When a Bot Cries Wolf: Exploring Voice AI Misuse through an Emergency Services Demo

Extended Abstract

Qizhou Wang
The University of Melbourne
Parkville, Australia

Hanxun Huang
The University of Melbourne
Parkville, Australia

Sarah Erfani
The University of Melbourne
Parkville, Australia

Christopher Leckie
The University of Melbourne
Parkville, Australia

Abstract

Recent voice AI agents can engage in multi-turn conversations with humans and demonstrate significantly improved capabilities. Despite a remaining gap in replicating human-like realism, they can still be used to disrupt services in communication systems, particularly those that rely on human operators. This demo illustrates this risk through a simulated emergency services scenario, showing how voice AI agents could conduct automated calls to interfere with critical service channels. We discuss the technical requirements and the feasibility of misusing such agents, and provide exploratory results on their task-oriented conversational capabilities. Based on preliminary observations, we highlight emerging risks and discuss potential indicators for identifying automated callers. Our goal is to raise awareness of this issue and provide an initial perspective on the capabilities of voice AI agents to inform future research and the development of effective safety mechanisms.

1 Introduction

Voice AI agents [2] are now capable of sustaining multi-turn interactions with coherent, task-oriented responses. While they remain imperfect in replicating human-like behaviour, such as disfluencies, emotional nuance, and handling interruptions, their semantic and logical capabilities are already sufficient to support such interactions with humans. These advances bring clear benefits to many real-world applications, but also introduce new risks of misuse.

Rather than providing productivity benefits, these systems could be exploited by malicious actors to impersonate human users. One concerning threat that was previously infeasible is the use of voice agents at scale to overwhelm communication channels via phone calls, effectively causing denial of service [3]. Call-based services are typically constrained by limited human resources, making them difficult to scale, while the rapid advancement of voice agents is making them increasingly difficult for humans to detect [7].

We explore this emerging risk through a simulated emergency services scenario, where service availability is critical and capacity is constrained by a limited number of trained human operators. We examine how voice AI agents could conduct automated calls to report false emergencies through real-time multi-turn interactions with human operators, thereby disrupting genuine incident reporting and exhausting service resources. Fig. 1 illustrates the comparison between human and voice agent scenarios. We conduct exploratory experiments to evaluate their logical and task-oriented conversational abilities. We further provide a set of practical indicators for detecting and defending against malicious voice agents. Our goal is to provide an initial perspective on these scenarios and their associated risks as voice agents continue to advance.

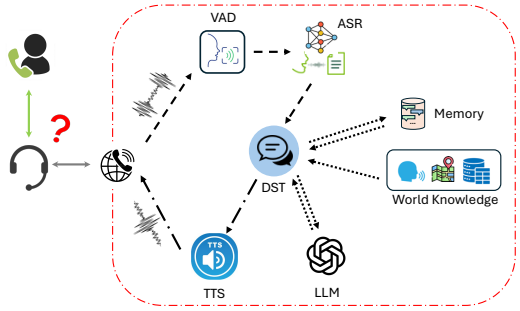

Figure 1: Overview of the voice agent simulation pipeline.

2 Voice Agents Impersonating Human Callers

Voice Agent Simulation Pipeline. As shown in Fig. 1, we adopt a cascaded stack of GenAI models (i.e., voice recognition → LLM reasoning → speech generation) rather than unified audio LLMs or agents, which allows greater control to ensure adherence to the target scenario. While unified audio agents offer advantages in latency and real-time duplex interaction, they are harder to control and may deviate from the intended task.

For input handling, a Voice Activity Detection (VAD) [5] module monitors the audio stream and detects human speech, which is passed to an Automatic Speech Recognition (ASR) [10] system to produce a text transcript. A Dialogue State Tracking (DST) module then determines whether to continue listening or generate a response based on the current dialogue state, while maintaining memory that stores conversational history and predefined context (e.g., identity and location). For response generation, the agent uses an LLM to produce contextually appropriate replies, interacting with this memory and retrieving relevant information from external knowledge sources when needed to ensure verifiability. The responses are then synthesised into speech using a text-to-speech (TTS) [1, 11] system.

Required Capabilities and Current Maturity. As shown in Tab. 1, imitating real human callers requires capabilities across three broad categories. First, *logical competence*, where the agent can maintain task-oriented coherence, track conversational context, respond appropriately, and avoid revealing its non-human identity. Second, *perceptual and generative capability*, which involves accurately capturing operator speech and producing realistic speech output. Third, *behavioural robustness*, where the agent can sustain interaction under dynamic conditions, including handling interruptions, expressing appropriate emotional cues, and adapting to unexpected conversational turns. Stronger capabilities make agents more difficult to distinguish from human users.

In practice, logical competence is largely well developed, with modern models performing strongly in task coherence and context

Capability	Requirement (related module)	Status
Logical competence	Task-oriented response (LLM)	Strong
	Context tracking (LLM, DST)	Strong
	Identity concealment (LLM)	Limited
	Information grounding (external knowledge)	Achievable
Perceptual & generative	Speech recognition (ASR)	Limited under noise
	Voice synthesis (TTS)	Realism gap remains
Behavioural robustness	Handling interruptions (DST)	Moderate
	Human imperfection and emotion (LLM, DST)	Limited
	Real-time interaction (all modules)	Optimisable

Table 1: Required capabilities and current maturity.

tracking, while identity concealment remains challenging without task-specific tuning. The primary limitations instead lie in perceptual and behavioural aspects, where realistic voice generation and modelling human-like imperfections, such as disfluencies, emotional variability, and natural interaction patterns, remain key gaps.

3 Empirical Study

Overview. Our evaluation focuses on the semantic and conversational abilities of the agents within the multi-turn interaction. Voice aspects (e.g., TTS and ASR) are treated as downstream components and left out of scope. Following the pipeline, we use 4 LLMs to drive the agent for evaluating: (1) *fidelity to reference caller*, and (2) *detectability under adversarial probing*. We employ Gemini-3-Pro as an automated evaluator. *Data Creation.* Synthetic script data are generated using Gemini-3-Flash with style conditioning, using real 911 call transcripts as prompt-level style references to produce realistic emergency-call transcripts with Australian-specific content. We generate 300 emergency call scripts covering diverse scenarios.

LLMs Used to Drive the Agent. We employ three pretrained LLMs of varying scales, namely Qwen3.5-4B [9], Qwen3.5-35B [9], and Gemini-2.5-Pro [4], together with a fine-tuned Qwen3.5-4B model to drive the voice agent implemented following the pipeline mentioned in Fig. 1. LoRA-based [6] SFT is used on 2700 gpt-oss-120b-generated [8] scripts style-conditioned by non-overlapping samples from the same corpus under the same approach.

Simulation. Each scenario is replayed turn by turn with identical operator inputs from the test script, ensuring that differences arise only from model responses. At each turn, the model conditions on the scenario, the current operator utterance, and its prior responses to generate the next turn, capturing compounding errors under realistic interaction settings.

Fidelity Evaluation. We evaluate four rubric dimensions per generated caller turn, each scored on a 1–5 scale by the judge [12]: (1) character consistency (emotion, tone, style), (2) information accuracy (no fabrication or over-sharing), (3) naturalness (human-like disfluency and pacing), and (4) responsiveness (direct and appropriate replies to operator queries). Each model generates 2,748 caller turns across 300 test conversations, with every turn evaluated on all four dimensions.

The judge receives the scenario, conversation context, reference script, and replayed script, and returns evaluation scores shown in Table 2. Gemini-2.5-Pro achieves the highest scores across all models, while the fine-tuned 4B model significantly outperforms its base model and matches the much larger Qwen3.5-35B. However, all non-fine-tuned models tend to generate longer responses, which may be favoured by the LLM judge but are not necessarily desirable for conversation. Our key finding is that attack-oriented fine-tuning substantially improves fidelity and task-oriented conversational ability, allowing a small LLM to match the fidelity of much larger open pretrained LLMs while substantially reducing detectability below that of state-of-the-art proprietary systems. While this requires technical effort, once achieved, it significantly reduces both the cost and scalability constraints of such attacks.

	Qwen3.5-4B FT	Qwen3.5-4B	Qwen3.5-35B	Gemini-2.5-Pro
Character consis.	2.84	2.70	2.87	**4.64**
Information acc.	3.36	2.80	3.48	**4.12**
Naturalness	2.91	2.30	2.95	**4.79**
Responsiveness	3.97	3.00	3.54	**4.72**
Average	3.27	2.70	3.21	**4.57**
Length (words)	**15.6**	83.3	103.6	45.4

Table 2: Fidelity results and average output lengths.

Model	Overall	AI Admission	Trivia	Small Talk
Qwen3.5-4B FT	1.4%	1.0%	3.6%	0.0%
Gemini-2.5-Pro	29.8%	28.9%	28.9%	31.2%
Qwen3.5-4B	39.8%	42.0%	33.8%	42.1%
Qwen3.5-35B	41.1%	36.9%	41.5%	44.6%

Table 3: Adversarial probing test results.

Adversarial Probing. We evaluate the detectability of voice agents under off-topic probing by injecting one of three probes at a fixed turn in each conversation: AI Admission ("Are you an AI agent?"), Trivia ("How far is Melbourne from LA?"), and Small Talk ("How is your day?"). The judge labels each response based on whether it reveals the non-human nature of the agent, where any direct engagement is considered indicative of exposure. As shown in Table 3, the fine-tuned model (Qwen3.5-4B FT) achieves a low exposure rate of 1.4%, compared to 29.8% for Gemini-2.5-Pro, 39.8% for Qwen3.5-4B, and 41.1% for Qwen3.5-35B, indicating that fine-tuning in domain-specific service settings makes the model much less likely to reveal its non-human identity.

4 Conclusion and Safety Considerations

We demonstrate how voice AI agents can automate malicious calls in human-facing systems, highlighting emerging risks as these technologies advance. We identify perceptual and behavioural realism as the key factor underlying high-fidelity, low-detectability attacks.

Safety Considerations. Based on these categories, we outline the following practical considerations for detecting and mitigating voice AI misuse. At the logical level, probing questions can be used to trigger revealing behaviours, while the plausibility of responses can be assessed in terms of wording and reasoning. At the perceptual and generative level, controlled noise can be introduced to trigger ASR failures without compromising intelligibility, while voice signals can be analysed for realism and cues of imperfect generation. In addition, audio deepfake detection models can be applied directly to the waveform to identify machine-generated speech. At the behavioural level, overall interaction realism can be assessed, while robustness can be evaluated under unexpected interaction scenarios. However, some assessments rely on human judgement, and not all technical methods are applicable across different settings. This highlights the need for machine-driven detection systems for voice AI misuse. We hope this work raises awareness of voice AI misuse and informs future research on detecting and mitigating these emerging risks.

References

[1] [n. d.]. ([n. d.]).

[2] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* (2024). https://arxiv.org/abs/2410.00037

[3] Federal Bureau of Investigation. 2021. Telephony Denial of Service Attacks Can Disrupt Emergency Call Center Operations. Public Service Announcement PSA210217, Internet Crime Complaint Center (IC3). https://www.ic3.gov/PSA/2021/PSA210217

[4] Gemini Team, Google. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[5] Google. [n. d.]. WebRTC Voice Activity Detector. https://webrtc.org/. Python interface: https://github.com/wiseman/py-webrtcvad.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

[7] Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE* 18, 8 (2023), e0285333. doi:10.1371/journal.pone.0285333

[8] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925* (2025).

[9] Qwen Team. 2026. Qwen3.5: Accelerating Productivity with Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 28492–28518.

[11] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. 2025. Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens. *arXiv preprint arXiv:2503.01710* (2025).

[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, Vol. 36.