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

Abstract

- 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 theorem**: concepts $\rightarrow$ directions; orthogonality encodes causal separability [1].
- **SAE objective**: 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. 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 $f : \mathbb{R}^d \rightarrow [0, 1]$ 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 $\rightarrow$ 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 $r \in \mathbb{R}^d$ 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**:

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

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.

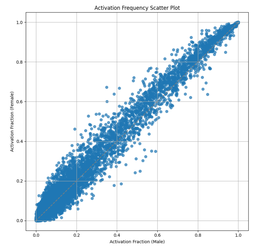


Figure 1. Gender Separation

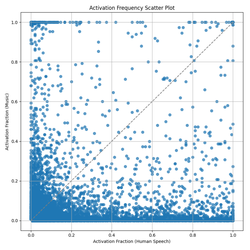


Figure 2. Music vs Speech

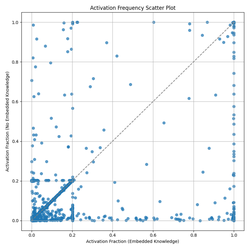


Figure 3. Knowledge Injection

Label Type	Count
Correct (Speech vs Music)	1,724
Incorrect	59
Other	217

Table 1. Music vs Speech Performance

Label Type	Count
Correct (Speaker Gender)	4421
Incorrect	5139
Other	440

Table 2. Gender Classification Performance

Steering via Activation Engineering

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^V$ be a transformer language model, where d is the dimensionality of hidden activations and V is the vocabulary size. At layer $l \in \{1, \ldots, L\}$, the hidden representation is $h^{(l)} \in \mathbb{R}^d$.

Activation steering modifies this representation via:

$$\tilde{h}^{(l)} = h^{(l)} + \alpha v$$

where $v \in \mathbb{R}^d$ is a concept direction and $\alpha \in \mathbb{R}$ is a scaling coefficient.

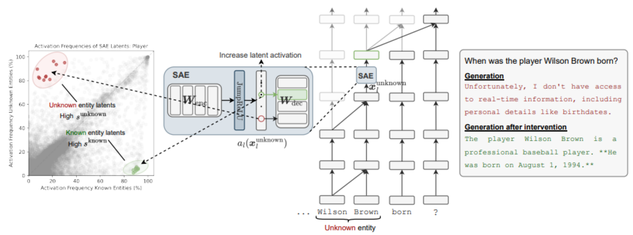
Let $a_{l,j}(x)$ be the activation of latent j at layer l for input x. Let $\mathbb{1}[\cdot]$ be the indicator function. Define:

$$f_{l,j}^{\text{known}} = \frac{1}{N^{\text{known}}} \sum_{i=1}^{N^{\text{known}}} \mathbb{1}[a_{l,j}(x_i^{\text{known}}) > 0]$$

$$f_{l,j}^{\text{unknown}} = \frac{1}{N^{\text{unknown}}} \sum_{i=1}^{N^{\text{unknown}}} \mathbb{1}[a_{l,j}(x_i^{\text{unknown}}) > 0]$$

Then the latent separation scores are:

$$s_{l,j}^{\text{known}} = f_{l,j}^{\text{known}} - f_{l,j}^{\text{unknown}}, \quad s_{l,j}^{\text{unknown}} = f_{l,j}^{\text{unknown}} - f_{l,j}^{\text{known}}$$



Further Work and Conclusion

- Identify causal directions relevant to uncertainty and AI safety.
- Bridge mechanistic interpretability with uncertainty quantification.
- Automate the process of finding interpretable latents in audio LLMs.

References

[1] Danny Hernandez, Tom Jones, Jack Clark, Catherine Olsson, Neel Nanda, Andy Mu, Nelson Conerly, Nova DasSarma, Nicholas Schiefer, Nicholas Joseph, Nelson Elhage, Dario Amodei, et al. The geometry of categorical and hierarchical concepts in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 28948–28963, 2022.