UNIVERSITY OF CAMBRIDGE

Uncertainty Quantification in Audio LLMs via Mechanistic Interpretability

Hung Phan Huy

University of Cambridge

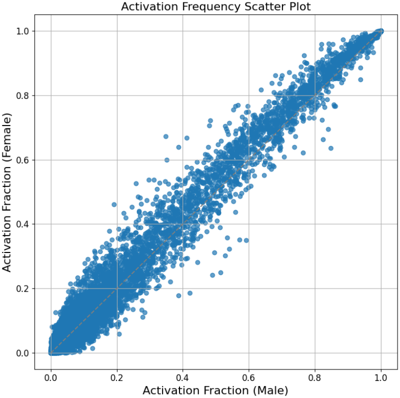
June 22, 2025

Introduction

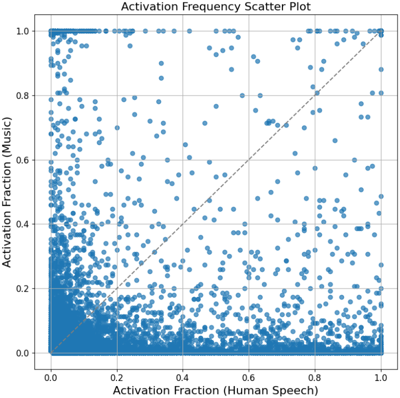
- **What is mechanistic interpretability?**
  The study of how neural networks compute their outputs by reverse-engineering their internal mechanisms.
- **What are Anthropic, OpenAI, and DeepMind doing?**
  - *Anthropic:* Published *Circuit Tracing*, *Towards Monosemanticity*, and *Scaling Monosemanticity*.
  - *OpenAI:* Published *Scaling Sparse Autoencoders*, *Extracting Concepts from GPT-4*, and more...
  - *DeepMind:* Published *GemmaScope* - open-source SAE checkpoints for community research
- **Aim of this project:**
  Extend this work by exploring concept geometry (e.g., music genre, speaker emotion, uncertainty) in Audio LLMs, and apply the findings to uncertainty quantification in audio tasks.

Progress

- Uncovered concepts that the model does and does not represent
- Demonstrated causal significance of features via model steering interventions



Gender Separation



Music vs Speech