Robust CAPTCHA Using Audio Illusions in the Era of Large Language Models: from Evaluation to Advances

Anonymous ACL submission

Abstract

CAPTCHAs are widely used by websites to block bots and spam by presenting challenges that are easy for humans but difficult for automated programs to solve. To improve accessibility, audio CAPTCHAs are designed to compliment visual ones. However the robustness of audio CAPTCHAs against advanced Large Audio Language Models (LALMs) and Automatic Speech Recognition (ASR) models remains unclear. In this paper, we introduce AI-CAPTCHA, a unified framework that offers (i) an evaluation framework, ACEVAL, which includes advanced LALM- and ASR-based solvers, and (ii) a novel audio CAPTCHA approach, ILLUSIONAUDIO, leveraging audio illusions. Through extensive evaluations of seven widely deployed CAPTCHAs, we show that most methods can be solved with high success rates by the advanced LALMs and ASR models, exposing critical security weaknesses. To address this, we design a new CAPTCHA approach, ILLUSIONAUDIO, which exploits perceptual illusion cues rooted in human auditory mechanisms. Extensive experiments demonstrate that our method defeats all tested LALMs- and ASR-based attacks while achieving a 100% human pass rate, significantly outperforming existing methods.

1 Introduction

Completely Automated Public Turing Tests to Tell Computers and Humans Apart (CAPTCHAs) are widely used to protect online services from automated bot attacks (Von Ahn et al., 2003; Gossweiler et al., 2009). Most deployed CAPTCHAs rely on vision-based challenges (Ding et al., 2025b,a), such as distorted text or image recognition. While effective, these visual challenges are inaccessible to people with visual impairments (PVIs) (Fanelle et al., 2020; Shirali-Shahreza and Shirali-Shahreza, 2007). To address this accessibility gap, audio CAPTCHAs were introduced as an alternative modality, enabling users to solve auditory challenges instead (Gao et al., 2010; Alnfiai, 2020).

Existing audio CAPTCHAs (Szegedy et al., 2017; He et al., 2016; Hunt, 2014; Fanelle et al., 2020; Labs, 2025) can be broadly categorized into two types: content-based and rule-based. Content-based audio CAPTCHAs (Abdullah et al., 2022; Aubry et al., 2025; Tam et al., 2008) require users to listen to spoken words or digits and transcribe the content (Fanelle et al., 2020). These schemes assume that speech perception and transcription are straightforward for humans but challenging for automated systems. To enhance robustness, many deployed content-based audio CAPTCHAs incorporate acoustic perturbations, such as background noise, distortions, or overlapping sounds. In contrast, rule-based audio CAPTCHAs (Fanelle et al., 2020; Labs, 2025) move beyond transcription, requiring users to follow audio instructions or identify specific sound events. This design tests higher-level auditory reasoning rather than raw speech recognition.

Existing audio CAPTCHAs effectively counter traditional bots (Fanelle et al., 2020). However, it is unclear whether their security assumptions remain valid given recent advances in Automatic Speech Recognition (ASR) and Large Audio Language Models (LALMs). Modern ASR systems achieve near-human performance on noisy and distorted speech, and LALMs further extend these capabilities by reasoning over complex audio inputs. These advances weaken the assumption that audio CAPTCHAs can reliably distinguish humans from automated solvers. Tasks such as transcribing corrupted speech and interpreting audio instructions may no longer separate human users from AI-driven systems. Therefore, systematic evaluation of current audio CAPTCHAs against advanced models is necessary. If performance is inadequate, new audio CAPTCHA designs should be developed.

To address the research gaps, we introduce AI-

CAPTCHA, a unified framework comprising two components: (1) ACEVAL, an evaluation framework for systematically assessing the robustness of audio CAPTCHA schemes against advanced LALM-based and ASR-based solvers, and (2) ILLUSIONAUDIO, a novel audio CAPTCHA design that leverages audio illusions.

ACEVAL incorporates two types of AI-driven solvers. The first is an LALM-based solver that directly reasons over audio input to generate answers, representing a new class of attacks against audio CAPTCHAs. The second is an ASR-based solver that follows a two-stage pipeline: audio is first transcribed by an ASR model, and the resulting transcript is then processed by a Large Language Model (LLM) to produce the final response. Both solvers ultimately produce an audio-based answer, rather than relying solely on ASR to transcribe the audio.

Using ACEVAL, we assess seven deployed audio CAPTCHA schemes, including four content-based (Geetest, Google, MTCaptcha, and Telephone audio CAPTCHAs) and three rule-based (Math, Character, and Arkose Labs audio CAPTCHAs). We assess their robustness against three advanced LALMs (Qwen-Audio-Chat, SeaLLMs-Audio-7B, and Qwen2-Audio-7B-Instruct) under zero-shot and chain-of-thought prompting, and two leading ASR models (GPT-4o-Transcript and GPT-4o-mini-Transcript) under prompt-guided and non-prompt-guided settings, with GPT-4o used for downstream reasoning. In addition, we conduct a user study with 63 human participants to assess the usability of these audio CAPTCHAs by measuring human success rates across multiple attempts.

Our evaluation yields two key findings. First, both LALM-based and ASR-based solvers achieve high bypass rates (41.67% and 49.99%, respectively) against existing deployed audio CAPTCHAs, highlighting their vulnerabilities to AI-driven attacks. Second, most existing audio CAPTCHA schemes impose substantial difficulty on human users, with an average first-attempt success rate of only 61.90%.

Based on these findings, we introduce ILLUSIONAUDIO, a new audio CAPTCHA scheme that leverages the sine-wave speech illusion—a perceptual phenomenon in which humans can recognize speech from sparse acoustic cues that lack explicit linguistic structure. While such signals remain intelligible to humans, they are challenging for LALM-based and ASR-based solvers to interpret. Beyond the illusion itself, ILLUSIONAUDIO incorporates additional mechanisms, such as irreversible audio transformations, to further resist automated solvers.

Extensive experiments using our evaluation framework demonstrate that ILLUSIONAUDIO defeats all tested LALM-based and ASR-based solvers, achieving a 0% bypass rate. A user study with 63 participants, including PVIs, shows that ILLUSIONAUDIO achieves a 100% first-attempt success rate, substantially outperforming existing audio CAPTCHA schemes in both security and usability.

In summary, we make the following contributions:

- We introduce AI-CAPTCHA, a unified framework comprising ACEVAL, an evaluation framework that employs advanced LALM-based and ASR-based solvers to assess audio CAPTCHA robustness, and ILLUSIONAUDIO, a novel audio CAPTCHA design based on audio illusions.
- Using ACEVAL, we conduct a comprehensive evaluation of seven widely deployed audio CAPTCHA schemes, revealing critical security vulnerabilities and usability limitations in existing designs.
- We propose ILLUSIONAUDIO, which leverages sine-wave speech illusions to create a perceptual gap between humans and AI. Experimental results demonstrate that ILLUSIONAUDIO achieves 0% bypass rate against all tested AI solvers while maintaining 100% first-attempt human success rate.

Ethical Considerations. Our research focuses on the security and usability of audio CAPTCHAs. All user studies were approved by our institutional review board and raise no ethical concerns. We develop AI-CAPTCHA to enhance web security by effectively distinguishing human users from bots. Details of our research and ethical declaration are provided on our website (Website, 2025): `https://sites.google.com/view/aicaptcha/`.

2 Background and Related Work

2.1 Audio CAPTCHAs

Audio CAPTCHAs are a class of challenge-response tests that leverage human auditory perception to defend against automated abuse (Gao et al., 2010; Fanelle et al., 2020; Alnfiai, 2020). They were introduced as an accessible alternative

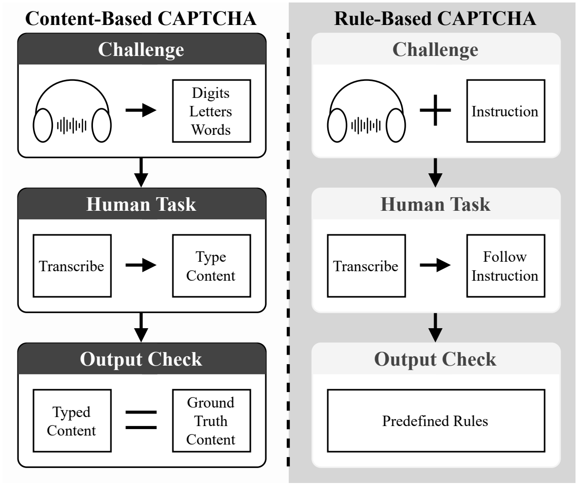


Figure 1: Architecture of content-based Audio CAPTCHAs and rule-based Audio CAPTCHAs.

to visual CAPTCHAs and have become an important component of usable web security (Saini and Bala, 2013). As illustrated in Figure 1, existing audio CAPTCHA schemes can be categorized into two types: content-based and rule-based. Content-based audio CAPTCHAs (Abdullah et al., 2022; Aubry et al., 2025; Tam et al., 2008) typically require users to transcribe spoken digits or letters. In contrast, rule-based audio CAPTCHAs (Fanelle et al., 2020; Labs, 2025) 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 (e.g., LALMs) has not been systematically examined.

2.2 Large Audio Language Models

Recent advances in LLMs have provided empirical evidence for scaling laws, reshaping how intelligent systems are trained and evaluated (Achiam et al., 2023; Team et al., 2023; Gu et al., 2024; Ge et al., 2023). By leveraging large-scale data and advanced training techniques (Liu et al., 2021), modern LLMs exhibit strong capabilities in perception and reasoning. Building on this foundation, LALMs extend LLMs to the audio domain by jointly modeling speech and text, enabling reasoning over complex audio inputs and multimodal instructions (Chu et al., 2023, 2024; Liu et al., 2025). Advanced LALMs, such as Qwen-Audio-Chat, SeaLLMs-Audio-7B, and Qwen2-Audio-7B-Instruct, have demonstrated strong performance on various audio understanding tasks (Wang et al., 2025, 2024; Yang et al., 2024), including audio question answering and audio-text reasoning.

2.3 Automatic Speech Recognition

ASR systems are designed to transcribe audio signals into text. Early approaches relied on modular pipelines combining Hidden Markov Models (HMMs) with Deep Neural Networks (DNNs) (Li, 2021; Nayeem et al., 2025). Subsequent research shifted toward end-to-end frameworks, including Connectionist Temporal Classification (CTC), attention-based encoder–decoder architectures, and transducer-based approaches (Bahdanau et al., 2014; Chorowski et al., 2015). Modern ASR systems leverage deep learning and large-scale datasets to achieve accurate and robust speech-to-text transcription. These systems can now handle a wide range of speech patterns and accents, making them effective at transcribing audio CAPTCHA challenges.

2.4 Audio Illusions

Audio illusions (Deutsch, 1974; Tiippana, 2014; Boebinger and Hicks, 2022) are perceptual phenomena in which human listeners experience sounds in ways that diverge from their physical acoustic structure. Such illusions arise from the human auditory system's ability to infer structure and meaning beyond explicit signal cues. Representative examples include the *McGurk effect*, where conflicting visual and auditory inputs alter perceived speech (Tiippana, 2014), and the *Shepard tone*, which creates the illusion of an endlessly rising pitch from a finite signal (Shepherd, 2017). Particularly relevant to our work is the *sine-wave speech illusion*, in which natural speech is reduced to sparse sinusoidal components. These signals remain intelligible to humans despite lacking the conventional spectral features that AI models typically rely on (Boebinger and Hicks, 2022). This perceptual asymmetry, where humans can interpret signals that AI systems cannot, enables the construction of audio CAPTCHAs that are solvable by humans yet challenging for automated solvers.

2.5 Related Work

Audio CAPTCHA Attacks. Prior work has demonstrated that traditional audio CAPTCHAs are vulnerable to ASR-based attacks (Tam et al., 2008; Sano et al., 2013). These attacks typically transcribe distorted speech using noise-robust ASR models. More recent work has shown that modern ASR systems can bypass many deployed audio CAPTCHAs despite noise injection (Abdullah et al., 2022; Aubry et al., 2025). However, these attacks only transcribe the audio rather than give the answer of audio CAPTCHAs. Moreover, they do not hold the reasoning capabilities needed to solve the rule-based CAPTCHAs. Therefore, the design of the solver in ACEVAL extends this line of research by evaluating the emerging threat of LALMs, which combine speech recognition with semantic reasoning, and add downstream reasoning function provided by LLMs for usual ASR systems.

Perceptual CAPTCHA Designs. Researchers have explored perceptual phenomena to create CAPTCHAs that exploit human cognitive advantages. In the visual domain, Illusion-CAPTCHA (Ding et al., 2025a) leverages visual illusions to create challenges that are easy for humans but difficult for vision models. Our work extends this concept to the audio domain, using sine-wave speech illusions to create a similar perceptual gap. Unlike adversarial perturbations that add imperceptible noise to confuse models (Zhang et al., 2021), our approach fundamentally transforms the signal in a way that preserves human intelligibility while removing features that AI models rely on.

Accessibility in CAPTCHAs. Audio CAPTCHAs were originally designed for accessibility, particularly for PVIs (Fanelle et al., 2020; Shirali-Shahreza and Shirali-Shahreza, 2007). However, studies have shown that many audio CAPTCHAs impose significant usability burdens on all users (Shi et al., 2020; Kulkarni and Fadewar, 2018). Our work aims to address both security and accessibility by designing an audio CAPTCHA that is both robust against AI attacks and easy for humans including PVIs to solve.

Given these advances in AI capabilities and the limitations of existing CAPTCHA designs, we next formalize the threat model that our approach addresses.

3 Threat Model

Audio CAPTCHAs are designed to prevent automated adversaries from bypassing access control mechanisms deployed by web services (Szegedy et al., 2017; He et al., 2016; Hunt, 2014; Fanelle et al., 2020; Labs, 2025). We consider an attacker whose goal is to automatically solve audio CAPTCHA challenges at scale to enable automated account creation, credential stuffing, or other abuse.

Attacker Capabilities. We assume a black-box setting where the attacker can obtain audio challenges through repeated interactions with the target website, a realistic assumption given that many CAPTCHA services are publicly accessible (Gao et al., 2010; Fanelle et al., 2020; Alnfiai, 2020). The attacker can leverage off-the-shelf ASR models and advanced LALMs to process and solve the audio challenges. We assume the attacker has knowledge of the general structure of the audio CAPTCHA (e.g., content-based vs. rule-based) but does not have access to the internal implementation or training data of the CAPTCHA generation system.

Out of Scope. We do not consider adaptive attacks where adversaries fine-tune models specifically on sine-wave speech data, though we discuss this as an important direction for future work in our limitations (Section 7). We also do not consider attacks that bypass the CAPTCHA system entirely (e.g., through human solving farms or browser automation vulnerabilities), as these are orthogonal to the audio CAPTCHA design.

To systematically evaluate audio CAPTCHA security under this threat model, we next introduce our evaluation framework and proposed defense.

4 AI-CAPTCHA Framework

We introduce AI-CAPTCHA, a unified framework comprising two components: (i) ACEVAL, an evaluation framework for systematically assessing the robustness of audio CAPTCHA schemes against modern LALM-based and ASR-based solvers; and (ii) ILLUSIONAUDIO, a novel audio CAPTCHA design that leverages audio illusion effects.

4.1 ACEVAL: Evaluation Framework

As shown in Figure 2a, ACEVAL employs two AI-driven solvers based on LALMs and ASR models.

LALM-based Solver. The LALM-based solver employs LALMs to solve audio CAPTCHA challenges in an end-to-end manner. Given an audio CAPTCHA, the solver feeds the raw audio signal into the LALM, which performs audio perception and semantic reasoning. For content-based audio CAPTCHAs, the LALM recognizes spoken content, salient acoustic cues, or contextual semantics. For rule-based audio CAPTCHAs, the model infers and applies the implicit rules specified by the challenge. The LALM produces a natural-language response describing its interpretation, from which the final answer is extracted automatically. We instantiate the LALM-based solver using three advanced models: Qwen-Audio-Chat, SeaLLMs-Audio-7B, and Qwen2-Audio-7B-Instruct.

ASR-based Solver. The ASR-based solver adopts a modular pipeline that decouples speech recognition from semantic reasoning. Given an audio CAPTCHA, the solver first transcribes the audio into text using an ASR model. We consider both prompt-guided and non-prompt-guided transcription settings, where prompts optionally provide task-specific hints to improve transcription accuracy. The resulting transcript is then passed to a downstream LLM (GPT-4o), which performs semantic parsing and reasoning. For content-based audio CAPTCHAs, the solver extracts the relevant spoken content; for rule-based schemes, it interprets the embedded rules and derives the answer. We employ two of the best performing ASR models by the time of writing this paper [1]: GPT-4o-Transcript and GPT-4o-mini-Transcript.

Using ACEval, we evaluate seven deployed audio CAPTCHA schemes and conduct a user study with 63 participants (Section 5). Our findings reveal that existing audio CAPTCHAs are simultaneously vulnerable to AI-driven solvers and difficult for human users. To address these limitations, we propose IllusionAudio, a new audio CAPTCHA approach based on audio illusions.

4.2 IllusionAudio: Our Method

As illustrated in Figure 2, IllusionAudio comprises three modules: Automated Audio Generation, Illusionary Audio Generation, and Audio CAPTCHA Generation. We describe each module below.

Automated Audio Generation. This module efficiently generates a dataset of short, intelligible audio samples from text prompts without manual annotation.

Starting from an initial prompt p_0, we iteratively generate candidate audio clips using a text-to-speech (TTS) model (hexgrad, 2024). At iteration t, we synthesize K candidate clips: $\mathcal{X}_t = \{f_{\text{TTS}}(p_t, \theta, \xi_k)\}_{k=1}^{K}$, where ξ_k are random seeds that ensure diversity and reproducibility. Clips exceeding 2 seconds are discarded. Each remaining clip is assigned an intelligibility score $s_t(\mathcal{X}_t^k) = s(\mathcal{X}_t^k; p_t)$, which evaluates whether the clip meets human listening requirements, including loudness and tonal quality. Each clip is also transcribed by an ASR model to produce $\hat{y}(\mathcal{X}_t^k)$. The transcription is compared with the prompt p_t, and pronunciation differences are used to iteratively refine the generation process.

Algorithm 1 Automated Audio Generation

```
Require: initial prompt $p_0$; candidates $K$; target size $N_{\text{target}}$; max duration $T_{\max}=2$ s; threshold $\tau$; refinement budget $R$.
$\mathcal{D} \leftarrow \varnothing$;   $t \leftarrow 0$;   $p_t \leftarrow p_0$
loop
  $\mathcal{X}_t \leftarrow \{ f_{\text{TTS}}(p_t; \theta, \xi_k) \}_{k=1}^{K}$
  $\mathcal{X}_t \leftarrow \{ x \in \mathcal{X}_t \mid T(x) \le T_{\max} \}$
  for all $x \in \mathcal{X}_t$ do
    $\hat{y}(x) \leftarrow g_{\text{ASR}}(x)$
    $s_t(x) \leftarrow s(x; p_t)$
  $\mathcal{X}_t^{\text{good}} \leftarrow \{ x \in \mathcal{X}_t \mid s_t(x) \ge \tau \}$
  $m_t \leftarrow N_{\text{target}} - |\mathcal{D}|$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \textsc{Take}(\mathcal{X}_t^{\text{good}}, m_t)$
  if $t < R$ then
    $\Phi_t \leftarrow \textsc{Feedback}\Big(\{(\hat{y}(x), x)\}_{x \in \mathcal{X}_t^{\text{good}}}\Big)$
    $p_{t+1} \leftarrow \mathcal{R}(p_t, \Phi_t)$
  else
    $p_{t+1} \leftarrow p_t$
  $t \leftarrow t + 1$;   $p_t \leftarrow p_{t+1}$
return $\mathcal{D}$
```

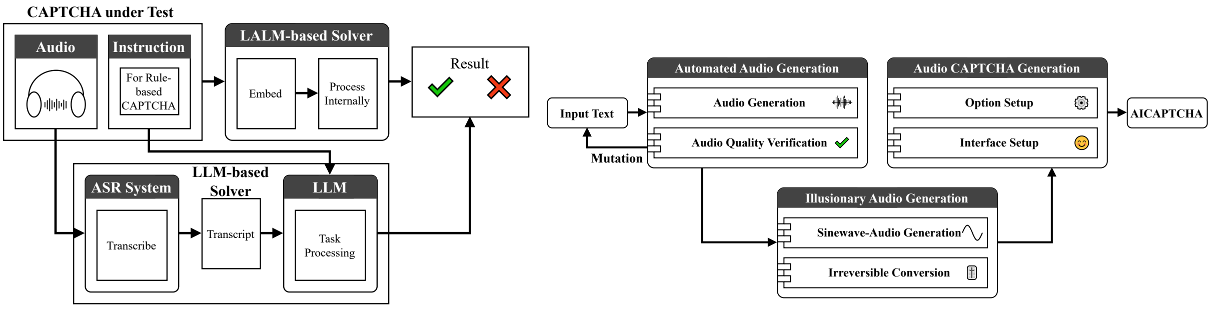
We retain clips satisfying $s_t(\mathcal{X}_t^k) \ge \tau$, where τ is an intelligibility threshold, and add them to the dataset $\mathcal{D}$. This process repeats until the target dataset size $|\mathcal{D}| \ge N_{\text{target}}$ is reached. The iterative refinement based on intelligibility scores and transcription feedback ensures high-quality audio samples.

Illusionary Audio Generation. We employ a sine-wave speech transformation that reduces natural speech to a small set of time-varying sinusoids while preserving the global temporal envelope and coarse formant trajectories. This representation remains intelligible to humans but significantly degrades AI model performance, creating a stable human–AI gap that our audio CAPTCHA exploits.

Formally, given an audio signal $x \in \mathbb{R}^L$, we

---

[1] According to OpenAI (OpenAI, 2025), the GPT-4o-Transcript model has lower word error rate and better language recognition accuracy than the original Whisper models. In addition, according to a third party benchmark (Dilmegani and Alper, 2025), GPT-4o-Transcript has the lowest WER and CER among all evaluated models.



(a) ACEVAL. (b) ILLUSIONAUDIO.

Figure 2: Overview of AI-CAPTCHA.

generate its sine-wave surrogate as:

$$\tilde{x} = \mathcal{S}_{\text{sine}}(x; \psi), \tag{1}$$

where $\mathcal{S}_{\text{sine}}$ denotes the sine-wave renderer with analysis–synthesis parameters $\psi = (\text{window size}, \text{hop length}, \text{number of formants})$. The resulting $\tilde{x}$ serves as our illusionary audio.

Although $\tilde{x}$ retains human intelligibility, deterministic sine-wave speech could potentially be inverted by specialized reconstruction attacks. To mitigate this risk, we apply a randomized *Irreversible Conversion* module that perturbs the signal through random downsampling:

$$\hat{x} = \mathcal{M}(\tilde{x}; \phi), \qquad \phi \sim \Pi, \tag{2}$$

where $\mathcal{M}$ applies downsampling to $\tilde{x}$, $\phi \in [0.5, 0.8]$ is the downsampling factor, and Π is a uniform distribution over this range. The information loss from downsampling prevents reconstruction of the original signal while preserving human intelligibility. The resulting non-invertible audio signals $\hat{x}$ form the illusionary audio dataset $\hat{\mathcal{D}}$ used for CAPTCHA generation.

Audio CAPTCHA Generation. Using the clean-audio dataset $\mathcal{D}$ and the illusionary dataset $\hat{\mathcal{D}}$, we adopt an option-based interaction where users select from a small set of audio clips rather than typing a transcript. This design, consistent with industry practice (e.g., Arkose Labs), improves accessibility for users with limited typing proficiency.

A naive multiple-choice formulation introduces three challenges: (1) *Guessability*: elevated success rates from random selection; (2) *Low-level heuristics*: solvers may exploit cues such as root mean square (RMS) amplitude; and (3) *Pattern leakage*: fixed option templates may become learnable.

To address these issues, we randomize both the task framing and option composition for each challenge. Users identify the illusionary audio that matches the linguistic content of a clean reference. Candidate options are dynamically sampled, and the unmodified clean audio is sometimes included to break amplitude-based heuristics. This design increases difficulty for automated solvers while keeping the task straightforward for humans.

Additionally, we support both full-length and segment-wise playback, allowing users to replay only relevant portions rather than entire clips, thereby improving usability.

5 Evaluation

Our evaluation addresses three research questions:

- **RQ1: Robustness:** How effectively can ILLUSIONAUDIO resist automated attacks from advanced LALM-based and ASR-based solvers?
- **RQ2: Usability:** Does ILLUSIONAUDIO remain user-friendly and easy to solve for human users?
- **RQ3: Ablation Study:** How do some important components of ILLUSIONAUDIO contribute to defending against automated attacks?

5.1 Experiment Setup

Baselines. We evaluate seven widely deployed audio CAPTCHAs: four content-based schemes (Geetest (Geetest, 2023), Google (Google, 2023), MTCaptcha (MTCaptcha, 2023), and Telephone-Audio (Lepture, 2023)) and three rule-based schemes (Math (Fanelle et al., 2020), Character (Fanelle et al., 2020), and Arkose Labs (Labs, 2025)):

- **Geetest**: Presents spoken numbers without background noise; users transcribe what they hear.
- **Google**: Uses distorted speech with background noise for transcription.
- **MTCaptcha**: Adds music as background noise to spoken numbers.

- **Telephone-Audio**: Simulates telephone-quality audio with background noise.
- **Math**: Embeds a spoken math problem that users must solve.
- **Character**: Asks users to count occurrences of a specific digit.
- **Arkose Labs**: Requires users to identify and classify audio content.

Table 4: Human solving attempts for IllusionAudio under different settings. IllusionAudio w/o Clean-Audio removes the clean reference audio, while IllusionAudio follows the default design with the clean reference played before the illusionary audio.

Method	Human Participants			
Attempt Times	One	Two	Three	>Three
IllusionAudio w/o Clean-Audio	0.00%	0.00%	3.33%	96.67%
IllusionAudio	100.00%	0.00%	0.00%	0.00%

We generate CAPTCHA samples using official open-source implementations with default configurations. To ensure fair and consistent evaluation, we collect 30 samples per scheme, yielding 210 audio CAPTCHA instances in total.

Metrics. We use two complementary metrics. To assess robustness, we report the *bypass rate*, which represents the percentage of CAPTCHA instances successfully solved by AI solvers. A lower bypass rate indicates stronger security. For usability, we report the *success rate*, which represents the percentage of human users who successfully solve the CAPTCHA on each attempt. Higher success rates indicate better accessibility. By comparing AI bypass rates with human success rates, we assess whether a scheme achieves the desired goal of being "AI-hard" yet "human-easy".

Computation Platform. All experiments are conducted on a workstation with the Intel Xeon 6 6787P CPU and two NVIDIA RTX 5090 GPUs.

5.2 Results

5.2.1 RQ1: Robustness of Our Method

We evaluate the robustness of IllusionAudio and seven widely deployed audio CAPTCHA schemes using IllusionAudioEval. The evaluation employs three LALM-based solvers: Qwen-Audio-Chat (Qwen), SeaLLMs-Audio-7B (SeaLLMs), and Qwen2-Audio-7B-Instruct (Qwen2), and two ASR-based solvers: GPT-4o-Transcript (GPT4o) and GPT-4o-mini-Transcript (GPT4o-mini), as described in Section 4.1. For fair comparison, we generate the same number of samples for IllusionAudio as for the baseline schemes.

Table 1 shows that all seven existing schemes exhibit high bypass rates under LALM-based and ASR-based solvers, indicating broad vulnerability to AI-driven attacks. Content-based CAPTCHAs such as Geetest and Google are easily bypassed by LALMs, with bypass rates near 100%. ASR-based solvers achieve similarly high bypass rates across most schemes.

In contrast, IllusionAudio achieves a 0% bypass rate against all tested LALM-based and ASR-based solvers, demonstrating strong robustness against automated attacks. This superior performance stems from the perceptual asymmetry created by sine-wave speech: the audio remains intelligible to humans but lacks the spectral features that AI models rely on for recognition (Section 2).

5.2.2 RQ2: Usability of Our Method

To assess usability, we conduct a user study with 240 audio CAPTCHA instances spanning seven baseline schemes and our method. We recruit 63 participants, including 36 PVIs and 27 sighted users. Each participant solves 12 distinct CAPTCHAs, yielding 756 total trials. We randomly partition the 240 CAPTCHAs into 20 batches of 12 items, with each batch evaluated by three participants [2]. For each CAPTCHA instance, we record the maximum number of attempts required by any evaluator as the measure of solving difficulty.

As shown in Table 2, most existing audio CAPTCHA schemes pose substantial usability challenges. While Geetest achieves a 100% first-attempt success rate, other content-based schemes perform worse: Google achieves only 23.33% and MTCaptcha 80%. Similar issues arise for rule-based schemes: Math attains 80%, but Character and Arkose Labs achieve only 50% and 76.67%, respectively. These results indicate that existing designs impose significant cognitive burdens, often requiring multiple attempts.

In contrast, IllusionAudio achieves a 100% first-attempt success rate, demonstrating superior usability alongside its strong security properties.

---

[2] More details about the design of the user study, such as the user interface design of the CAPTCHA instances and the information of the recruited participants can be found in Appendix A.

Table 1: AI bypass rates of ILLUSIONAUDIO and existing audio CAPTCHAs against LALM-based and ASR-based solvers. Lower bypass rate indicates stronger security.

Method		**LALM**						**ASR/ASR-LLM**			
Prompt-Mode		**Zero-Shot**			**Chain-of-Thought**			**Non-Prompt-Guide**		**Prompt-Guide**	
Model-Name		**Qwen**	**SeaLLMs**	**Qwen2**	**Qwen**	**SeaLLMs**	**Qwen2**	**GPT4o-mini**	**GPT4o**	**GPT4o-mini**	**GPT4o**
Content-based Audio CAPTCHA	**Geetest**	93.33%	96.66%	100.00%	96.66%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Google	76.67%	80.00%	96.67%	83.33%	80.00%	96.67%	70.00%	80.00%	83.33%	80.00%
	MTCaptcha	46.66%	16.66%	30.00%	43.33%	13.33%	33.33%	10.00%	10.00%	16.66%	16.66%
	Telephone	6.67%	3.33%	6.67%	10.00%	3.33%	3.33%	10.00%	10.00%	20.00%	30.00%
Rule-based Audio CAPTCHA	**Math**	23.30%	0.00%	0.00%	63.33%	10.00%	3.33%	13.33%	20.00%	23.33%	23.33%
	Character	16.67%	16.67%	86.67%	26.67%	16.67%	86.67%	93.33%	93.33%	93.33%	96.66%
	Arkoselabs	13.33%	3.33%	10.00%	46.66%	3.33%	6.66%	23.33%	23.33%	30.00%	30.00%
Our Method	**ILLUSIONAUDIO**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 2: Human success rates for ILLUSIONAUDIO and existing audio CAPTCHAs by number of attempts. Higher first-attempt success rate indicates better usability.

Method		**Human Participants**			
Attempt Times		**One**	**Two**	**Three**	**> Three**
Content-based Audio CAPTCHA	**Geetest**	100.00%	0.00%	0.00%	0.00%
	Google	23.33%	33.33%	26.67%	16.67%
	MTCaptcha	80.00%	20.00%	0.00%	0.00%
	Telephone	23.33%	36.67%	30.00%	10.00%
Rule-based Audio CAPTCHA	**Math**	80.00%	10.00%	10.00%	0.00%
	Character	50.00%	16.67%	33.33%	0.00%
	Arkoselabs	76.67%	6.67%	3.33%	13.33%
Our Method	**ILLUSIONAUDIO**	100.00%	0.00%	0.00%	0.00%

5.3 RQ3: Ablation Study

We study the impact of two key design choices in ILLUSIONAUDIO: the irreversible conversion module and the clean reference audio (Section 4.2).

5.3.1 Impact of Irreversible Conversion Module

We compare ILLUSIONAUDIO with a variant where the irreversible conversion module is disabled. To evaluate robustness against lightweight heuristic attacks, we additionally test an RMS-based solver that exploits amplitude patterns.

?? reports the results. ILLUSIONAUDIO remains robust against LALM-based and ASR-based attacks regardless of whether irreversible conversion is enabled (0% bypass rate in both cases). However, robustness against RMS-based attacks critically depends on this module: without it, the RMS-based attack achieves 100% bypass rate by exploiting amplitude cues; with irreversible conversion enabled, the attack completely fails (0% bypass rate). These results demonstrate that the irreversible conversion module is essential for eliminating low-level amplitude patterns that simple heuristic solvers could exploit.

5.3.2 Impact of Clean Reference Audio

We compare our default setting, where clean reference audio is played before the illusionary audio, with a variant that presents only the illusionary audio.

Table 4 shows a stark contrast. Without the clean reference, 96.67% of participants required more than three attempts, and none succeeded within the first two attempts. With the clean reference, all participants (100%) solved the task on their first attempt. These results confirm that the clean reference audio is critical for human usability, serving as a perceptual "priming" cue that enables listeners to decode the sine-wave speech illusion.

6 Conclusion

In this paper, we present ACEVAL, a novel method to assess audio CAPTCHAs in the AI era. Through extensive evaluation of seven widely deployed audio CAPTCHA schemes, we show that existing designs are vulnerable to AI-driven solvers like Large Audio Language Models and Automatic Speech Recognition systems, which easily bypass them with high success rates. Additionally, we observe significant usability challenges, as many audio CAPTCHAs are not only easily solved by AI but are also difficult for humans, especially those with disabilities. These findings reveal the limitations of current systems, which either fail to thwart AI solvers or impose substantial cognitive burdens on users. In response to these challenges, we propose ILLUSIONAUDIO, which exploits the sine-wave speech illusion, a perceptual phenomenon in which speech is encoded into sparse sinusoidal components. While such signals remain intelligible to human listeners, they are much more challenging for AI solvers to interpret due to their lack of conventional acoustic features that AI models typically rely on. By leveraging these perceptual cues, ILLUSIONAUDIO creates a significant perceptual gap between humans and machines, ensuring that only humans, regardless of background, can reliably solve the challenge.

7 Limitations

While our results demonstrate the robustness and usability of AI-CAPTCHA under the evaluated settings, several limitations should be acknowledged.

User Study Scale. Although we recruited 63 participants, including both PVIs and sighted users, the sample size limits the statistical power of our analysis. Future studies with larger and more diverse populations would enable more rigorous validation across demographic subgroups.

Controlled Environment. Our experiments were conducted in a controlled setting that does not fully capture real-world usage conditions. Factors such as background noise, audio playback devices, hearing ability variations, and network latency may influence both usability and security. Extending the evaluation to more realistic deployment scenarios would provide a more comprehensive assessment.

Adaptive Attacks. While ILLUSIONAUDIO defeats all tested off-the-shelf AI solvers, we did not evaluate against adversaries who specifically fine-tune models on sine-wave speech data. An attacker with access to sine-wave speech examples could potentially train specialized recognition models, leading to an arms race between attack and defense. Possible countermeasures include periodically rotating transformation parameters, combining sine-wave speech with other perceptual phenomena, or incorporating behavioral signals beyond audio recognition. Investigating such adaptive attacks and their mitigations is an important direction for future work.

We view these limitations as natural directions for future research and believe that addressing them will strengthen the practical deployment of perceptually grounded audio CAPTCHA systems.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mrim Alnfiai. 2020. A novel design of audio captcha for visually impaired users. *International Journal of Communication Networks and Information Security*, 12(2):168–179.

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Dana Boebinger and Jarrod Hicks. 2022. How do illusions trick the brain? - MIT McGovern Institute. `https://mcgovern.mit.edu/2022/05/13/use-your-illusion/`. [Accessed 30-09-2025].

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *ArXiv*, abs/1506.07503.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Diana Deutsch. 1974. An auditory illusion. *Nature*, 251(5473):307–309.

Cem Dilmegani and Şevval Alper. 2025. Speech-to-text benchmark: Deepgram vs. whisper in 2026. `https://research.aimultiple.com/speech-to-text/`. [Accessed 02-10-2025].

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

Ziqi Ding, Shangzhi Xu, Wei Song, and Yuekang Li. 2025b. Ngcaptcha: A captcha bridging the past and the future.

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

Haichang Gao, Honggang Liu, Dan Yao, Xiyang Liu, and Uwe Aickelin. 2010. An audio captcha to distinguish humans from computers. In *2010 third international symposium on electronic commerce and security*, pages 265–269. IEEE.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, and 1 others. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568.

Geetest. 2023. float — gt4.geetest.com. https://gt4.geetest.com/demov4/voice-float-en.html. [Accessed 01-10-2025].

Google. 2023. Recaptcha demo — google.com. https://www.google.com/recaptcha/api2/demo. [Accessed 01-10-2025].

Rich Gossweiler, Maryam Kamvar, and Shumeet Baluja. 2009. What's up captcha? a captcha based on image orientation. In *Proceedings of the 18th international conference on World wide web*, pages 841–850.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

hexgrad. 2024. Kokoro TTS - a Hugging Face Space by hexgrad. https://huggingface.co/spaces/hexgrad/Kokoro-TTS. [Accessed 02-10-2025].

Earl B Hunt. 2014. *Artificial intelligence*. Academic Press.

Sushama Kulkarni and Hanumant Fadewar. 2018. Audio captcha techniques: A review. In *Proceedings of the Second International Conference on Computational Intelligence and Informatics: ICCII 2017*, pages 359–368. Springer.

Arkose Labs. 2025. Arkose matchkey (captcha software) | arkose labs. https://www.arkoselabs.com/arkose-matchkey/. [Accessed 30-09-2025].

Lepture. 2023. Github - lepture/captcha: A captcha library that generates audio and image captchas. https://github.com/lepture/captcha. [Accessed 01-10-2025].

Jinyu Li. 2021. Recent advances in end-to-end automatic speech recognition. *ArXiv*, abs/2111.01690.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. 2025. Seallms-audio: Large audio-language models for southeast asia. https://github.com/DAMO-NLP-SG/SeaLLMs-Audio.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

MTCaptcha. 2023. Demo | mtcaptcha code generator — service.mtcaptcha.com. https://service.mtcaptcha.com/mtcv1/demo/index.html. [Accessed 01-10-2025].

Md. Nayeem, Md Shamse Tabrej, Kabbojit Jit Deb, Shaonti Goswami, and Md. Azizul Hakim. 2025. Automatic speech recognition in the modern era: Architectures, training, and evaluation. *ArXiv*, abs/2510.12827.

OpenAI. 2025. Speech-to-text model powered by gpt-4o. https://platform.openai.com/docs/models/gpt-4o-transcribe. [Accessed 02-10-2025].

Baljit Singh Saini and Anju Bala. 2013. A review of bot protection using captcha for web security. *IOSR Journal of Computer Engineering*, 8(6):36–42.

Shotaro Sano, Takuma Otsuka, and Hiroshi G Okuno. 2013. Solving google's continuous audio captcha with hmm-based automatic speech recognition. In *International Workshop on Security*, pages 36–52. Springer.

John Shepherd. 2017. The musical coding of ideologies. In *Whose Music?*, pages 69–124. Routledge.

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

Mohammad Shirali-Shahreza and Sajad Shirali-Shahreza. 2007. Captcha for blind people. In *2007 IEEE international symposium on signal processing and information technology*, pages 995–998. IEEE.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Jennifer Tam, Jiri Simsa, Sean Hyde, and Luis Ahn. 2008. Breaking audio captchas. *Advances in Neural Information Processing Systems*, 21.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kaisa Tiippana. 2014. What is the mcgurk effect?

Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. 2003. Captcha: Using hard ai problems for security. In *International conference on the theory and applications of cryptographic techniques*, pages 294–311. Springer.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025. When audio and text disagree: Revealing text bias in large audio-language models. *arXiv preprint arXiv:2508.15407*.

Our Website. 2025. AICAPTCHA. `https://sites.google.com/view/aicaptcha/`. [Accessed 02-10-2025].

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Hongting Zhang, Qiben Yan, Pan Zhou, and Xiao-Yang Liu. 2021. Generating robust audio adversarial examples with temporal dependency. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3167–3173.

A Appendix

A.1 CAPTCHA Interface

The interface of our CAPTCHA system is designed with multiple features to enhance user interaction and accessibility. It supports both random multiple-choice and single-choice options, requiring users to listen to all available choices before making their selection. Additionally, users have the ability to listen to the options sequentially or replay individual options separately. To ensure a smooth experience, a brief training session is provided to introduce basic concepts, such as sine waves. This system advances traditional CAPTCHA designs by engaging deeper levels of human perceptual processing, making it more effective and interactive. Figure 3 shows the interface of our CAPTCHA. The design of our CAPTCHA system incorporates key features that enhance user interaction, accessibility, and security. The system supports both random multiple-choice and single-choice options, requiring users to listen to all available choices before selecting an answer. Users can replay or listen to options sequentially, providing greater control over the interaction and reducing frustration. Beyond these interaction mechanisms, the interface is carefully designed to balance usability with security constraints. Visual elements are kept minimal to avoid introducing unnecessary cognitive load, allowing users to focus primarily on auditory perception. Clear and intuitive controls are provided to guide users through the listening and selection process, ensuring that the system remains accessible to users with varying levels of technical proficiency. To further enhance robustness against automated attacks, the system enforces controlled interaction flows. For example, users are required to complete the audio playback before submitting a response, preventing automated agents from skipping directly to an answer. The randomization of option order and playback sequences further increases resistance to pattern-based attacks while preserving a consistent user experience.

A.2 Participants Information

Our recruitment process involved thorough documentation of participant information to ensure transparency and ethical use of data. Each participant was required to provide informed consent for their data to be used specifically for the research experiments presented in this paper. As part of the recruitment process, participants were fully informed about the scope of the study and the intended use of their data, and they agreed to participate under these terms. Recruitment was conducted through online video chats, during which we explained the study and monitored participants' behavior throughout the experiment. This approach allowed for real-time interaction, providing valuable insights into their actions and responses.

Throughout both the recruitment and experimental processes, we adhered to all ethical guidelines, prioritizing participant privacy and confidentiality. Personal data was anonymized, and identifying information was kept separate from the experiment data. Participants were informed that they could withdraw from the study at any time without facing any negative consequences, and were assured that their participation was entirely voluntary. Each participant voluntarily chose to take part in the study, with no financial compensation or incentives offered. Their involvement was driven solely by their consent to contribute to the research, ensuring there was no coercion or expectation of rewards for participation.

As shown in Figure 4, the participant pool exhibits a balanced gender distribution, with male participants accounting for 50.8% and female participants comprising 49.2% of the total sample. In terms of age, participants span a wide range, with the majority concentrated between 40 and 60 years


Illusion-based Audio CAPTCHA

QUESTION

Listen to the original audio, then select which options are sine-wave speech version of it.

Select one or more options

Figure 3: Interface of our CAPTCHA.

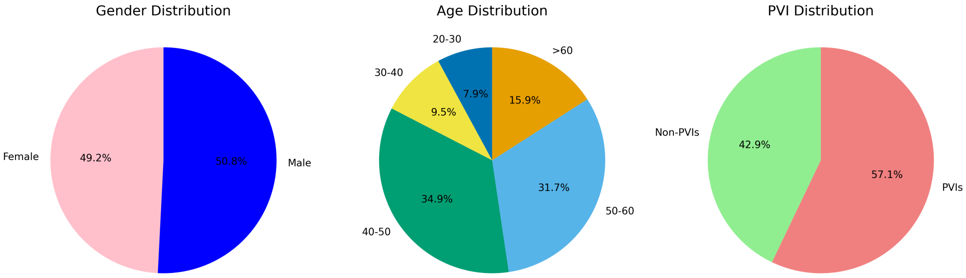


Figure 4: Demographic distributions of participants.

old. Specifically, individuals aged 40–50 represent the largest group at 34.9%, followed by those aged 50–60 at 31.7%. Younger participants aged 20–30 and 30–40 constitute 7.9% and 9.5% of the sample, respectively, while participants over 60 years old account for 15.9%. Regarding prior exposure to similar systems, 57.1% of participants fall into the category labeled as PVIs, while the remaining 42.9% are categorized as Non-PVIs, indicating a diverse range of participant backgrounds relevant to the study. Furthermore, participants are from multiple countries, but most choose not to disclose their country of origin. Therefore, we will not show their origin at this stage. Overall, this demographic composition ensures a heterogeneous participant population, supporting the generalizability of our experimental findings.