THE ACOUSTIC STALEMATE: BENCHMARKING LARGE LANGUAGE MODELS VIA AUDIO-BASED CAPTCHAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio CAPTCHAs represent one of the most challenging real-world tests for AI systems, requiring robust auditory perception and logical inference under diverse acoustic degradation. However, existing benchmarks largely focus on vision tasks, leaving a critical gap in systematic evaluation of audio-language models. Thus, we introduce AU-BENCH, the first interactive, open-source benchmark specifically designed to systematically evaluate the auditory perception and reasoning capabilities of large language models through 18 diverse audio CAPTCHA tasks spanning Content-based and Rule-based paradigms. We propose a dual-gradient complexity framework comprising Audio-Temporal Reasoning Depth (ATRD) to quantify the cognitive, multi-step reasoning burden, and Acoustic Perceptual Load (APL) to formally measure physical signal degradation through SNR, reverberation, and spectral masking density. Our fully interactive web-based testing platform systematically evaluates state-of-the-art native multimodal large language models and cascaded ASR-LLM systems under both zero-shot and Chain-of-Thought settings. Comprehensive experiments reveal a sharp capability asymmetry: current foundation models match or exceed human baselines on clean, single-step transcription, yet fall substantially short on degraded and affective tasks, while still achieving non-trivial success rates overall. This emerging proficiency serves as a critical security warning for the automated bypassing of audio CAPTCHAs. By exposing these highly distinct failure topologies, AU-BENCH aims to shift the research frontier from conversationally fluent assistants toward robust, perceptually grounded agents capable of navigating real-world web interactions.

1 INTRODUCTION

The rapid advancement of large language model (LLM) agents Ye et al. (2025); He et al. (2024) has accelerated their integration into real-world web environments, enabling unprecedented levels of automated navigation and interaction. Consequently, traditional human-centric verification mechanisms, such as CAPTCHAs, are facing a profound security threat. Originally designed to block automated abuse on high-stakes platforms such as e-commerce sites and secure authentication portals, these defenses are increasingly vulnerable to advanced AI systems. As LLM-driven agents develop stronger reasoning and multimodal perception capabilities, their potential to autonomously bypass or solve CAPTCHAs is evolving from a theoretical concern into a pressing real-world vulnerability.

State-of-the-art Multimodal LLMs (MLLMs) and Automatic Speech Recognition (ASR) systems such as Gemini-3.1-Pro Google DeepMind (2026), Mimo-V2.5 Xiaomi (2026), Whisper-1 OpenAI (2022) and GPT-4o OpenAI (2024) perform strongly on visual and audio recognition tasks Luo et al. (2026); Wu et al. (2026b), yet the audio modality remains critically under-explored. Audio CAPTCHAs Fanelle et al. (2020); Ding et al. (2025) demand not only robust acoustic perception under degradation but sophisticated temporal reasoning, and there is no reliable assessment of whether current models can emulate human-level auditory perception and reasoning on CAPTCHA sloving.

To close this gap, we introduce AU-BENCH, the first interactive, open-source benchmark for audio perception and reasoning in MLLMs. It comprises 18 audio CAPTCHA tasks across two paradigms: *Content-based* tasks Abdullah et al. (2022); Aubry et al. (2025) probing low-level signal decoding and robust transcription, and *Rule-based* tasks Fanelle et al. (2020) demanding higher-order reasoning such as spatial audio inference and complex logical instruction following. All tasks are delivered

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.

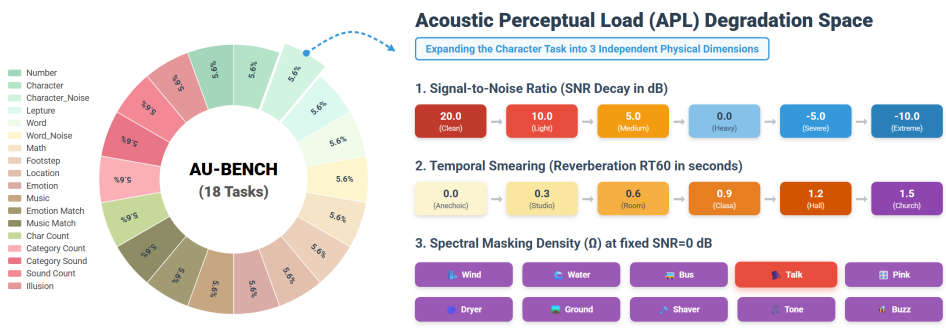

Figure 1: Au-Bench data distribution and Acoustic Perceptual Load Degradation Space

through a fully interactive web-based testing platform that mirrors the interaction patterns of real-world CAPTCHA systems, requiring models to process acoustic information and reason temporally.

Our core methodological contributions center on a dual-gradient evaluation framework. We introduce **Audio-Temporal Reasoning Depth (ATRD)**, which adapts reasoning depth taxonomies Luo et al. (2026) to the audio-temporal domain by quantifying task complexity as the minimal number of atomic reasoning and interaction steps. Complementing this cognitive dimension, we propose **Acoustic Perceptual Load (APL)**, which quantifies physical signal difficulty through three rigorously formalized acoustic parameters: signal-to-noise ratio (SNR), reverberation index ($\mathcal{R}_T$, measured via RT_{60}), and spectral masking density (Ω_T). Together, ATRD and APL form a two-dimensional difficulty matrix that allows us to pinpoint whether a model is "confused" or "deaf".

We systematically evaluate state-of-the-art native MLLMs, including Gemini-3.1-pro, Gemini-3.5-flash, and Mimo-V2.5, alongside cascaded ASR-LLM pipelines (Whisper-1 paired with GPT-4o), under both zero-shot and CoT prompting regimes. Our experiments reveal critical findings: (1) models match or exceed humans on clean transcription but substantially underperform on tasks requiring operation under acoustic degradation; (2) CoT prompting yields inconsistent benefits, improving reasoning-heavy tasks but occasionally degrading performance on perception-heavy ones; (3) cascaded ASR-LLM systems show severe vulnerability on tasks requiring fine-grained acoustic feature extraction (e.g., emotion, footstep direction), where critical prosodic and spatial information is lost in the transcription bottleneck. Across the benchmark, models approach human accuracy on clean transcription, but humans remain substantially more robust on degraded and affective tasks.

In summary, we make the following contributions:

- We propose AU-BENCH, the first interactive, open-source benchmark specifically designed to systematically evaluate the auditory perception and reasoning capabilities of multimodal large language models through 18 diverse audio CAPTCHA tasks.
- We introduce a dual-gradient complexity framework comprising ATRD to quantify the cognitive, multi-step reasoning burden, and APL to formally measure physical signal degradation through SNR, reverberation index, and masking density.
- We construct a fully interactive web-based testing platform and systematically evaluate state-of-the-art native MLLMs and cascaded ASR-LLM under both zero-shot and CoT settings, exposing massive performance gaps compared to humans.

2 Related Work

2.1 Advances and Challenges in CAPTCHA Security

CAPTCHA systems have evolved from simple text-based challenges Google (2025a); Bursztein et al. (2011); Shi et al. (2020) to visual-based variants Google (2025b); hcaptcha (2025). Prior research has extensively analyzed the security vulnerabilities of visual CAPTCHAs against both traditional computer vision attacks Hossen et al. (2020); Plesner et al. (2024) and modern deep learning methods Sivakorn et al. (2016a;b). However, the security properties of audio-based CAPTCHAs Abdullah et al. (2022); Ding et al. (2026) remain comparatively understudied. As illustrated in Fig. 8 of Appendix A, such audio-based variants can be categorized into two types: content-based and

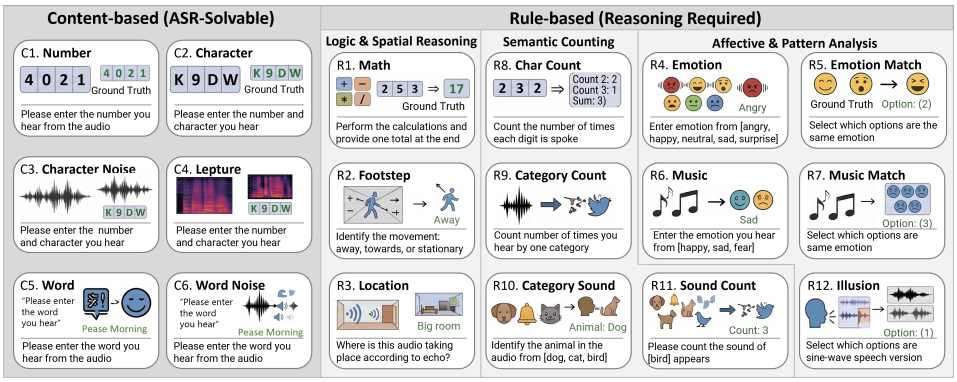

Figure 2: AU-BENCH samples across Content- and Rule-based paradigms.

rule-based. Content-based audio CAPTCHAs Abdullah et al. (2022); Aubry et al. (2025) typically require users to transcribe spoken digits or letters. In contrast, rule-based audio CAPTCHAs Fanelle et al. (2020) require users to reason over both audio content and textual instructions, shifting the challenge from simple transcription toward higher-level audio reasoning. While these designs improve resistance to traditional automated attacks, their effectiveness against advanced AI-driven solvers has not been systematically examined.

2.2 CAPTCHA BENCHMARKS AND MODELS

Advances in deep learning have significantly improved CAPTCHA recognition. Early methods relied on CNNs and RNNs for feature extraction Thobhani et al. (2020); Hu et al. (2018); Derea et al. (2023). Recently, systemic benchmarks Luo et al. (2026); Wu et al. (2026a;b) have been proposed to evaluate the performance of LLMs. Open CaptchaWorld Luo et al. (2026) proposed a web-based benchmark designed to assess whether agents can autonomously solve modern CAPTCHAs through perception, reasoning, and multi-step interaction. MirrorCAPTCHA Wu et al. (2026a) is a curated benchmark of real-world CAPTCHAs that are challenging for agents yet easy for humans. MCA-Bench Wu et al. (2026b) evaluates vision-language models against synthetic, homogeneous CAPTCHA puzzles. Unlike visual CAPTCHAs where static images provide instant, parallel access to information, audio CAPTCHAs introduce unique challenges: temporal sequentiality (information unfolds over time), perceptual degradation (noise, reverberation, masking), and the need for sustained attention across audio streams.

3 BENCHMARK

Following Fanelle et al. (2020), Fig. 2 taxonomizes the 18 tasks into two paradigms: Content-based tasks test low-level acoustic decoding and robust transcription under noise. Rule-based tasks demand higher-order reasoning, forcing agents to synthesize acoustic physics, affective prosody, and logical instructions, which exposes the developmental asymmetries in modern architectures.

3.1 DATA CONSTRUCTION

Audio samples are constructed through three pipelines: *human recording* from the Emotional Speech Dataset (ESD) zho (2022) for authentic prosody, *TTS synthesis* (Google Cloud Text-to-Speech) for standardized content-based stimuli, and *algorithmic acoustic simulation* using DSP (dynamic volume enveloping, time-variant low-pass filtering, parametric delay lines) to engineer spatial soundscapes rather than relying on black-box generative models. All tasks are delivered through a web interface (HTML/CSS/JavaScript) that mirrors real-world CAPTCHA interaction patterns.

3.2 AUDIO-TEMPORAL REASONING DEPTH

To better characterize the cognitive difficulty of puzzles in AU-BENCH, we introduce a new metric called *Audio-Temporal Reasoning Depth*, which adapts the CAPTCHA Reasoning Depth framework Luo et al. (2026) to the audio-temporal domain. Unlike traditional classifications that group

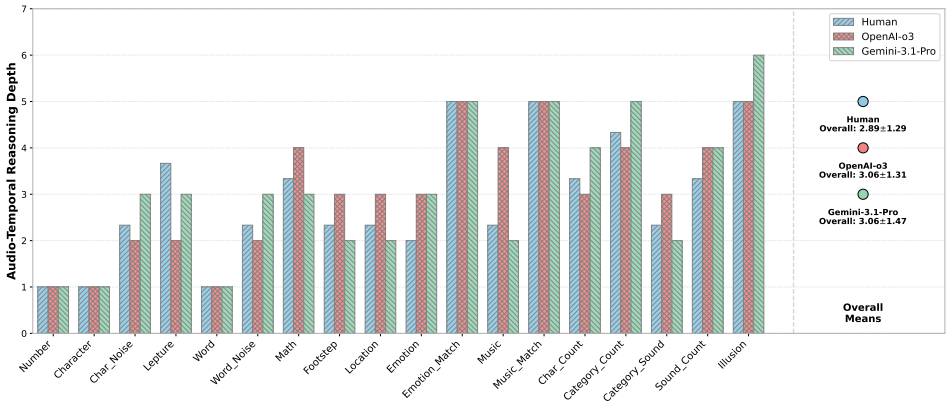

Figure 3: Audio-Temporal Reasoning Depth Estimation by Human and Reasoning Models.

puzzles merely by degradation type (e.g., clean vs. noisy), ATRD offers a task-agnostic measure of complexity that aligns more closely with the multi-step nature of agent reasoning. We define ATRD as the minimal number of atomic reasoning or decision-making steps required by a human or a model to arrive at a correct solution, where each step involves interpreting auditory content, planning a subgoal, or executing a discrete interaction.

Formally, let an audio CAPTCHA be defined as a task T requiring a sequence of operations. We define the Audio-Temporal Reasoning Depth $D_{\text{audio}}(T)$ as:

$$D_{\text{audio}}(T) = \sum_{i=1}^{N} \mathbb{I}[s_i \in \mathcal{S}_T] \tag{1}$$

where $\mathcal{S}_T$ is the set of atomic steps needed to solve T, s_i is an atomic reasoning or interaction step from a predefined checklist $\mathcal{C}$, and $\mathbb{I}[\cdot]$ is the indicator function. Each s_i contributes 1 unit of depth if the step is observed during the solution process. The checklist $\mathcal{C}$ comprises categories including **Auditory Perception** (e.g., speech detection, speaker identification, content transcription, emotion recognition, acoustic event detection, music classification, degradation type perception) and **Cognitive Reasoning** (e.g., cross-modal alignment, temporal ordering, logical inference, elimination, goal decomposition, short-term memory, spatial reasoning, semantic understanding).

For instance, a puzzle that asks the user to "transcribe the numbers" typically requires only two steps: first, perceive the spoken digits, and second, output the content. In contrast, an *Emotion Match* CAPTCHA may require identifying the affective prosody of the main audio, memorizing it, sequentially processing three distinct option audios, and executing a similarity comparison in the latent emotional space, leading to a significantly higher reasoning depth.

To measure this across the benchmark, we conducted a rigorous human annotation study where participants were instructed to decompose their process into a sequence of atomic steps. We constructed heuristic rules to guide the annotators and ensure consistency. The median number of steps across three independent annotators was taken as the ground-truth ATRD per puzzle. Fleiss' κ across annotators exceeds 0.75, indicating substantial agreement. To better compare the reasoning depth, we also prompted OpenAI-o3 and Gemini-3.1-Pro with the same heuristic rules to estimate the reasoning depth of each CAPTCHA type. Fig. 3 illustrates this distribution.

Different Reasoning Depth Estimates Between Humans and Models. Humans and models estimate reasoning depth differently in Fig. 4: for *Music Match*, a human grasps the holistic "sad" vibe in one step, whereas OpenAI-o3 oversegments the task into granular steps ("extracting chord progression," "mapping tempo to emotion," "holding acoustic profile in memory"), inflating its estimate. Models thus over-think by atomizing tasks while humans rely on holistic psychoacoustic priors, which must be closed for human-like audio agents. Finally, ATRD scores exhibit a natural hierarchical structure that we use for stratified analysis: *Shallow Perception* (ATRD $\in \{1, 2\}$), *Intermediate Reasoning* (ATRD $\in \{3, 4\}$), and *Deep Reasoning* (ATRD $= 5$).

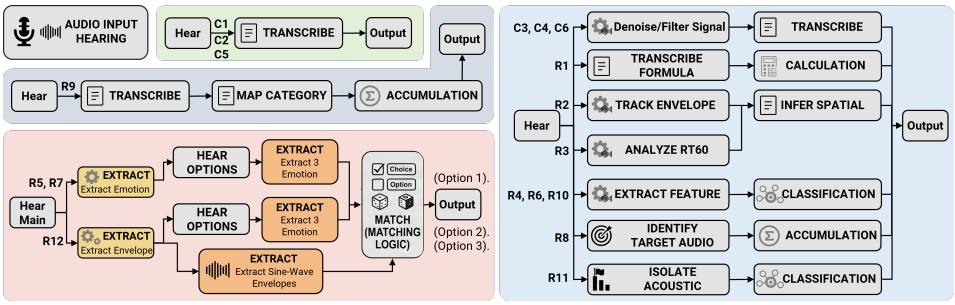

Figure 4: Audio-Temporal Reasoning Depth scores and representative reasoning chains for all 18 tasks in AU-BENCH, systematically sorted by paradigm category (Content vs. Rule) and task ID.

3.3 ACOUSTIC PERCEPTUAL LOAD

While ATRD quantifies the cognitive reasoning complexity, it assumes the signal is perfectly perceivable. To measure the physical difficulty of *hearing* the signal itself, we introduce the Acoustic Perceptual Load. APL quantifies the severity of acoustic degradation and informational overlap that obscures the raw signal. We formalize APL for a given task T as:

$$APL(T) = \alpha \cdot \mathcal{D}(SNR_T) + \beta \cdot \mathcal{R}_T + \gamma \cdot \Omega_T, \quad \alpha + \beta + \gamma = 1. \tag{2}$$

Here $\mathcal{D}(\cdot)$ denotes a monotonically decreasing penalty function and the three terms are normalized to $[0, 1]$ before combination so that the sum is unit-free: $\mathcal{D}(SNR_T) = \sigma(-SNR_T/6)$ (a logistic in dB), $\mathcal{R}_T$ is the RT_{60} value normalized by its 1.5 s ceiling, and $\Omega_T \in [0, 1]$ by construction (Eq. 7). The weights α, β, γ are calibrated from the single-axis ablations in Table 2: for each constituent we regress human Pass@1 against the corresponding normalized term along its ablation axis (the SNR, reverberation, and masking blocks) and normalize the absolute slopes to sum to one, yielding $\alpha = 0.56$, $\beta = 0.15$, $\gamma = 0.29$ (per-block R^2 ranging 0.72–0.93). A consistency check against the joint 6×6 degradation matrix of Table 3 is reported in Appendix A. The three physical constituents are strictly formulated as follows:

Signal-to-Noise Ratio (SNR_T): Quantifies the relative background noise intensity (in dB):

$$SNR_T = 10 \log_{10} \left(\frac{\sum_n x_{clean}^2[n]}{\sum_n (x_{distorted}[n] - x_{clean}[n])^2} \right) \tag{3}$$

where $x_{clean}[n]$, $x_{distorted}[n]$ denote discrete-time clean and distorted audio signals, respectively.

Reverberation Time ($\mathcal{R}_T$): Measures spatial smearing by quantifying the standard RT_{60} (in seconds). We extract the spatial Room Impulse Response (RIR), denoted as $h(t)$, via frequency-domain deconvolution. The energy decay curve is then computed using Schroeder's backward integration:

$$EDC(t) = \int_t^{\infty} h^2(\tau) d\tau \tag{4}$$

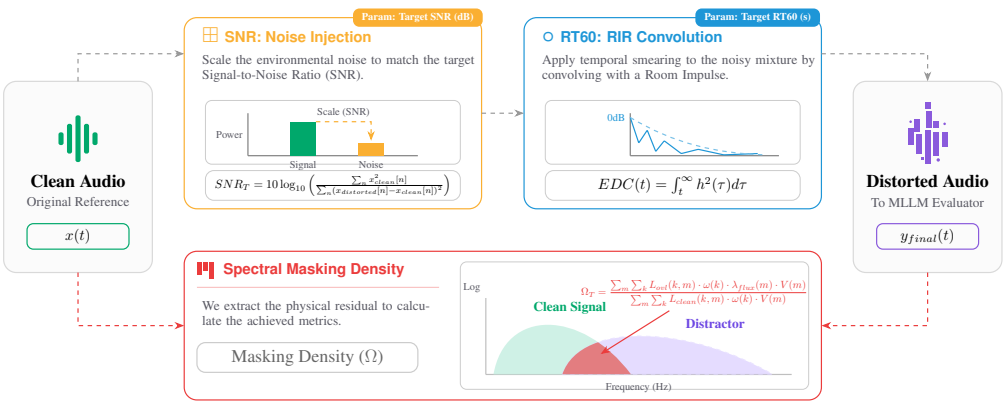

Figure 5: Comprehensive Quantitative Simulation Pipeline for Acoustic Perceptual Load. The pipeline programmatically degrades clean audio through mathematically rigorous acoustic physics to construct challenging distorted audio. Spectral Masking Density (Ω) quantifies the frequency-domain overlap between clean audio and distorted audio.

$\mathcal{R}_T$ is strictly defined as the time required for the logarithmic decay curve, $10\log_{10}(EDC(t))$, to decrease by 60 dB. Following acoustic standards, we estimate RT_{60} using the T_{20} method. We perform linear regression on the energy decay curve over the interval $L_{20} = [-5\text{ dB}, -25\text{ dB}]$ to find decay slope S (in dB/s), and extrapolate our parameter as:

$$\mathcal{R}_T = -\frac{60}{S} \quad \text{where} \quad S = \frac{d}{dt}\left[10\log_{10}(EDC(t))\right]\Big|_{L_{20}} \tag{5}$$

Masking Density (Ω_T): Measures spectral interference by quantifying the Temporal-Flux Weighted Masking Density. We align the physical metric with AI feature-drop mechanisms. We first extract the log-compressed Short-Time Fourier Transform (STFT) magnitudes for both the clean signal and noise, denoted as $L(k,m) = \log(1 + |S(k,m)|)$. The transient nature of the distractor is then computed using the normalized Spectral Flux penalty $\lambda_{flux}(m)$ as:

$$c_1 + c_2 \frac{\sum_k \left|L_{noise}(k,m) - L_{noise}(k,m-1)\right|}{\max_m\left(\sum_k \left|L_{noise}(k,m) - L_{noise}(k,m-1)\right|\right)}, \tag{6}$$

where c_1 and c_2 are scaling constants (e.g., 0.2 and 1.5). This dynamic multiplier heavily penalizes unpredictable, non-stationary distractors (e.g., babble or water noise) that actively disrupt the model's self-attention mechanism, while discounting highly stationary interference. We evaluate the masking effect strictly during active speech frames using a Voice Activity Detection (VAD) mask, $V(m) \in \{0,1\}$, and apply a frequency importance weighting function, $\omega(k)$, to prioritize high-frequency consonant bands. We extrapolate our final parameter Ω_T as:

$$\Omega_T = \frac{\sum_m \sum_k L_{ovl}(k,m) \cdot \omega(k) \cdot \lambda_{flux}(m) \cdot V(m)}{\sum_m \sum_k L_{clean}(k,m) \cdot \omega(k) \cdot V(m)}, \tag{7}$$

where $L_{ovl}(k,m)$ denotes the spectral overlap, defined as $\min\left(L_{clean}(k,m), L_{noise}(k,m)\right)$.

4 Experiments

This section analyzes experimental results on Au-Bench, comparing model performance across varying cognitive reasoning depths and acoustic perceptual loads. We highlight the distinct strengths and limitations of MLLM and ASR-LLM under different task paradigms, and discuss the implications for developing more robust, perceptually grounded audio-language verification systems.

4.1 Set Up

We compare native MLLMs (Gemini-3.1-pro, Gemini-3.5-flash, Mimo-V2.5) and cascaded ASR-LLM pipelines (Whisper-1, GPT-4o) against a Human baseline across 18 audio CAPTCHA tasks. Inputs are structured as raw acoustic signals, paired with task instructions. All model evaluations are conducted programmatically via a multi-agent framework with temperature $\tau = 0.0$. We perform exactly 100 independent trials for each CAPTCHA task, reporting the final Pass@1 accuracy as the primary evaluation metric. With $n = 100$ trials per cell, the 95% Wilson confidence interval on Pass@1 is approximately ± 0.10 at $p = 0.5$, scaling down to ± 0.06 near the extremes ($p = 0.1$ or 0.9); performance differences within this band are therefore conservatively treated as statistical noise. Human accuracy was collected from 20 self-reported normal-hearing adults via the same interface and stimuli. Each completed 5 stratified trials per task, giving $n = 100$ pooled evaluations per category to match the model budget; we report mean accuracy with 95% Wilson intervals.

4.2 ATRD Results and Analysis

4.2.1 Overall Model Performance

Table 1 presents the full Pass@1 accuracy results across all 18 tasks, 6 model configurations, and 4 setting combinations. Human performance serves as the reference baseline. Several macro-level patterns emerge. First, current models are reasonably competent at pure transcription but struggle profoundly with fine-grained acoustic reasoning. For instance, the best-performing model, Gemini-3.1-pro, achieves near-perfect zero-shot scores on simple transcription (Number: 1.0, Character:

Table 1: Performance of LLMs on AU-BENCH. This table presents Pass@1 accuracy on 18 CAPTCHA task types across native MLLMs and cascaded ASR-LLM systems.

Method		MLLM						ASR/ASR-LLM				Human
Prompt-Mode		Zero-Shot			Chain-of-Thought			Non-Prompt-Guide		Prompt-Guide		
	Model-Name	Mimo V2.5	Gemini-3.5 3.5-flash	Gemini-3.1 3.1-pro	Mimo V2.5	Gemini 3.5-flash	Gemini 3.1-pro	Whisper-1	GPT4o	Whisper-1	GPT4o	
Content-based	Number	0.91	1	1	0.91	0.95	0.95	0.9	0.99	0.58	0.91	0.97
	Character	0.58	0.69	0.82	0.38	0.77	0.7	0.68	0.72	0.28	0.41	0.96
	Character*	0.57	0.7	0.83	0.38	0.67	0.68	0.62	0.67	0.4	0.41	0.97
	Lepture	0	0	0.1	0	0	0.07	0	0.01	0	0	0.18
	Word	0.64	0.73	0.76	0.44	0.56	0.77	0.61	0.5	0.14	0.21	0.56
	Word*	0.53	0.6	0.73	0.58	0.68	0.76	0.49	0.44	0.18	0.17	0.49
Rule-based	Math	0.97	0.92	0.98	0.98	1	0.95	0.89	0.99	0.79	0.96	0.92
	Footstep	0.62	0.38	0.68	0.73	0.39	0.57	0.22	0.08	0.21	0.15	0.95
	Location	0.62	0.56	0.75	0.64	0.55	0.93	0.03	0.04	0.31	0.38	0.94
	Emotion	0.23	0.59	0.27	0.34	0.47	0.42	0.26	0.19	0.19	0.25	0.90
	Music	0.89	0.87	0.83	0.85	0.85	0.96	0.16	0.05	0.21	0.07	0.96
	Emotion Match	0.35	0.37	0.66	0.5	0.34	0.69	0.39	0.5	0.31	0.33	0.91
	Music Match	0.9	0.92	1	0.95	0.82	0.98	0.38	0.33	0.39	0.4	0.94
	Char Count	0.84	0.53	0.83	0.75	0.2	0.5	0.97	0.71	0.79	0.75	0.96
	Category Count	0.85	0.23	0.82	0.78	0.26	0.92	0.92	0.79	0.84	0.75	0.95
	Category Sound	0.65	0.57	0.77	0.68	0.54	0.96	0.45	0.52	0.50	0.41	0.97
	Sound Count	0.55	0.26	0.74	0.65	0.3	0.96	0.07	0.05	0.08	0.05	0.92
	Illusion	0.45	0.47	0.82	0.62	0.46	0.78	0.52	0.37	0.35	0.33	0.88

0.82) but drops substantially on affective or spatial tasks (Footstep: 0.68, Emotion: 0.27). This asymmetry reveals a fundamental observation: current models are reasonably competent at "hearing words" but struggle profoundly with "understanding sound".

In addition, providing task instructions to cascaded transcription encoders (e.g., Whisper-1) paradoxically *reduces* accuracy relative to unguided transcripts. For example, instruction-guided performance drops precipitously on Character ($0.68 \rightarrow 0.28$) and Word ($0.61 \rightarrow 0.14$). This indicates that instruction-conditioned transcription introduces a systematic bias or hallucinations that the downstream reasoning LLM fails to recover from.

4.2.2 MLLM VS. CASCADED SYSTEMS

A key architectural comparison emerges between native MLLMs and cascaded ASR-LLM systems. On Content-based tasks, the performance gap is relatively narrow: both architectures achieve comparable accuracy on Number and Word tasks. However, on Rule-based tasks involving **acoustic feature extraction beyond transcription**, cascaded systems dramatically underperform. For example, on Footstep (which requires tracking dynamic volume and frequency envelopes to infer spatial movement), cascaded systems achieve only 0.08$\sim$0.22, compared to 0.62$\sim$0.73 for native MLLMs. Similarly, on Location (which requires RT_{60}-based spatial size estimation), cascaded systems score below 0.04 while MLLMs reach 0.93 and 0.75 for Gemini-3.1-pro CoT and zero-shot.

This pattern reveals the transcription bottleneck: ASR systems strip away critical prosodic, spectral, and temporal acoustic features that are essential for higher-order audio reasoning. When the task requires reasoning about *how* something sounds rather than *what* was said, the cascaded architecture suffers catastrophic information loss. This finding has practical implications for the design of audio CAPTCHA systems: including tasks that require non-lexical acoustic reasoning (spatial, affective, temporal) can serve as a strong defense against ASR-based solvers.

4.2.3 EFFECT OF CHAIN-OF-THOUGHT PROMPTING

The impact of CoT prompting is heterogeneous across task types and model architectures. For tasks requiring multi-step logical operations such as Math and Category Count, CoT consistently improves accuracy: Gemini-3.1-pro rises from 0.82 to 0.92 on Category Count (two-proportion z-test, $p = 0.036$). On Illusion, the change from 0.82 to 0.78 is not significant ($p = 0.48$), within the ± 0.10 sampling noise at $n = 100$. This aligns with the intuition that explicit intermediate reasoning steps help primarily where the reasoning burden, rather than perception, is the bottleneck.

However, for perception-heavy tasks, CoT yields mixed results. On Character (MLLMs), CoT *decreases* accuracy for Mimo-V2.5 (0.58 falls to 0.38, $p = 0.005$), suggesting that forcing verbal reasoning about acoustic percepts may interfere with direct signal processing when the model's audio encoder and language decoder operate under a shared computational budget. We hypothesize that CoT shifts attention from low-level acoustic features to high-level semantic processing, which is counterproductive for tasks where precise phonemic discrimination is paramount.

Table 2: Impact of APL on Multimodal Agent Performance. We evaluate the Character CAPTCHA task across three independent degradation dimensions: SNR, Reverberation Time (RT_{60}), and Spectral Masking Density (Ω). The baseline clean condition is marked with an asterisk (*).

Acoustic Condition		MLLM (Zero-Shot)			ASR (Non-Prompt-Guide)		Human
Dimension	**Parameter**	**Mimo V2.5**	**Gemini-3.5 flash**	**Gemini-3.1 pro**	**Whisper-1**	**GPT-4o**	
SNR Decay (dB)	20.0 (Clean)*	0.60	0.70	0.83	0.65	0.72	0.97
	10.0 (Light)	0.45	0.68	0.82	0.62	0.68	0.96
	5.0 (Medium)	0.30	0.60	0.80	0.50	0.55	0.94
	0.0 (Heavy)	0.10	0.35	0.65	0.22	0.25	0.88
	-5.0 (Severe)	0.01	0.02	0.08	0.00	0.01	0.68
	-10.0 (Extreme)	0.00	0.00	0.00	0.00	0.00	0.25
Reverberation RT60 (s)	0.0 (Anechoic)*	0.88	0.92	0.90	0.71	0.79	0.96
	0.3 (Studio)	0.51	0.69	0.80	0.67	0.89	0.95
	0.6 (Room)	0.45	0.75	0.90	0.28	0.76	0.93
	0.9 (Classroom)	0.24	0.59	0.65	0.53	0.76	0.88
	1.2 (Hall)	0.31	0.75	0.85	0.41	0.49	0.82
	1.5 (Church)	0.16	0.42	0.75	0.38	0.54	0.76
Masking Density Ω (at SNR=0)	Bus ($\Omega = 0.43$)	0.49	0.81	0.89	0.6	0.89	0.96
	Dryer ($\Omega = 0.53$)	0.37	0.74	0.88	0.55	0.67	0.93
	Ground ($\Omega = 0.45$)	0.16	0.78	0.94	0.5	0.34	0.97
	Pink ($\Omega = 0.60$)	0.24	0.63	0.73	0.5	0.62	0.89
	Shaver ($\Omega = 0.56$)	0.26	0.67	0.74	0.5	0.61	0.88
	Talk ($\Omega = 0.44$)	0.47	0.76	0.87	0.36	0.45	0.94
	Tone ($\Omega = 0.46$)	0.39	0.81	0.91	0.85	0.71	0.95
	Water ($\Omega = 0.55$)	0.34	0.68	0.79	0.7	0.82	0.92
	Wind ($\Omega = 0.41$)	0.38	0.67	0.72	0.85	0.88	0.94
	Buzz ($\Omega = 0.44$)	0.53	0.86	0.89	0.65	0.99	0.97

4.2.4 Perception-Reasoning Bottleneck Analysis

By jointly analyzing ATRD and APL scores across all task-model pairs, we characterize distinct failure topologies. High-ATRD, low-APL tasks (e.g., Emotion Match, Illusion) reveal *cognitive failures*: models correctly perceive the acoustic content but fail at multi-step reasoning (e.g., comparing emotions across multiple audio samples). Conversely, high-APL, low-ATRD tasks (e.g., Lepture, Character_Noise) reveal *perceptual failures*: models cannot reliably extract the target signal from degraded audio, even when the required reasoning is trivial (simple transcription).

The most challenging quadrant, i.e., high-ATRD, high-APL, is currently unpopulated in our benchmark (our APL scores are deliberately stratified by task design). We identify this as an important direction for future benchmark expansion, as real-world audio environments (e.g., crowded public spaces, emergency alerts) simultaneously impose high perceptual and cognitive loads.

4.3 APL Results and Analysis

To systematically disentangle the perceptual bottlenecks of current multimodal systems, Table 2 presents a fine-grained ablation study on the Character CAPTCHA task across three independent dimensions of APL: SNR, Reverberation Time (RT_{60}), and Spectral Masking Density (Ω).

Temporal Smearing via Reverberation In the spatial dimension (RT_{60}), increasing reverberation induces temporal smearing, effectively erasing the transient boundaries between phonemes. Traditional cascaded ASR systems (Whisper-1) and early MLLMs (Mimo-V2.5) exhibit extreme fragility here. In contrast, Gemini-3.1-pro demonstrates exceptional spatial robustness (> 1.50 s), matching human-level tolerance even in highly echoic "Church" environments. This suggests that advanced, heavily scaled native audio encoders have learned to implicitly perform dereverberation, a capability wholly absent in standard transcription models.

The "Cocktail Party" Masking Problem To further probe perceptual limitations, we fix the SNR at 0.0 dB and vary the spectral masking density through different real-world distractors. The results in Table 2 prove that *not all noise is created equal*. Under low-frequency, sparse masking (Wind), models like Gemini-3.1-pro maintain relatively high accuracy (0.72). However, when exposed to severe spectral overlap (Talk/Babble noise), the architectures diverge sharply: native MLLMs remain comparatively robust (Gemini-3.1-pro retains 0.87), whereas cascaded ASR systems collapse (Whisper-1 falls to 0.36). Conversely, Humans retain a 0.94 accuracy under the same conditions via top-down cognitive unmasking and binaural processing to isolate the target speaker. This sophisticated psychoacoustic mechanism remains difficult for current MLLMs architectures to replicate.

Coupled Degradation Landscapes across Distractor Types Across the $RT_{60} \times SNR$ accuracy plane in Fig. 6, performance declines roughly monotonically toward the highly degraded region.

However, the exact failure boundary—where Pass@1 drops below 0.5—is strongly dictated by the distractor's spectral geometry. Distractors that heavily overlap the speech band (Talk/Babble, Buzz) push this boundary toward higher SNRs, collapsing models early; conversely, spectrally sparse distractors (Wind, Tone) preserve a performance plateau much deeper into the noise space.



Figure 6: Comprehensive APL Degradation Matrix across Distractor Types

To understand the physical mechanism driving this collapse, Fig. 7 plots the masking-density (Ω) counterpart for the exact same acoustic grid. Crucially, reading the two figures together reveals a striking geometric alignment: the regions of empirical model failure in Fig. 6 coincide almost exactly with the $\Omega \gtrsim 0.5$ isolines mapped in Fig. 7. This one-to-one correspondence demonstrates that Ω, despite being computed purely from the physical clean and distractor spectra (Eq. 7), perfectly recovers the empirically measured difficulty ordering without relying on any model-specific parameters. Ultimately, this serves as direct evidence that our APL framework captures an intrinsic perceptual property of the acoustic signal rather than overfitting a particular AI architecture.

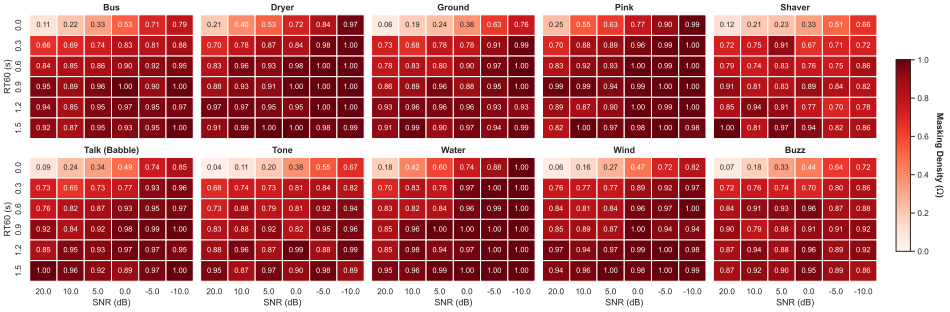

Figure 7: Masking Density Matrix across Distractor Types

5 Conclusion

Limitations. We note several limitations of the current benchmark. First, the high-ATRD $\times$ high-APL quadrant is deliberately unpopulated, so AU-BENCH does not yet probe the most realistic "crowded public space" regime that couples heavy perceptual load with heavy reasoning. Second, model results are a snapshot of commercial API versions whose behavior may drift over time. Third, the APL weights are fitted on a single task (Character); their generalization to other tasks is assumed rather than verified. Finally, because per-stimulus pairing was not retained across prompting regimes, the CoT comparisons above rely on aggregate two-proportion tests rather than paired tests such as McNemar.

Summary. We introduced AU-BENCH, an interactive benchmark designed to systematically evaluate AI agents on complex audio CAPTCHA tasks. By organizing 18 diverse tasks across Content and Rule paradigms, we establish a dual-gradient evaluation framework quantified by two core dimensions: the cognitive-level ATRD and the physical-level APL. Comprehensive evaluation of state-of-the-art foundation models reveals a sharp asymmetry: current agents match or exceed human baselines on clean transcription, yet fall substantially short under both physical acoustic degradation and multi-step temporal reasoning, while displaying highly distinct failure topologies. We hope AU-BENCH serves as a pivotal catalyst, shifting the research frontier from conversationally fluent assistants toward robust, perceptually grounded agents capable of navigating the full spectrum of real-world web interactions.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used generative AI tools to assist with implementing methods, executing and automating evaluation tasks, code development and debugging, result analysis, literature summarization, figure preparation, language polishing, and manuscript editing. Text-to-speech synthesis (Google Cloud TTS) was used to generate the standardized content-based audio stimuli, while all acoustic degradations were produced by deterministic DSP simulation. Generative AI was *not* used to formulate or prove mathematical claims, propose or refine hypotheses, or design the research methodology. All AI-assisted content was independently reviewed and verified by the authors, who take full responsibility for the final paper.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This research introduces AU-BENCH, a benchmark designed to systematically evaluate the auditory perception and reasoning capabilities of large language models through the lens of audio CAPTCHAs. We recognize the inherent dual-use nature of this work: evaluating the capacity of AI agents to solve CAPTCHAs could theoretically be misused to inform automated bypass mechanisms. However, our primary objective is strictly defensive. By systematically exposing the distinct failure topologies of state-of-the-art multimodal foundation models and highlighting the fragility of cascaded ASR-LLM pipelines under physical acoustic degradation, this benchmark is intended to proactively guide the development of more robust, perceptually secure verification systems against future automated web agents.

For the human baseline evaluations and the annotation process used to establish the Audio-Temporal Reasoning Depth scores, we prioritized participant privacy and data security. All participants were fully informed about the nature of the annotation tasks prior to their involvement. The collected response data was strictly limited to task performance metrics and was fully anonymized. No personally identifiable information or sensitive biometric data was collected, processed, or stored at any point during the interactive testing phases.

Regarding dataset release and data sources, the audio materials constructed for this benchmark rely on either openly licensed datasets (e.g., the Emotional Speech Dataset) or algorithmically synthesized and heavily degraded audio generated via DSP simulations. We are committed to responsible disclosure; the open-source release of the AU-BENCH platform and its associated datasets will be distributed under appropriate academic licenses to ensure the materials are used for security evaluation and model robustness research rather than malicious exploitation. The authors declare no conflicts of interest or potentially biasing sponsorships related to this work.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We have made comprehensive efforts to ensure the reproducibility of the AU-BENCH benchmark and all experimental evaluations. Section 3.2 describes the data-construction pipelines (human recording, TTS synthesis, and DSP-based acoustic degradation), Section 3.3 defines the Audio-Temporal Reasoning Depth metric and its annotation protocol, and Section 3.4 gives the exact parameterizations of SNR, reverberation (RT_{60}), and masking density (Ω). Section 4.1 states the evaluation protocol, including the model versions, the $n = 100$ trials per task, the temperature $\tau = 0.0$ setting, and the Wilson confidence-interval analysis. Furthermore, the theoretical formulations and exact mathematical parameterizations for the physical degradation dimensions (SNR, reverberation RT_{60}, and masking density Ω) are strictly defined in Section 3.4 and further visualized in the Appendix.

REFERENCES

Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393.

Hadi Abdullah, Aditya Karlekar, Saurabh Prasad, Muhammad Sajidur Rahman, Logan Blue, Luke A Bauer, Vincent Bindschaedler, and Patrick Traynor. Attacks as defenses: Designing robust audio captchas using attacks on automatic speech recognition systems. *arXiv preprint arXiv:2203.05408*, 2022.

Paul Aubry, Juliette Devoivre, Damien Carron, Simon Fernandez, Andrzej Duda, and Maciej Korczyński. Bypassing audio recaptcha with automatic speech recognition models. In *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 1–5. IEEE, 2025.

Elie Bursztein, Matthieu Martin, and John Mitchell. Text-based captcha strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security*, pp. 125–138, 2011.

Zaid Derea, Beiji Zou, Asma A Al-Shargabi, Alaa Thobhani, and Amr Abdussalam. Deep learning based captcha recognition network with grouping strategy. *Sensors*, 23(23):9487, 2023.

Ziqi Ding, Gelei Deng, Yi Liu, Junchen Ding, Jieshan Chen, Yulei Sui, and Yuekang Li. Illusioncaptcha: A captcha based on visual illusion. In *Proceedings of the ACM on Web Conference 2025*, pp. 3683–3691, 2025.

Ziqi Ding, Yunfeng Wan, Wei Song, Yi Liu, Gelei Deng, Nan Sun, Huadong Mo, Jingling Xue, Shidong Pan, and Yuekang Li. Robust captcha using audio illusions in the era of large language models: from evaluation to advances. *arXiv preprint arXiv:2601.08516*, 2026.

Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das. Blind and human: Exploring more usable audio {CAPTCHA} designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pp. 111–125, 2020.

Google. recaptcha v1. `https://developers.google.com/recaptcha/docs/versions#recaptcha_v1_-_shut_down`, 2025a.

Google. recaptcha v2. `https://developers.google.com/recaptcha/docs/versions#recaptcha_v2_im_not_a_robot_checkbox`, 2025b.

Google DeepMind. Gemini: Our mission is to build ai responsibly to benefit humanity. `https://deepmind.google/models/gemini/`, 2026. Accessed: 2026-07-01.

hcaptcha. hcaptcha. `https://www.hcaptcha.com`, 2025.

Yanheng He, Jiahe Jin, Shijie Xia, Jiadi Su, Runze Fan, Haoyang Zou, Xiangkun Hu, and Pengfei Liu. Pc agent: While you sleep, ai works–a cognitive journey into digital world. *arXiv preprint arXiv:2412.17589*, 2024.

Md Imran Hossen, Yazhou Tu, Md Fazle Rabby, Md Nazmul Islam, Hui Cao, and Xiali Hei. An object detection based solver for {Google's} image {reCAPTCHA} v2. In *23rd international symposium on research in attacks, intrusions and defenses (RAID 2020)*, pp. 269–284, 2020.

Yu Hu, Li Chen, and Jun Cheng. A captcha recognition technology based on deep learning. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 617–620. IEEE, 2018.

Yaxin Luo, Zhaoyi Li, Jiacheng Liu, Jiacheng Cui, Xiaohan Zhao, and Zhiqiang Shen. Open captchaworld: A comprehensive web-based platform for testing and benchmarking multimodal llm agents. *Advances in Neural Information Processing Systems*, 38, 2026.

OpenAI. Whisper-1. `https://openai.com/index/whisper/`, 2022. Accessed: 2026-07-01.

OpenAI. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2026-07-01.

Andreas Plesner, Tobias Vontobel, and Roger Wattenhofer. Breaking recaptchav2. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1047–1056. IEEE, 2024.

Chenghui Shi, Shouling Ji, Qianjun Liu, Changchang Liu, Yuefeng Chen, Yuan He, Zhe Liu, Raheem Beyah, and Ting Wang. Text captcha is dead? a large scale deployment and empirical study. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 1391–1406, 2020.

Suphannee Sivakorn, Iasonas Polakis, and Angelos D Keromytis. I am robot: (deep) learning to break semantic image captchas. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 388–403, 2016a.

Suphannee Sivakorn, Jason Polakis, and Angelos D Keromytis. I'm not a human: Breaking the google recaptcha. *Black Hat*, 14:1–12, 2016b.

Alaa Thobhani, Mingsheng Gao, Ammar Hawbani, Safwan Taher Mohammed Ali, and Amr Abdussalam. Captcha recognition using deep learning with attached binary images. *Electronics*, 9 (9):1522, 2020.

Xiangyu Wu, Yuwei Hu, Tianyu Cui, Yueying Tian, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yang Yang, and Jianfeng Lu. Mirrorcaptcha: Wild captcha, wild distribution, wild web-based platform meet multimodal llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31001–31017, 2026a.

Zonglin Wu, Yule Xue, Yaoyao Feng, Xiaolong Wang, and Yiren Song. Mca-bench: A multimodal benchmark for evaluating captcha robustness against vlm-based attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38039–38047, 2026b.

Xiaomi. Mimo-v2.5. `https://mimo.xiaomi.com/mimo-v2-5/`, 2026. Accessed: 2026-07-01.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, et al. Mobile-agent-v3: Fundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*, 2025.

A CONTENT-BASED AND RULE-BASED AUDIO CAPTCHAS

As illustrated in Fig. 8, audio CAPTCHAs can be fundamentally categorized into two distinct paradigms based on interaction mechanisms and cognitive requirements: Content- and Rule-based:

- Content-based Audio CAPTCHAs primarily evaluate low-level auditory perception and robust transcription capabilities. In this conventional paradigm, the challenge consists solely of an audio clip containing spoken digits, letters, or words. The required task is straightforward transcription, and the system verifies the response via a direct string comparison against a predefined ground truth.
- Rule-based Audio CAPTCHAs are designed to assess higher-order cognitive reasoning and instruction following. In this advanced paradigm, the challenge couples an acoustic signal with a specific logical instruction. The user must first process the audio and then actively apply the instruction to deduce the correct answer, which is validated against predefined logical rules rather than a static transcript.

This structural evolution essentially shifts the defense mechanism from simple acoustic signal decoding to complex, multi-step auditory reasoning. Consequently, these architectural differences naturally expose qualitatively different failure modes in modern AI systems (as detailed in Table 1). Content-based tasks typically collapse when the acoustic channel is degraded, exposing a *perceptual limit*. In contrast, Rule-based tasks frequently fail even under clean audio conditions when models cannot logically chain the textual instruction with the perceived acoustic content, thereby exposing a distinct *reasoning limit*.

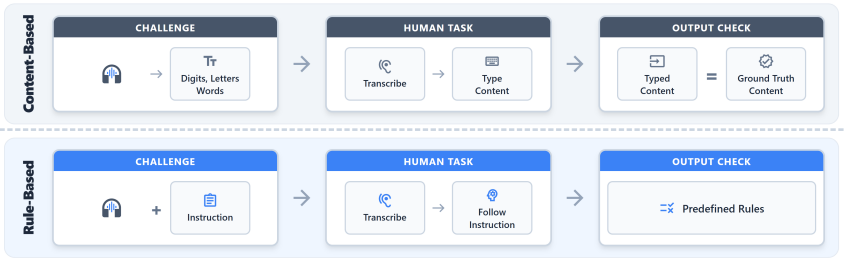

Figure 8: Content-based Audio CAPTCHAs and Rule-based Audio CAPTCHAs.

Orthogonalization of the Degradation Axes Table 3 presents a comprehensive 3D ablation matrix of Acoustic Perceptual Load, systematically orthogonalizing and jointly varying Reverberation Time (RT_{60}) and SNR. This design uniquely illustrates the coupled degradation effects across native MLLMs, cascaded ASR systems, and human baseline, as well as the dynamic physical impact on *Spectral Masking Density* (Ω). Crucially, Ω values here represent the full degraded mixture; thus, reverberation adds time–frequency overlap that elevates Ω ($0.65 \sim 0.77$) relative to the anechoic, single-distractor values in Table 2 ($0.41 \sim 0.60$), rendering the two scales not directly comparable. Our observations confirm that these physical degradation axes are interactive rather than strictly additive. For instance, while dropping the SNR to 0.0 dB under anechoic conditions ($RT_{60} = 0.0$ s) incurs an accuracy penalty of only 0.18 ($0.83 \rightarrow 0.65$) for Gemini-3.1-pro, imposing the identical 20.0 dB drop in a highly reverberant "Church" environment ($RT_{60} = 1.5$ s) triggers a severe collapse, dropping accuracy by 0.52 ($0.82 \rightarrow 0.30$). As an independent robustness check, jointly regressing human Pass@1 on the three normalized constituents over the 36 cells of this table yields $\alpha = 0.58$, $\beta = 0.15$, $\gamma = 0.28$ ($R^2 = 0.93$). This close agreement (within 0.02 of main-text weights in §3.4) confirms that the APL weights reflect a stable property of human perception rather than an artifact of any single table.

In stark contrast, the human baseline exhibits remarkable resilience across this compounded degradation landscape. Even in the most extreme boundary condition—combining maximum reverberation ($RT_{60} = 1.5$ s) with severe amplitude masking (-10.0 dB SNR)—humans retain a measurable Pass@1 accuracy of 0.08, whereas every evaluated MLLM and cascaded ASR system completely fails (scoring exactly 0.00). This qualitative gap highlights a fundamental architectural limitation: current models lack the top-down cognitive linguistic priors and psychoacoustic binaural unmasking

mechanisms that humans inherently leverage to disentangle target speech from compounded spatial and spectral interference.

Table 3: Comprehensive 3D Ablation Matrix of Acoustic Perceptual Load.

Acoustic Condition			MLLM (Zero-Shot)			ASR (Non-Prompt-Guide)		Human
Reverberation RT_{60} (s)	**SNR Decay (dB)**	**Masking (Ω)**	**Mimo V2.5**	**Gemini-3.5 flash**	**Gemini-3.1 pro**	**Whisper-1**	**GPT-4o**	
0.0 (Anechoic)	20.0 (Clean)*	0.73	0.60	0.70	0.83	0.65	0.72	0.97
	10.0	0.72	0.45	0.68	0.82	0.62	0.68	0.96
	5.0	0.65	0.30	0.60	0.80	0.50	0.55	0.94
	0.0	0.66	0.10	0.35	0.65	0.22	0.25	0.88
	-5.0	0.71	0.01	0.02	0.08	0.00	0.01	0.68
	-10.0	0.74	0.00	0.00	0.00	0.00	0.00	0.25
0.3 (Studio)	20.0 (Clean)*	0.73	0.55	0.68	0.85	0.58	0.69	0.96
	10.0	0.71	0.40	0.65	0.77	0.50	0.62	0.95
	5.0	0.65	0.25	0.55	0.79	0.38	0.48	0.93
	0.0	0.66	0.08	0.30	0.62	0.15	0.18	0.85
	-5.0	0.72	0.00	0.01	0.05	0.00	0.00	0.65
	-10.0	0.75	0.00	0.00	0.00	0.00	0.00	0.22
0.6 (Room)	20.0 (Clean)*	0.72	0.48	0.64	0.84	0.48	0.65	0.95
	10.0	0.72	0.32	0.60	0.67	0.40	0.55	0.93
	5.0	0.66	0.18	0.48	0.78	0.28	0.40	0.90
	0.0	0.66	0.05	0.25	0.63	0.08	0.12	0.82
	-5.0	0.73	0.00	0.00	0.01	0.00	0.00	0.60
	-10.0	0.75	0.00	0.00	0.00	0.00	0.00	0.18
0.9 (Classroom)	20.0 (Clean)*	0.72	0.38	0.58	0.80	0.38	0.55	0.92
	10.0	0.72	0.25	0.52	0.69	0.30	0.45	0.90
	5.0	0.66	0.12	0.40	0.72	0.18	0.30	0.87
	0.0	0.65	0.02	0.18	0.47	0.04	0.08	0.78
	-5.0	0.73	0.00	0.00	0.00	0.00	0.00	0.55
	-10.0	0.76	0.00	0.00	0.00	0.00	0.00	0.15
1.2 (Hall)	20.0 (Clean)*	0.72	0.28	0.50	0.83	0.25	0.42	0.89
	10.0	0.72	0.15	0.42	0.62	0.18	0.32	0.86
	5.0	0.67	0.08	0.30	0.72	0.10	0.20	0.82
	0.0	0.66	0.00	0.12	0.36	0.02	0.04	0.73
	-5.0	0.73	0.00	0.00	0.00	0.00	0.00	0.50
	-10.0	0.76	0.00	0.00	0.00	0.00	0.00	0.12
1.5 (Church)	20.0 (Clean)*	0.72	0.18	0.42	0.82	0.15	0.30	0.85
	10.0	0.69	0.08	0.35	0.80	0.10	0.22	0.82
	5.0	0.66	0.04	0.22	0.63	0.05	0.12	0.78
	0.0	0.66	0.00	0.08	0.30	0.00	0.02	0.68
	-5.0	0.74	0.00	0.00	0.00	0.00	0.00	0.45
	-10.0	0.77	0.00	0.00	0.00	0.00	0.00	0.08

Physical Visualization of the Failure Surface Fig. 9 visualizes the destructive mechanics of the Acoustic Perceptual Load on the raw audio waveform through a 6×6 matrix of controlled acoustic degradations. Along the horizontal axis (SNR Decay), decreasing the signal-to-noise ratio from a clean reference (20.0 dB) to extreme noise (-10.0 dB) progressively induces amplitude masking; the distinct dynamic envelope of the target speech is steadily engulfed by the additive noise floor, culminating in a solid acoustic block at -10.0 dB where intensity contrasts are completely obliterated. Orthogonally, along the vertical axis (Reverberation RT_{60}), increasing reverberation from an anechoic recording (0.0 s) to church-scale smearing (1.5 s) introduces severe temporal degradation. The sharp transient boundaries essential for phoneme demarcation transition into diffuse, exponentially decaying tails, causing adjacent syllables to bleed into one another. The bottom-right quadrant highlights the compounded destructive extreme: a signal stripped of both its amplitude contrast and temporal boundaries. This simultaneous obliteration of acoustic features physically explains the catastrophic failure of cascaded ASR systems documented in our quantitative evaluations.

Fig. 10 visualizes the acoustic substrate underlying these quantitative trends by rendering the log-Mel spectrograms across the corresponding 6×6 $RT_{60} \times SNR$ grid. These spectrogram panels clearly reveal the distinct time-frequency signatures of each degradation dimension. Along the horizontal axis (SNR Decay), as the SNR decreases from 20.0 dB to -10.0 dB, the fundamental formant structure and harmonic content are progressively buried by broadband noise energy. This vertical masking effect ultimately renders phonemic boundaries completely indistinguishable at extreme noise levels. Orthogonally, along the vertical axis (Reverberation RT_{60}), increasing reverberation introduces severe temporal smearing, which visually manifests as the horizontal elongation of spectral events. At church-scale reverberation ($RT_{60} = 1.5$ s), these formant transitions are stretched so significantly that they overlap and bleed across adjacent syllable boundaries. Ultimately, the

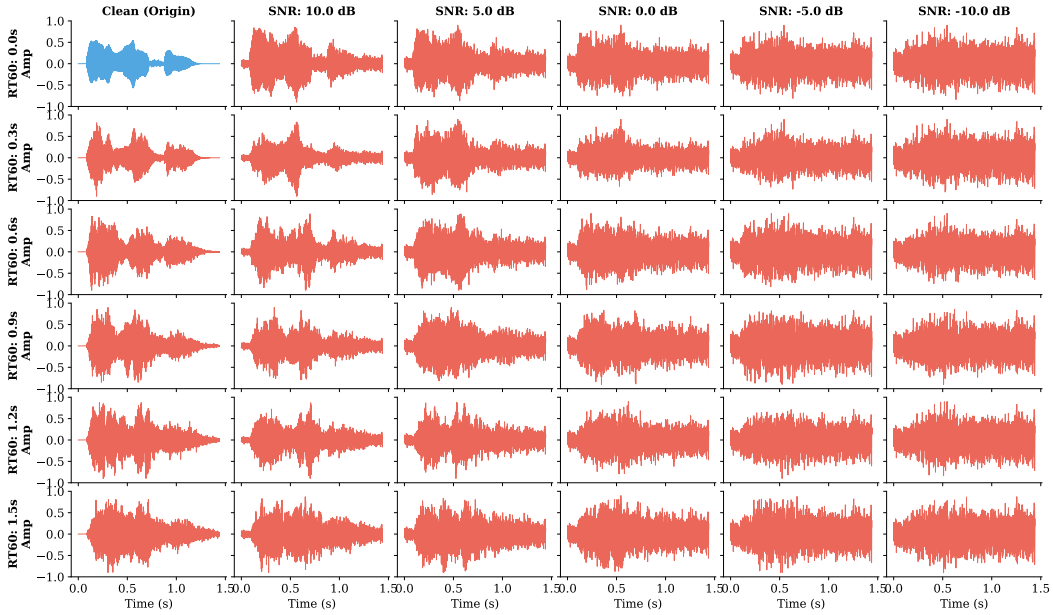

Figure 9: APL-Coupled Waveform Visualization across SNR and Reverberation Dimensions.

bottom-right quadrant demonstrates the catastrophic coupling of these effects, physically explaining the model failures observed in our quantitative APL evaluations.

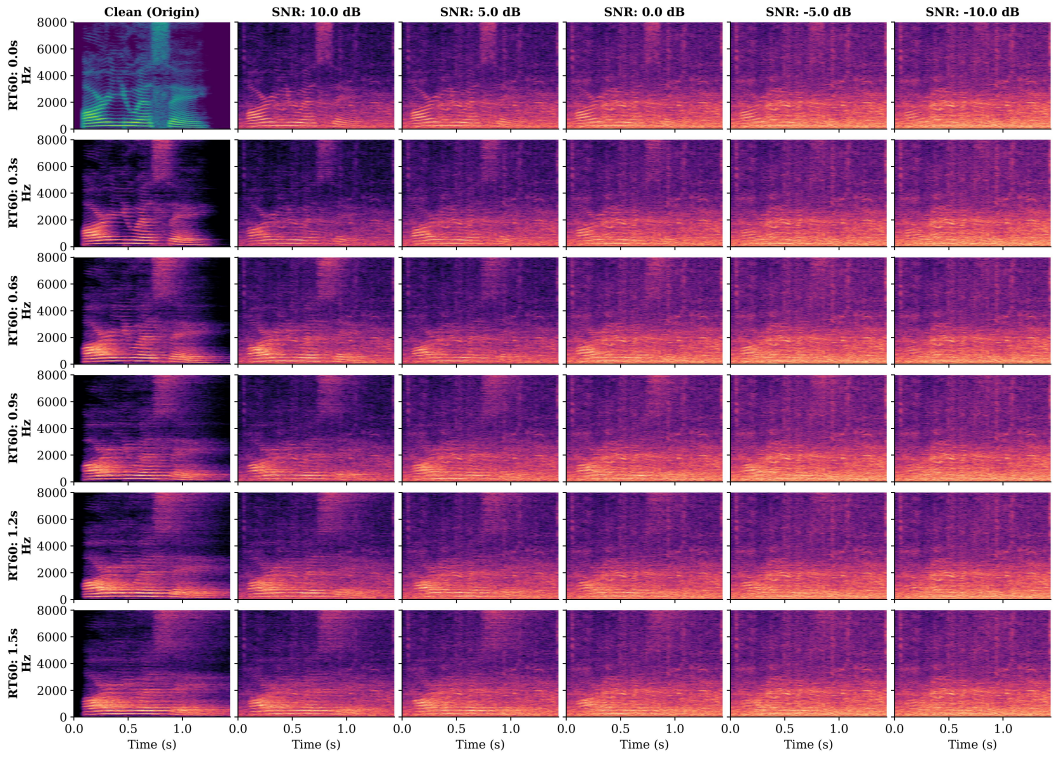

Figure 10: APL-Coupled Spectrogram Visualization across Three Degradation Dimensions.

Fig. 11 isolates the frequency selectivity of environmental distractors by comparing their log-Mel spectrograms at a fixed SNR of 0 dB. By analyzing four canonical noise categories, the spectrograms

vividly reveal the frequency-selective nature of acoustic masking. **Wind** (Low Ω) concentrates its energy below 500 Hz, leaving the upper formants of the speech signal clearly visible and intact. **Water** (Medium Ω) introduces a broadband but sparse energy distribution across 0–4 kHz that only partially obscures mid-frequency formants. In contrast, **Bus** noise (High Ω) applies dense low-to-mid frequency masking (50–2000 Hz) that directly overlaps with the fundamental frequency and the first two formants of speech. Finally, **Talk** or Babble noise (Severe Ω) creates a near-complete spectral overlap with the critical target speech band (300–3400 Hz), producing the highest masking density and the greatest perceptual difficulty. This precise frequency-domain behavior accounts directly for why the models' failure boundaries in Fig. 6 align so closely with the Ω contours mapped in Fig. 7: the more a distractor overlaps the specific formant regions that carry phonemic identity, the earlier and steeper the model's collapse.

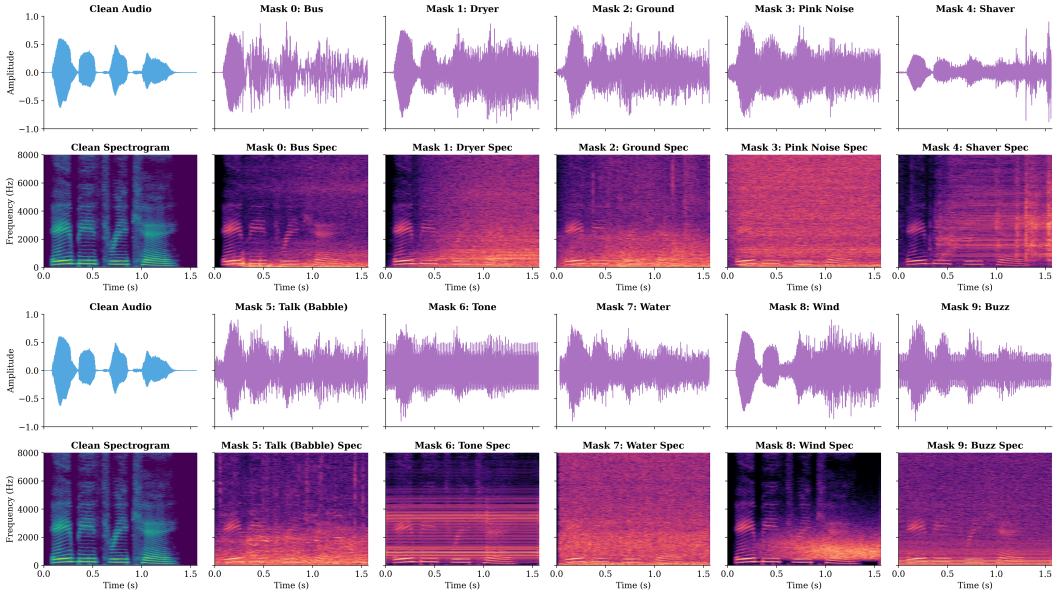

Figure 11: Spectral Masking Density Analysis across Noise Types.