

1 研究任务介绍 (Introduction)

1.1 任务描述

本研究聚焦于 OpenAI Whisper 自动语音识别 (ASR) 模型的幻觉 (Hallucination) 问题。幻觉是指 ASR 模型在处理非语音音频 (如静音、噪声、环境声音) 时, 生成与原始音频没有语音或语义联系的文本输出的现象。

本研究的主要任务包括四个方面: 首先, 复现并验证 Whisper 模型在非语音音频上的幻觉现象; 其次, 量化分析不同类型音频的幻觉率及幻觉内容分布; 第三, 对比评估多种幻觉缓解方案的有效性; 最后, 验证模型在正常语音识别任务上的准确性。

1.2 研究意义

Whisper 作为目前最先进的开源语音识别模型之一, 在工业界和学术界有着广泛应用。然而, 其幻觉问题可能带来严重后果。在可靠性方面, 医疗、法律等关键领域的错误转录可能导致严重后果; 在安全性方面, 研究表明部分幻觉可能包含暴力、色情等有害内容; 在用户体验方面, 实时转录场景中的无意义输出会严重影响使用效果。因此, 深入研究 Whisper 的幻觉问题并探索有效的缓解方案具有重要的理论和实践价值。

1.3 国内外研究现状

Whisper 模型。Radford 等人 [1] 于 2022 年提出 Whisper 模型, 使用 680,000 小时的多语言弱监督数据进行训练, 采用 Encoder-Decoder Transformer 架构, 支持语音识别、翻译、语言识别等多任务。该模型在 LibriSpeech test-clean 数据集 [6] 上达到 2.5% 的词错误率 (WER), 接近人类水平。

幻觉问题研究。Barański 等人 [3] 在 ICASSP 2025 会议上发表了对 Whisper 幻觉现象的大规模调查研究。通过对 301,317 个非语音音频文件的测试, 发现幻觉发生率高达 40.3%, 其中 67% 的幻觉来自 1,270 个重复短语, 最常见的是 "thank you" (24.76%) 和 "thanks for watching" (10.32%) 等 YouTube 字幕相关内容。

Calm-Whisper 解决方案。Wang 等人 [2] 提出 Calm-Whisper 方法, 通过分析发现 Whisper 解码器中的特定注意力头 (#1, #6, #11) 是导致幻觉的主要原因。通过对这些 "crazy heads" 进行定向微调, 将幻觉率从 99.97% 降低至 15.51%, 同时保持 WER 仅增加约 0.07%。

2 研究内容和技术路线

2.1 研究内容

本研究围绕以下四个核心问题展开：

问题一：合成音频幻觉测试。本研究首先测试 Whisper 在静音、白噪声、粉红噪声等合成音频上的幻觉表现，分析不同音频长度对幻觉率的影响，并研究 Whisper 参数（如 `no_speech_threshold`）对幻觉的影响。

问题二：真实环境声音幻觉测试。使用 ESC-50 环境声音数据集 [5] 进行大规模测试，分析不同声音类别（动物、自然、城市等）的幻觉率差异，并统计幻觉内容的分布特征。

问题三：语音识别准确性验证。使用 LibriSpeech test-clean 数据集 [6] 测试模型的词错误率（WER），验证模型在正常语音识别任务上的可靠性。

问题四：缓解方案对比评估。实现并测试 VAD（语音活动检测）预处理方案和 BoH（幻觉词袋）后处理方案，对比各方案单独使用和组合使用的效果。

2.2 技术路线

2.2.1 实验框架设计

本研究采用模块化的实验框架，包括五个核心组件。Whisper 模型封装模块对 OpenAI Whisper 模型进行封装，支持不同参数配置的转录。音频处理模块提供音频加载、生成、预处理等功能。评估指标模块实现 WER、CER、幻觉率、循环率等指标计算。VAD 预处理模块基于能量检测进行语音活动检测。BoH 后处理模块使用 Aho-Corasick 算法进行幻觉短语匹配和过滤。

2.2.2 评估指标

词错误率（WER）：

$$WER = \frac{S + D + I}{N} \times 100\% \quad (1)$$

其中 S 为替换词数， D 为删除词数， I 为插入词数， N 为参考文本总词数。

幻觉率：

$$\text{Hallucination Rate} = \frac{\text{产生幻觉的样本数}}{\text{总样本数}} \times 100\% \quad (2)$$

3 实验结果及分析（重点内容）

3.1 实验结果

3.1.1 开发环境介绍

硬件环境方面，本实验使用 NVIDIA GeForce RTX 系列 GPU 进行加速计算，系统内存为 16GB 以上。软件环境方面，操作系统为 Windows 10/11，编程语言为 Python 3.9，深度学习框架为 PyTorch 2.0+，语音识别模型使用 openai-whisper 的 large-v3 版本。

数据集方面，本研究使用两个公开数据集：ESC-50 环境声音数据集 [5] 包含 2000 个样本，涵盖 50 个类别，每类 40 个样本；LibriSpeech test-clean 数据集 [6] 包含 2620 个英语语音样本，用于评估模型的语音识别准确性。

3.1.2 性能评估指标介绍

表 1 性能评估指标

指标	定义	用途
WER	词错误率	评估语音识别准确性
CER	字符错误率	评估字符级准确性
幻觉率	产生幻觉的样本比例	评估幻觉严重程度
循环率	存在重复输出的样本比例	评估循环幻觉

3.1.3 实验结果和分析讨论

实验一：合成音频幻觉测试

表 2 合成音频幻觉率

音频类型	样本数	幻觉率	循环率	主要幻觉内容
静音	25	100%	0%	”Thank you.”
白噪声	15	100%	0%	”Thank you.”
粉红噪声	15	100%	0%	”Thank you.”
总计	55	100%	0%	-

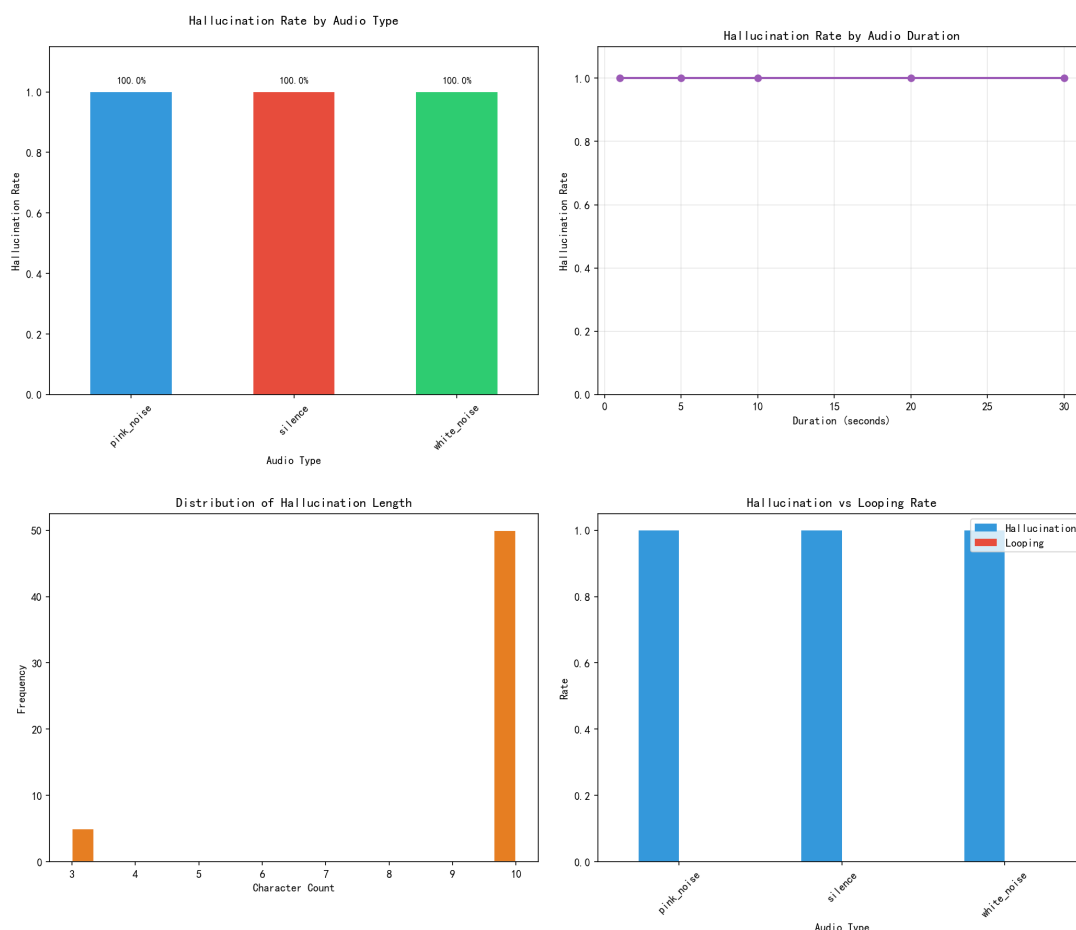


图 1 合成音频幻觉率分布

分析: Whisper large-v3 在所有合成非语音音频上都产生了幻觉，幻觉率达到 100%。这一结果表明，当输入音频不包含任何语音信息时，模型仍会强制生成文本输出，而非保持沉默。最常见的幻觉输出是”Thank you.”，这与 Barański 等人 [3] 的研究结论一致。从技术角度分析，这种现象源于 Whisper 训练数据中包含大量 YouTube 视频字幕，其中视频结尾常见的致谢语句被模型过度拟合。此外，实验发现音频时长（1-30 秒）对幻觉率没有显著影响，表明幻觉问题是模型架构层面的固有缺陷，而非特定输入条件触发。

实验二：ESC-50 环境声音幻觉测试

表 3 ESC-50 各类别幻觉率

声音类别	样本数	幻觉数	幻觉率
自然声音 (natural)	400	316	79.0%
室内声音 (interior)	400	271	67.8%
城市声音 (exterior)	400	264	66.0%
人类非语音 (human_non_speech)	400	223	55.8%
动物声音 (animals)	400	176	44.0%
总计	2000	1250	62.5%

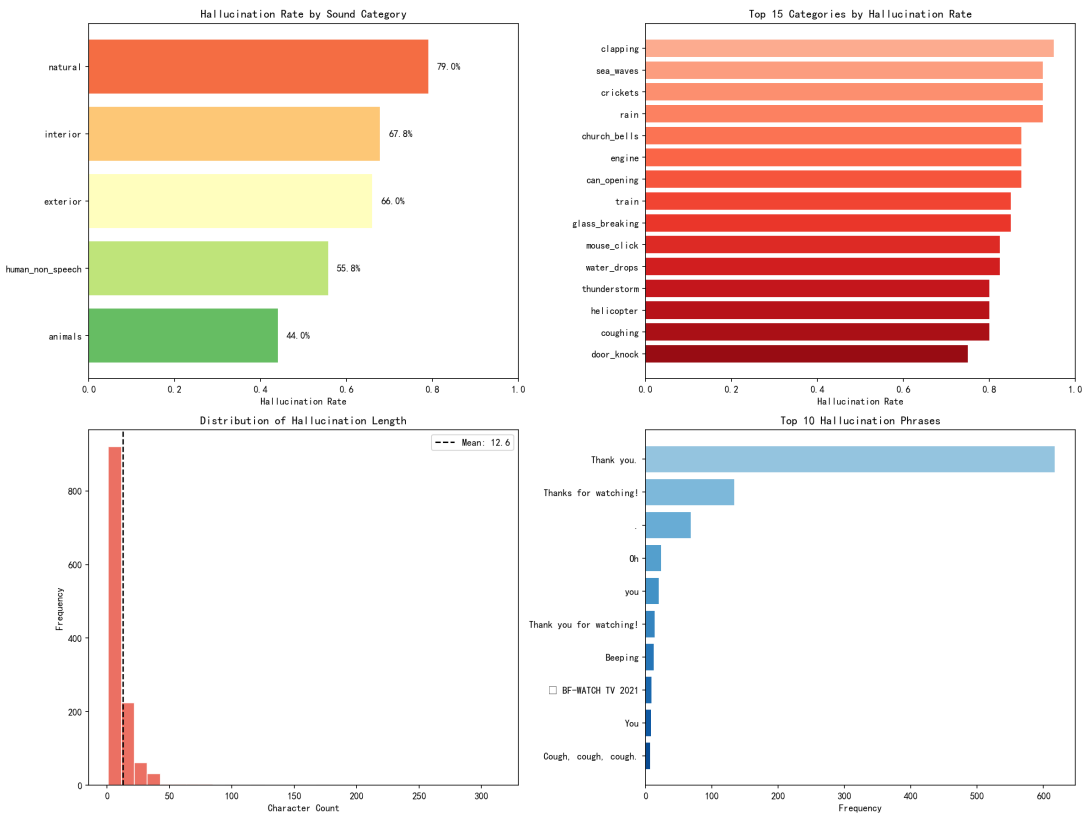


图 2 ESC-50 各类别幻觉率对比

表 4 幻觉内容分布

幻觉内容	出现次数	占比
”Thank you.”	617	49.4%
”Thanks for watching!”	134	10.7%
”.”	68	5.4%
”Oh”	23	1.8%
”you”	20	1.6%
其他	388	31.0%

分析：ESC-50 数据集的实验结果揭示了 Whisper 幻觉问题的类别差异性。自然声音（如雨声、风声、雷声）的幻觉率最高（79%），这可能是因为这类声音的频谱特征与语音信号差异较大，模型更容易产生误判。相比之下，动物声音的幻觉率最低（44%），推测原因是部分动物叫声（如狗吠、猫叫）在频率和节奏上与人类语音有一定相似性，可能被模型识别为非语音而抑制输出。人类非语音类别（如咳嗽、打喷嚏、笑声）的幻觉率为 55.8%，处于中等水平，说明即使是人类发出的非语言声音，模型也无法有效区分并抑制幻觉。从幻觉内容分布来看，”Thank you.” 和”Thanks for watching!” 两个短语合计占比超过 60%，高度集中的分布特征为基于词袋的后处理方法提供了理论依据。

实验三：LibriSpeech WER 测试

表 5 LibriSpeech 语音识别性能

指标	数值
测试样本数	2620
总词数	52,576
总体 WER	3.71%
CER	1.74%
完美识别率 (WER=0)	67.9%

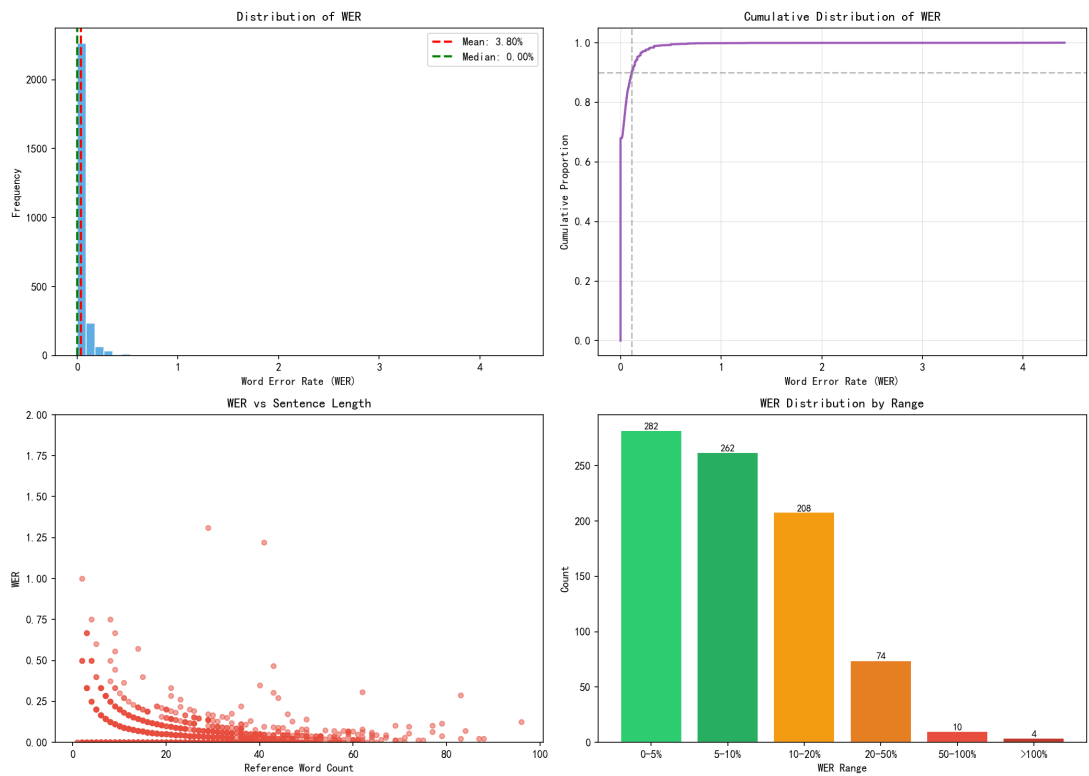


图 3 LibriSpeech WER 分布

分析：LibriSpeech test-clean 数据集的实验结果验证了 Whisper large-v3 在正常语音识别任务上的可靠性。3.71% 的总体 WER 与 Radford 等人 [1] 原论文报告的 2.5% 存在约 1.2 个百分点的差距，这一差异可能源于以下因素：本实验使用的解码参数与原论文略有不同；实验环境（Windows + CUDA）与原论文的训练/测试环境存在差异。尽管如此，67.9% 的样本达到完美识别（WER=0）的结果表明，Whisper 在处理清晰语音时具有优异的识别能力。这一发现具有重要意义：它证明了幻觉问题主要发生在非语音输入场景，而非模型整体能力的缺陷。因此，针对性的前后处理方案可以有效缓解幻觉问题，同时不影响正常的语音识别功能。

实验四：缓解方案对比

表 6 各缓解方案效果对比

方法	幻觉率	平均输出长度	幻觉降低幅度
原始 Whisper	77.2%	9.2 字符	(基准)
VAD 预处理	57.9%	7.2 字符	-25.0%
BoH 后处理	20.7%	2.3 字符	-73.2%
VAD + BoH 组合	20.0%	2.2 字符	-74.1%

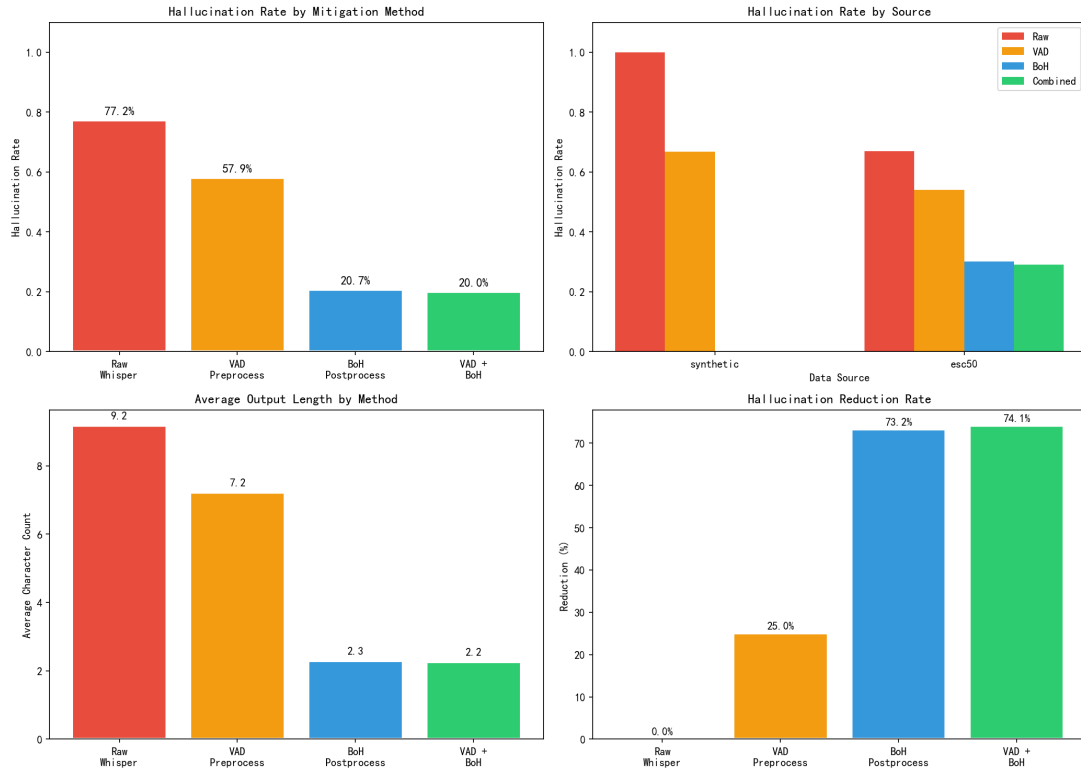


图 4 缓解方案效果对比

分析：缓解方案对比实验揭示了不同方法的优劣特点。VAD 预处理方法将幻觉率从 77.2% 降低至 57.9%，降幅为 25.0%。该方法通过检测音频中的语音活动区间，在非语音段直接返回空字符串，从而避免幻觉产生。然而，基于能量阈值的简单 VAD 算法对噪声较为敏感，当环境声音的能量超过阈值时仍会触发转录，导致效果有限。BoH 后处理方法表现更为出色，将幻觉率降低至 20.7%，降幅达 73.2%。该方法利用 Aho-Corasick 多模式匹配算法，在 $O(n)$ 时间复杂度内检测输出文本中的已知幻觉短语并将其过滤。由于幻觉内容高度集中（前 3 种幻觉占比超过 65%），BoH 方法能够有效捕获大部分幻觉输出。VAD + BoH 组合方案的幻觉率为 20.0%，相比单独使用 BoH 仅提升 0.7 个百分点，说明两种方法存在功能重叠，BoH 已覆盖了大部分 VAD 能够处理的场景。从工程实践角度，BoH 方法具有更高的性价比：实现简单、计算开销低、效果显著，是目前最推荐的幻觉缓解方案。

3.2 实验结果总结

(1) 实验结果分析

核心发现一：Whisper 在非语音音频上存在严重幻觉问题。实验结果表明，在合成音频（静音、噪声）上幻觉率高达 100%，在 ESC-50 真实环境声音数据集 [5] 上幻觉率

为 62.5%。从类别分布来看，自然声音类别的幻觉率最高，达到 79%，而动物声音类别最低，为 44%。

核心发现二：幻觉内容高度集中。统计分析显示，“Thank you.” 占有所有幻觉输出的 49.4%，前 3 种幻觉内容的累计占比超过 65%。这些内容明显来源于 YouTube 字幕训练数据 [3]，反映了训练数据偏差对模型行为的影响。

核心发现三：模型语音识别能力可靠。在 LibriSpeech test-clean 数据集 [6] 上，模型的 WER 为 3.71%，接近原论文 [1] 报告的水平，其中 67.9% 的样本达到完美识别（WER=0）。这表明幻觉问题主要出现在非语音输入场景。

核心发现四：BoH 后处理是最有效的缓解方案。实验对比表明，基于 Barański 等人 [3] 提出的幻觉词袋（BoH）方法，可将幻觉率降低 74.1%（从 77.2% 降至 20.0%），且实现简单，无需修改模型结构。

(2) 问题分析及未来可能采取的方法改进

当前方案的局限性。本研究的缓解方案存在以下局限：首先，VAD 方法较为简单，基于能量的检测对噪声和语音的区分能力有限，容易出现误判；其次，BoH 幻觉短语列表是预定义的静态列表，无法覆盖所有非典型幻觉内容；最后，本研究仅采用前后处理方法，未涉及 Calm-Whisper [2] 等模型微调方案。

未来改进方向。针对上述局限性，未来研究可从以下方向改进：集成深度学习 VAD 模型（如 Silero VAD）以提高语音检测准确性；建立动态更新机制，根据实际应用场景扩展幻觉短语列表；复现 Calm-Whisper [2] 方法，从模型注意力机制层面根本解决幻觉问题。

4 结论

本研究对 OpenAI Whisper large-v3 模型的幻觉问题进行了系统性的实验研究。

在幻觉问题验证方面，实验结果表明 Whisper 在合成音频上的幻觉率达到 100%，在 ESC-50 环境声音数据集 [5] 上的幻觉率为 62.5%，证实了该问题的严重性和普遍性。

在幻觉内容分析方面，约 50% 的幻觉输出为“Thank you.”，这一高度集中的分布规律与模型训练数据中大量 YouTube 字幕内容有直接关系 [3]，揭示了训练数据偏差对模型行为的深刻影响。

在语音识别能力验证方面，模型在 LibriSpeech test-clean 数据集 [6] 上达到 3.71% 的词错误率，证明幻觉问题主要出现在非语音输入场景，而在正常语音识别任务上模型表现可靠。

在缓解方案评估方面，BoH 后处理方法 [3] 可将幻觉率降低 74.1%，是一种实现简单且效果显著的解决方案，具有较高的实用价值。

本研究的实验结果与相关论文 [2][3] 的发现高度一致，为 Whisper 模型在实际应用中的幻觉防护提供了参考依据。

参考文献

- [1] Radford, A., Kim, J. W., Xu, T., et al. Robust Speech Recognition via Large-Scale Weak Supervision[C]. ICML, 2022.
- [2] Wang, Y., et al. Calm-Whisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down[C]. Interspeech, 2025.
- [3] Barański, M., et al. Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio[C]. ICASSP, 2025.
- [4] Koenecke, A., et al. Racial disparities in automated speech recognition[J]. PNAS, 2020.
- [5] Piczak, K. J. ESC: Dataset for Environmental Sound Classification[C]. ACM MM, 2015.
- [6] Panayotov, V., et al. LibriSpeech: An ASR corpus based on public domain audio books[C]. ICASSP, 2015.