rely on large audio-language models as monolithic judges, which can be costly to run and brittle in practice, and they frequently inherit text-centric safety formulations that do not align with audio-specific risk taxonomies (e.g., non-speech sound events and voice–content compositional risks). To address these gaps, we propose AUDIOGUARD, a unified audio guardrail that decouples audio-native cue detection from semantic safety guard: *SoundGuard* extracts audio-native risk signals such as unsafe sound events and speaker identities directly from waveforms, *ContentGuard* performs automatic speech recognition (ASR) followed by *TextGuard* for guardrail over transcripts, and finally integrates both to produce interpretable, scenario-specific guardrail decisions.

Our evaluation separates two regimes. Results on AUDIOSAFETYBENCH measure a *matched-policy* setting because AUDIOGUARD is trained to implement the same policy space used to annotate the benchmark; these results demonstrate policy specialization rather than architecture-independent generalization. We therefore use four independently constructed external benchmarks—Nemotron-Audio, Jailbreak-AudioBench, Omni-SafetyBench, and AdvWave—as the primary evidence of out-of-distribution generalization. On these external benchmarks, AUDIOGUARD attains an average accuracy of 0.876, compared with 0.836 for Gemini 3 and 0.823 for GPT-Audio. We further validate the contributions of *SoundGuard* and *TextGuard*, study detector-oriented safety–utility metrics, and analyze cross-lingual transfer. Under the interfaces available to users, AUDIOGUARD also has lower observed deployment wall-clock latency than the proprietary Gemini 3 and GPT-Audio APIs; because these measurements are not compute-matched, we do not interpret them as intrinsic computational-efficiency comparisons.

We summarize the main **contributions** below:

- **Red-teaming driven audio risk discovery.** We conduct large-scale red teaming on audio-capable AI systems and voice generation pipelines, and develop a comprehensive, policy-grounded audio risk taxonomy that captures audio-native sound events, voice–content compositional risks, impersonation/voice cloning misuse, and multilingual scenario-specific violations.
- **Audio safety benchmark.** We develop **AUDIOSAFETYBENCH**, a comprehensive benchmark considering diverse threat models against *audio-input* and *audio-output* safety, built upon policy-grounded taxonomy and equipped with sliceable metadata across language, celebrity, child voice, and application scenarios.
- **Unified audio guardrail.** We develop AUDIOGUARD, a flexible guardrail that combines *SoundGuard* (audio-native cue detection), *ContentGuard* (ASR + *TextGuard*), and a compositional mechanism for scenario-specific guardrail across threat models.
- **Comprehensive evaluation and findings.** We distinguish matched-policy evaluation on AUDIOSAFETYBENCH from out-of-distribution evaluation on four independently constructed external benchmarks. AUDIOGUARD achieves an external-only average accuracy of 0.876 versus 0.836 for Gemini 3 and 0.823 for GPT-Audio, with paired uncertainty analysis for the headline comparison. We additionally report detector-oriented safety–utility metrics, component-wise analyses, cross-lingual transfer, and observed deployment wall-clock latency under the interfaces available to users.

2 RELATED WORK

Recent progress in **large audio-language models (LALMs)** extends foundation models beyond text to process speech, general sound events, and music, enabling applications such as voice chat, audio question answering, and multi-turn spoken dialogue. Representative systems include SpeechGPT (Zhang et al., 2023a), SALMONN (Tang et al., 2023), Qwen2-Audio (Chu et al., 2024), and Audio Flamingo 3 (Goel et al., 2025). In parallel, the community has developed standardized evaluations for audio-centric instruction following and dialogue understanding, such as AIR-Bench (Yang et al., 2024) and ADU-Bench (Gao et al., 2024). While these efforts rapidly expand the capabilities and deploy-

Table 1: Comparison between AUDIOSAFETYBENCH and other audio safety benchmarks.

Benchmark	Diverse Threat Models	Input & Output Safety	Audio-Specific Risk Taxonomy	Non-Speech Sound Events	Diverse Voices	Multi-Lingual
Nemotron-Content-Safety-Audio	✗	✗	✗	✗	✓	✗
Jailbreak-AudioBench	✗	✗	✗	✗	✓	✗
Omni-SafetyBench	✗	✗	✗	✗	✗	✗
AdvWave	✗	✗	✗	✓	✗	✗
AUDIOSAFETYBENCH (Ours)	✓	✓	✓	✓	✓	✓

ment surface of audio-enabled AI, they do not yet provide a unified, policy-grounded framework for defining and measuring audio safety risks across realistic threat models. In addition to audio understanding, modern speech generation has advanced rapidly, making high-fidelity **text-to-speech** and **voice cloning** widely accessible. Neural codec language models such as VALL-E (Wang et al., 2023b) and its cross-lingual extension VALL-E X (Zhang et al., 2023b), as well as large-scale generative speech models like Voicebox (Le et al., 2023), demonstrate strong in-context speaker conditioning, editing, and multilingual synthesis. These advances motivate safety evaluation beyond transcript-level policy violations to include voice impersonation, speaker-attribute risks (e.g., child voice), and voice–content compositional harms.

Guardrail models provide a flexible way to improve foundation-model safety without modifying the underlying model (Inan et al., 2023; Chi et al., 2024; Zeng et al., 2024; Li et al., 2024; Han et al., 2024; Ghosh et al., 2024; Padhi et al., 2024; Chen et al.). However, most existing guardrails are *text-only* and therefore overlook audio-specific risks (e.g., non-speech harmful sound events, speaker attributes such as child voice, and voice impersonation or voice–content compositional harms). A straightforward alternative is to use large audio-language models as end-to-end safety judges (e.g., GPT-Audio, Gemini, Audio Flamingo (Goel et al., 2025), Qwen3-Omni (Xu et al., 2025)), but such monolithic approaches can be expensive to deploy, sensitive to prompting, and—more importantly—often inherit text-centric safety formulations that do not align with audio-native risk taxonomies. In contrast, AUDIOGUARD offers a unified and efficient audio safety guardrail by decomposing the task into complementary components: *SoundGuard* performs waveform-level audio-native cue detection, *ContentGuard* applies *ASR* followed by *TextGuard* for policy-grounded semantic moderation, and an interpretable integration module composes both signals to produce scenario-specific decisions across threat models.

3 AUDIOSAFETYBENCH

In this section, we introduce AUDIOSAFETYBENCH, a comprehensive audio safety benchmark under controlled, deployment-motivated settings. We first present the motivation for audio-centered safety evaluation in Section 3.1, then describe the end-to-end dataset development pipeline for AUDIOSAFETYBENCH in Section 3.2, including taxonomy design, audio synthesis, red-teaming driven instance generation, hard-benign balancing, and multilingual augmentation. Finally, we summarize benchmark statistics and position AUDIOSAFETYBENCH against existing audio safety benchmarks in Section 3.3. An overview of AUDIOSAFETYBENCH is provided in Figure 1.

3.1 MOTIVATION

As audio-capable AI is increasingly used across diverse interaction scenarios (e.g., voice assistants, TTS/voice cloning, audio synthesis, and interactive voice agents), standardized audio safety evaluation must capture the key factors that drive real-world failures: *audio-native* cues (e.g., non-speech harmful sound events), *speaker-aware* risks (e.g., child voice and impersonation), and *scenario-dependent* policies that differ across audio-input moderation and audio-output generation. Existing audio safety benchmarks typically cover a single threat model or primarily convert text-centric risks into speech, limiting their ability to diagnose these audio-specific vulnerabilities. AUDIOSAFETYBENCH is built to fill this gap by providing a policy-grounded taxonomy and a standardized benchmark for both *audio-input* and *audio-output* safety across diverse threat models.

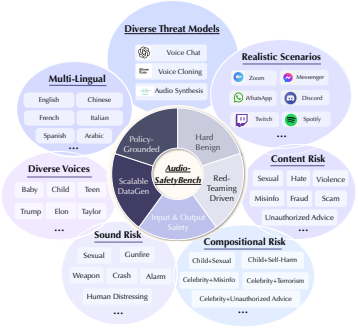

Figure 1: AUDIOSAFETYBENCH

This distinction is central to our threat model: *SoundGuard* detects acoustic evidence, while the scenario-specific policy layer determines whether that evidence is safety-relevant given the protected party, semantic context, and deployment action. Representative threat-model examples are summarized in Appendix Table 3.

3.2 DEVELOPMENT OF AUDIOSAFETYBENCH

Policy-grounded audio risk taxonomy from real-world scenarios. We first construct an audio-specific, policy-grounded risk taxonomy by distilling common safety requirements across realistic audio interaction scenarios. Concretely, we collect and consolidate publicly available safety policies from 20+ audio-related platforms spanning podcast sharing (e.g., Spotify), conferencing (e.g., Zoom), messaging/voice chat (e.g., WhatsApp, Discord), and live streaming (e.g., Twitch). We then parse these policy documents with an LLM-assisted pipeline to extract enforceable rules and map them into a unified, hierarchical taxonomy of audio risks. Beyond transcript-level violations, the taxonomy explicitly includes: **(i) non-speech harmful sound events** (e.g., weapon handling cues, sexual sounds, distress screams), **(ii) impersonation and voice-cloning misuse** (e.g., restricted public-figure voices), and **(iii) voice–content compositional risks** where the audio unsafety depends jointly on voice and semantic content (e.g., child voice $+$ sexual content). This taxonomy provides a unified label space across applications and serves as the blueprint for benchmark instance construction and evaluation.

Audio synthesis and control over voice conditions. To cover audio-native and speaker-aware risks beyond transcripts, we construct audio instances with explicit control over *sound events*, *speaker identity*, and *spoken content*. For non-speech harmful sound events, we curate and scrape representative audio clips (e.g., weapon handling cues, explosion-like sounds, distress screams, sexual sounds) from public audio sources and sound-effect libraries, and optionally mix them with speech to create realistic mixed-signal contexts. For impersonation and voice-cloning misuse, we collect reference voice samples of public figures/celebrities, perform voice cloning to obtain a target voice, and synthesize policy-violation utterances via TTS under that cloned identity. For speaker-attribute and compositional risks (e.g., child voice $+$ sexual content), we similarly construct voice-conditioned audio by controlling speaker attributes (child vs. adult; target identity when applicable) and generating matched transcripts that realize the intended risk category. To make the spoken content better reflect real-world usage, we use LLMs to generate scenario-grounded utterances conditioned on safety policies and typical interaction patterns from the same 20+ platforms used for policy scraping, spanning podcast sharing, conferencing, messaging/voice chat, and live streaming.

Red-teaming driven instance construction across diverse threat models. Guided by the taxonomy, we conduct large-scale red teaming to construct *challenging* unsafe audio instances across multiple deployment-relevant threat models, including: (1) audio-input moderation for audio-language model I/O, (2) audio-input moderation for voice-cloning pipelines (screening user audio before cloning), (3) audio-output safety for TTS systems (screening synthesized speech before delivery), and (4) interactive voice agents, where both incoming user audio and outgoing agent audio must be monitored. To ensure the resulting unsafe samples are *non-trivial* (i.e., not easily rejected by default platform safeguards or detected by simple heuristics), we apply a multi-stage filtering procedure: we first generate candidate instances and then retain only those that (i) successfully pass basic TTS/voice-cloning generation constraints and produce natural, intelligible audio, (ii) are not flagged by off-the-shelf transcript-only moderation under benign transcription errors, and (iii) remain clearly policy-violating under the intended threat model when reviewed against our taxonomy by human check.

Hard benign instances for false-positive stress testing. Audio safety systems must avoid over-blocking benign content that contains ambiguous cues (e.g., sirens in movies or sensitive keywords in educational contexts). To stress-test false positives, we curate a substantial set of *hard benign* instances that are semantically safe but contain potentially triggering audio cues or keywords. In the English content split, we use a balanced set of 4,275 unsafe and 4,275 hard-benign examples. Many hard-benign items are deliberately constructed as minimal-difference counterfactuals that preserve speakers, acoustic cues, and surrounding context while changing the policy-relevant semantics; they are therefore substantially harder than ordinary benign audio. These cases support explicit measurement of the safety–utility trade-off rather than only aggregate accuracy.

Multilingual augmentation. To reflect global deployment, we augment speech-involved audio with multilingual coverage by constructing instances across *17* languages. For each instance, we preserve

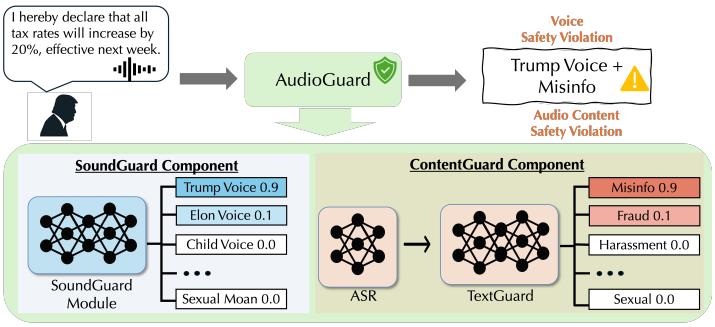

Figure 2: **AudioGuard framework overview.** Given an input audio waveform, *SoundGuard* detects audio-native safety cues directly from the signal (e.g., speaker identity, child voice, gunshot, sexual sounds) and outputs sound risk scores. In parallel, *ContentGuard* transcribes the audio via ASR and leverages *TextGuard* to predict the content risk scores (e.g., misinformation, fraud, harassment, sexual content). AUDIOGUARD provides the end-to-end audio guardrail with interpretable predictions under diverse audio threat models.

the underlying risk intent and threat-model scenario while varying the linguistic realization of the spoken content. We translate the source transcript using the Google Translate API and re-synthesize the corresponding speech under the same voice condition (e.g., celebrity/child voice when applicable), yielding multilingual audio pairs with matched semantics and controllable speaker attributes.

3.3 ANALYSIS OF AUDIOSAFETYBENCH

AUDIOSAFETYBENCH presents a large-scale benchmark for *audio-input* and *audio-output* safety, constructed via policy grounding and red-teaming driven instance generation. It contains **10K+** labeled audio instances spanning **17** languages and **50+** speaker identities, with rich, sliceable metadata covering diverse voice conditions (including celebrity/impersonation and child voice), non-speech harmful sound events, and multiple real-world audio application scenarios. The benchmark is also designed to support fine-grained diagnosis across risk categories, languages, speaker conditions, and audio provenance.

Table 1 situates AUDIOSAFETYBENCH relative to existing audio safety benchmarks. Prior datasets typically focus on a single threat model or a narrow class of attacks, and often lack either (i) coverage of both input and output safety, (ii) an audio-specific risk taxonomy beyond transcript-level violations, (iii) explicit non-speech sound-event evaluation, (iv) diverse voice conditions, or (v) multilingual breadth. In contrast, AUDIOSAFETYBENCH jointly covers all of these dimensions, enabling comprehensive, controlled evaluation of audio safety systems across threat models.

4 AUDIOGUARD

We present AUDIOGUARD, a unified audio safety guardrail that composes audio-native cue detection with transcript-based semantic moderation. We first provide an overview of the framework and its design principles in Section 4.1, then detail the model architecture, including *SoundGuard*, *ContentGuard* (ASR+*TextGuard*), and the compositional integration in Section 4.2. Finally, we describe the training and configuration for each component and the overall system in Section 4.3.

4.1 MOTIVATION AND OVERVIEW

To safeguard audio safety under diverse threat models, guardrails offer a flexible layer that can be deployed alongside heterogeneous AI systems with minimal impact on benign utility. To develop an audio safety guardrail, a straightforward approach is to fine-tune a single audio-language model as a monolithic judge. In practice, this design is difficult to train and deploy: (1) it requires large-scale audio supervision that jointly annotates audio-native cues (e.g., non-speech sound events, speaker attributes) and semantic policy violations, which is expensive and hard to curate at scale; and (2) it is inflexible across heterogeneous threat models. For example, in voice-cloning pipelines the primary requirement may be to detect restricted speaker identities, whereas in voice chat the dominant risks

are transcript-level policy violations; forcing a single model to cover all cases often depends on prompt-based task re-framing, which can introduce distribution shifts and brittle performance.

To address these challenges, we propose AUDIOGUARD, a dual-path audio safety guardrail that explicitly decomposes safety reasoning into *audio-native* and *semantic* channels and then composes them into a scenario-specific decision (Figure 2). Given an input audio waveform x, AUDIOGUARD produces (i) **sound/voice risk scores** via *SoundGuard*, (ii) **content risk scores** via *ContentGuard* (*automatic speech recognition (ASR)* $\rightarrow$ *TextGuard*), and (iii) a final guardrail action for the target threat model (e.g., allow or block). This decomposition directly targets two practical failure modes: (1) critical risks are often *not* recoverable from transcripts alone (e.g., non-speech harmful sound events or speaker identity/child voice), and (2) real policies are frequently *compositional* and scenario-dependent (e.g., *public-figure voice* $+$ *misinformation*), requiring joint reasoning over both audio-native and semantic signals.

4.2 AUDIOGUARD MODEL

We detail the three components of AUDIOGUARD: *SoundGuard* for audio-native cue detection, *ContentGuard* for transcript-based semantic moderation, and a compositional integration module that maps both signals to scenario-specific guardrail actions.

SoundGuard: audio-native cue detection. *SoundGuard* takes an input waveform x and predicts audio-native cues that are directly observable from the signal, including *speaker attributes* (e.g., child voice, celebrity) and *non-speech sound events* (e.g., gunfire/explosion-like sounds, distress screams, sexual sounds). Formally, SoundGuard outputs a multi-label score vector $\mathbf{s} = \mathrm{SoundGuard}(x) \in [0,1]^{K_s}$, where each dimension corresponds to a sound/voice cue in our taxonomy (or a calibrated grouping of cues), and s_k denotes the confidence of cue k.

ContentGuard: ASR + transcript-based safety reasoning. *ContentGuard* targets policy-grounded semantic violations expressed in speech. It first transcribes audio with an ASR model $\hat{t} = \mathrm{ASR}(x)$, and then applies *TextGuard* to the transcript to predict content risk scores $\mathbf{c} = \mathrm{TextGuard}(\hat{t}) \in [0,1]^{K_c}$, where each dimension corresponds to a policy-grounded content risk category (e.g., fraud, harassment, sexual content, misinformation). In practice, TextGuard can be instantiated as a lightweight classifier or an instruction-tuned LM fine-tuned to output structured risk labels/scores.

Compositional integration for scenario-specific decisions. AUDIOGUARD combines $\mathbf{s}$ and $\mathbf{c}$ into a unified, interpretable guardrail decision via compositional integration. The integration is configured per threat model to reflect scenario-specific constraints (e.g., voice cloning/TTS vs. voice chat vs. voice agents). We represent each policy rule r as a conjunction of threshold tests over sound/voice cues and content risks: $\phi_r(\mathbf{s}, \mathbf{c}) = \Big(\bigwedge_{k \in \mathcal{S}_r} s_k \geq \tau_k\Big) \wedge \Big(\bigwedge_{\ell \in \mathcal{C}_r} c_\ell \geq \tau_\ell\Big)$, where $\mathcal{S}_r \subseteq \{1, \ldots, K_s\}$ indexes the subset of SoundGuard cues used by rule r (e.g., *public-figure voice*, *child voice*, *gunfire*), $\mathcal{C}_r \subseteq \{1, \ldots, K_c\}$ indexes the subset of TextGuard content risks used by rule r (e.g., *misinformation*, *sexual content*), and $\{\tau_k\}, \{\tau_\ell\}$ are rule-specific thresholds. Triggered rules are mapped to actions (e.g., ALLOW, BLOCK, REVIEW) via a priority-ordered rule list. This decomposition provides: (i) **interpretability**, since decisions are attributable to specific triggered cues/risks; and (ii) **flexibility**, since policies can be updated or specialized to new threat models by changing rule definitions and thresholds without retraining SoundGuard or TextGuard.

4.3 TRAINING OF AUDIOGUARD

We next describe how we train the two learnable components of AUDIOGUARD (SoundGuard and TextGuard), and how we configure the model for efficiency and flexibility.

Training SoundGuard. SoundGuard is implemented as a lightweight audio classifier on top of a pretrained encoder. Concretely, we adopt the SpeechBrain ECAPA-TDNN speaker encoder (`speechbrain/spkrec-ecapa-voxceleb`) as the backbone feature extractor and attach an MLP prediction head for multi-label cue classification. During training, we keep the backbone fixed (unless otherwise noted) and fine-tune the MLP head (and optionally the last encoder block) to predict audio-native cues. We train SoundGuard as a multi-label predictor over audio-native cues using a mixture of (i) curated non-speech sound-event clips and mixtures (speech+event), and (ii) voice-conditioned data for speaker attributes (e.g., celebrity/impersonation condition and

Table 2: **End-to-end audio-safety accuracy.** Results on AUDIOSAFETYBENCH are a matched-policy evaluation because AUDIOGUARD is trained to implement the benchmark's policy space; the four external benchmarks are independently constructed and are used as the primary evidence of out-of-distribution generalization. Brackets show 95% stratified bootstrap confidence intervals from 10,000 example-level resamples. "Ext. Avg." averages the four external benchmarks. Latency is observed deployment wall-clock time under the interfaces available to users and is *not* compute-matched across local and proprietary API systems.

Method	AUDIOSAFETYBENCH (matched policy)			External benchmarks				Ext. Avg.	Wall-clock latency (s)
	Speech	Non-Speech	ElevenLabs	Nemotron	Jailbreak	Omni	AdvWave		
Gemma 3n	0.194 [0.169, 0.221]	0.379 [0.346, 0.412]	0.427 [0.394, 0.461]	0.686 [0.655, 0.716]	0.829 [0.804, 0.852]	0.462 [0.429, 0.496]	0.692 [0.661, 0.722]	0.667	0.823
Qwen3-Omni	0.245 [0.218, 0.274]	0.607 [0.574, 0.639]	0.362 [0.330, 0.395]	0.723 [0.693, 0.752]	0.682 [0.650, 0.713]	0.536 [0.502, 0.569]	0.872 [0.849, 0.893]	0.703	1.834
Audio Flamingo 3	0.235 [0.208, 0.264]	0.597 [0.564, 0.629]	0.103 [0.084, 0.124]	0.800 [0.773, 0.825]	0.430 [0.397, 0.464]	0.550 [0.516, 0.583]	0.780 [0.752, 0.807]	0.640	1.042
Gemini 3	0.357 [0.326, 0.389]	0.749 [0.719, 0.777]	0.732 [0.701, 0.761]	0.900 [0.879, 0.919]	0.862 [0.838, 0.884]	0.680 [0.648, 0.711]	0.900 [0.879, 0.919]	0.836	3.245
GPT-Audio	0.289 [0.260, 0.319]	0.720 [0.689, 0.749]	0.400 [0.367, 0.434]	0.894 [0.872, 0.914]	0.829 [0.804, 0.852]	0.652 [0.619, 0.684]	0.918 [0.899, 0.935]	0.823	2.542
AUDIOGUARD	**0.832** [0.807, 0.855]	**0.876** [0.853, 0.897]	**0.890** [0.868, 0.910]	**0.913** [0.893, 0.931]	**0.893** [0.871, 0.913]	**0.743** [0.713, 0.771]	**0.953** [0.938, 0.966]	**0.876**	1.423

child voice). Given cue labels $\mathbf{y}^{(s)} \in \{0,1\}^{K_s}$, we optimize the binary cross-entropy objective: $\mathcal{L}_{\text{sound}} = \sum_{k=1}^{K_s} \text{BCE}(s_k, y_k^{(s)})$.

Training TextGuard. TextGuard is implemented by fine-tuning a lightweight instruction-tuned LLM (Gemma-3-it) for policy-grounded transcript moderation. Starting from a pretrained Gemma-3 checkpoint, we fine-tune the model to predict a SAFElabel or a fine-grained unsafe risk category from our taxonomy, given text input. Concretely, we attach a small classification head (linear layer) on top of the final hidden state and optimize cross-entropy over the target label space. Training supervision is derived from our policy-grounded taxonomy, and based on that we develop text safety data following the construction pipeline of Poly-Guard (Kang et al., 2025a).

Because ASR transcripts can exhibit distribution shift relative to clean text (e.g., recognition errors and punctuation loss), we improve robustness via transcript noise augmentation. During training and evaluation, we inject realistic perturbations by passing text through a *TTS→ASR* round-trip to simulate recognition artifacts. This encourages TextGuard to remain stable under transcription noise and reduces brittleness to ASR-specific artifacts.

End-to-end efficiency. By decoupling waveform-level cue detection from transcript-level semantic moderation, AUDIOGUARD avoids relying on a large audio-LLM as a monolithic judge. The resulting guardrail is efficient, *SoundGuard* and *ContentGuard* can run in parallel, and modular, allowing components to be swapped or improved independently. Moreover, AUDIOGUARD is flexible across threat models: deployments can enable only the necessary submodules (e.g., *SoundGuard*-only for speaker-identity screening in voice cloning) to meet latency and cost constraints.

5 EVALUATION RESULT

We separate matched-policy performance from out-of-distribution generalization. **(1)** On AUDIOSAFETYBENCH, AUDIOGUARD is evaluated in the policy space used for its training and therefore measures policy specialization. **(2)** On four independently constructed external benchmarks, AUDIOGUARD obtains the strongest external-only average among the evaluated systems. **(3)** Detector-oriented metrics expose the asymmetric costs of missed unsafe audio and over-blocked benign audio. **(4)** Non-speech and voice–content compositional risks remain challenging for monolithic audio-LLM judges, while TextGuard exhibits strong cross-lingual transfer across 17 languages.

5.1 EVALUATION SETUP

Datasets. We evaluate on AUDIOSAFETYBENCH and four complementary external audio safety benchmarks. AUDIOSAFETYBENCH contains three controlled, deployment-motivated splits designed to isolate key audio safety vulnerabilities: **Speech** (speech inputs with *voice–content compositional* risks), **Non-Speech** (audio-native harmful sound events), and **ElevenLabs Red-Teaming** (a separate commercial voice-cloning pipeline that captures service-specific artifacts, speaker characteristics,

and impersonation-oriented test cases). The first two splits emphasize controlled counterfactual diagnosis rather than ecological realism. For external benchmarks, **Jailbreak-AudioBench** (Cheng et al., 2025) evaluates audio jailbreak attacks; **Nemotron-Content-Safety-Audio** (Ghosh et al., 2025) provides a content-safety test set with speaker and violation metadata; **Omni-SafetyBench** (Pan et al., 2025) includes multimodal safety evaluations with an audio subset covering diverse attack types; and **AdvWave** (Kang et al., 2025b) contains adversarial audio attacks via wave-based perturbations.

Baselines. We compare AUDIOGUARD against advanced audio-LLMs used as end-to-end guardrails via prompting, including Gemma 3n (Google DeepMind, 2025), Qwen3-Omni (Xu et al., 2025), Audio Flamingo 3 (Goel et al., 2025), Gemini, and GPT-Audio. For each audio-LLM baseline, we prompt the model to act as a safety judge given an input waveform and to output structured predictions aligned with our taxonomy. Concretely, the prompt enumerates all relevant risk categories and requires the model to produce both (i) **audio-native sound/voice cues** (e.g., non-speech harmful sound events, child voice, celebrity/impersonation) and (ii) **content risk categories** based on the spoken semantics (when speech is present). We use a fixed output schema (e.g., `Sound: <labels>` and `Content: <labels>`) and evaluate baselines using the same joint sound+content accuracy metric as AUDIOGUARD.

Comparison scope and output parsing. This comparison reflects a practical deployment setting—a purpose-built, supervised guardrail versus general audio LLMs used as prompted safety judges—and is *not* a supervision-matched architectural comparison. Because AUDIOSAFETYBENCH and AUDIOGUARD share a policy taxonomy, AUDIOSAFETYBENCH results should be read as matched-policy evaluation. We separately report external benchmarks to assess transfer beyond this construction pipeline. Structured-output parsing failures are tracked separately from semantic safety errors; a formatting failure is not treated as evidence that the underlying semantic judgment was unsafe.

5.2 MAIN RESULTS

Table 2 reports end-to-end joint sound+content accuracy. On the four independently constructed external benchmarks, AUDIOGUARD achieves an external-only average of 0.876, compared with 0.836 for Gemini 3 and 0.823 for GPT-Audio, with the highest point estimate on each benchmark. Across all evaluated examples, the average improvement over Gemini 3 is 0.131 with a paired 95% bootstrap CI of $[0.115, 0.147]$ ($p < 0.001$). On the matched-policy AUDIOSAFETYBENCH splits, AUDIOGUARD obtains 0.832/0.876/0.890 on Speech/Non-Speech/ElevenLabs, respectively. Detector-oriented metrics, hard-benign false-positive behavior, the noise/overlap stress test, statistical details, and the scope of the wall-clock latency comparison are provided in Appendix B.1.

5.3 FINE-GRAINED DIAGNOSIS ON AUDIO RISKS

While overall results in Table 2 establish consistent gains, AUDIOSAFETYBENCH is designed to enable *risk-category-wise diagnosis* of audio-specific vulnerabilities that are often obscured by aggregate metrics. Accordingly, Figures 3–4 present two representative category-wise results: (i) severe *voice–content compositional* harms in the Speech split, and (ii) *non-speech* harmful sound-event slices where transcript-only reasoning is unavailable.

Severe voice–content compositional vulnerabilities. Figure 3 reports joint sound+content accuracy on challenging cases where a high-risk *voice attribute* (child voice or celebrity/impersonation) co-occurs with a semantic risk category (e.g., Sexual, Self-Harm, Criminal, Misinformation). Across these slices, end-to-end audio-LLM judges may correctly identify the semantic risk from the transcript yet often fail to reliably detect the voice attribute, leading to low *joint* accuracy under compositional policies. The gap is especially apparent in celebrity/impersonation combinations (e.g., celebrity+misinformation or celebrity+terrorism), where correctly moderating requires *simultaneous* speaker-aware detection and content moderation. In contrast, AUDIOGUARD remains strong across all combinations, reflecting the benefit of explicitly separating waveform-level cue detection (*SoundGuard*) from semantic moderation (*ContentGuard*) and composing them via interpretable rules aligned with the threat model.

Non-speech harmful sound events remain challenging for monolithic judges. Figure 4 evaluates representative non-speech sound-event slices. Since these clips contain no spoken content, transcript-based reasoning provides little signal. We observe that audio-LLM baselines can be brittle in this

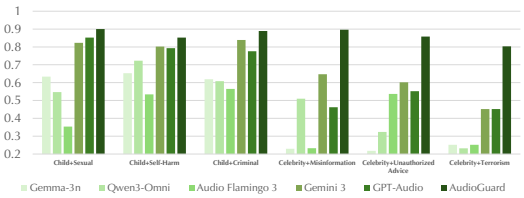

Figure 3: **Severe voice–content compositional slices in AUDIOSAFETYBENCH (Speech).** Joint sound+content accuracy on representative child/celebrity voice and semantic-risk combinations.

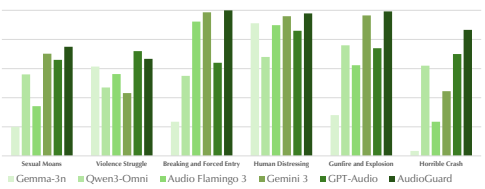

Figure 4: **Non-speech harmful sound-event slices in AUDIOSAFETYBENCH (Non-Speech).** Sound-event accuracy on representative audio-native risks where transcript-only reasoning is unavailable.

setting, likely due to limited calibration for acoustically diverse events and the absence of linguistic context. AUDIOGUARD achieves consistently higher accuracy across event types, indicating that a dedicated waveform-level detector trained on audio-native cues is critical for capturing these risks.

5.4 ABLATION STUDIES

We ablate AUDIOGUARD by evaluating the two core components, *TextGuard* for transcript-level semantic moderation and *SoundGuard* for waveform-level cue detection.

TextGuard improves policy-grounded semantic moderation. Table 5 in Appendix B reports *content-only* accuracy with 95% bootstrap confidence intervals. ContentGuard reaches 0.936 on AUDIOSAFETYBENCH Speech and 0.959 on ElevenLabs, and an average of 0.929 across the reported content-observable splits/benchmarks. On several external benchmarks, ContentGuard alone has a higher point estimate than the full compositional system; this is expected when the benchmark primarily tests semantic content rather than audio-native cues and underscores the value of enabling only the components required by a deployment's threat model.

SoundGuard is necessary for audio-native and speaker-aware risks. Table 6 in Appendix B reports sound-only accuracy with 95% bootstrap confidence intervals. SoundGuard reaches 0.887 on Speech, 0.876 on Non-Speech, and 0.918 on ElevenLabs. These marginals satisfy the required consistency relation that end-to-end joint correctness cannot exceed sound-only correctness on the corresponding split. The results show that audio-native cues (e.g., non-speech sound events and speaker attributes) benefit from dedicated waveform-level supervision, while the policy layer determines whether a detected cue is actually safety-relevant.

Cross-lingual generalization of TextGuard. Table 7 in Appendix B shows that TextGuard trained with *English-only* supervision still generalizes strongly to non-English safety classification. Despite never seeing labeled training data in other languages, TextGuard consistently improves F1 over the base Gemma-1B across all 17 languages, indicating that policy-grounded decision boundaries learned from English transfer to multilingual settings. This suggests that, in practice, high-quality supervision in a single language can already yield meaningful multilingual robustness, reducing the annotation burden for deploying transcript-based guardrails globally.

Limitations and conclusions. We introduce a policy-grounded audio risk taxonomy, AUDIOSAFETYBENCH, and the modular AUDIOGUARD guardrail. The results should be interpreted with several scope constraints. First, AUDIOSAFETYBENCH is a matched-policy evaluation for AUDIOGUARD, so its larger in-domain margins demonstrate policy specialization rather than architecture-independent generalization; the four independently constructed external benchmarks provide the primary evidence of transfer. Second, much of AUDIOSAFETYBENCH uses controlled synthesis, translation, and re-synthesis to create precisely labeled counterfactuals. This supports diagnosis but limits ecological validity: accents, low-quality microphones, emotional speech, codec artifacts, naturally overlapping speakers, and broader in-the-wild conditions remain incompletely covered. Third, the reported latency is deployment wall-clock latency under non-compute-matched interfaces and should not be interpreted as intrinsic model efficiency. Finally, proprietary API baselines are version-sensitive and some audio assets require restricted handling, which limits exact third-party reproduction unless versions, manifests, predictions, and release constraints are carefully documented. Within these boundaries, the external-benchmark, detector-oriented, uncertainty, and component analyses support the value of separating audio-native cue detection from transcript-level semantic moderation and composing them through scenario-specific policy.

AI USE STATEMENT

Generative AI tools were used to assist with editing (including grammar, spelling, and wording), data processing and filtering, and implementation of standard methods. All AI-assisted outputs used in this work were reviewed by the authors. The authors take responsibility for the final content of the paper, including all text, claims, analyses, code, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies safety risks in audio-capable AI systems and therefore includes potentially sensitive categories such as impersonation, child-like voices, sexual content, and harmful non-speech events. Audio-native attributes are treated as contextual signals rather than intrinsically unsafe characteristics; final safety decisions depend on semantic content and deployment policy. To reduce dual-use and privacy risks, the release excludes raw scraped reference audio and other assets that could facilitate voice cloning, and raw audio is distributed only when redistribution is permitted. Restricted examples are represented through releasable metadata, labels, stable identifiers or hashes, and evaluation code. We further report false-positive behavior on benign and non-target cases and discuss remaining representation and fairness limitations in Appendix C.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, benchmark construction procedure, evaluation protocol, and operating metrics in the main paper, with detailed main-result analysis and component-level results in Appendix B.1 and Appendix B.

REFERENCES

Amazon Web Services. Amazon polly api reference. `https://docs.aws.amazon.com/polly/latest/dg/API_Reference.html`, 2026.

Elias Bassani and Ignacio Sanchez. Guardbench: A large-scale benchmark for guardrail models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18393–18409, 2024.

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. In *The Thirteenth International Conference on Learning Representations*.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models, 2025. URL `https://arxiv.org/abs/2501.13772`.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024.

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. Qwen2-audio technical report, 2024. URL `https://arxiv.org/abs/2407.10759`.

ElevenLabs. Text to speech — elevenlabs documentation. https://elevenlabs.io/docs/overview/capabilities/text-to-speech, 2026.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models, 2024. URL https://arxiv.org/abs/2412.05167.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5992–6026, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.306. URL https://aclanthology.org/2025.naacl-long.306/.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025. URL https://arxiv.org/abs/2507.08128.

Google Cloud. Cloud text-to-speech documentation. https://docs.cloud.google.com/text-to-speech/docs, 2026.

Google DeepMind. google/gemma-3n-e2b. https://huggingface.co/google/gemma-3n-E2B, 2025. Hugging Face model repository. Accessed: 2026-01-27.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Mintong Kang, Zhaorun Chen, Chejian Xu, Jiawei Zhang, Chengquan Guo, Minzhou Pan, Ivan Revilla, Yu Sun, and Bo Li. Polyguard: Massive multi-domain safety policy-grounded guardrail dataset. *arXiv preprint arXiv:2506.19054*, 2025a.

Mintong Kang, Chejian Xu, and Bo Li. Advwave: Stealthy adversarial jailbreak attack against large audio-language models. *ICLR*, 2025b.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023. URL https://arxiv.org/abs/2306.15687.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*, 2023.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Microsoft. Text to speech documentation - azure ai speech. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-text-to-speech, 2026.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian. *arXiv preprint arXiv:2412.07724*, 2024.

Leyi Pan, Zheyu Fu, Yunpeng Zhai, Shuchang Tao, Sheng Guan, Shiyu Huang, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Felix Henry, Aiwei Liu, and Lijie Wen. Omni-safetybench: A benchmark for safety evaluation of audio-visual large language models, 2025. URL https://arxiv.org/abs/2508.07173.

PlayHT. Playht text to speech api: Quickstart. https://docs.play.ht/reference/api-getting-started, 2026.

Resemble AI. Resemble documentation: Welcome. https://docs.resemble.ai/welcome, 2026.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2023. URL https://arxiv.org/abs/2310.13289.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023a.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023b. URL https://arxiv.org/abs/2301.02111.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension, 2024. URL https://arxiv.org/abs/2402.07729.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023a. URL https://arxiv.org/abs/2305.11000.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling, 2023b. URL https://arxiv.org/abs/2303.03926.

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

A More Related Work

Safety datasets such as DecodingTrust (Wang et al., 2023a), HarmBench (Mazeika et al., 2024), AdvBench (Zou et al., 2023), ToxicChat (Lin et al., 2023), BeaverTails (Ji et al., 2024), HarmfulQA (Bhardwaj & Poria, 2023), JailbreakBench (Chao et al., 2024), GuardBench (Bassani & Sanchez, 2024), and Poly-Guard (Kang et al., 2025a) are predominantly *text-only*. Simply converting these prompts into speech (via TTS) does not meaningfully introduce *audio-native* safety challenges, and thus fails to stress-test audio safety guardrails beyond transcript-level moderation. Several recent benchmarks begin to study safety in audio settings, including Omni-SafetyBench (Pan et al., 2025), Jailbreak-AudioBench (Cheng et al., 2025), and AdvWave (Kang et al., 2025b). While valuable, they are often scoped to specific attack families or spoken renditions of text-centric risks, and do not comprehensively cover audio-specific vulnerabilities such as non-speech harmful sound events, speaker-attribute risks (e.g., child voice), voice impersonation/voice cloning, and voice–content compositional harms. In contrast, AudioSafetyBench is constructed from large-scale red teaming to enable comprehensive *audio-input* and *audio-output* safety evaluation, with a policy-grounded taxonomy and a standardized protocol that supports fine-grained slicing across language, celebrity, child voice, and diverse audio application scenarios.

B Additional Evaluation Results

B.1 Detailed Main-Result Analysis

Table 2 reports end-to-end joint sound+content accuracy. The AudioSafetyBench columns are a matched-policy evaluation and show how effectively AudioGuard implements the target policy space. The four external benchmarks provide the more direct test of out-of-distribution generalization: AudioGuard reaches an external-only average of 0.876, compared with 0.836 for Gemini 3 and 0.823 for GPT-Audio, and has the highest point estimate on each of the four external benchmarks. Across all evaluated examples, the average accuracy improvement over the strongest end-to-end baseline, Gemini 3, is 0.131 with a paired 95% bootstrap confidence interval of $[0.115, 0.147]$ ($p < 0.001$). We use paired tests because methods are evaluated on the same examples and avoid interpreting unsupported benchmark-specific point-estimate differences as statistically meaningful.

On AudioSafetyBench, AudioGuard achieves 0.832 on Speech, 0.876 on Non-Speech, and 0.890 on the ElevenLabs red-teaming split. These larger in-domain margins should not be interpreted as architecture-independent superiority: the benchmark and guardrail share the same policy space. The ElevenLabs split nevertheless provides a complementary distribution shift because it uses a separate commercial voice-cloning pipeline rather than the controlled synthesis pipeline used for the core benchmark.

Detector-oriented operating characteristics. Accuracy alone can hide the asymmetric costs of missed unsafe audio and false alarms. Appendix Table 4 therefore reports unsafe recall/TPR, hard-benign FPR, precision, and F1 at the default operating threshold. Unsafe recall ranges from 0.900 to 0.940 across the three AudioSafetyBench splits, while hard-benign FPR ranges from 0.160 to 0.236. The highest FPR, 23.6%, occurs on the deliberately difficult Speech hard-benign subset, where benign examples differ minimally from unsafe counterparts while preserving speakers, acoustic cues, and context. On standard benign Nemotron-Audio data, the observed FPR is 3.2%, indicating that the hard-benign subset is a stress test rather than an estimate of the typical benign false-positive rate.

Because the integration layer is threshold-based, deployments can move the operating point according to the relative cost of false negatives and false positives. We therefore treat the default threshold as one operating point rather than a universally optimal choice.

Controlled distribution-shift stress test. Controlled synthesis enables precise counterfactual labels but does not reproduce all naturally occurring acoustics. Under an added noise/overlap stress test, AudioGuard decreases from 0.832 to 0.793 while remaining above Gemini 3 and GPT-Audio in the same test. This result, together with the ElevenLabs split and the four external benchmarks, provides complementary evidence under distribution shift, but it does not cover the full range of accents, low-quality microphones, emotional speech, codec artifacts, or multi-speaker conversations.

Table 3: **When audio-native cues become safety-relevant.** Acoustic cues are contextual signals; the policy action depends on deployment context and semantic composition.

Scenario	Cue and context	Potential risk / action
Voice-cloning service	Public-figure voice in a cloning request	Unauthorized impersonation or fraud; block/review
TTS / voice-agent output	Public-figure voice + misinformation	Increased deceptive credibility; review/block
TTS / voice-agent output	Child-like voice + sexual content	Child-safety concern; block
Audio moderation	Gunfire, distress, or sexual sounds in policy-violating generated audio	Harmful audio event; review/block
News / entertainment	The same acoustic cues in benign context	Contextually benign; allow

Table 4: **Detector-oriented metrics at the default operating threshold.** TPR is unsafe recall; FPR is measured on the corresponding hard-benign subset.

Split	Acc.	TPR	FPR	Prec.	F1
Speech	0.832	0.900	0.236	0.792	0.843
Non-Speech	0.875	0.930	0.180	0.838	0.882
ElevenLabs	0.890	0.940	0.160	0.855	0.895

Observed deployment latency. The latency measurements in Table 2 are not compute-matched: AUDIOGUARD and open models are measured in local execution, whereas proprietary Gemini 3 and GPT-Audio are accessed through APIs that include network transmission, queueing, and service overhead. We therefore report these numbers only as *deployment wall-clock latency under the interfaces available to users*. Under that definition, AUDIOGUARD is observed at 1.423s per sample versus 3.245s for Gemini 3 and 2.542s for GPT-Audio; we do not use these measurements to claim intrinsic computational efficiency over all audio-LLM baselines.

C IMPACT STATEMENT

This work aims to improve the safety of audio-capable AI systems by providing a standardized benchmark (AUDIOSAFETYBENCH) and an efficient, interpretable guardrail (AUDIOGUARD) for both audio-input and audio-output settings. By explicitly modeling audio-native cues (e.g., non-speech harmful sound events and speaker attributes) and voice–content compositional risks, our approach can help developers better evaluate and mitigate safety failures in real-world voice assistants, TTS/voice-cloning services, and interactive voice agents.

Potential positive impacts include reducing harmful or policy-violating audio generation, improving detection of impersonation and child-safety risks, and enabling more transparent, diagnosis-driven safety improvements through slice-based evaluation. Potential risks include dual use: the benchmark taxonomy and red-teaming insights could reveal failure modes, and voice-related assets could be misused to facilitate impersonation. Public-figure and child-like voices are therefore treated only as contextual cues whose safety relevance depends on deployment context and semantic composition, not as intrinsically unsafe attributes.

Our release policy excludes raw scraped reference audio and other assets that could facilitate voice cloning, and distributes raw audio only when redistribution is permitted. For restricted examples, we provide releasable metadata, labels, stable identifiers/hashes, and evaluation code without distributing the underlying cloning-enabling audio. We also report false-positive behavior on benign/non-target cases and encourage fairness analysis across under-represented accents, dialects, and speaker groups. The applicable institutional ethics/IRB approval or exemption status, together with source provenance, licensing, and redistribution status, should be documented alongside the released artifacts.

C.1 UNCERTAINTY, REPRODUCIBILITY, AND RESPONSIBLE RELEASE

All confidence intervals in the end-to-end, content-only, and sound-only tables use 10,000 stratified example-level bootstrap resamples. Because methods are evaluated on the same examples, the headline comparison additionally uses paired bootstrap resampling: AUDIOGUARD improves over Gemini 3 by 0.131 in average end-to-end accuracy, with paired 95% CI $[0.115, 0.147]$ and $p <$

Table 5: **Content-only accuracy with 95% bootstrap confidence intervals.** For AUDIOSAFETYBENCH, Non-Speech is omitted because transcript-level content is not observable. Intervals use 10,000 stratified example-level bootstrap resamples.

Method	Speech	ElevenLabs	Nemotron	Jailbreak	Omni	AdvWave	Avg.
Gemma 3n	0.801 [0.774,0.826]	0.828 [0.803,0.851]	0.752 [0.722,0.780]	0.881 [0.858,0.902]	0.598 [0.565,0.630]	0.768 [0.739,0.796]	0.771 [0.760,0.782]
Qwen3-Omni	0.824 [0.798,0.847]	0.623 [0.590,0.655]	0.789 [0.761,0.815]	0.756 [0.727,0.784]	0.662 [0.630,0.693]	0.910 [0.890,0.928]	0.761 [0.749,0.772]
Audio Flamingo 3	0.523 [0.489,0.557]	0.345 [0.314,0.378]	0.842 [0.817,0.865]	0.516 [0.482,0.550]	0.632 [0.599,0.664]	0.824 [0.798,0.848]	0.614 [0.601,0.627]
Gemini 3	0.934 [0.915,0.951]	0.902 [0.880,0.922]	0.938 [0.920,0.953]	0.911 [0.890,0.929]	0.781 [0.753,0.807]	0.942 [0.924,0.957]	0.901 [0.892,0.910]
GPT-Audio	0.846 [0.821,0.869]	0.876 [0.853,0.897]	0.932 [0.913,0.948]	0.887 [0.864,0.907]	0.758 [0.729,0.785]	0.951 [0.935,0.964]	0.875 [0.865,0.884]
ContentGuard (Ours)	**0.936** [0.917,0.952]	**0.959** [0.944,0.971]	**0.948** [0.932,0.961]	**0.936** [0.918,0.951]	**0.824** [0.798,0.847]	**0.968** [0.954,0.978]	**0.929** [0.921,0.936]

Table 6: **Sound-only accuracy on AUDIOSAFETYBENCH with 95% bootstrap confidence intervals.**

Model	Speech	Non-Speech	ElevenLabs
Gemma 3n	0.238 [0.210,0.268]	0.379 [0.346,0.412]	0.511 [0.477,0.545]
Qwen3-Omni	0.306 [0.276,0.337]	0.607 [0.574,0.639]	0.448 [0.415,0.482]
Audio Flamingo 3	0.302 [0.272,0.333]	0.597 [0.564,0.629]	0.181 [0.156,0.208]
Gemini 3	0.420 [0.388,0.453]	0.749 [0.719,0.777]	0.800 [0.772,0.826]
GPT-Audio	0.353 [0.322,0.385]	0.720 [0.689,0.749]	0.495 [0.461,0.529]
SoundGuard (Ours)	**0.887** [0.865,0.907]	**0.876** [0.853,0.897]	**0.918** [0.899,0.935]

0.001. Benchmark-specific superiority claims should be interpreted only when supported by the corresponding paired analysis.

For third-party verification, the release package is designed to include canonical evaluation manifests, stable example identifiers, all releasable per-example predictions, split-generation code, metric and bootstrap scripts, random seeds, preprocessing settings, prompts and thresholds, model/API versions, evaluation dates, and hashes of the prediction files used to generate the reported tables. Proprietary API results should be interpreted as version- and date-specific because the underlying services can change over time.

For data sources, release metadata records source provenance, licensing information, and redistribution status. Raw scraped reference audio and assets that could facilitate voice cloning are excluded from public distribution; restricted examples are represented by metadata, labels, stable identifiers/hashes, and evaluation code where redistribution of the raw audio is not permitted.

Table 7: **Multilingual F1 of TextGuard.** Per-language F1 on AUDIOSAFETYBENCH (17 languages), comparing Base Gemma-1B vs. our finetuned TextGuard.

Language	Base	TextGuard	Language	Base	TextGuard
en	0.657	0.862	nl	0.466	0.618
es	0.523	0.689	cs	0.447	0.595
fr	0.547	0.692	ar	0.431	0.584
de	0.538	0.678	zh-cn	0.462	0.621
it	0.512	0.653	ja	0.428	0.573
pt	0.505	0.641	hu	0.439	0.589
pl	0.493	0.627	ko	0.436	0.587
tr	0.471	0.612	hi	0.415	0.562
ru	0.458	0.603			