A Multimodal Verifier for Trustworthy AI in High-Risk Environments

Part of ITEA 4 CLEAR, an International Industry-Academia Research Project

Paria Sarzaeim · Supervisor: Dr. Akramul Azim · Faculty of Engineering and Applied Science, Ontario Tech University
Paria.Sarzaeim@ontariotechu.net

Introduction

AI is used in high-risk fields such as healthcare [1], cybersecurity [2] and law [3]. It now also helps take calls, audit them and write police reports [4, 5].

Who checks that what it writes is true?

- Evidence is mixed e.g. there are accented, noisy calls [6], notes and scene photos.
- Errors e.g. a missed emergency or a false statement in a legal record are costly [7].
- Private data must stay in the agency, so models run locally.

Use Cases and Modalities

Table 1. Five use cases studied and the evidence each one involves.

	Speech	Text	Images
Call handling	✓	✓	–
Call auditing	✓	✓	–
Incident reports	✓	✓	✓
Policy guide	–	✓	–
Orchestrator	✓	✓	✓

Research Questions

- **RQ1.** Does changing the format of the data (e.g. hearing the call vs. reading the call transcript) affect final decisions?
- **RQ2.** How are AI-generated claims pointed to evidence e.g. timestamps, image regions, and speakers?
- **RQ3.** How do the verification methods for AI-generated claims handle conflict detection between sources across modalities?
- **RQ4.** How can we generalize verification methods across different modalities and AI models?

Spot the Error

Which sentence in this AI-drafted report would you trust?

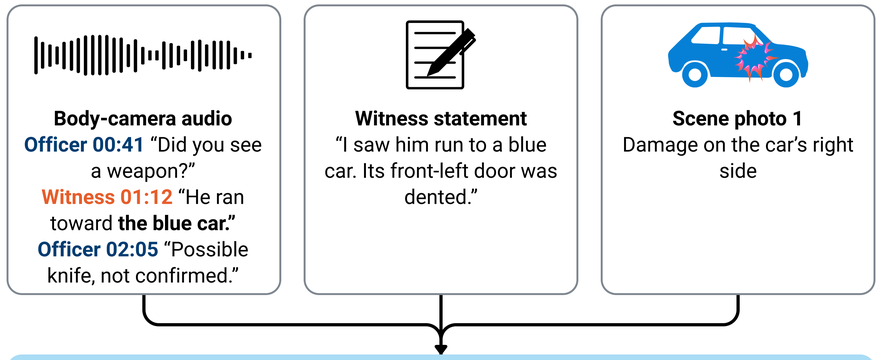


Figure 1. Proposed verifier: each AI-written sentence is checked against audio, text and image evidence; wrong speakers and source conflicts are flagged; unsure cases go to a person [8-10].

How we will test the verifier

- **Data:** Synthetic incidents (audio, statements, photos) where every true, false and missing fact is known.
- **Baselines:** A text-only checker [8], a large multimodal model as judge, and self-checking.
- **Metrics:** Label and speaker accuracy, conflict detection, and error rate at a fixed hand-over level [11].

Experimental Setup

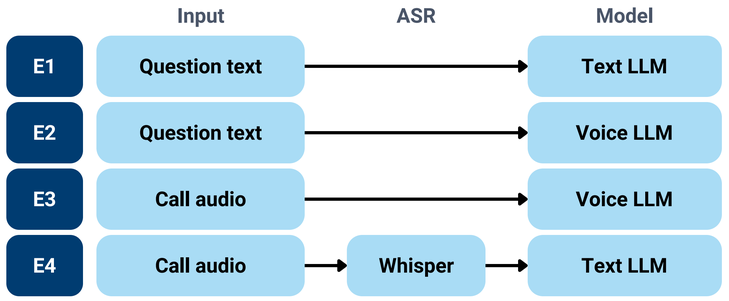


Figure 2. Experiment 1: four conditions on 1,526 synthetic calls (regular, noisy, 7 accents) with the models in Table 2. E4 uses Whisper large-v3-turbo [12-14].

Table 2. Models in Experiment 1 (all open and run locally).

Voice LLM (hears the call)	Text model it is built on
Ultravox v0.5	Llama 3.1 8B
Qwen2.5-Omni 7B	Qwen2.5 7B
Qwen2.5-Omni 3B	Qwen2.5 3B
Voxtral Small 24B	Mistral Small 24B
Voxtral Mini 3B	*Not openly available*

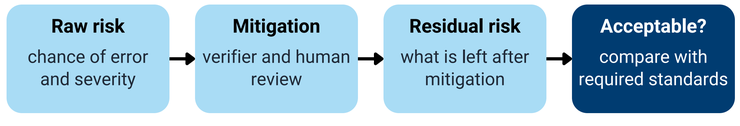
Where We Are

- Prepared 425 questions with answer keys
- Prepared 1,526 synthetic calls in 7 accent groups
- Conducted speech-to-text and set up 5 voice models
- Performed initial literature scan and gap analysis

→ Next: run all experiments and build the verifier

Road Ahead

From accuracy to acceptable risk



Agencies must follow privacy and AI standards that limit data access and may set how accurate an AI tool must be [15, 16].

Future work: a probabilistic model of residual risk after mitigation, mapped to these requirements (e.g., ISO/IEC 23894, ISO/IEC 42001, Ontario's MFIPPA).

References

[1] Topol, Eric J. "High-performance medicine: the convergence of human and artificial intelligence." Nature medicine 25.1 (2019): 44-56.
[2] Xu, HanXiang, et al. "Large language models for cyber security: A systematic literature review." ACM Transactions on Software Engineering and Methodology 35.11 (2026): 1-61.
[3] Dahl, Matthew, et al. "Large legal fictions: Profiling legal hallucinations in large language models." Journal of Legal Analysis 16.1 (2024): 64-93.
[4] Adams, Ian T., et al. "No man's hand: Artificial intelligence does not improve police report writing speed." Journal of Experimental Criminology (2024).
[5] Sarzaeim, Paria, et al. "A systematic review of using machine learning and natural language processing in smart policing." Computers 12.12 (2023): 255.
[6] Koenecke, Allison, et al. "Racial disparities in automated speech recognition." Proceedings of the national academy of sciences 117.14 (2020): 7684-7689.
[7] Ji, Ziwei, et al. "Survey of hallucination in natural language generation." ACM computing surveys 55.12 (2023): 1-38.
[8] Tang, Liyan, Philippe Laban, and Greg Durrett. "Minicheck: Efficient fact-checking of llms on grounding documents." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.
[9] Yang, J., et al. "OmniHallu: Unified hallucination detection for cross-modal comprehension and generation in multimodal large language models." arXiv preprint arXiv:2609.11244 (2026).
[10] Thorne, James, et al. "FEVER: a large-scale dataset for fact extraction and VERification." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.
[11] Angelopoulos, Anastasios, et al. "Conformal risk control." International conference on learning representations. Vol. 2024. 2024.
[12] Cuervo, Santiago, et al. "Closing the gap between text and speech understanding in llms." International Conference on Learning Representations. Vol. 2026. 2026.
[13] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International conference on machine learning. PMLR, 2023.
[14] Sanabria, Ramon, et al. "The edinburgh international accents of english corpus: Towards the democratization of english asr." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
[15] ISO/IEC 23894:2023. Artificial intelligence - Guidance on risk management.
[16] ISO/IEC 42001:2023. Artificial intelligence - Management system.