Uncertainty Quantification in Audio LLMs via Mechanistic Interpretability

Hung Phan Huy[1] Brian (Guangzhi) Sun[1] Phil Woodland[1]

[1]Department of Engineering, University of Cambridge

Objectives

- Extend linear interpretability tools from text to audio LLMs by training Sparse Autoencoders (SAEs).
- Discover sparse latent features encoding model uncertainty.
- Extract causal directions Δh for uncertainty; activation editing to address identifying low-confidence frames.

Background

- **Linear representation hypothesis** [4]: concepts → directions; orthogonality encodes causal separability.
- **SAE objective** [5]: Let $h \in \mathbb{R}^d$ be a hidden-state sample. The SAE learns an encoder $E \in \mathbb{R}^{k \times d}$ and decoder $D \in \mathbb{R}^{d \times k}$ by solving:

$$\min_{E,D} \sum_{h \sim \mathcal{D}} \underbrace{\|h - D\,\sigma(Eh)\|_2^2}_{\text{reconstruction loss}} + \lambda \underbrace{\|\sigma(Eh)\|_1}_{\text{sparsity penalty}}, \tag{1}$$

  where σ is a monotone element-wise nonlinearity (e.g., ReLU).
- **Sparse Dictionary Features**: Define the latent code:

$$z := \sigma(Eh) \quad \Rightarrow \quad h \approx \hat{h} = Dz = \sum_{j=1}^{k} z_j d_j,$$

  where $d_j := D_{\cdot j}$ is the j-th column of D. The L_1-penalty promotes sparse activations z_j, yielding interpretable basis vectors.

Are LLMs Aware of Their Uncertainty?

LLMs exhibit both **epistemic** (knowledge-based) and **aleatoric** (inherent) uncertainty [2]. A token t is labeled epistemic for a smaller model M_s if:

$$H_{M_s}(t|x) > \epsilon \quad \text{and} \quad H_{M_l}(t|x) < \delta$$

where H is predictive entropy and M_l is a larger, more capable model.

Supervised: Train linear probes $f : \mathbb{R}^d \to [0, 1]$ on activations from M_s to predict when M_l is confident ($H_{M_l}(t|x) < \delta$). *Result:* AUC > 0.9, generalizes across domains (e.g. Wikipedia → Code).

Unsupervised (ICLT): For top-k completions $\{t_i\}$, insert into context:

$$x' = x + t_i + x \quad \Rightarrow \quad \min_i H_{M_s}(t|x') \ll H_{M_s}(t|x)$$

Entropy drops more for epistemic cases, revealing in-context "suggestibility".

Conclusion: LLMs internally encode type-specific uncertainty signals. These can be extracted to detect epistemic uncertainty and reduce hallucinations.

Concept Representation in LLMs

In large language models (LLMs), hidden activations at layer l and token position i are vectors $x_i^{(l)} \in \mathbb{R}^d$. Many abstract concepts (e.g., refusal, toxicity, truthfulness) can be represented by directions $r \in \mathbb{R}^d$ [1] such that the inner product $r^\top x_i^{(l)}$ indicates the degree to which the concept is present.

Concepts are often encoded linearly, enabling direct manipulation:

- **Addition:** $x_i^{(l)} \leftarrow x_i^{(l)} + \alpha r$ (strengthen concept)
- **Subtraction:** $x_i^{(l)} \leftarrow x_i^{(l)} - \alpha r$ (suppress concept)
- **Ablation:** $x_i^{(l)} \leftarrow x_i^{(l)} - (r^\top x_i^{(l)}) r$ (remove concept)

Multiple concept directions may form a **cone** [6]:

$$\mathcal{C} = \left\{ \sum_{i=1}^{k} \lambda_i b_i \;\middle|\; \lambda_i \geq 0 \right\}$$

where all $r \in \mathcal{C}$ express the same high-level behavior. This captures the geometric complexity of conceptual representations in LLMs.

Findings So Far

- We see that audio LLMs do not represent gender.
- We can try to identify a "concept cone" for human speech, music, etc.
- We can isolate features relating to uncertainty and AI safety.

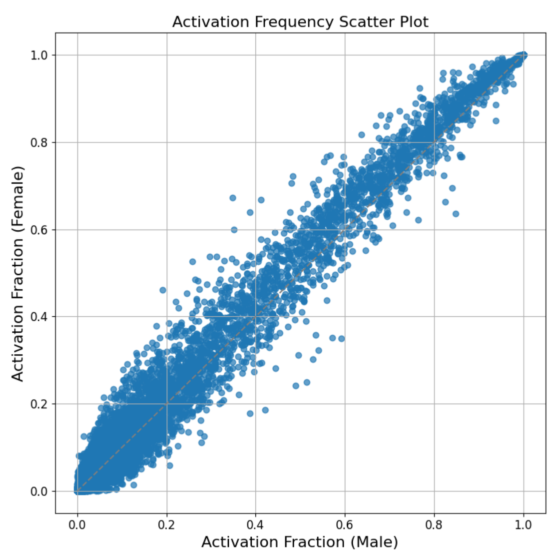


Figure 1. Gender Separation

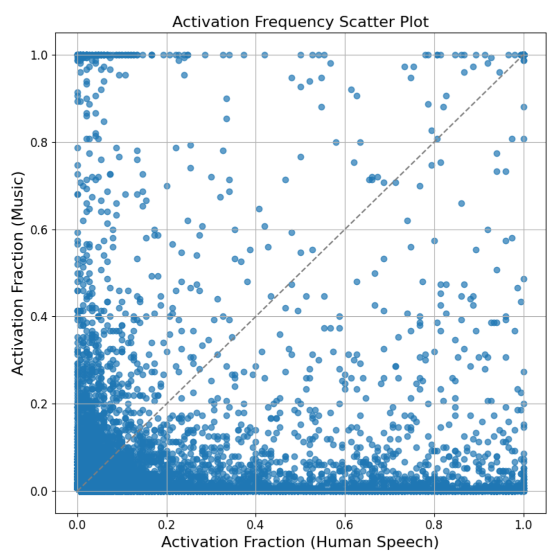


Figure 2. Music vs Speech

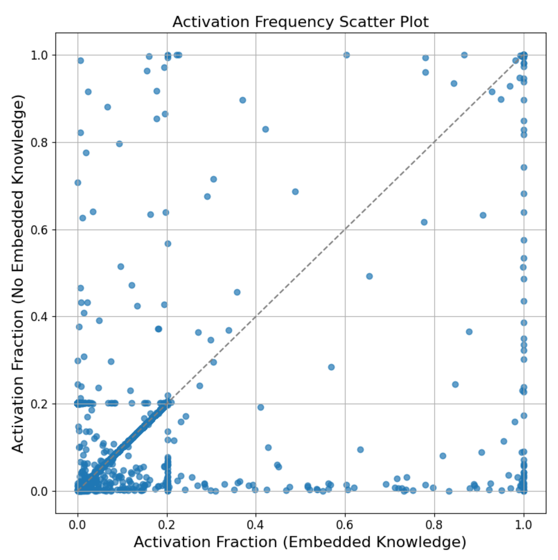


Figure 3. Knowledge Injection

Label Type	Percentage
Correct (Speaker Gender)	44.2%
Incorrect	51.4%
Other	4.4%

Table 1. Gender Classification Performance

Label Type	Percentage
Correct (Speech vs Music)	86.2%
Incorrect	3.0%
Other	10.9%

Table 2. Music vs Speech Performance

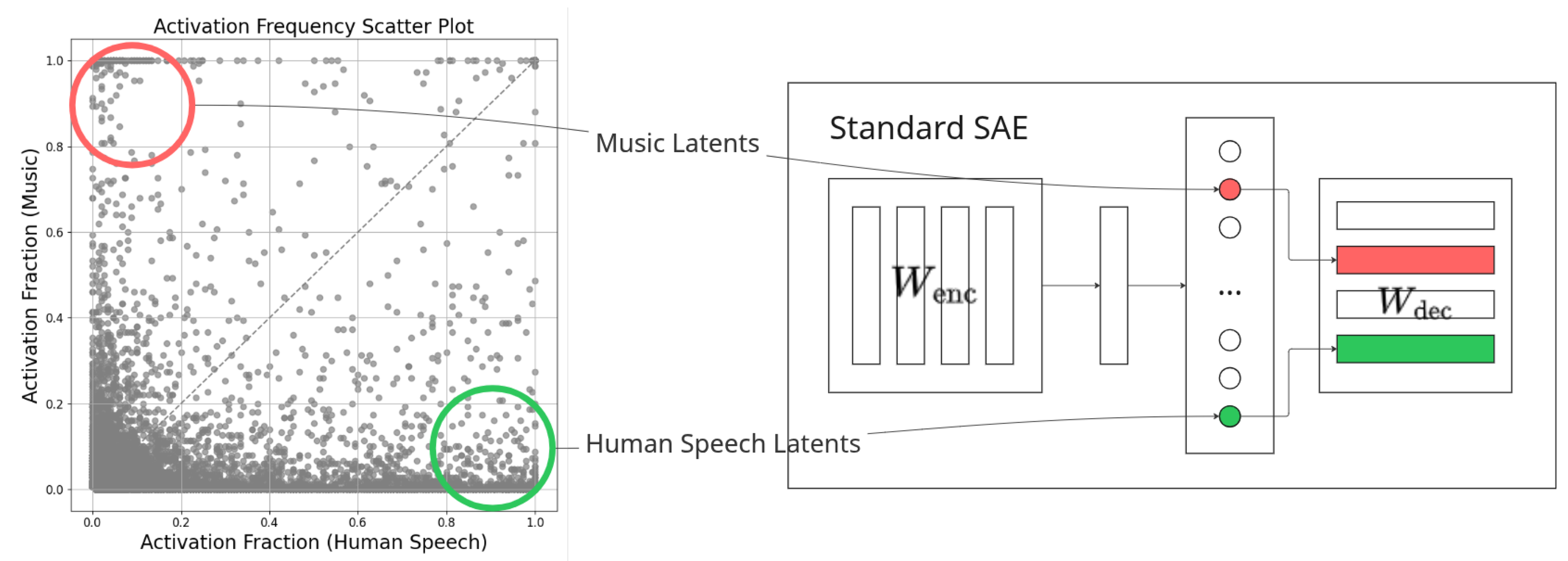
Steering via Activation Engineering

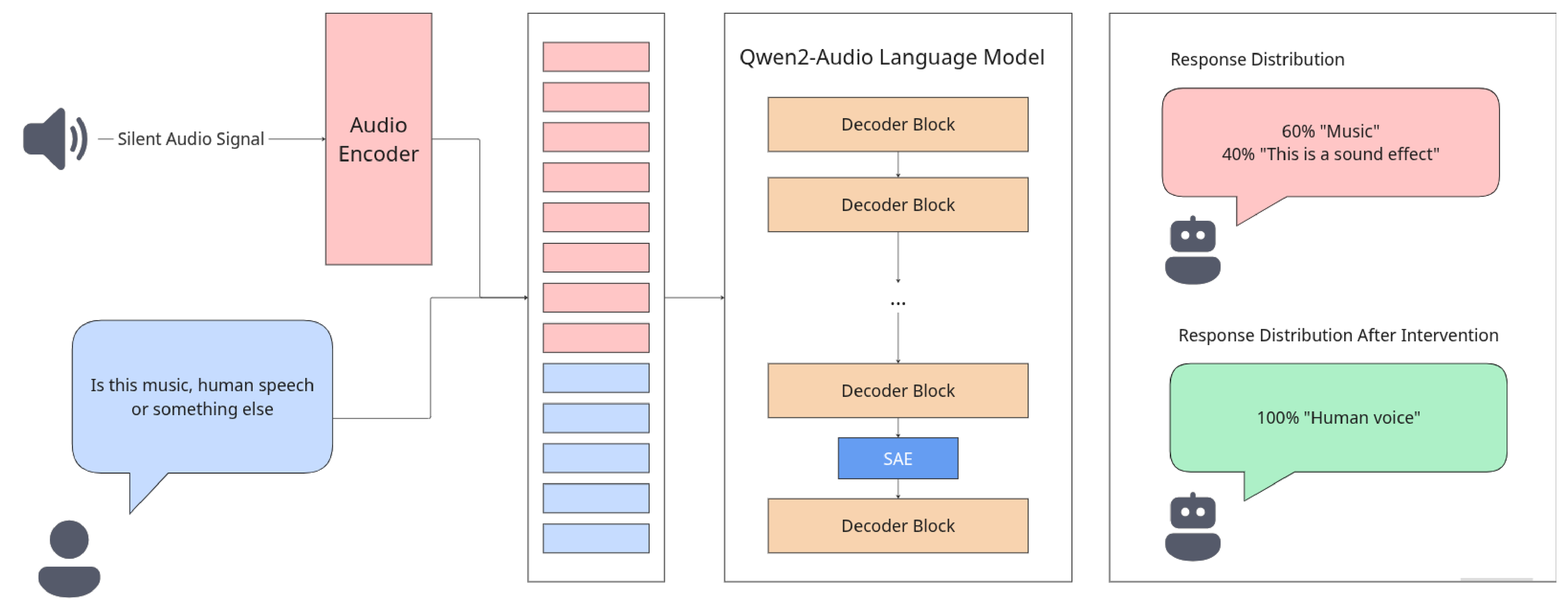
The dataset is partitioned into two disjoint subsets $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$, each containing N^α and N^β samples, respectively. Let $a_{l,j}(x)$ denote the activation of latent unit j at layer l for input x, and let $\mathbb{1}[\cdot]$ be the indicator function. Define:

$$f_{l,j}^\alpha = \frac{1}{N^\alpha} \sum_{i=1}^{N^\alpha} \mathbb{1}[a_{l,j}(x_i^\alpha) > \theta], \quad f_{l,j}^\beta = \frac{1}{N^\beta} \sum_{i=1}^{N^\beta} \mathbb{1}[a_{l,j}(x_i^\beta) > \theta]$$

Then the latent separation scores [3] are:

$$s_{l,j}^\alpha = f_{l,j}^\alpha - f_{l,j}^\beta, \quad s_{l,j}^\beta = f_{l,j}^\beta - f_{l,j}^\alpha$$





Conclusion and Further Work

- Identified causal directions for human speech recognition.
- Plan to explore causal links related to model uncertainty.
- Aim to automate discovery of interpretable latents in audio LLMs.

References

[1] A. T. et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

[2] G. A. et al. Distinguishing the knowable from the unknowable with language models. *ICML 2024*, 2024. URL https://arxiv.org/abs/2402.03563.

[3] J. F. et al. Do i know this entity? knowledge awareness and hallucinations in language models. 2024. URL https://arxiv.org/abs/2411.14257. Accepted at ICLR 2025.

[4] K. P. et al. The linear representation hypothesis and the geometry of large language models. . accepted at ICML 2024.

[5] L. G. et al. Scaling and evaluating sparse autoencoders. . URL https://arxiv.org/abs/2406.04093.

[6] T. W. et al. The geometry of refusal in large language models. . URL https://arxiv.org/abs/2502.17420.