利用音频错觉的鲁棒 CAPTCHA：大语言模型时代的评估与进展

Ziqi Ding YunfengWan Wei Song YiLiu Gelei Deng Nan Sun Huadong Mo Jingling Xue Shidong Pan Yuekang Li

arXiv:2601.08516v1 [cs.SD] 13 Jan 2026

摘要

验证码被网站广泛用于阻止机器人和垃圾信息，其通过呈现对人类而言简单、但对自动化程序而言难以解决的挑战来实现这一目的。为提升可访问性，音频验证码被设计用于补充视觉验证码。然而，音频验证码抵御先进大型音频语言模型（LALM）和自动语音识别（ASR）模型的能力仍不明确。本文提出 AI-CAPTCHA，这是一个统一框架，包括：（i）评估框架 ACEVAL，其中集成了先进的基于 LALM 和 ASR 的求解器；以及（ii）一种利用音频错觉的新型音频验证码方法 ILLUSIONAUDIO。通过对七种广泛部署的验证码进行全面评估，我们表明，大多数方法都能被先进的大型音频语言模型和自动语音识别模型以较高成功率破解，暴露出严重的安全性弱点。为解决这一问题，我们设计了一种新的验证码方法 ILLUSIONAUDIO，该方法利用根植于人类听觉机制的感知错觉线索。大量实验表明，我们的方法能够击败所有经过测试的基于大型音频语言模型和自动语音识别的攻击，同时实现 100% 的人类通过率，显著优于现有方法。

1 引言

区分计算机与人类的全自动公开图灵测试（验证码）被广泛用于保护在线服务免受自动化机器人攻击（Von Ahn 等人，2003年；Gossweiler 等人，2009年）。大多数已部署的验证码依赖视觉挑战（Ding 等人，2025b，a），例如扭曲文本或图像识别。尽管这些视觉挑战有效，但视力障碍人士（视障人士）无法使用它们（Fanelle 等人，2020年；Shirali-Shahreza 和 Shirali-Shahreza，2007年）。为弥合这一无障碍性差距，音频验证码应运而生，作为一种替代方案模态，使用户能够改为解决听觉挑战（Gao 等人，2010年；Alnfiai，2020年）。

现有的音频验证码（Szegedy 等，2017；He 等，2016；Hunt，2014；Fanelle 等，2020；Labs，2025）大致可分为两类：基于内容的和基于规则的。基于内容的音频验证码（Abdullah 等，2022；Aubry 等，2025；Tam 等，2008）要求用户听取口语词语或数字，并转录其中的内容（Fanelle 等，2020）。这类方案假设人类能够轻松感知语音并进行转录，而自动化系统则难以做到。为增强鲁棒性，许多已部署的基于内容的音频验证码会加入声学扰动，例如背景噪声、失真或重叠声音。相比之下，基于规则的音频验证码（Fanelle 等，2020；Labs，2025）不再局限于语音转写，而是要求用户遵循音频指令或识别特定的声音事件。这种设计考查的是更高层次的听觉推理，而非单纯的语音识别。

现有音频验证码能够有效抵御传统机器人（Fanelle 等人，2020年）。然而，随着自动语音识别（ASR）和大型音频语言模型（LALM）近年来不断发展，其安全性假设是否依然成立尚不明确。现代自动语音识别系统在含噪失真语音上的表现已接近人类水平，而大型音频语言模型则通过对复杂音频输入进行推理，进一步拓展了这些能力。这些进展削弱了这样一种假设：音频验证码能够可靠地区分人类与自动化求解器。转录受损语音、理解音频指令等任务，或许已无法再将人类用户与人工智能驱动的系统区分开来。因此，有必要针对先进模型对当前音频验证码进行系统评估。如果其表现不足，就应开发新的音频验证码设计。

为弥补这些研究空白，我们提出了人工智能-

验证码，这是一个由两个部分组成的统一框架：(1) ACEVAL：一个评估框架，用于系统评估音频验证码方案抵御先进的大型音频语言模型和基于自动语音识别的求解器的稳健性；(2) ILLUSIONAUDIO：一种新型音频验证码设计，利用音频错觉实现。

ACEVAL 集成了两类人工智能驱动的求解器。第一类是基于 LALM 的求解器，可直接对音频输入进行推理并生成答案，代表了针对音频验证码的一类新型攻击。第二类是基于 ASR 的求解器，采用两阶段流程：首先由 ASR 模型转录音频，然后将生成的转录文本交由大语言模型（LLM）处理，以得出最终回答。这两类求解器最终都会生成基于音频的答案，而非仅依靠 ASR 转录音频。

我们使用 ACEVAL 评估了七种已部署的音频验证码方案，包括四种基于内容的方案（极验、谷歌、MTCaptcha 和电话音频验证码）以及三种基于规则的方案（数学、字符和 Arkose Labs 音频验证码）。我们评估了这些方案在零样本和思维链提示下抵御三种先进大型音频语言模型（Qwen-Audio-Chat、SeaLLMs-Audio-7B 和 Qwen2-Audio-7B-Instruct）的鲁棒性；还评估了它们在提示引导和非提示引导设置下抵御两种领先的 ASR 模型（GPT-4o-Transcript 和 GPT-4o-mini-Transcript）的鲁棒性，并使用 GPT-4o 进行下游推理。此外，我们还对 63 名人类参与者开展了用户研究，通过测量他们多次尝试的成功率，评估这些音频验证码的可用性。

我们的评估得出了两项关键发现。首先，基于大型音频语言模型和基于自动语音识别的求解器在攻击现有已部署的音频验证码时，均实现了较高的绕过率（分别为 41.67% 和 49.99%），凸显了这些验证码易受人工智能驱动的攻击。其次，大多数现有音频验证码方案都给人类用户带来了相当大的难度，首次尝试的平均成功率仅为 61.90%。

基于这些发现，我们推出 ILLUSIONAUDIO，这是一种新型音频验证码方案，利用正弦波语音错觉——一种感知现象，人类能够从缺乏明确语言结构的稀疏声学线索中识别语音。虽然这类信号对人类来说仍然可以理解，但基于大型音频语言模型和基于自动语音识别的求解器却难以解读。除了错觉本身，ILLUSIONAUDIO 还加入了不可逆音频变换等机制，以进一步抵御自动化求解器。

利用我们的评估框架开展的大量实验表明，ILLUSIONAUDIO 能够击败所有受测的基于大型音频语言模型和基于自动语音识别的求解器，绕过率为 0%。对 63 名参与者（包括视障人士）进行的用户研究显示，ILLUSIONAUDIO 的首次尝试成功率达到 100%，在安全性和可用性方面均显著优于现有音频验证码方案。

总之，我们的贡献如下：

- 我们提出了 AI-CAPTCHA，这是一种统一框架，由 ACEVAL 和 ILLUSIONAUDIO 构成：ACEVAL 是一个评估框架，采用先进的基于大型音频语言模型和基于自动语音识别的求解器来评估音频验证码的鲁棒性；ILLUSIONAUDIO 则是一种基于音频错觉的新型音频验证码设计。
- 我们使用 ACEVAL 对七种广泛部署的音频验证码方案进行了全面评估，揭示了现有设计中严重的安全漏洞和可用性限制。
- 我们提出了 ILLUSIONAUDIO，利用正弦波语音错觉在人类与人工智能之间创造感知差距。实验结果表明，面对所有经过测试的人工智能求解器，ILLUSIONAUDIO 的绕过率为 0%，同时人类首次尝试成功率保持在 100%。

伦理考量。 我们的研究聚焦于音频验证码的安全性和可用性。所有用户研究均经机构审查委员会批准，不存在伦理问题。我们开发 AI-CAPTCHA，旨在通过有效区分人类用户和机器人来增强网络安全。关于本研究及伦理声明的详细信息，请访问我们的网站（我们的网站，2025）：https://sites.google.com/view/aicaptcha/。

2 背景与相关工作

2.1 音频验证码

音频验证码是一类利用人类听觉感知来抵御自动化滥用的挑战—响应测试（Gao 等人，2010；Fanelle 等人，2020年；Alnfiai，2020年）。它们作为视觉验证码的一种无障碍替代方案应运而生

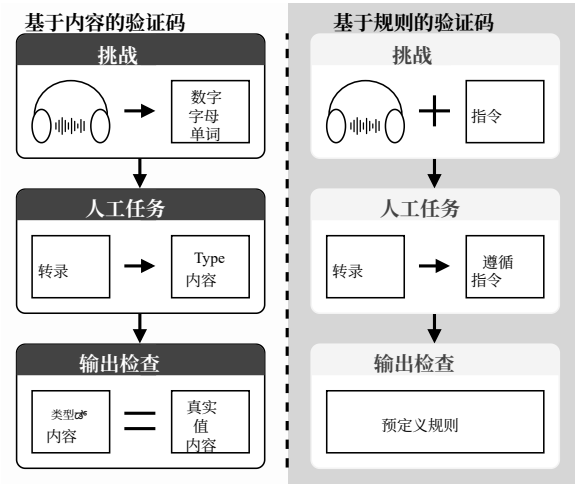


图1：基于内容的音频验证码和基于规则的音频验证码的架构。

，如今已成为易用网络安全的重要组成部分（Saini 和 Bala，2013）。如图1所示，现有音频验证码方案可分为两类：基于内容的和基于规则的。基于内容的音频验证码（Abdullah 等人，2022年；Aubry 等人，2025；Tam 等人，2008）通常要求用户转录口述数字或字母。相比之下，基于规则的音频验证码（Fanelle 等人，2020年；Labs，2025）要求用户结合音频内容和文本指令进行推理，将挑战从简单的语音转写转向更高层次的音频推理。尽管这些设计提高了对传统自动化攻击的抵御能力，但其抵御先进人工智能驱动的求解器（例如大型音频语言模型）的效果尚未得到系统性研究。

2.2 大型音频语言模型

大语言模型的最新进展为缩放定律提供了实证依据，重塑了智能系统的训练与评估方式（Achiam 等，2023年；Team 等，2023年；Gu 等，2024年；Ge 等，2023年）。借助大规模数据和先进的训练技术（Liu 等，2021年），现代大语言模型在感知和推理方面展现出强大的能力。在此基础上，大型音频语言模型通过联合建模语音和文本，将大语言模型扩展到音频领域，从而能够对复杂音频输入和多模态指令进行推理（Chu 等，2023年，2024年；Liu 等，2025）。Qwen-Audio-Chat、SeaLLMs-Audio-7B 和 Qwen2-Audio-7B-Instruct 等先进的大型音频语言模型，在各种音频理解任务上均展现出强劲表现（Wang 等，2025，2024年；Yang 等，2024年），包括音频问答和音频文本推理。

2.3 自动语音识别

自动语音识别系统旨在将音频信号转录为文本。早期方法依赖于将隐马尔可夫模型（HMM）与深度神经网络（DNN）结合的模块化流程（Li，2021年；Nayeem 等，2025）。随后，研究转向端到端框架，包括连接时序分类（CTC）、基于注意力的编码器–解码器架构以及基于转换器的方法（Bahdanau 等，2014年；Chorowski 等，2015年）。现代自动语音识别系统利用深度学习和大规模数据集，实现准确且稳健的语音转写。这些系统如今能够处理广泛的语音模式和口音，因此能够有效转录音频验证码挑战。

2.4 音频错觉

音频错觉（Deutsch，1974年；Tiippana，2014年；Boebinger 和 Hicks，2022年）是指一种感知现象：人类听众体验到的声音与其物理声学结构存在偏差。这类错觉源于人类听觉系统能够超越明确的信号线索，对结构和意义进行推断。具有代表性的例子包括麦格克效应，其中相互冲突的视觉和听觉输入会改变感知到的语音（Tiippana，2014年），以及谢泼德音调，它会让人产生有限信号中的音高不断上升、仿佛没有尽头的错觉（Shepherd，2017）。与我们的工作尤其相关的是正弦波语音错觉，在这种错觉中，自然语音被还原为稀疏正弦分量。尽管缺少人工智能模型通常依赖的传统频谱特征，人类仍然能够理解这些信号（Boebinger 和 Hicks，2022年）。这种感知不对称性，即人类能够解读而人工智能系统无法解读的信号，使得构造出由人类可解、但对自动化求解器构成挑战的音频验证码成为可能。

2.5 相关工作

音频验证码攻击。先前研究表明，传统音频验证码容易受到基于 ASR 的攻击（Tam 等人，2008；Sano 等人，2013）。这类攻击通常使用抗噪自动语音识别来转录失真语音

模型。近期研究表明，尽管注入了噪声，现代自动语音识别系统仍能绕过许多已部署的音频验证码（Abdullah 等人，2022年；Aubry 等人，2025）。然而，这些攻击只是转录音频，并不会给出音频验证码的答案。此外，它们不具备解决基于规则的验证码所需的推理能力。因此，ACEVAL 中求解器的设计将这条研究路线进一步拓展，通过评估大型音频语言模型这一新兴威胁。这类模型将语音识别与语义推理相结合，并为常规自动语音识别系统补充了由大语言模型提供的下游推理功能。

感知型验证码设计。研究人员探索了感知现象，以利用人类的认知优势设计验证码。在视觉领域，Illusion-CAPTCHA（Ding 等人，2025a）利用视觉错觉设计出对人类来说容易、但对视觉模型来说困难的挑战。我们的工作将这一概念扩展到音频领域，利用正弦波语音错觉制造类似的感知差距。与通过添加人眼难以察觉的噪声来混淆模型的对抗扰动（Zhang 等人，2021）不同，我们的方法从根本上改变信号，同时保留人类能够理解的特性，并去除人工智能模型所依赖的特征。

验证码的无障碍性。音频验证码最初是为无障碍需求设计的，尤其是面向视障人士（Fanelle 等人，2020年；Shirali-Shahreza 和 Shirali-Shahreza，2007）。然而，研究表明，许多音频验证码给所有用户都带来了显著的使用负担（Shi 等人，2020；Kulkarni 和 Fadewar，2018）。我们的工作旨在兼顾安全性和无障碍性，设计一种既能有效抵御人工智能攻击、又便于包括视障人士在内的人群解答的音频验证码。

鉴于人工智能能力的这些进展以及现有验证码设计的局限性，接下来我们将正式定义本方法所针对的威胁模型。

3 威胁模型

音频验证码旨在防止自动化对手绕过网络服务部署的访问控制机制（Szegedy 等，2017年；He 等，2016年；Hunt，2014年；Fanelle 等，2020年；Labs，2025）。我们考虑这样一种攻击者：其目标是自动大规模解决音频验证码挑战，从而实现自动创建账户、撞库攻击或其他滥用行为。

攻击者能力。我们假设攻击者处于黑盒设置中，可以通过与目标网站反复交互来获取音频挑战。鉴于许多验证码服务均可公开访问，这一假设符合实际情况（Gao 等人，2010年；Fanelle 等人，2020年；Alnfiai，2020年）。攻击者可以利用现成的自动语音识别模型和先进的大型音频语言模型来处理并解答音频挑战。我们假设攻击者了解音频验证码的一般结构（例如基于内容或基于规则），但无法访问验证码生成系统的内部实现或训练数据。

范围之外。我们不考虑对手专门针对正弦波语音数据微调模型的自适应攻击，不过我们在局限性（第 7 节）中将其列为未来工作的重要方向。我们也不考虑完全绕过验证码系统的攻击（例如通过人工解题农场或浏览器自动化漏洞），因为这些攻击与音频验证码设计并不相关。

为了在此威胁模型下系统地评估音频验证码的安全性，接下来我们将介绍评估框架和所提出的防御方案。

4 AI-CAPTCHA 框架

我们介绍 AI-CAPTCHA，这是一个统一框架，由两个部分组成：(i) ACEVAL，一种评估框架，用于系统地评估音频验证码方案抵御现代基于大型音频语言模型和基于自动语音识别的求解器的鲁棒性；(ii) ILLUSIONAUDIO，一种利用听觉错觉效果的新型音频验证码设计。

4.1 ACEVAL：评估框架

如图2a所示，ACEVAL 采用了两种由人工智能驱动的求解器，分别基于大型音频语言模型和自动语音识别模型。**基于 LALM 的求解器**。基于 LALM 的求解器利用大型音频语言模型，以端到端的方式解决音频验证码挑战。给定一个音频验证码，求解器会将原始音频信号输入大型音频语言模型，由其执行音频感知和语义推理。对于基于内容的音频验证码，大型音频语言模型会识别语音内容、显著声学线索或上下文语义。对于基于规则的音频验证码，模型会推断并应用挑战中指定的隐含规则。大型音频语言模型会生成一段

描述其解读的自然语言响应，并从中自动提取最终答案。我们采用三种先进模型来实例化基于 LALM 的求解器：Qwen-Audio-Chat、SeaLLMs-Audio-7B 和 Qwen2-Audio-7B-Instruct。

基于 ASR 的求解器。 基于 ASR 的求解器采用模块化流程，将语音识别与语义推理分离。给定一个音频验证码，求解器首先使用 ASR 模型将音频转录为文本。我们考虑提示引导和非提示引导两种转录设置，其中提示可选择性地提供特定任务的提示信息，以提高转录准确率。随后，将生成的转录文本传递给下游大语言模型（GPT-4o），由其进行语义解析和推理。对于基于内容的音频验证码，求解器会提取相关口语内容；对于基于规则的方案，则会解读其中嵌入的规则并推导答案。我们采用了撰写本文时性能最佳的两种 ASR 模型 [1]：GPT-4o-Transcript 和 GPT-4o-mini-Transcript。

我们使用 ACEVAL 评估了七种已部署的音频验证码方案，并对 63 名参与者开展了用户研究（第 5 节）。研究结果表明，现有音频验证码既容易受到人工智能驱动的求解器攻击，又令人工用户难以应对。为解决这些局限性，我们提出了 ILLUSIONAUDIO，这是一种基于音频错觉的新型音频验证码方法。

4.2 ILLUSIONAUDIO：我们的方法

如图 2 所示，ILLUSIONAUDIO 包含三个模块：自动音频生成、幻觉音频生成和音频验证码生成。下面将分别介绍各个模块。

自动音频生成。该模块无需人工标注，即可根据文本提示高效生成由简短、清晰易懂的音频样本组成的数据集。

从初始提示 p_0出发，我们使用文本转语音（TTS）模型迭代生成候选音频片段（hexgrad，2024年）。在第 t次迭代中，我们合成 K 个候选片段：$\mathcal{X}_t = \{f_{\text{TTS}}(p_t,\ \theta,\ \xi_k)\}_{k=1}^{K}$，其中 ξ_k 是用于确保多样性和可复现性的随机种子。超过 2 秒的片段将被丢弃。为每个剩余片段指定一个可懂度分数 $s_t(\mathcal{X}_t^k) = s(\mathcal{X}_t^k;\ p_t)$，用于评估该片段是否符合人类聆听要求，包括响度和音调质量。每个片段还会由 ASR 模型转录，生成 $\hat{y}(\mathcal{X}_t^k)$。随后将转录文本与提示 p_t进行比较，并根据发音差异迭代优化生成流程。

算法 1自动音频生成

要求：初始提示 p_0；候选项 K；目标大小 N_{target}；最长时长 $T_{\max}$=2 秒；阈值 τ；细化预算 R。

$\mathcal{D} \leftarrow \varnothing$；　$t \leftarrow 0$；　$p_t \leftarrow p_0$

循环

　$\mathcal{X}_t \leftarrow \{ f_{\text{TTS}}(p_t;\ \theta, \xi_k) \}_{k=1}^{K}$

　$\mathcal{X}_t \leftarrow \{ x \in \mathcal{X}_t \mid T(x) \le T_{\max} \}$

　对于所有 $x \in \mathcal{X}_t$ **执行**

　　$\hat{y}(x) \leftarrow g_{\text{ASR}}(x)$

　　$s_t(x) \leftarrow s(x; p_t)$

　$\mathcal{X}_t^{\text{good}} \leftarrow \{ x \in \mathcal{X}_t \mid s_t(x) \ge \tau \}$

　$m_t \leftarrow N_{\text{target}} - |\mathcal{D}|$

　$\mathcal{D} \leftarrow \mathcal{D} \cup$ 取值$(\mathcal{X}_t^{\text{good}},\ m_t)$

　如果 $t < R$ **那么**

　　$\Phi_t \leftarrow$ 反馈$\Big(\{(\hat{y}(x),\ x)\}_{x \in \mathcal{X}_t^{\text{good}}}\Big)$

　　$p_{t+1} \leftarrow \mathcal{R}(p_t,\ \Phi_t)$

　else

　　$p_{t+1} \leftarrow p_t$

　$t \leftarrow t+1$；　$p_t \leftarrow p_{t+1}$

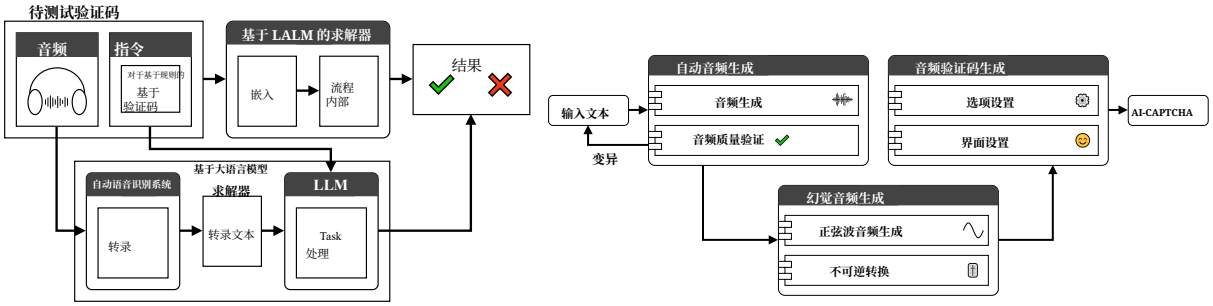
返回 $\mathcal{D}$

我们保留满足 $s_t(\mathcal{X}_t^k) \ge \tau$ 的片段，其中 τ 是可懂度阈值，并将其添加到数据集 $\mathcal{D}$ 中。此流程会重复进行，直到达到目标数据集规模 $|\mathcal{D}| \ge N_{\text{target}}$。基于可懂度分数和转录反馈的迭代优化，确保了音频样本的高质量。

幻觉音频生成。 我们采用正弦波语音变换，将自然语音简化为一组少量时变正弦波，同时保留全局时间包络和粗略的共振峰轨迹。这种表示对人类而言仍然清晰可懂，但会显著降低人工智能模型性能，从而形成稳定的人类与人工智能差距，而我们的音频验证码正是利用了这一点。

形式化地说，给定一个音频信号 $x \in \mathbb{R}^L$，我们

[1]根据 OpenAI（OpenAI，2025）的说法，GPT-4o-Transcript 模型的词错误率更低，语言识别准确率也高于原始 Whisper 模型。此外，根据一项第三方基准测试（Dilmegani 和 Alper，2025），GPT-4o-Transcript 在所有受评模型中的 WER 和 CER 均为最低。



(a) ACEVAL。 （b）ILLUSIONAUDIO。

图 2：AI-CAPTCHA 概述。

将其生成正弦波替代信号：

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

其中，$\mathcal{S}_{\text{sine}}$ 表示采用分析-合成参数ψ =（窗口大小，跳跃长度，共振峰数量）的正弦波渲染器。

生成的 $\tilde{x}$ 即作为我们的幻觉音频。

尽管 $\tilde{x}$ 仍具有人类可理解性，但确定性的正弦波语音可能会受到专门的重建攻击而被逆向还原。为降低这一风险，我们采用随机化的不可逆转换模块，通过随机下采样来扰动信号：

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

其中，$\mathcal{M}$ 对 $\tilde{x}$进行下采样，$\phi \in [0.5, 0.8]$ 为下采样因子，而 Π 是在此范围内的均匀分布。下采样造成的信息损失可防止原始信号被重建，同时保留人类可理解性。由此生成的不可逆音频信号$\hat{x}$ 构成用于验证码生成的幻觉音频数据集 $\mathcal{D}$。

音频验证码生成。 我们使用干净音频数据集 $\mathcal{D}$ 和幻觉音频数据集 $\mathcal{D}$，采用选项式交互，让用户从少量音频片段中进行选择，而非输入转录文本。这种设计符合行业惯例（例如 Arkose Labs），能提升打字能力有限用户的可访问性。

简单的多项选择题形式会带来三项挑战：(1)可猜测性：随机选择也可能获得较高的成功率；(2)低级启发式：求解器可能利用均方根（RMS）振幅等线索；(3)模式泄露：固定的选项模板可能被学习。

为应对这些问题，我们会针对每个挑战随机设置任务表述方式和选项构成。用户需要识别出与干净参考音频的语言内容相匹配的幻觉音频。候选选项会动态采样，有时还会加入未经修改的干净音频，以破除基于振幅的启发式方法。这种设计提高了自动化求解器的难度，同时让人类能够轻松完成任务。

此外，我们同时支持完整播放和分段播放，用户可以只重放相关片段，而不必播放整个音频片段，从而提升可用性。

5 评估

我们的评估围绕三个研究问题展开：

- **RQ1：鲁棒性：** ILLUSIONAUDIO 能在多大程度上抵御来自先进的大型音频语言模型（LALM）求解器和基于自动语音识别的求解器的自动化攻击?
- **研究问题2：可用性：**ILLUSIONAUDIO 对人类用户而言是否仍然友好且易于解答?
- **研究问题3：消融研究：**ILLUSIONAUDIO 的哪些重要组件有助于抵御自动化攻击?

5.1 实验设置

基线方法。我们评估了七种广泛部署的音频验证码：四种基于内容的方案（极验（极验，2023年）、谷歌（谷歌，2023年）、MTCaptcha（MTCaptcha，2023年）和电话音频（Lepture，2023年））以及三种基于规则的方案（数学（Fanelle 等人，2020年）、字符（Fanelle 等人，2020年）和 Arkose Labs（Labs，2025）：

- **极验**：呈现不带背景噪声的口述数字；用户转录听到的内容。
- **谷歌**：使用带有背景噪声的失真语音进行转录。
- **MTCaptcha**：在口述数字中加入音乐作为背景噪声。

• **Telephone-Audio**：模拟带有背景噪声的电话音质音频。

• **数学**：嵌入一道语音数学题，用户必须解答。

• **字符**：要求用户统计特定数字出现的次数。

• **Arkose Labs**：要求用户识别并分类音频内容。

表 4：不同设置下人类参与者解决 ILLUSIONAUDIO 的尝试次数。ILLUSIONAUDIO w/o Clean- Audio 移除了干净参考音频，而 ILLU- SIONAUDIO 遵循默认设计，先播放干净参考音频，再播放幻觉音频。

方法	人类参与者			
尝试次数	One	二	三	>三
ILLUSIONAUDIO（无干净音频）	0.00%	0.00%	3.33%	96.67%
ILLUSIONAUDIO	100.00%			

我们使用官方开源实现及默认配置生成验证码样本。为确保评估公平且一致，我们为每种方案收集 30 个样本，总计得到 210 个音频验证码实例。

指标。我们使用两项互补指标。为评估稳健性，我们报告绕过率，即人工智能求解器成功解答的验证码实例所占的百分比。绕过率越低，安全性越强。为评估可用性，我们报告成功率，即人类用户每次尝试成功解答验证码的百分比。成功率越高，无障碍性越好。通过比较人工智能绕过率与人类成功率，我们评估该方案是否实现了“难倒人工智能、方便人类”的预期目标。

计算平台。所有实验均在一台配备 Intel Xeon 6 6787P CPU 和两块 NVIDIA RTX 5090 GPU 的工作站上进行。

5.2 结果

5.2.1 RQ1：我们方法的稳健性

我们使用 IllusionAudioEval 评估 ILLUSIONAUDIO以及七种广泛部署的音频验证码方案的鲁棒性。评估采用三种基于 LALM 的求解器：Qwen-Audio- Chat（通义千问）、SeaLLMs-Audio-7B（SeaLLMs）和 Qwen2-Audio-7B-Instruct（通义千问2），以及两种基于自动语音识别的求解器：GPT-4o-Transcript（GPT4o）和 GPT-4o-mini-Transcript（GPT4o-mini），具体如第4.1节所述。为确保公平比较，我们为 ILLU- SIONAUDIO 生成与基线方案相同数量的样本。

表1 显示，在基于大型音频语言模型和基于自动语音识别的求解器面前，现有的七种方案都表现出较高的绕过率，表明它们普遍容易受到人工智能驱动的攻击。极验和谷歌等基于内容的验证码很容易被大型音频语言模型绕过，绕过率接近 100%。基于自动语音识别的求解器在大多数方案上也取得了同样高的绕过率。

相比之下，ILLUSIONAUDIO 针对所有测试的大型音频语言模型和基于自动语音识别的求解器，的绕过率均为 0%，展现出对自动化攻击的强大抵御能力。这一优异表现源于正弦波语音所造成的感知不对称性：音频对人类而言仍然清晰可辨，但缺少人工智能模型识别语音所依赖的频谱特征（第 2节）。

5.2.2 RQ2：我们的方法的可用性

为评估可用性，我们开展了一项用户研究，使用涵盖七种基线方案和我们的方法的 240 个音频验证码实例。我们招募了 63 名参与者，其中包括 36 名视障人士和 27 名明眼用户。每位参与者解答 12 个不同的验证码，共进行 756 次试验。我们将 240 个验证码随机划分为 20 批，每批 12 个项目，并由三名参与者评估每批 [2]。对于每个验证码实例，我们记录任一评估者所需的最大尝试次数，作为解题难度的衡量指标。

如表2所示，大多数现有的音频验证码方案都面临显著的可用性挑战。极验的首次尝试成功率达到 100%，而其他基于内容的方案表现较差：谷歌仅为 23.33%，MTCaptcha 为 80%。基于规则的方案也存在类似问题：数学验证码的成功率为 80%，而字符验证码和 Arkose Labs 分别仅为 50% 和 76.67%。这些结果表明，现有设计给用户带来了沉重的认知负担，往往需要多次尝试。

相比之下，ILLUSIONAUDIO 实现了 100% 的首次尝试成功率，展现出卓越的可用性，同时具备强大的安全性。

[2]关于用户研究设计的更多详情，例如验证码实例的用户界面设计以及招募参与者的信息，请参见附录 A。

表 1：ILLUSIONAUDIO 与现有音频验证码针对基于 LALM 和基于 ASR 的求解器的人工智能绕过率。绕过率越低，安全性越强。

方法		LALM						ASR/自动语音识别大语言模型			
提示模式		零样本			思维链			非提示引导		提示引导	
模型名称		Qwen	SeaLLMs	通义千问2	Qwen	SeaLLMs	通义千问2	GPT-4o mini	GPT-4o	GPT-4o mini	GPT-4o
基于内容音频验证码	极验	93.33%	96.66%	100.00%	96.66%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	谷歌	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%
	电话	6.67%	3.33%	6.67%	10.00%	3.33%	3.33%	10.00%	10.00%	20.00%	30.00%
基于规则音频验证码	Math	23.30%	0.00%	0.00%	63.33%	10.00%	3.33%	13.33%	20.00%	23.33%	23.33%
	字符	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%
我们的方法	ILLUSIONAUDIO	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

表 2：按尝试次数统计的 ILLUSIONAUDIO 与现有音频验证码的人类成功率。首次尝试成功率越高，表示可用性越好。

方法		人类参与者			
尝试次数		One	Two	三个	>三个
基于内容音频验证码	极验	100.00%	0.00%	0.00%	0.00%
	谷歌	23.33%	33.33%	26.67%	16.67%
	MTCaptcha	80.00%	20.00%	0.00%	0.00%
	电话	23.33%	36.67%	30.00%	10.00%
基于规则音频验证码	Math	80.00%	10.00%	10.00%	0.00%
	字符	50.00%	16.67%	33.33%	0.00%
	Arkoselabs	76.67%	6.67%	3.33%	13.33%
我们的方法	幻觉音频	100.00%			

5.3 研究问题3：消融研究

我们研究了 IllusionAudio中的两个关键设计选择所产生的影响：不可逆转换模块和干净参考音频（第 4.2节）。

5.3.1 不可逆转换模块的影响

我们将 ILLUSIONAUDIO 与禁用不可逆转换模块的变体进行比较。为评估其抵御轻量级启发式攻击的鲁棒性，我们还测试了一种利用振幅模式的基于均方根的求解器。

表3 报告了结果。无论是否启用不可逆转换，ILLUSIONAUDIO 对基于大型音频语言模型和基于自动语音识别的攻击均保持鲁棒（两种情况下绕过率均为 0%）。然而，其抵御基于均方根的攻击的鲁棒性则关键取决于该模块：禁用该模块时，基于均方根的攻击利用振幅线索实现了 100% 的绕过率；启用不可逆转换后，该攻击则彻底失效（绕过率为 0%）。这些结果表明，不可逆转换模块对于消除简单启发式求解器可能利用的低级振幅模式至关重要。

5.3.2 干净参考音频的影响

我们将默认设置（先播放干净的参考音频，再播放幻觉音频）与仅播放幻觉音频的变体进行比较。

表格4 呈现出鲜明的对比。没有干净参考音频时，96.67%的参与者需要尝试超过三次，且没有人在前两次尝试内成功。提供干净参考音频后，所有参与者（100%）都在第一次尝试时完成了任务。这些结果证实，干净参考音频对于人类的可用性至关重要，它作为一种感知“启动”线索，使听者能够解码正弦波语音错觉。

6 结论

本文提出 ACEVAL，这是一种用于评估人工智能时代音频验证码的新方法。通过对七种广泛部署的音频验证码进行全面评估，我们发现，现有设计容易受到大型音频语言模型和自动语音识别系统等人工智能驱动的求解器攻击，这些求解器能够轻松绕过验证，且成功率很高。此外，我们还观察到显著的可用性挑战：许多音频验证码不仅很容易被人工智能破解，对人类来说也很难，尤其是对残障人士而言。这些发现揭示了当前系统的局限性：它们要么无法抵御人工智能求解器，要么给用户带来沉重的认知负担。为应对这些挑战，我们提出 ILLUSIONAUDIO，该方法利用正弦波语音错觉这一感知现象，将语音编码为稀疏正弦分量。尽管人类听众仍能理解此类信号，但人工智能求解器却难以解读，因为这类信号缺少人工智能模型通常依赖的传统声学特征。借助这些感知线索，ILLUSIONAUDIO在人类与机器之间形成显著的感知差距，确保无论背景如何，只有人类才能可靠地完成挑战。

表 3：ILLUSIONAUDIO 在不同设置下的鲁棒性。ILLUSIONAUDIO w/o 不可逆转换表示禁用不可逆转换模块（即保留正弦波变换，但移除基于下采样的转换），而 ILLUSIONAUDIO 则遵循默认设计，启用不可逆转换。

方法	LALM						ASR				RMS
提示模式	零样本			思维链			非提示引导		提示引导		
模型名称	通义千问	SeaLLMs	通义千问2	通义千问	SeaLLMs	通义千问2	GPT-4o mini	GPT-4o	GPT-4o mini	GPT-4o	
ILLUSIONAUDIO 不含不可逆转换	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
ILLUSIONAUDIO	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

7 局限性

尽管我们的结果表明，在所评估的设置下，AI-CAPTCHA 具备稳健性和可用性，但仍应承认其存在若干局限性。

用户研究规模。 尽管我们招募了 63 名参与者，包括视障人士和明眼用户，但样本量限制了分析的统计功效。未来针对规模更大、更多元化人群开展的研究，将有助于在不同人口统计学亚组中进行更严格的验证。

受控环境。 我们的实验是在受控环境中进行的，无法充分反映真实世界的使用条件。背景噪声、音频播放设备、听力差异和网络延迟等因素，可能会影响可用性和安全性。将评估扩展到更贴近现实的部署场景，有助于进行更全面的评估。

自适应攻击。 虽然 ILLUSIONAUDIO 能够击败所有经过测试的现成 AI 求解器，但我们并未评估针对专门使用正弦波语音数据微调模型的对手。能够获取正弦波语音示例的攻击者，可能会训练专用的识别模型，从而引发攻击与防御之间的军备竞赛。可能的应对措施包括定期轮换变换参数、将正弦波语音与其他感知现象结合，或纳入音频识别之外的行为信号。研究此类自适应攻击及其缓解措施，是未来工作的重要方向。

我们将这些局限性视为未来研究的自然方向，并相信解决这些问题将有助于加强以感知为基础的音频验证码系统的实际部署。

参考文献

Hadi Abdullah、Aditya Karlekar、Saurabh Prasad、Muhammad Sajidur Rahman、Logan Blue、Luke A Bauer、VincentBindschaedler 和 Patrick Traynor。2022年。《以攻击作为防御：利用对自动语音识别系统的攻击设计稳健的音频验证码》。*arXiv* 预印本 *arXiv:2203.05408*。

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。2023年。GPT-4 技术报告。*ArXiv* 预印本 *arXiv:2303.08774*。

Mrim Alnfiai。2020年。为视障用户设计的一种新型音频验证码。《国际通信网络与信息安全期刊》，12(2):168–179。

Paul Aubry、Juliette Devoivre、Damien Carron、Simon Fernandez、Andrzej Duda 和 Maciej Korczy´nski。2025。利用自动语音识别模型绕过音频 reCAPTCHA。载于*2025 IEEE* 欧洲安全与隐私研讨会专题讨论会（*Eu-roS&PW*），第1–5页。IEEE。

Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。2014年。通过联合学习对齐与翻译的神经机器翻译。*CoRR*，abs/1409.0473。

Dana Boebinger 和 Jarrod Hicks。2022年。错觉如何欺骗大脑？——麻省理工学院 McGovern 研究所。https://mcgovern.mit.edu/2022/05/13/use-your-illusion/。[访问日期：30-09-2025]。

Jan Chorowski、Dzmitry Bahdanau、Dmitriy Serdyuk、Kyunghyun Cho 和 Yoshua Bengio。2015年。基于注意力的语音识别模型。*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 和 Jingren Zhou。2024年。通义千问2-Audio技术报告。*arXiv*预印本 *arXiv:2407.10759*。

YunfeiChu、JinXu、Xiaohuan Zhou、Qian Yang、Shiliang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou。2023年。Qwen-Audio：通过统一的大规模音频语言模型推进通用音频理解。*ArXiv* 预印本 *ArXiv:2311.07919*。

Diana Deutsch。1974年。听觉错觉。《自然》，251(5473)：307–309。

CemDilmeganiand ,SevvalAlper.2025. 语音转文本基准测试：Deepgram 与 whisper 在 2026 年的比较。https://research.aimultiple.com/speech-to-text/。[访问日期：02-10-2025]。

Ziqi Ding、Gelei Deng、Yi Liu、Junchen Ding、Jieshan Chen、Yulei Sui 和 Yuekang Li。2025a。《视觉错觉验证码：一种基于视觉错觉的验证码》。载于*2025* 年 *ACM* 网络会议论文集，第 3683–3691 页。

Ziqi Ding、Shangzhi Xu、Wei Song 和 Yuekang Li。2025b。Ngcaptcha：连接过去与未来的验证码。

Valerie Fanelle、Sepideh Karimi、Aditi Shah、Bharath Subramanian 和 Sauvik Das。2020年。盲人与人类：探索更易用的音频 {验证码}设计。载于第十六届可用隐私与安全研讨会（*SOUPS 2020*年），第 111–125 页。

Haichang Gao、Honggang Liu、Dan Yao、Xiyang Liu 和 Uwe Aickelin。2010年。一种用于区分人类与计算机的音频验证码。载于*2010*年第三届国际电子商务与安全研讨会，第 265–269 页。IEEE。

Yingqiang Ge、Wenyue Hua、Kai Mei、Juntao Tan、Shuyuan Xu、Zelong Li、Yongfeng Zhang 和另外 1 位作者。2023年。OpenAGI：当大语言模型遇上领域专家。神经信息处理系统进展，36：5539–5568。

极验。2023年。浮动 — gt4.geetest.com。https://gt4.geetest.com/demov4/voice-float-en.html。[访问日期：01-10-2025]。

谷歌。2023年。reCAPTCHA演示 — google.com。https://www.google.com/recaptcha/api2/demo. [访问日期01-10-2025]. Rich Gossweiler、Maryam Kamvar与Shumeet Baluja。2009年。《验证码怎么了？一种基于图像方向的验证码》。载于第*18*届万维网国际会议论文集，第841–850页。Jiawei Gu、Xuhui Jiang、Zhichao Shi、Hexiang Tan、Xuehao Zhai、Chengjin Xu、Wei Li、Yinghan Shen、Shengjie Ma、Honghao Liu等1人。2024年。大语言模型评审综述。 *ArXiv*预印本*arXiv:2411.15594*. Kaiming He、Xiangyu Zhang、Shaoqing Ren与Jian Sun。2016年。面向图像识别的深度残差学习。载于 *IEEE*计算机视觉与模式识别会议论文集，第770–778页。hexgrad。2024年。Kokoro文本转语音——hexgrad的Hugging Face Space。
h
ttps://huggingface.co/spaces/hexgrad/Kokoro-TTS. [访问日期02-10-2025]. Earl B Hunt。2014年。 人工智能。学术出版社。Sushama Kulkarni与Hanumant Fadewar。2018年。音频验证码技术：综述。载于 第二届计算智能与信息学国际会议论文集：*ICCII 2017*，第359–368页。施普林格。

Arkose Labs。2025。Arkose matchkey（验证码软件） | Arkose Labs。 https://www.arkoselabs.com/arkose-matchkey/. [访问日期30-09-2025].

Lepture。2023年。Github - lepture/captcha：一个可生成音频和图像验证码的验证码库。https://github.com/lepture/captcha。 [访问日期：01-10-2025]。

李金宇。2021年。端到端自动语音识别的最新进展。*ArXiv*，abs/2111.01690。

刘超群、Mahani Aljunied、陈贵珍、Hou Pong Chan、徐维文、容瑜和张文轩。2025。SeaLLMs-Audio：面向东南亚的大型音频语言模型。https://github.com/DAMO-NLP-SG/SeaLLMs-Audio。

刘家昌、DinghanShen、张一哲、Bill Dolan、Lawrence Carin和陈为柱。2021年。什么样的上下文示例适合GPT-3？ *arXiv*预印本 *arXiv:2101.06804*。

MTCaptcha. 2023年。演示 |mtcaptcha 验证码生成器— service.mtcaptcha.com。https://service.mtcaptcha.com/mtcv1/demo/index.html。 [访问于 01-10-2025]。

Md. Nayeem、Md Shamse Tabrej、Kabbojit Jit Deb、Shaonti Goswami 和 Md. Azizul Hakim。2025年。现代：架构、训练与评估中的自动语音识别。*ArXiv*，abs/2510.12827。OpenAI。2025。由 gpt-4o 驱动的语音转文本模型。
h
t
t
p
s://
p
l
atform.openai.com/docs/models/gpt-4o-transcribe。[访问日期：02-10-2025]。Baljit Singh Saini 和 Anju Bala。2013。使用验证码实现网络安全机器人防护的综述。 *IOSR*计算机工程期刊，8(6)：36–42。佐野翔太郎、大塚拓真和奥野博。2013。使用基于隐马尔可夫模型的自动语音识别破解谷歌的连续音频验证码。在国际安全研讨会，第 36–52 页。施普林格。John Shepherd。2017。意识形态的音乐编码。在 《谁的音乐？》，第 69–124 页。劳特利奇。石成辉、纪守领、刘千钧、刘长昌、陈岳峰、何源、刘哲、Raheem Beyah 和王挺。2020年。文本验证码已死？大规模部署与实证研究。在*2020*年*ACM SIGSAC*计算机与通信安全会议论文集，第1391–1406 页。Mohammad Shirali-Shahreza 和 Sajad Shirali- Shahreza。2007。面向盲人的验证码。在 *2007*年*IEEE*信号处理与信息技术国际研讨会，第995–998 页。IEEE。

Christian Szegedy、Sergey Ioffe、Vincent Vanhoucke 和 Alexander Alemi。2017。Inception-v4、Inception-ResNet 以及残差连接对学习的影响。在 *AAAI*人工智能会议论文集，第 31 卷。

Jennifer Tam、Jiri Simsa、Sean Hyde 和 Luis Ahn。2008年。破解音频验证码。神经信息处理系统进展，21。

Gemini团队、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 及其他1位作者。2023年。Gemini：一系列能力出众的多模态模型。*arXiv*预印本 *arXiv:2312.11805*。

Kaisa Tiippana。2014年。什么是麦格克效应?

Luis Von Ahn、Manuel Blum、Nicholas J Hopper 和 John Langford。2003年。验证码：利用高难度人工智能问题保障安全性。载于国际密码学理论与应用会议，第294–311页。施普林格。

Bin Wang、Xunlong Zou、Geyu Lin、Shuo Sun、Zhuohan Liu、Wenyu Zhang、Zhengyuan Liu、AiTi Aw 和 Nancy F Chen。2024年。AudioBench：音频大语言模型的通用基准测试。*arXiv*预印本 *arXiv:2406.16020*。

Cheng Wang、Gelei Deng、Xianglin Yang、Han Qiu 和 Tianwei Zhang。2025年。当音频和文本不一致时：揭示大型音频语言模型中的文本偏差。*arXiv*预印本 *arXiv:2508.15407*。

我们的网站。2025年。AI-CAPTCHA。https://sites.google.com/view/aicaptcha/。 [访问日期：02-10-2025]。

Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、Xiaohuan Zhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou 和另外 1 人。2024年。Air-bench：通过生成式理解对大型音频语言模型进行基准测试。*arXiv*预印本 *arXiv:2402.07729*。

Hongting Zhang、Qiben Yan、Pan Zhou 和 Xiao-Yang Liu。2021年。利用时间依赖生成鲁棒的音频对抗样本。载于第二十九届国际人工智能联合会议论文集，第 3167–3173 页。

A 附录

A.1 验证码界面

我们的验证码系统界面设计了多项功能，以提升用户交互体验和无障碍性。系统支持随机多选和单项选择选项，要求用户在选择前听完所有可选项。此外，用户可以依次收听各个选项，也可以单独重播选项。为确保体验流畅，系统提供简短的培训课程，介绍正弦波等基本概念。该系统通过激发更深层次的人类感知处理，推进了传统验证码设计，使其更加有效且具有交互性。图 3展示了我们的验证码界面。我们的验证码系统设计融入了多项关键功能，以提升用户交互体验、无障碍性和安全性。系统支持随机多选和单项选择选项，要求用户在选择答案前听完所有可选项。用户可以重播选项，也可以依次收听，从而更好地掌控交互过程并减少挫败感。除了这些交互机制外，界面还经过精心设计，在可用性与安全性限制之间取得平衡。视觉元素尽量精简，以避免造成不必要的认知负荷，让用户能够主要专注于听觉感知。界面提供清晰直观的控件，引导用户完成收听和选择流程，确保不同技术熟练程度的用户都能使用该系统。为进一步增强系统抵御自动化攻击的能力，系统对交互流程实施控制。例如，用户必须完成音频播放后才能提交回答，以防自动化代理直接跳过播放环节并提交答案。选项顺序和播放序列的随机化进一步增强了系统抵御基于模式的攻击的能力，同时保持一致的用户体验。

A.2 参与者信息

我们的招募流程包括对参与者信息进行详尽记录，以确保透明度和数据使用符合伦理规范。每位参与者都必须提供知情同意，允许将其数据专门用于本文介绍的研究实验。作为招募流程的一部分，我们向参与者充分说明了研究的范围及其数据的预期用途，他们也同意按照这些条款参与研究。招募通过在线视频聊天进行，期间我们向参与者解释研究内容，并在整个实验过程中观察其行为。这种方式支持实时互动，为了解他们的行动和反应提供了宝贵的见解。

在招募和实验流程的整个过程中，我们始终遵守所有伦理准则，



图 3：我们的验证码界面。

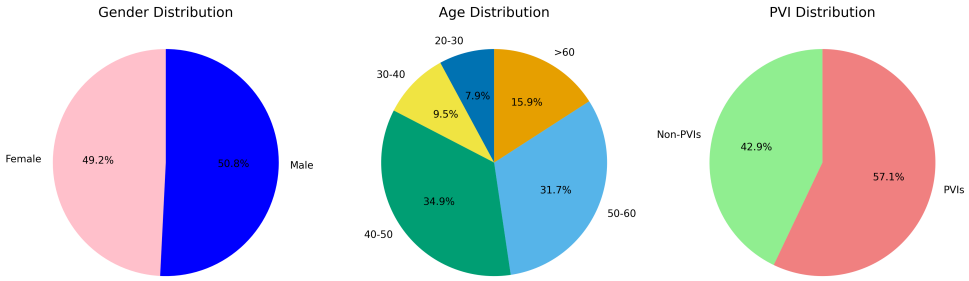


图 4：参与者的人口统计分布。

并将参与者的隐私与保密性放在首位。个人数据均经过匿名化处理，身份识别信息与实验数据分开保存。我们告知参与者，他们可以随时退出研究，且不会因此受到任何负面影响，并向他们保证参与完全出于自愿。每位参与者均自愿选择参加研究，且未获得任何经济补偿或激励。他们参与研究完全是出于同意为这项研究作出贡献，确保不存在任何强迫，也不抱有参与即可获得奖励的预期。

如图 4所示，参与者群体呈现出均衡的性别分布，其中男性参与者占样本总数的 50.8%，女性参与者占 49.2%。在年龄方面，参与者的年龄跨度较大，且大多数集中在 40 至 60 岁之间。具体而言，40–50 岁的参与者占比最高，为 34.9%；其次是 50–60 岁的参与者，占 31.7%。20–30 岁和 30–40 岁的年轻参与者分别占样本的 7.9% 和 9.5%，而 60 岁以上的参与者占 15.9%。关于此前接触类似系统的情况，57.1% 的参与者属于标记为视障人士的类别，其余 42.9% 被归类为非 PVIs，表明参与者背景具有多样性，与本研究相关。此外，参与者来自多个国家，但大多数人选择不透露其原籍国。因此，我们现阶段将不展示其来源。总体而言，这种人口构成确保了参与者群体的异质性，有助于提升实验结果的普遍适用性。