Cascaded to End-to-End: New Safety, Security, and Evaluation Questions for Audio Language Models

Luxi He* **Xiangyu Qi** * **Inyoung Cheong**
Prateek Mittal **Danqi Chen** **Peter Henderson**
Princeton Language and Intelligence (PLI), Princeton University
{luxihe, xiangyuqi}@princeton.edu

1 Introduction

A growing number of large language models (LLMs) now process audio input alongside text, commonly known as Audio LMs. Although this development is not entirely new—voice assistants have existed for some time—audio was historically processed by language models through a **cascaded** pipeline approach until recently. In this pipeline, audio is first transcribed into text by a separate automatic speech recognition (ASR) model, and then the text is fed into an LLM. Some voice assistants, like Amazon's Alexa, Apple's Siri, ChatGPT's early voice mode [1], and others, use such a cascaded pipeline. The cascaded pipeline is limited, though. The transcription step discards rich information in the audio input, such as the speaker's intonation, pronunciation, the presence of multiple speakers, background scene information, and more. Since the LLM does not have access to this information in the cascaded system, it cannot incorporate it into any downstream decision-making and processing. This consideration has driven a shift toward **end-to-end** Audio LMs, such as GPT-4o [2]. These models process audio inputs by directly accessing rich audio features rather than relying on text tokens transcribed by an intermediary ASR model. In this perspective paper, we first underscore novel safety and security risks that could be introduced by the transition from cascaded to end-to-end Audio LMs. We then highlight tensions and gaps in current end-to-end Audio LM evaluation protocols between open and closed-source models. We hope our work spurs a re-alignment in open-source Audio LM safety, security, and capability evaluations.

2 Expanded Safety and Security Risks in End-to-End Audio LMs

End-to-end Audio LMs can directly access rich audio features, making them more amenable to solving a broader range of tasks—a key goal for general-purpose systems. However, direct access to audio leaks sensitive information to the model and expands the model's risk profile.

Sociotechnical Safety Challenges. Audio recordings of human speech contain rich information, much of which might be considered sensitive. Whether correctly or not, a large swath of researchers have tried to infer sensitive attributes from audio data [3], including: identity [4]; demographic attributes like age [5], gender [6], race and ethnicity [7], and socioeconomic status [8]; and psychophysiological traits like emotions [9], personality [10], intoxication [11], and mental health [12].

There is a risk, then, that end-to-end Audio LMs may implicitly use these features in their processing of user requests or make **unintended inferences** about the speaker, leading to allocative and representational harms [2, 13]. Consider, for example, if a model is asked to find information about potential jobs, but infers the speaker's demographics and adjusts its recommendations towards stereotypical biases. Or a model infers that a user may be vulnerable in some way and manipulates them. While some have begun examining these risks in closed source settings [14], or have examined biases arising from ASR errors across disparate groups [15], there is still a dearth of open-source evaluations examining these harms in end-to-end Audio LMs.

*Equal Contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024), EvalEval Workshop

Audio LMs could also be explicitly abused for **harmful inference** since they leak potentially sensitive information. If Audio LMs retain the few-shot prompting and adaptation capabilities of text-based LMs, they may enable a wide range of surveillance or privacy-violating uses with relative ease. This capability, a byproduct of shifting toward end-to-end Audio LMs, may come into tension with several laws globally. The EU AI Act explicitly prohibits emotion recognition in educational and professional settings and still classifies it as high risk in scenarios where it is not outright prohibited [16]. The collection and the use of biometric information—which could include voice data used to make these inferences, or even potentially identifying features stored within Audio LMs—might be subject to the Biometric Information Privacy Act (BIPA) of Illinois [17] and the EU General Data Protection Regulation (GDPR) [18]. Without appropriate safeguards and evaluations, developers and users of Audio LMs may find themselves facing legal repercussions.

Security Risks of Adversarial Attacks. End-to-end Audio-LMs are also more vulnerable to audio-domain adversarial attacks than classical cascaded architectures. For the latter, audio adversarial attacks can, at most, trick the ASR module into generating incorrect transcriptions [19], and the impact of attacks is no greater than that of purely textual adversarial attacks. However, end-to-end Audio-LMs, like vision language models, make possible strong gradient-based adversarial attacks [20]. Consequently, downstream behaviors of the Audio LM can be directly manipulated by adversarially crafted audio inputs. This vulnerability can lead to targeted attacks on the model's output [21] or jailbreaking behaviors [20].

3 Gaps in Current Audio-LMs Evaluation

The expanded risks introduced by end-to-end Audio-LMs raise new evaluation questions.

How should benchmarks align on safety and capability evaluations? There is an ongoing tension in Audio LM safety versus capability evaluation. Some major Audio LM evaluation benchmarks, such as AIR-Bench [22] and AudioBench [23], reward improved ability to identify sensitive features from audio, including gender, age, and emotion. Model developers may then explicitly optimize for performance on these tasks. Qwen-2-Audio [24], for example, explicitly includes emotion recognition in its training process. In contrast, some closed-source tech companies with proprietary models have adopted more cautious measures to mitigate legal risks. OpenAI, for example, states that GPT-4o is designed to avoid inferring a speaker's race or socioeconomic status and to refrain from prompting emotional reliance [14]. This discrepancy creates an unsettling outcome: proprietary evaluations test that models *do not engage* in behaviors with some undesirable legal and ethical implications, while open source evaluations test that models *do well* in those tasks.

What legitimate use cases genuinely benefit from the end-to-end approach and justify its additional risks? General-purpose audio capabilities may be desirable. But Audio LM benchmark creators may want to revisit which safe capabilities actually outweigh the increased safety and security risks from an end-to-end system. For conversation-oriented tasks, if the primary intended functions of an Audio LM are already satisfied by a cascaded model, why should we even deploy an end-to-end alternative? If the goal is merely to build an AI assistant capable of responding to voice commands, a cascaded model with a good ASR module may already suffice. In such scenarios, providing the AI assistant with additional audio features that could reveal the speaker's identity and emotional state does not directly contribute to a better response to commands while increasing another risk.

4 Conclusion

Including rich audio features is motivated by capturing paralinguistic information, but is that fundamentally at odds with the risks of introducing sensitive information? This is a fundamental tension that current evaluation frameworks insufficiently address. Notably, some developers report performance on tasks like accent, gender, and emotion recognition [25, 23]. These incentivized benchmark tasks may result in models that bear increased risks in their use of sensitive information. Currently, closed-source models like GPT-4o have spent more effort in benchmarking their Audio LM's potential risks, covering categories like speaker identification, ungrounded inference and sensitive trait attribution, and generating problematic speech content (relevant for audio output models) [14]. However, none of these (or equivalents) are openly available for other model creators. In all, as we shift from cascading to end-to-end Audio LMs, it's time to make sure that capability and safety evaluations are aligned.

References

[1] OpenAI. Chatgpt can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/, 2023.

[2] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[3] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pages 242–258, 2020.

[4] David Karpey and Mark Pender. Customer identification through voice biometrics, July 19 2016. US Patent 9,396,730.

[5] Seyed Omid Sadjadi, Sriram Ganapathy, and Jason W Pelecanos. Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5040–5044. IEEE, 2016.

[6] Adrian P Simpson. Phonetic differences between male and female speech. *Language and linguistics compass*, 3(2):621–640, 2009.

[7] Xingyu Chen, Zhengxiong Li, Srirangaraj Setlur, and Wenyao Xu. Exploring racial and gender disparities in voice biometrics. *Scientific Reports*, 12(1):3723, 2022.

[8] Sei Jin Ko, Melody S Sadler, and Adam D Galinsky. The sound of power: Conveying and detecting hierarchical rank through voice. *Psychological Science*, 26(1):3–14, 2015.

[9] Nicholas Cummins, Maximilian Schmitt, Shahin Amiriparian, Jarek Krajewski, and Björn Schuller. "you sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3806–3809. IEEE, 2017.

[10] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 17–20, 2010.

[11] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski, Felix Weninger, and Florian Eyben. Medium-term speaker states—a review on intoxication, sleepiness and the first challenge. *Computer Speech & Language*, 28(2):346–374, 2014.

[12] Nik Wahidah Hashim, Mitch Wilkes, Ronald Salomon, Jared Meggs, and Daniel J France. Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice*, 31 (2):256–e1, 2017.

[13] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, 2023.

[14] Gpt-4o system card. URL https://openai.com/index/gpt-4o-system-card/.

[15] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

[16] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[17] 740 ilcs 14/1 et seq.

[18] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj, 2016.

[19] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

[20] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023.

[21] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023.

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

[23] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models, 2024. URL https://arxiv.org/abs/2406.16020.

[24] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[25] 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. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.