Demystifying Representation Alignment in Multilingual and Multimodal Aspects of Large Audio Language Models

Anonymous ACL submission

Abstract

The mechanistic interpretability of large language models (LLMs) has facilitated advancements in controllable generation, knowledge editing, model stitching, and other foundational techniques. However, the behavior of LLMs in multimodal multilingual contexts remains largely unexplored, despite their increasing complexity. This paper investigates how large audio language models (LALMs) process and represent language, modality, and speaker demography. Through a series of experiments, we analyze the latent representations extracted from diverse input cases using two state-of-the-art open-weight LALMs: Ultravox 0.5 and Qwen2 Audio. Our study examines patterns in these representations to uncover the processing mechanisms of LALMs across seven languages and two modalities (text and speech). Additionally, we explore paralinguistic speech features such as gender, age, and accents, as well as acoustic features arising from variations in the recording setup. By bridging the gap in interpretability LALMs, this work provides insights into their behavior and lays the groundwork for future research in this critical area.

1 Introduction

Investigating the internal mechanisms of large language models (LLMs) has deepened our interpretability of how they learn and process inputs (Zhang et al., 2025b; Wilie et al., 2025; Zhao et al., 2024; Gurnee and Tegmark, 2024). Previous studies have employed techniques, such as probing (Gurnee and Tegmark, 2024; Azaria and Mitchell, 2023), sparse autoencoders (Kang et al., 2025; Ghilardi et al., 2024), and agnostic-specific neurons (Mondal et al., 2025; Tang et al., 2024; Bhattacharya and Bojar, 2023) to uncover LLM behavior across different input scenarios. These investigations have revealed insights into the range of mechanistic behaviors, from LLM ability to represent specific downstream tasks (Gurnee and Tegmark, 2024; Azaria and Mitchell, 2023) to broader phenomena like representation alignment across controlled parameters (Zhou et al., 2025). Furthermore, recent studies have also leveraged these insights into novel advancements in LLM capabilities. Mechanistic insights have enabled methods such as controllable text generation (Liu et al., 2024a,b), knowledge editing (Meng et al., 2023), and model stitching (Moschella et al., 2023) to get more precise control, adaptation, and modularity in LLMs. These marks a significant step toward more interpretable and adaptable LLM.

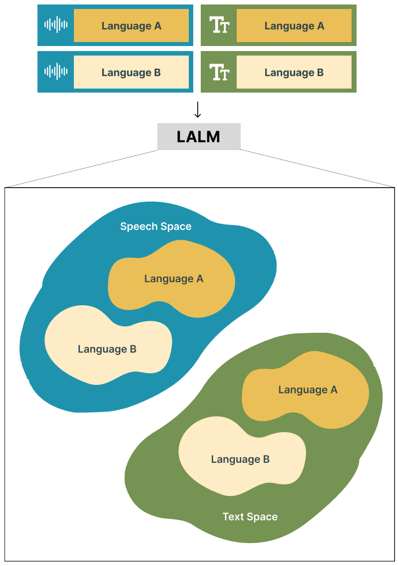


Figure 1: Our findings suggest that language clusters are present within each modality's representational space. These clusters emerge in the early and late layers in text space, and persist throughout all layers in speech space.

As LLMs increase in complexity, their growing processing steps present challenges in interpreting their mechanistic processes, stemming not only from architectural advancements but also from the complexity of inputs, which now extend beyond text-based reasoning to task-specific contexts (Li et al., 2025) and multimodal settings e.g. audio language models (ALMs) (Chu et al., 2024) and visual language models (VLMs) (Bai et al., 2025). While existing research has identified patterns such as attention mechanisms (Yan et al., 2025), knowledge storage (Cao et al., 2024), and token representations (Wu et al., 2025), achieving a sufficient level of interpretability remains a significant hurdle. Substantial progresses have been made in understanding how LLMs process textual information (Ryan et al., 2024; Wilie et al., 2025), the same level of interpretability, however, does not fully translate to multimodal LLMs , which must integrate information from various modalities such as image, audio, and video, adding layers of complexity (Yin et al., 2024).

To bridge the gap in interpretability of the internal mechanisms of multimodal LLMs, our study explores the mechanistic behavior of large audio language models (LALMs), focusing on how representation alignment occurs across different input cases. In this study, we investigate representation alignment across three key aspects in LALMs, i.e., modality, language, and speaker demography. By analyzing activation values in response to specific inputs and comparing them across a series of experiments, we aim to investigate how these features are represented in LALMs across layers. Our work contributes to a deeper interpretability of the mechanistic behavior of LALMs, paving the way for further research into the interpretability of LALMs.

Our study provides key insights into the representation alignment of multilingual LALMs:

- The capability of LALMs to process inputs is reflected in their representation patterns.
- We show that the textual semantic alignment – both within and across languages – is retained after adding the speech modality support to language models.
- Within the speech modality, we conclude that the representation is clustered semantically rather than based on the speaker demography indicating that LALMs tend to be robust to variation in paralinguistic features.
- We scrutinize the alignment across modalities and find that there is no semantic alignment emerged between parallel speech and textual representations.
- Our findings reveal different behaviors in multilingual processing between text and speech: while cross-language semantic alignment is emerged in text space, it is not present in speech space.

2 Related Works

Latent Representations Across Languages, Modalities, and Speakers. LLMs exhibit structured latent activation patterns that vary across languages and modalities. In the multilingual setting, some neurons are shared across languages while others are language-specific, though this neuron sharing does not necessarily align with linguistic similarity (Wang et al., 2024b). LLMs contain both language-specific and language-agnostic regions, with the former predominantly located in the early and late layers of the model (Zhao et al., 2024; Tang et al., 2024). As training progresses and model capacity increases, semantically equivalent inputs across different languages tend to converge within a shared latent space (Chang et al., 2022; Wilie et al., 2025; Zeng et al., 2025). Initially, knowledge is grounded in a dominant language, but the model gradually constructs language-specific knowledge systems as exposure to new languages increases (Zhao et al., 2024; Chen et al., 2025).

Beyond language, latent representations for different input modalities, such as audio and image, also require specialized encoders to interface with the language model. Multimodal LLMs (MLLMs) use modality-specific adapters like Whisper for speech (Radford et al., 2022) and ViT for images (Dosovitskiy et al., 2021), as common sense and multimodal data reside in distinct embedding spaces (Wang et al., 2024a). While speaker embeddings in speech processing have shown strong clustering behavior by speaker identity (Horiguchi et al., 2025; Ashihara et al., 2024), their integration into LALMs remains an open question. Existing works in speech representation often rely on supervised models to extract speaker-specific features (Zhang et al., 2025a), but analogous mechanisms in LALMs are yet to be systematically explored.

Language Model Mechanistic. Transformer-based language models have been shown to encode knowledge by projecting activations linearly across various output settings, including binary (Olah et al., 2020), continuous (Gurnee and Tegmark, 2024), and task-specific outputs (Nanda et al., 2023). These models also contain knowledge neurons, units whose activations are positively correlated with specific factual expressions, enabling targeted knowledge editing (Dai et al., 2022). While internal state analysis has contributed significantly to our mechanistic intepretability of LLMs, most

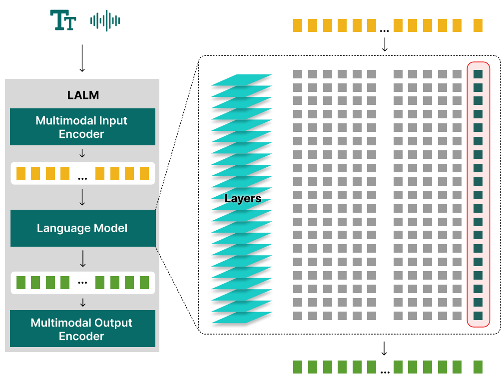


Figure 2: The process of extracting LALM representations by recording activation values from each layer of the LALM in response to the last token of an input (shown in red .).

existing studies have focused on narrow, task-specific scenarios (Olah et al., 2020; Nanda et al., 2023; Gurnee and Tegmark, 2024; Ji et al., 2024). A limited number of works have investigated representation alignment across different input cases, and these are confined to text-based LLMs (Tang et al., 2024; Zhou et al., 2025; Wilie et al., 2025). In this work, we extend internal state analysis and representation alignment to the multimodal text-speech domain to gain insight into the mechanistic behavior and representation alignment in LALMs.

3 Methods

3.1 Activation Values Extraction

Large Audio Language Models. Multimodal LLMs generally consist of three main components: a multimodal input encoder, a feature-fusion language model, and a multimodal output decoder (Wang et al., 2024a). As shown in Figure 2, we focus exclusively on the language model component since our aim is to analyze the representation alignment in LALMs. To understand the mechanisms by which LALMs process inputs, we utilize the activation values from each layer to observe patterns in input processing. We extract the activation values produced by the LALMs for each input and use these latents as the primary objects of observation in our experiments. This extraction process is performed for every input case used in the subsequent experiments.

Extracting Representation. Since inputs have varying lengths, they are translated into a varying number of tokens, leading to varying sizes of latents produced by LALMs. This variation becomes even more significant with the presence of multimodal properties, such as text and speech, because the tokenization processes differ: text is tokenized using textual units (Kudo and Richardson, 2018; Bostrom and Durrett, 2020), while speech uses audio frames (Radford et al., 2022). To address this, we standardize the latent size by taking only the activation values produced by the last token inputted into the LALMs for each input because all the LALMs we use are based on transformer decoder models (Vaswani et al., 2017). By using this approach, differences in language latent sizes caused by varying numbers of tokens are eliminated. To minimize variation of latent size further, we use the output produced by the last sublayer of the i-th layer for a given input to represent activation values at layer i. Using this approach, the latent size varies only with the number of neurons in the last sublayer of the i-th layer. Each input will then produce n Layer Latents, where n is the number of layers in the language model component.

Task	Scope
Modality	Text, Speech Audio
Gender	Male, Female
Language	English, French, German, Chinese, Japanese, Indonesian, Vietnamese
Accent (English)	British, American, Scottish, Northern Irish, Irish, Indian, Welsh, Canadian, South African, Australian, New Zealand

Table 1: Feature set used in our experiments.

3.2 Utilizing the Activation Values

Given a layer latent ($\mathbf{z}_i \in \mathbb{R}^n$), where n represents the number of neurons in the last sublayer of layer i, our objective is to visualize and analyze the internal representations of a LALM. Due to the high dimensionality of $\mathbf{z}_i$, direct visualization is infeasible. To address this, we employ dimensionality reduction techniques to map $\mathbf{z}_i$ onto a 2D Cartesian plane. Specifically, we utilize t-SNE for dimensionality reduction, which preserves the similarity between points and maintains local structures. Let $\tilde{\mathbf{z}}_i \in \mathbb{R}^2$ denote the 2D embedding of the high-dimensional latent vector $\mathbf{z}_i$, where the t-SNE algorithm maps each $\mathbf{z}_i$ to a 2D vector $\tilde{\mathbf{z}}_i$. The set of embedded vectors $\tilde{\mathbf{z}}_i$ can be visualized to gain insights into the internal representational structure learned by the LALM, with clustering or separation in the 2D space potentially reflecting semantic, syntactic, or task-relevant groupings encoded by the model. To complement this qualita-

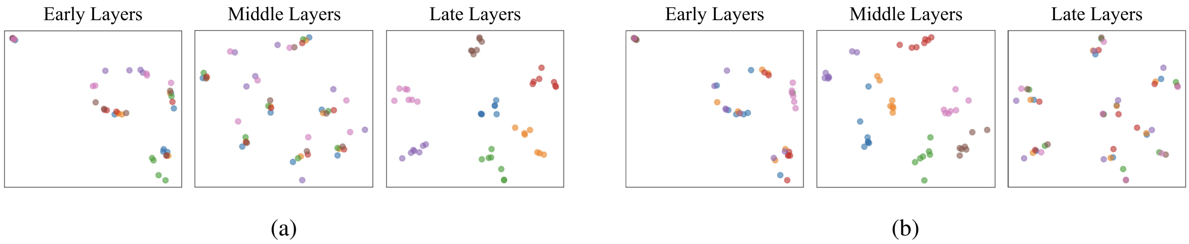


Figure 3: Representation of texts in 7 languages and 7 sample semantics (context) from Ultravox 0.5 LLaMA 3.1 8B across layers. Points are colored by (a) language and (b) semantics, revealing language-specific clusters in the final layers, semantic clusters in middle layers, and a blend of both language and semantics in early layers.

tive analysis, we perform quantitative evaluations of the clustering and separation patterns observed in the 2D embedding space. First, we compute the Euclidean similarity by measuring the Euclidean distance between any pair of embedded vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, serving as a proxy for assessing similarity in the original high-dimensional space. Next, we evaluate the silhouette score of the local clusters formed in the 2D projection, which quantifies how well each point fits within its cluster compared to others, thereby reflecting cluster compactness and separability. As an additional analysis, we may also compute the Euclidean distances between centroids of well-formed clusters to help quantify the degree of separation between distinct internal representation groups.

First Layer				
	A1	**A2**	**B1**	**B2**
A1		0.75	1.38	1.41
A2	0.75		1.45	1.49
B1	1.38	1.45		0.37
B2	1.41	1.49	0.37	
Last Layer				
	A1	**A2**	**B1**	**B2**
A1		7.46	19.13	18.81
A2	7.46		19.90	18.82
B1	19.13	19.90		7.08
B2	18.81	18.82	7.08	

Table 2: Distance between two samples of context (denoted by letters) in two linguistic structures (denoted by numbers) from Ultravox 0.5 LLaMA 3.2 1B. Texts with the same semantic meaning are closer to each other even when presented in different structures. Cells with same context are colored blue, while cells with different context are colored green.

4 Experiment Details

Language Model. We use three state-of-the-art LALMs publicly available on HuggingFace: Ultravox 0.5 LLaMA 3.2 1B[1], Ultravox 0.5 LLaMA 3.1 8B[1] (Grattafiori et al., 2024), and Qwen2 Audio 7B (Chu et al., 2024).

No Modification. We conduct analysis on the latent across 3 features: language, modality, and speaker demography for input of speech utterance (Table 1). All audio used in each experiment is speech audio, which means that each audio file has a transcript. We do not modify any model processes, alter model structures, or manipulate activation values. The only variable we change is the input cases, which lead to different internal states. We feed different input cases into the models, extract the corresponding activation values, and analyze them as they are.

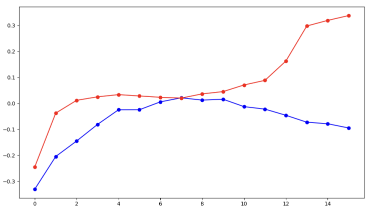
Dataset. Since our experiments involve multiple input features, we use several datasets to simulate diverse input scenarios while controlling certain parameters. The datasets used in this study include Common Voice 4 (Ardila et al., 2020), CoVoST 2 (Wang et al., 2021), CVSS 2 (Jia et al., 2022), M-Vicuna (Tang et al., 2024), VCTK (Yamagishi et al., 2019), and PAWS (Zhang et al., 2019).

[1]https://www.ultravox.ai/

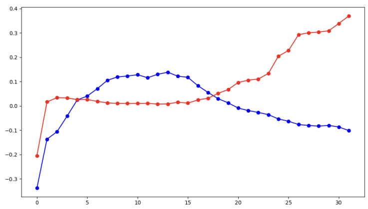
5 Results

5.1 Semantic Alignment in Text Modality

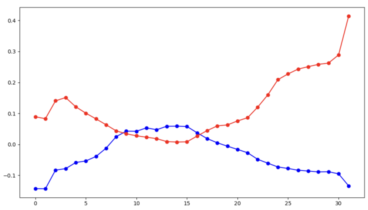
Monolingual Semantic Alignment. Before being processed by LLMs, all inputs are decomposed into tokens through a tokenization and embedding lookup process that transforms text into a vector format. This process not only converts the input into a form that the model can understand but also maps semantically similar words to nearby positions in the embedding space (Peng et al., 2024). The effect of this semantic alignment is especially noticeable in the early layers. Texts with the same meaning, even if they have different sentence structures, tend to have significantly closer vector representations compared to inputs from different contexts (Table 2). As the model moves through deeper layers, the



(a) Ultravox 0.5 LLaMA 3.2 1B

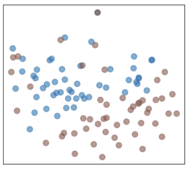


(b) Ultravox 0.5 LLaMA 3.1 8B

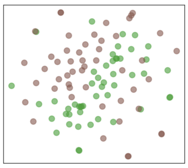


(c) Qwen2 Audio 7B

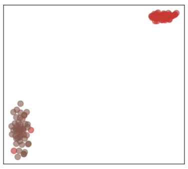
Figure 4: Silhouette scores (y-axis) for language (red) and semantic clusters (blue) across layers (x-axis) based on representations extracted from 3 different models. The results show that language clusters emerge in the early and late layers, while semantic clusters are prominent in the middle layers.



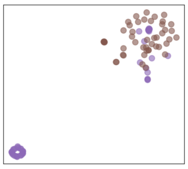
(a)



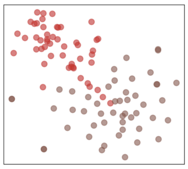
(b)



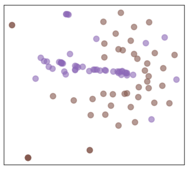
(c)



(d)



(e)



(f)

Figure 5: Representations in the first layer extracted from Ultravox 0.5 LLaMA 3.1 8B for multilingual text pairs: (a) English-German, (b) English-French, (c) English-Japanese, and (d) English-Chinese; and from Qwen2 Audio 7B for (e) English-Japanese and (f) English-Chinese.

distance between these representations increases, reflecting a divergence in how the inputs are processed. This phenomenon occurs across all text inputs, as cross-context texts also show increasing separation. However, semantically similar texts, although they become more distant, still remain closer to each other than to representations from different contexts.

Multilingual Semantic Alignment. During text processing in LALM, we observed several interesting patterns when controlling the language of the inputs. Late layers tend to distinctly cluster text inputs from the same language together (Figure 3a), while early layers form clusters based on a combination of semantics and language (Figure 3). In contrast, the middle layers focus on language-agnostic processing i.e. semantic processing, as semantic clusters form during this stage before being separated again in the later layers (Figure 3b). These clusters present in all LALMs we use in this experiment (Figure 4). This pattern aligns with recent research identifying language-processing areas in LLMs (Tang et al., 2024; Zhao et al., 2024; Wilie et al., 2025), which suggests that the early and late layers play a key role in handling language-specific information.

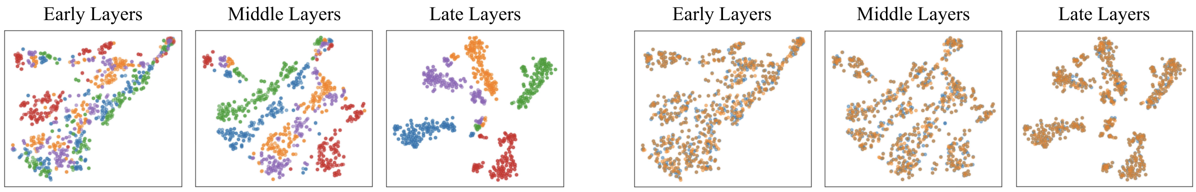
Several first layers show low silhouette scores, before increasing significantly afterward. We tested several multilingual texts and found the reason why this happens. It is because there are differences in the text embeddings inputted into the models. Inputs from languages that share similar linguistic structures often come in a similar space (Figure 5a, 5b). In contrast, inputs from totally different languages are represented as distinct clusters, showing that processing those languages needs separate processing spaces (Figure 5c, 5d). However, it seems to depend on the data the model is trained on, as shown in Figure 5e and 5f: the representation of English-Chinese and English-Japanese seems relatively closer, although still separated. The higher percentage of Chinese and Japanese data on QwenLM (compared to LLaMA) makes representations in the first layers tend to be closer to each other.

Text Alignment Key Insight

Mechanistics of text processing in language model are preserved even when speech modality support is added to the model. The linguistic and semantic phenomena we observed in LALMs align with those found in text-only LLMs.

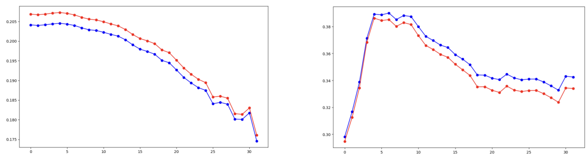
5.2 Semantic Alignment in Speech Modality

Audio Robustness. We found that LALMs demonstrate semantic clustering behavior when given speech input. As the process moves to deeper



(a) (b)

Figure 6: Representation across layers Ultravox 0.5 LLaMA 3.1 8B of speech inputs from 2 controlled recording devices with 5 sample English transcripts, colored by (a) semantic (each color denotes each speech transcript) and (b) recording devices (each color denotes each recording device). This image suggests LALM clusters inputs by their transcripts, and differences in recording setup do not affect the representation much as they are distributed evenly.



(a) (b)

Figure 7: Clustering performance (y-axis) is plotted across model layers (x-axis). Red denotes raw speech input, while blue denotes normalized speech. Normalization affects models differently: Ultravox 0.5 LLaMA 3.1 8B (a) shows weaker clustering overall, whereas Qwen2 Audio 7B (b) demonstrates better representation clusters.

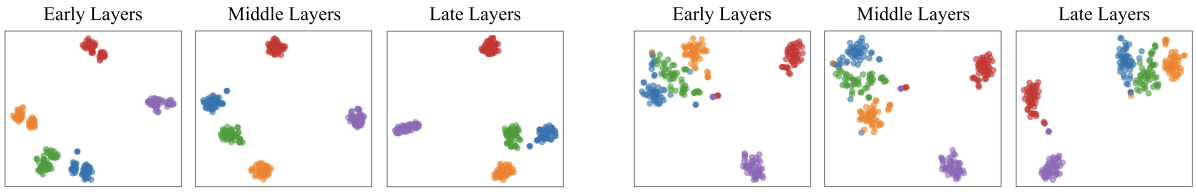
	$c_1mic_1p_1$	$c_1mic_2p_1$	$c_1mic_1p_2$	$c_2mic_1p_1$
$c_1mic_1p_1$		6.61	19.05	31.93
$c_1mic_2p_1$	6.61		20.86	31.21
$c_1mic_1p_2$	19.05	20.86		31.69
$c_2mic_1p_1$	31.93	31.21	31.69	

Table 3: The distance between representations Ultravox 0.5 LLaMA 3.2 1B in a sample layer under the controlled context (denoted as c_x) in a controlled recording setup (denoted as mic_x) spoken by a controlled speaker (denoted as p_x) shows that the differences between microphones result in the least representation divergence (cell colored green).

layers, the model attempts to tightly cluster inputs with the same transcript, and we can see semantic clusters emerge in the late layers (Figure 6a). In our experiment, where we controlled the recording devices, we observed that variations in recording devices introduced minimal amount of divergence. The differences in the recording setup did not introduce much divergence, as the representations overlapped with each other in an evenly distributed manner (Figure 6b).

We also found that differences in the recording setup resulted in the least amount of divergence in representation compared to inputs with different contexts and speakers (Table 3). However, cross-context and cross-speaker representations varied: in some cases, cross-context inputs were more distant from each other, while in other cases, cross-speaker inputs were. We also tested simple preprocessing of the speech before feeding it into the LALM, where we normalized the speech tensor to the range $[-1, 1]$ under varying recording device conditions. We found that this preprocessing had differing effects on the models: the LLaMA-based model produced poorer clusters, while the Qwen model produced better clusters (Figure 7). This highlights differences in capabilities for processing acoustic features in speech. These clusters emerge not only from real speech recordings, but also computer-generated speeches. However, in the case of unified computer-generated speech, clustering tends to be relatively better in the early layers. In contrast, multi-speaker speech, whether real or synthetic, often shows overlapping representations in the early layers. This suggests that speaker embeddings in speech recordings may influence, or even distort, the representations of the speech content.

Monolingual Semantic Alignment. As shown in Figure 6 and Figure 9, LALMs clusters speeches with similar semantic meaning together. Processing of speech in LALMs prioritizes understanding of the speech i.e. semantic meaning rather than paralinguistic features that come with the speech. As a result, they tend to separate semantically different speech from the same speaker more strongly than semantically similar speech from different speakers. This may be due to the fact that current LALMs do not yet support the processing of paralinguistic speech features. Nevertheless, their effectiveness



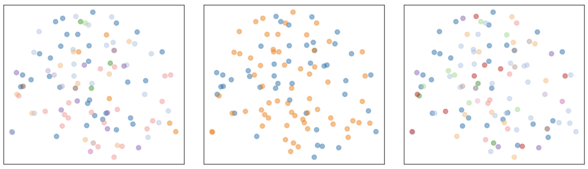
(a) (b)

Figure 8: Representations across layers of multilingual speech extracted from (a) Qwen2 Audio 7B and (b) unified Ultravox 0.5 LLaMA 3.1 8B show distinct speech clusters across all layers during multilingual speech processing, indicating separate processing spaces for each language.

in clustering speech by semantic meaning suggests the existence of a well-defined semantic space, similar to that observed in text processing.

Multilingual Semantic Alignment. We found that multilingual speech inputs are represented as tightly clustered based on their language. Unlike in multilingual text processing (Figure 3), semantically similar clusters across languages do not emerge in the middle layers for speech (Figure 8). Instead, language-based clusters are present from the beginning to the end of processing. Different types of clustering behavior emerge across models. LLaMA-based models tend to group similar language (Figure 8b), such as French, Spanish, and German, into overlapping clusters, while distinct languages like Chinese and Japanese form separate clusters. In contrast, Qwen-based models represent each language in distinct, non-overlapping clusters (Figure 8a). These language clusters remain stable throughout the processing layers. Due to data limitations, all tested speech samples were recorded using different devices and in varied environments. However, since previous experiments suggest that such differences have minimal effect on the representations, we can reasonably conclude that language-based clustering also emerges in multilingual speech processing, just as it does in multilingual text processing. This phenomenon further suggests that current speech processing in LALMs is primarily capable of capturing "what" is being said, stopping at understanding literal speech content, without fully modeling the real semantics of the speech.

Speaker Demography. Since none of the LALMs in our experiment natively support paralinguistic features, this limitation is evident in the absence of meaningful clusters based on speaker demographics such as age, accent, and gender, even under controlled contexts and recording devices



(a) (b) (c)

Figure 9: Representations of speech inputs with the same transcript in a sample layer of Qwen2 Audio 7B, colored by (a) accent, (b) gender, and (c) age, show no meaningful clustering based on paralinguistic features. Each color denotes a different category.

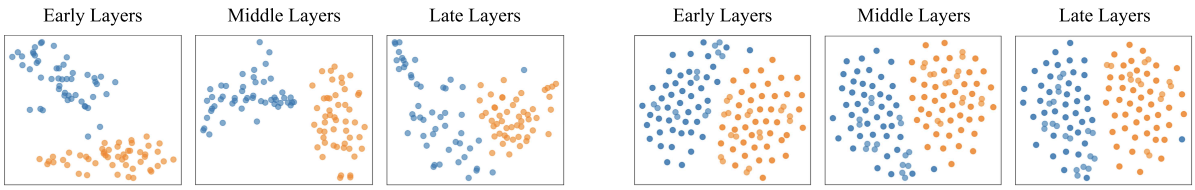
(Figure 9). We also conducted experiments under more controlled settings, combining multiple paralinguistic features (e.g., gender within accent, age within gender), and similarly observed no meaningful clustering. This suggests that speech representations in LALMs are more strongly influenced by "what" is said rather than "how" it is said.

Speech Alignment Key Insight

Mechanistics of speech processing in LALMs are primarily designed to capture the literal content ("what" is being said), but they often struggle to fully model the real semantic meaning or capture paralinguistic audio features.

5.3 Speech-Text Semantic Alignment

Monolingual Semantic Alignment. In LALMs, text and speech inputs are processed in distinctly separate representational spaces. This distinction becomes particularly evident through a series of controlled experiments involving various types of speech: controlled recording setups, varying recording conditions, multilingual pair text-speech, and computer-generated speech (Figure 10). These experiments consistently show that the representational space for each modality forms reliably and



(a) (b)

Figure 10: Representations across layers extracted from Ultravox 0.5 LLaMA 3.2 1B for (a) text with computer-generated speech and (b) text with real speech recording, showing separation in processing space from the beginning to the end of processing.

independently across all layers, indicating robust and modality-specific semantic encoding. This separation between modalities is not unexpected, as each input type undergoes different encoding before being fed into the model. As a result, the semantic meaning of inputs tends to cluster within its own modality space across layers, despite the underlying semantic similarity.

Multilingual Semantic Alignment. In multilingual contexts, both text and speech inputs form clusters during processing. Figure 11 illustrates the interaction between multilingual text–speech pairs in LALMs. We observe that language-specific clusters emerge in both the early and late layers. However, their representations pose different dynamics. In text, these language clusters tend to dissolve in the middle layers, where semantically similar texts form shared clusters (in Figure 3b). In contrast, multilingual processing in the speech modality remains confined within language-specific clusters throughout the entire pipeline (in Figure 8).

Speech-Text Alignment Key Insight

Mechanistics text and speech inputs in LALMs are present in separate representational spaces, meaning there is currently no cross-modal semantic alignment, which hinders their ability to fully connect meaning across modalities.

6 Discussion

We demonstrate that semantically identical audio samples could occupy distinct regions in the representational space (Figure 8). This variation arises from differences in audio features, specifically acoustic and paralinguistic elements, that accompany the speech signal (Table 3). Our findings indicate that current LALMs are not yet capable of fully accounting for these variations, as evidenced by the absence of consistent clustering within their

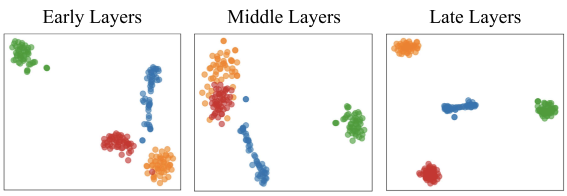


Figure 11: Representations of English text (orange), English speech (blue), Japanese text (red), and Japanese speech (green) extracted from Ultravox 0.5 LLaMA 3.1 8B across layers.

semantic spaces (Figure 6b, 9). These features may influence, or even distort, the representation of speech content. However, the extent to which these representations can vary while still being considered semantically equivalent remains an open question. Understanding this boundary is crucial for improving alignment and robustness in speech processing tasks. This is particularly relevant for multilingual scenarios, where differences in recording setups and language can introduce additional variation in the representational space.

7 Conclusion

Our study provides foundational insights into the representation of LALMs, demonstrating how parallel semantic clusters exist in different representational spaces and revealing the potential for cross-space mapping (Figure 1). The results suggest that LALMs can encode semantically equivalent inputs in distinct representational spaces while still maintaining the ability to organize their semantics. For future work, we encourage researchers to build upon these findings by exploring real-world multilingual environments, expanding the scope to include a broader range of linguistic phenomena and use cases. The journey to better understand and control semantic representations is still in its early stages, and we hope our study inspires others to contribute to this exciting field.

Limitations

Due to computing constraints, we were only able to analyze representations in three LALMs. Larger LALMs may have greater capacity to represent input features and could reveal additional patterns beyond those observed in the smaller models we used. To enable more understanding about multilingual speech processing, future work should employ a set of speakers delivering parallel multilingual transcripts in controlled recording setups. This would allow for more consistent cross-language comparisons.

Ethical Consideration

All language models and datasets used in our experiments are publicly available, primarily sourced from Hugging Face. We ensured compliance with the licenses and usage policies associated with each resource. No proprietary or private data was used, and all experiments were conducted with the intention of promoting open and reproducible research. AI assistants such as ChatGPT were used as productivity tools to support ideation, code debugging, and refining explanations. Their use was limited to non-generative support and did not replace original research, critical analysis, or authorship. All final decisions, implementations, and evaluations were conducted by the authors.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Takanori Ashihara, Takafumi Moriya, Shota Horiguchi, Junyi Peng, Tsubasa Ochiai, Marc Delcroix, Kohei Matsuura, and Hiroshi Sato. 2024. Investigation of speaker representation for target-speaker speech processing. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 423–430.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Sunit Bhattacharya and Ondřej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *Computing Research Repository*, abs/2411.17401.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiawei Chen, Wentao Chen, Jing Su, Jingjing Xu, Hongyu Lin, Mengjie Ren, Yaojie Lu, Xianpei Han, and Le Sun. 2025. The rise and down of babel tower: Investigating the evolution process of multilingual code large language model. In *The Thirteenth International Conference on Learning Representations*.

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. 2024. Qwen2-audio technical report.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Davide Ghilardi, Federico Belotti, Marco Molinari, and Jaehyuk Lim. 2024. Accelerating sparse autoencoder training via layer-wise transfer learning in large language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 530–550, Miami, Florida, US. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Shota Horiguchi, Takafumi Moriya, Atsushi Ando, Takanori Ashihara, Hiroshi Sato, Naohiro Tawara, and Marc Delcroix. 2025. Guided speaker embedding. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.

Hao Kang, Tevin Wang, and Chenyan Xiong. 2025. Interpret and control dense retrieval with sparse latent features. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 700–709, Albuquerque, New Mexico. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kun Li, Xinwei Chen, Tianyou Song, Chengrui Zhou, Zhuoran Liu, Zhenyan Zhang, Jiangjian Guo, and Qing Shan. 2025. Solving situation puzzles with large language model and external reformulation.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024a. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yi Liu, Xiangyu Liu, Xiangrong Zhu, and Wei Hu. 2024b. Multi-aspect controllable text generation with disentangled counterfactual augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9231–9253, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting* of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech 2021*, pages 2247–2251.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Computing Research Repository*, abs/2408.01319.

Weixuan Wang, Barry Haddow, Wei Peng, and Alexandra Birch. 2024b. Sharing matters: Analysing neurons across languages and tasks in llms. Workingpaper, ArXiv.

Bryan Wilie, Samuel Cahyawijaya, Junxian He, and Pascale Fung. 2025. High-dimensional interlingual representations of large language models.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng YAN. 2025. Towards semantic equivalence of tokenization in multimodal LLM. In *The Thirteenth International Conference on Learning Representations*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Shaotian Yan, Chen Shen, Wenxiao Wang, Liang Xie, Junjie Liu, and Jieping Ye. 2025. Don't take things out of context: Attention intervention for enhancing chain-of-thought reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Ke Zhang, Junjie Li, Shuai Wang, Yangjie Wei, Yi Wang, Yannan Wang, and Haizhou Li. 2025a. Multi-level speaker representation for target speaker extraction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025b. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. Linguistic minimal pairs elicit linguistic similarity in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888, Abu Dhabi, UAE. Association for Computational Linguistics.