Geometric Alignment without Functional Equivalence: A Layer-wise Analysis of the Speech–Text Modality Gap

Ming-Hao Hsu[1] **Xiaohai Tian**[2] **Jun Zhang**[2] **Lu Lu**[2] **Yuxuan Wang**[2] **Zhizheng Wu**[1,3,4]*

Abstract

End-to-end speech large language models (LLMs) often answer the same question less accurately from speech than from text, even when the spoken content matches the text prompt. Existing analyses study this modality gap through the geometry of representations and measure how closely speech representations align with text. Geometric alignment gives the gap a measurable form, but it does not show whether the model uses aligned representations in the same way. This work connects representation geometry with answer formation and asks whether geometric alignment implies functional equivalence. We compare cross-layer Centered Kernel Alignment (CKA) with layer-wise answer margins for four open-weight speech LLMs on SpeechMMLU and VoiceBench BBH. In Qwen2.5-Omni and MiniCPM-o, speech and text representations align in middle layers while their late-layer answer margins still differ, so mid-layer alignment does not imply functional equivalence. Qwen2-Audio instead shows weak alignment, which indicates a different failure mode. An attention intervention at fixed input and token count identifies a causal effect of late attention allocation on answer-code separation, a quantity distinct from answer correctness. Compressing the speech signal or its token sequence repairs a subset of the errors that occur only with speech.

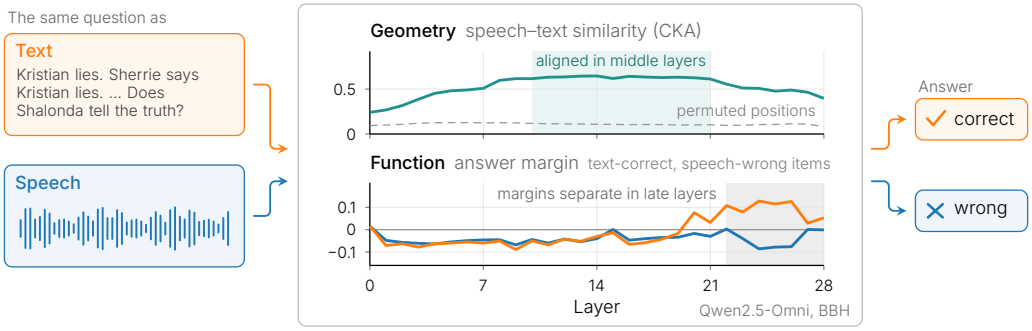


Figure 1: **Geometric alignment without functional equivalence.** The same question enters a speech LLM as text and as speech. Speech–text CKA is high in middle layers, yet the answer margins of the two inputs separate in late layers. Curves use Qwen2.5-Omni on BBH, with CKA from Appendix A.2 and margins averaged over the 233 items answered correctly only from text.

*[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen. [2]ByteDance. [3]City University of Macau. [4]Amphion Technology Co., Ltd. Correspondence to: Zhizheng Wu <wuzhizheng@cuhk.edu.cn>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Speech-based assistants aim to make the reasoning capabilities of language models available through spoken interaction [Chu et al., 2023, Zhang et al., 2023, Ghosh et al., 2025]. End-to-end speech large language models (LLMs) pursue this goal by passing speech representations directly into an LLM, so that a spoken question can draw on the same knowledge and reasoning as a written one. When an answer depends on linguistic content alone, the spoken and written versions of a question carry the same information, and the model should reach the same answer from either. Yet speech LLMs answer such questions less accurately from speech than from text (Figure 2), a difference known as the modality gap [Chen et al., 2026, Fang et al., 2025b, Wang et al., 2026a, Hsu et al., 2026b]. The reasoning ability that the model already shows on text is therefore not fully available through speech.

A common way to study this gap is through the geometry of representations. Speech and text reach the LLM through different input pathways, and prior work measures how closely the resulting representations align inside the model [Mousavi et al., 2025, Xiang et al., 2025], while related analyses describe modality-specific regions in vision-language models [Liang et al., 2022]. This view gives the gap a measurable form and motivates aligning speech representations more closely with text. Yet geometric similarity describes what representations look like, and it does not by itself show how the model uses them to reach an answer [Ding et al., 2021, You, 2025, Park et al., 2024].

A complementary line of work studies how a language model forms its answer across layers. Layer-wise readouts such as the logit lens and the tuned lens show that a final answer is typically preceded by intermediate evidence formation and late-layer sharpening [nostalgebraist, 2020, Belrose et al., 2023, Lad et al., 2024]. This view follows the answer directly. For speech, however, layer-wise analyses have so far traced how acoustic features become semantic representations in recognition and classification systems [Glazer et al., 2026, Ma et al., 2026]. It therefore remains unclear whether speech that has become geometrically aligned with text also goes on to form answers in the same way.

This work connects the two views to study where the modality gap arises inside the model (Figure 1). We ask whether geometric alignment implies functional equivalence, meaning that the model reaches the same answer from speech as from text. Specifically, we present the same questions as speech and as text and measure where the two representations become similar across layers with Centered Kernel Alignment (CKA), a measure of similarity between two sets of representations. At each layer, we also track how strongly the model favors the correct answer over competing ones. We then edit attention inside the model while holding the input fixed, which tests whether late-layer answer formation depends causally on how attention is allocated over the speech input.

We study four open-weight speech LLMs, Qwen2.5-Omni [Xu et al., 2025], MiniCPM-o [OpenBMB, 2025], Qwen2-Audio [Chu et al., 2024], and LLaMA-Omni [Fang et al., 2025a], on Speech-MMLU [Hendrycks et al., 2021, Team et al., 2025] and VoiceBench BIG-Bench Hard (BBH) [Chen et al., 2026, Suzgun et al., 2023]. The results show that geometric alignment does not imply functional equivalence. In Qwen2.5-Omni and MiniCPM-o, speech and text representations become aligned in middle layers, yet late layers still separate the correct answer from its competitors differently for the two inputs. In Qwen2-Audio, the alignment itself is weak, so the gap takes different forms across models. In Qwen2.5-Omni, the attention intervention shows that late-layer attention allocation causally affects how strongly the model separates candidate answers, irrespective of which answer is correct, and compressing the speech signal or its token sequence repairs a subset of speech-only errors. These findings suggest that work on the modality gap should examine late-layer answer formation alongside representation alignment.

2 Related Work

Speech LLMs and the modality gap. Speech processing has moved from cascades of automatic speech recognition (ASR) and LLMs toward end-to-end speech LLMs [Zhang et al., 2023, Chu et al., 2023, Fang et al., 2025a]. Recent surveys organize this space into adapter-based models that map speech features into an LLM input space [Chu et al., 2024, OpenBMB, 2025, Hsu et al., 2026a] and more unified systems that tokenize or model speech directly [Ghosh et al., 2025, Yang et al., 2025, Cui et al., 2025]. These systems perform well on recognition, translation, and instruction following. Yet speech-input reasoning still lags behind matched text-input reasoning [Chen et al., 2026, Fang

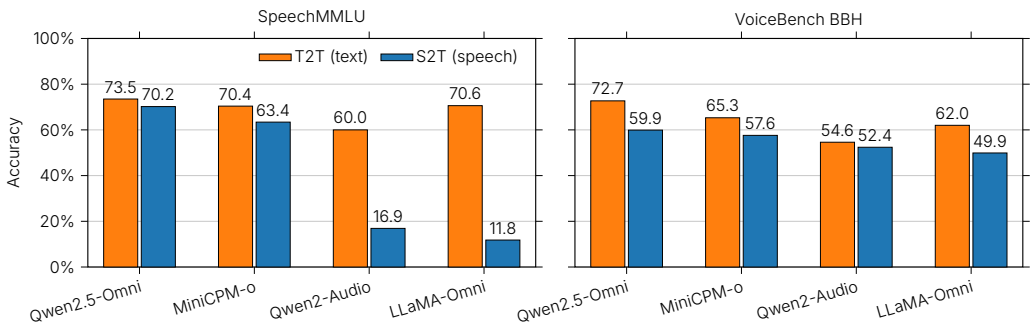


Figure 2: **The speech–text modality gap.** In the benchmark overview, speech-to-text (S2T) accuracy is below matched text-to-text (T2T) accuracy for all four speech LLMs on both SpeechMMLU and VoiceBench BBH.

et al., 2025b, Wang et al., 2026a, Hsu et al., 2026b], and work on closing this gap targets model-level score improvement [Wang et al., 2026a]. In contrast, we study where the speech and text routes differ within a model, using paired questions whose answers are determined by linguistic content.

Geometric analyses of the modality gap. In vision-language models, Liang et al. [2022] show that different modalities occupy separate cones of a shared embedding space. For speech, ALAS [Mousavi et al., 2025] measures latent speech–text alignment layer by layer, and Xiang et al. [2025] analyze speech–text representational mismatch empirically. Both motivate stronger feature alignment. Yet analyses of cosine similarity [You, 2025], of linear representations [Park et al., 2024], and of how similarity measures track functional behavior [Ding et al., 2021] caution that geometric similarity alone does not determine how a model behaves. We instead test whether speech that is geometrically aligned with text also supports the same late-layer answer formation.

Layer-wise analysis of answer formation. The logit lens [nostalgebraist, 2020] and the tuned lens [Belrose et al., 2023] decode intermediate hidden states into predictions, and stage-wise analyses show that final-layer answers are often preceded by intermediate evidence formation and late residual sharpening [Lad et al., 2024]. For speech, mechanistic studies trace how acoustic features evolve into semantic representations in ASR or classification systems [Glazer et al., 2026, Ma et al., 2026]. We bring this perspective to speech reasoning, comparing matched speech and text inputs with cross-layer CKA, decision-position attention, and answer-margin trajectories.

Speech redundancy and compression. Speech models face a density mismatch: speech token rates can be much higher than semantic word rates [Zuo et al., 2025, Hsu et al., 2025], and transformer speech stacks contain substantial redundancy [Dorszewski et al., 2025]. Prior compression work often treats this redundancy mainly as an efficiency issue [Liu et al., 2025]. In contrast, we use compression as a diagnostic probe, comparing which speech-input errors temporal compaction and token merging repair (Section 4.6), and we examine attention allocation alongside answer separation.

3 Experimental Setup and Diagnostics

Task and behavioral gap. We compare the same semantic query under text-to-text (T2T) and speech-to-text (S2T) prompting. For paired items $\mathcal{D}$, let y_i be the reference answer and $\hat{y}_i^T, \hat{y}_i^S$ the parser-normalized text and speech answers. The dataset-level modality gap is the difference in exact-match accuracy:

$$\Delta_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \left[\mathbf{1}(\hat{y}_i^T = y_i) - \mathbf{1}(\hat{y}_i^S = y_i) \right]. \tag{1}$$

A positive value means matched speech input underperforms matched text input.

Models, datasets, and scope. We study Qwen2.5-Omni-7B [Xu et al., 2025], MiniCPM-o 2.6 [OpenBMB, 2025], Qwen2-Audio-7B-Instruct [Chu et al., 2024], and LLaMA-Omni [Fang

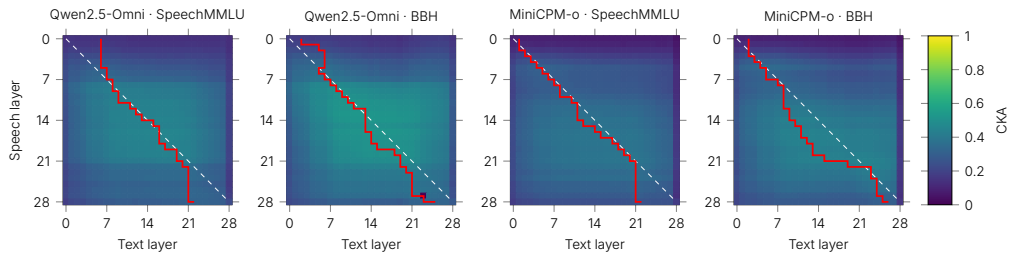


Figure 3: **Speech and text align in a broad middle-layer band.** Each heatmap shows cross-layer CKA between S2T layers on the y-axis and T2T layers on the x-axis after dynamic time warping (DTW) token alignment, with brighter colors denoting higher CKA. The red path marks the best-matching text layer for each speech layer, and the dashed diagonal marks equal layer indices.

et al., 2025a]. Figure 2 reports their T2T and S2T accuracies. Layer 0 denotes the embedding output, and layers 1–28 denote successive Transformer-block outputs for Qwen2.5-Omni and MiniCPM-o. SpeechMMLU [Hendrycks et al., 2021, Team et al., 2025] adapts MMLU to paired text and audio questions, while VoiceBench BBH [Chen et al., 2026] adapts BIG-Bench Hard [Suzgun et al., 2023] to speech reasoning tasks. Both benchmarks are semantics-matched: the answer is determined by linguistic content rather than speaker identity, emotion, prosody, or interaction structure. We additionally use MMSU [Wang et al., 2026b] as a speech-native boundary check for Qwen2.5-Omni and MiniCPM-o. MMSU includes tasks where stress, intonation, pauses, emotion, or background audio can be answer-relevant. It tests the scope of conclusions drawn from semantics-matched questions.

ASR-to-LLM baseline. To compare the direct speech route with a transcription interface, we evaluate an ASR-to-LLM pipeline that transcribes the audio and prompts the same downstream LLM with the transcript. For multiple-choice settings, we preserve or restore explicit question and option markers when present, because answer formatting is part of the task interface.

Representation similarity. We compare speech and text hidden states after aligning their different sequence lengths. Hidden states are extracted from the last user turn. Let $H^S \in \mathbb{R}^{T_s \times D}$ and $H^T \in \mathbb{R}^{T_t \times D}$ be speech and text hidden states at a selected base layer. We compute cosine similarities $M_{i,j}$ and use Dynamic Time Warping (DTW) to find a monotonic path from speech frames to text tokens. Because each audio clip is paired with the same ordered text prompt, monotonic speech–text alignment is a reasonable diagnostic assumption for this setting. We estimate this token correspondence once per item and freeze it across CKA layer pairs. The main CKA analyses use base layer 10 for SpeechMMLU and 20 for BBH. Both come from the stable region of a sweep over base layers 0, 5, 10, and 20, in which base layer 0 is the most sensitive (Appendix A.1). When multiple speech frames align to the same text token, we mean-pool those frames, producing paired matrices $X, Y \in \mathbb{R}^{N \times D}$ for each example. Our main similarity metric is linear CKA [Kornblith et al., 2019], computed on the aligned speech matrix X and text matrix Y:

$$\mathrm{CKA}(X, Y) = \frac{\langle HKH, HLH \rangle_F}{\|HKH\|_F \|HLH\|_F}, \quad K = XX^\top,\ L = YY^\top.$$

Here H is the centering matrix and N is the number of aligned token positions within the example. Metrics are computed per example and averaged over the fixed evaluation subset (Appendix I.1). We also report standardized L2 distance and layer-wise mean token norm as controls for geometric proximity and activation scale.

Scope of the DTW alignment. The frozen path is a fixed measurement frame for the geometry analysis. It relates residual-stream positions, without assuming that deep states contain only local information. A path that is uninformative at depth would misalign rows and lower CKA toward the position-permutation baseline, so it cannot create the middle-layer band. The best-matching-layer curve drawn on a CKA heatmap is a separate row-wise argmax, not the token-DTW path. DTW is used only for geometry, and the margin readouts and attention edits use no cross-modal token alignment.

Table 1: **Behavioral gap for Qwen2.5-Omni.** Subtable (a) gives benchmark-overview accuracy with deltas in points relative to direct S2T, where SpMMLU denotes SpeechMMLU. Subtable (b) gives the BBH categories Hyperbaton, Navigate, Sports, and Web-of-lies in the 1,024-token diagnostic, where the gap is T2T minus S2T.

(a) Behavior and ASR control

Setting	SpMMLU	Δ	BBH	Δ
Gold T2T	73.49	+3.25	72.7	+12.8
Direct S2T	70.24	base	59.9	base
ASR-to-LLM	59.23	−11.0	67.4	+7.5

(b) BBH category gap

Setting	Hyp.	Nav.	Sport	WoL
Gold-text T2T	75.2	58.0	64.8	89.6
Direct S2T	72.0	58.0	56.8	52.0
Gap	+3.2	+0.0	+8.0	+37.6

Answer formation. We characterize answer formation using layer-wise candidate margins. The *decision position* is the final valid prompt position after applying the chat template and its assistant-generation prefix. Readouts at this position measure next-token candidate preference. Let $z^c_{i,\ell}$ be the readout score for candidate answer c at item i and layer ℓ. For the paired BBH readout, intermediate states pass through the model's final RMS normalization (RMSNorm) and the shared vocabulary head, so hidden-state magnitude is divided out before logits are formed. The final layer uses native logits without a second normalization. Our primary margin is $m^p_{i,\ell} = q^{y_i}_{i,\ell} - \max_{c \neq y_i} q^c_{i,\ell}$, where $q = \mathrm{softmax}(z)$ over the task's answer candidates at temperature one. Candidate probabilities express relative preference within the answer set and remain sensitive to logit scale. We retain the logit margin $m^z_{i,\ell} = z^{y_i}_{i,\ell} - \max_{c \neq y_i} z^c_{i,\ell}$ as a complementary coordinate. Appendix B.1 specifies candidate surface forms and the separate free-generation evaluation that assigns correctness groups.

Attention. We measure attention dispersion at the final prompt query with entropy, peak mass, and token coverage, and edit it at fixed input to test its effect on answer separation (Appendix C.3).

4 Empirical Analysis

We first describe the behavioral gap and then follow the two views from the Introduction: where speech and text representations align, and where their answer formation diverges. Unless otherwise stated, the detailed diagnostics use Qwen2.5-Omni, with additional model diagnostics in the Appendix. The benchmark overview, 1,024-token diagnostic, and shared-instruction evaluation use different generation settings, summarized in Table 10.

4.1 The Behavioral Gap Is Task-Dependent

Figure 2 shows that S2T accuracy is below matched T2T accuracy for all four models, with a gap size that depends on the model and benchmark. Table 1 details the gap for Qwen2.5-Omni. On BBH, an ASR-to-LLM cascade improves over direct S2T by 7.5 percentage points but remains 5.3 points below gold-text T2T in the benchmark overview, while on SpeechMMLU the cascade is worse than direct S2T. A Whisper-large-v2 audit of 200 items per benchmark gives a mean item word error rate (WER) of 4.65% on BBH and 3.28% on SpeechMMLU (Table 12). The cascade comparison combines transcription effects with changes in the downstream input interface, so the accuracy differences do not separate those effects causally. The BBH categories are adjective ordering (Hyperbaton), navigation, sports plausibility, and truth-chain inference (Web-of-lies), which asks the model to follow truth relations over a chain of statements. The gap is highly task-dependent, with Web-of-lies driving most of the aggregate difference [Hsu et al., 2026b]. The comparison reverses when acoustic cues are relevant to the answer. On the 2,420-item MMSU reasoning subset, deterministic rescoring gives an S2T advantage over text-only prompting of 34.3 points for Qwen2.5-Omni and 26.9 points for MiniCPM-o, with contributions from acoustic cues and answer-formatting behavior (Table 15). The findings below therefore concern semantics-matched questions.

4.2 Speech and Text Align in Middle Layers

The cross-layer CKA heatmaps in Figure 3 show three regimes along the speech layers, which we call Phases I to III. In early layers, speech representations are not close to any text layer. In middle

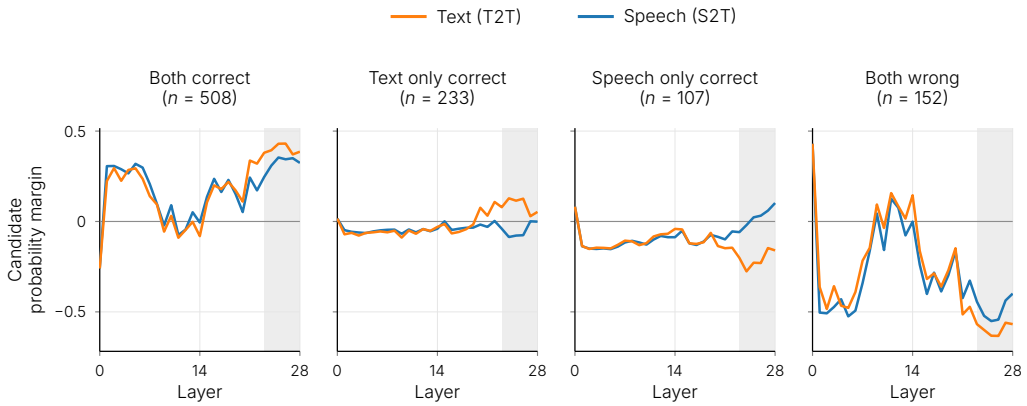


Figure 4: **Answer margins separate in late layers.** Curves show the mean correct-minus-competing candidate probability for Qwen2.5-Omni on 1,000 BBH items, grouped by free-generation correctness, and shading marks the late window of layers 23–28. In this window, the text margin is larger for text-only-correct items and the speech margin is larger for speech-only-correct items.

layers, similarity rises and forms a broad band rather than a sharp diagonal, so several text layers have similarly high CKA with the same speech layer. On SpeechMMLU, each speech layer aligns with roughly six text layers near its peak CKA value for both Qwen2.5-Omni and MiniCPM-o. We use *Semantic Smearing* as a descriptive name for this broad band and for the attention dispersion examined in Section 4.4. In late layers, consecutive speech layers often share the same best-matching text layer (Figure 8). The band persists without pooling. When each text token is paired with one matched native speech frame, avoiding interpolation and temporal averaging, the middle-layer alignment lift and band width exceed their text-position-permutation baselines for both models on both benchmarks, with 95% intervals above zero (Table 2a). In a separate exploratory comparison, calibrating speech hidden states to match text means and standard deviations reduces BBH accuracy from a 62.3% baseline to 46.8% at the input layer and 61.2% at the last layer.

4.3 Answer Margins Diverge in Late Layers

We next examine how the two routes separate the correct answer from competing candidates. Figure 4 shows candidate-probability margins for the paired BBH readout, with items grouped by the correctness of free generation. We compute a matched-instance late-margin difference over the final six layers, $\mathcal{L}_{\text{late}} = \{23, \ldots, 28\}$. This fixed window covers roughly the last 20% of the 28-layer backbone and lies within the CKA late-phase window of layers 22–28:

$$\delta_i^p = \frac{1}{|\mathcal{L}_{\text{late}}|} \sum_{\ell \in \mathcal{L}_{\text{late}}} \left(m_{i,\ell}^{p,\text{T2T}} - m_{i,\ell}^{p,\text{S2T}} \right). \tag{2}$$

The mean is +0.1351 for the 233 text-only-correct items and −0.2301 for the 107 speech-only-correct items. The difference between these group means, which we call the crossover contrast, is +0.3652 with a category-by-group paired-bootstrap 95% interval of [0.2867,0.4426]. The contrast is also positive for full-vocabulary probability margins and signed logit margins (Table 2b). In both groups, the route that answers correctly has the larger late-layer margin. These comparisons are conditioned on generation outcomes, so they locate where the routes differ without establishing a cause. Additional projected-logit and probe plots appear in Appendix B.2.

4.4 Late Attention Allocation Causally Affects Answer Separation

Speech and text also differ in how attention is allocated at the decision position. Across every head of Qwen2.5-Omni on 100 SpeechMMLU items, the median speech/text ratio of the tokens needed to cover 90% of the attention is 2.20× in layers 7–21 and 3.74× in layers 23–28, and the speech-minus-text coverage difference is positive for 78.4% and 86.1% of head–layer pairs (Table 7). Figure 5 shows selected examples. Dispersion is a route-level property: within speech, mean attention entropy provides no clear separation of correct and incorrect items (Appendix C.3).

Table 2: **Controls for the two central findings.** Subtable (a) gives the middle-layer CKA lift A and band width $W_{.95}$ without pooling, each minus its position-permutation baseline, with 95% bootstrap intervals (Appendix A.2). Subtable (b) gives the mean late T2T-minus-S2T margin of Qwen2.5-Omni on BBH for the two discordant groups in three margin coordinates, with the crossover contrast and its 95% interval.

(a) Alignment without pooling

Model	Benchmark	ΔA	$\Delta W_{.95}$
Qwen2.5-Omni	SpeechMMLU	.0952 [.0867, .1034]	7.56 [6.88, 8.22]
Qwen2.5-Omni	BBH	.0981 [.0925, .1038]	5.64 [5.07, 6.21]
MiniCPM-o	SpeechMMLU	.1343 [.1250, .1438]	5.17 [4.52, 5.80]
MiniCPM-o	BBH	.0916 [.0840, .0993]	5.47 [4.98, 5.95]

(b) Late-margin crossover

Margin	Only T2T correct	Only S2T correct	Crossover [95% CI]
Candidate probability	+.1351	−.2301	+.3652 [.2867, .4426]
Full-vocabulary probability	+.0569	−.0527	+.1096 [.0876, .1322]
Signed logit	+.8991	−.8906	+1.7897 [1.4431, 2.1331]

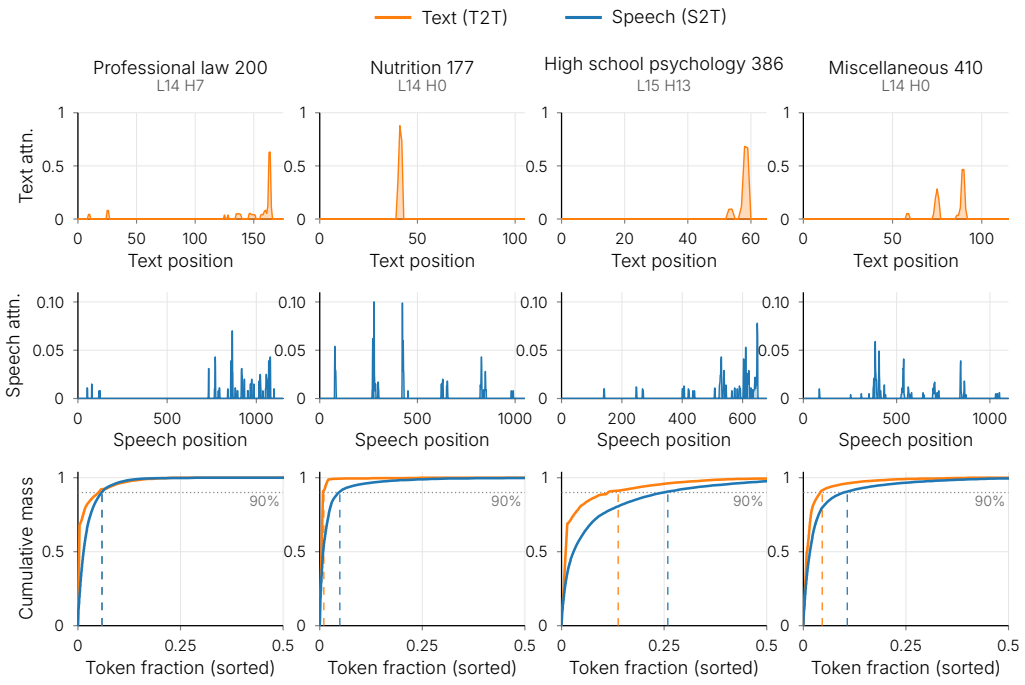


Figure 5: **Speech attention is more dispersed than text attention.** Each column shows a selected decision-position attention distribution for Qwen2.5-Omni on SpeechMMLU, labeled by subject, item, layer (L), and head (H). The speech row uses a ten-times smaller vertical scale, and dashed lines mark the sorted-token fraction that covers 90% of the mass.

Attention weights alone do not establish causal use [Serrano and Smith, 2019], so we test whether attention allocation affects answer separation with an intervention [Vig et al., 2020, Shim et al., 2026] that edits attention while the input and token count stay fixed. We encode each task's two answers as X and Y and counterbalance the answer-to-label mapping and the order of the two mapping instructions. We then sharpen decision-position speech attention in layers 23–28 while preserving each head's total attention mass on the speech span (Appendix C.3). The change in the absolute two-choice logit gap, $\Delta|z_X - z_Y|$, measures how strongly either code is preferred, irrespective of correctness. Figure 6 shows that late speech sharpening increases this gap in all four

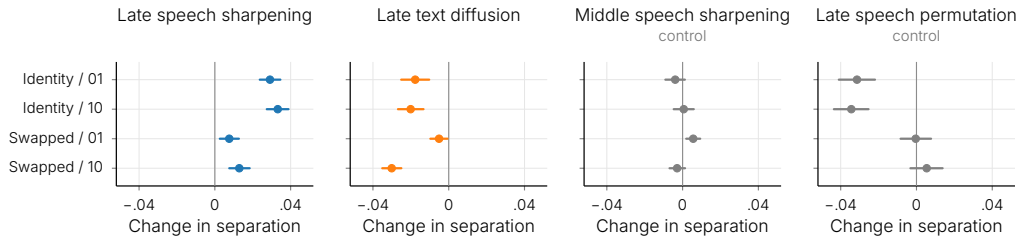


Figure 6: **Sharpening late speech attention increases answer-code separation.** Points show the mean change in $|z_X - z_Y|$ for Qwen2.5-Omni on 1,000 BBH items with 95% bootstrap intervals, under two answer-to-code mappings and two rule orders. The input and token count are fixed, and the gap measures separation irrespective of which code is correct.

Table 3: **Two failure modes across models.** Gaps are T2T-minus-S2T accuracy points in the benchmark overview, approximate for LLaMA-Omni. Mid-band describes the middle-layer CKA band.

Model	SpeechMMLU gap	BBH gap	Mid-band	Mode
Qwen2.5-Omni	+3.25	+12.8	clear	Decision Instability
MiniCPM-o	+7.02	+7.7	clear	Decision Instability
LLaMA-Omni	$\approx$+58.8	$\approx$+12.1	broad	Intermediate
Qwen2-Audio	+43.07	+2.2	absent	Representation Failure

conditions, while the reciprocal diffusion of text attention decreases it in all four. Among the controls, equal-depth middle-layer sharpening has a 95% interval entirely above zero in one condition, and location permutation in none. The conditions reuse the same items, and baseline semantic predictions vary under answer-label swapping, failing the prespecified 75% agreement criterion in three of four modality/order comparisons. The causal interpretation therefore concerns answer-code separation.

4.5 The Gap Takes Two Forms Across Models

The four-model comparison reveals different combinations of geometric alignment and answer behavior. For Qwen2.5-Omni and MiniCPM-o, speech reaches a clear middle-layer alignment band with text (Figure 3), while the late answer margins depend on the route (Figure 4 and Appendix E). We call this *Decision Instability*. Qwen2-Audio instead has weak cross-modal CKA without a clear high-similarity middle band, a pattern we call *Representation Failure*. LLaMA-Omni, reported in the Appendix, shows broad similarity with less layer-specific correspondence. Table 3 summarizes these descriptive patterns, which suggest different targets for subsequent interventions: representation formation when alignment is weak and late answer separation when middle-layer alignment is present.

4.6 Compression Repairs a Subset of Speech-Only Errors

We finally ask which speech errors compression can repair, comparing temporal compaction of the waveform with token merging inside the model on items that gold-text prompting answers correctly and direct speech answers incorrectly (Table 4). In the overview protocol, tempo 1.2× repairs 61 of 178 such items, including 50 of the 123 with ASR-correct transcripts, silence trimming repairs 68, and token merging repairs 26. Within this protocol, temporal compaction repairs 2.35 to 2.62 times as many items as token merging. In the shared-instruction protocol, the three interventions repair 113, 104, and 99 of 233 items, with paired difference intervals that include zero (Appendix D.3). Under both protocols, compaction and merging each repair a subset of these errors, while their relative advantage is protocol-dependent. Token merging is an inference-only control on audio tokens after the early projection phase (Appendix D.2). It changes the merged states as well as their count, so it does not isolate sequence length. A text-side proxy that repeats each word also lowers T2T accuracy (Table 14). The comparisons identify distinct targets for testing repairs: cross-modal representation formation, allocation of attention within the speech span, and late answer separation.

Table 4: **Compression repairs a subset of speech-only errors on BBH.** Gap repaired counts corrected answers among the text-correct, speech-wrong items of Qwen2.5-Omni under each protocol, and ΔAcc. is the aggregate accuracy change in points.

		Overview protocol		Shared-instruction protocol	
Intervention	**Target**	**Gap repaired**	**ΔAcc.**	**Gap repaired**	**ΔAcc.**
Tempo 1.2×	audio duration	61/178 (34.3%)	+2.2	113/233 (48.5%)	+2.2
Silence trim	pauses	68/178 (38.2%)	+2.0	104/233 (44.6%)	+1.3
Token merging	speech tokens	26/178 (14.6%)	+0.5	99/233 (42.5%)	−1.5

Which intervention improves generated answers remains an empirical question for each model and task.

5 Discussion

What the evidence establishes. The crossover contrast localizes natural outcome differences to late-layer answer margins, and the attention edit shows that late attention allocation changes answer-code separation at fixed input and token count. These are different estimands, so their ratio is not a fraction of the gap explained by attention. The edit establishes the sensitivity of answer-code separation, and it leaves accuracy repair, semantic direction, and other models untested.

Alternative explanations. Less reasoning-heavy speech training data could shape a weaker speech route and is compatible with the patterns we report. Our analysis localizes where a fixed model diverges, whereas identifying the upstream cause requires checkpoints trained on matched data mixtures. We therefore do not attribute the gap uniquely to architecture or to residual acoustic information. Whether late speech states retain acoustic or phonetic variation that interferes with answer formation remains a hypothesis, and direct layer-wise feature probing is the test that would discriminate it.

6 Conclusion and Limitations

We study the speech–text modality gap by connecting two views of it: how similar speech and text representations are, and how the model forms its answer from them. Geometric alignment does not imply functional equivalence. In Qwen2.5-Omni and MiniCPM-o, speech representations become geometrically close to text in middle layers while still failing to form stable late-layer answer margins, whereas in Qwen2-Audio the alignment itself is weak. A fixed-input attention edit in Qwen2.5-Omni shows that late attention allocation causally affects answer-code separation, a quantity distinct from answer correctness. Temporal compaction and token merging each repair a subset of speech-only errors, with relative repair rates that vary across evaluation protocols. These results suggest that evaluations of the modality gap should distinguish geometric alignment, late-layer answer margins, and generated-answer correctness, and that interventions should target representation formation when alignment is weak and late answer separation when alignment is present.

Limitations. The scope is semantics-matched reasoning over synthesized prompts and accessible open-weight speech LLMs around the 7B scale. Speech-native tasks can require acoustic cues, as MMSU shows. The phase description is model-dependent, and the intervention is limited to Qwen2.5-Omni on BBH. Differences in training data may also contribute to the observed model differences. The diagnostics do not identify which acoustic or phonetic features contribute to the speech errors. Natural conversational speech, larger models, SpeechMMLU compaction, and post-compaction Phase-II/III diagnostics remain open tests.

Acknowledgments and Disclosure of Funding

The work of Zhizheng Wu is supported by the National Natural Science Foundation of China under Grant 62376237. This work is supported by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044. The authors have no financial relationships relevant to this work with any entity other than their listed affiliations.

References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14: 378–398, 2026. doi: 10.1162/tacl.a.628. URL https://aclanthology.org/2026.tacl-1.18/.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. URL https://arxiv.org/abs/2311.07919.

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. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.682. URL https://aclanthology.org/2025.acl-long.682/.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pages 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0c0bf917c7942b5a08df71f9da626f97-Paper.pdf.

Teresa Dorszewski, Albert Kjøller Jacobsen, Lenka Tětková, and Lars Kai Hansen. How redundant is the transformer stack in speech representation models? In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10887866.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, pages 57607–57624, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/90d1fc07f46e31387978b88e7e057a31-Paper-Conference.pdf.

Yuanbo Fang, Haoze Sun, Jun Liu, Tao Zhang, Zenan Zhou, Weipeng Chen, Xiaofen Xing, and Xiangmin Xu. S2SBench: A benchmark for quantifying intelligence degradation in speech-to-speech large language models. *arXiv preprint arXiv:2505.14438*, 2025b. URL https://arxiv.org/abs/2505.14438.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pages 41819–41886. Curran Associates, Inc., 2025. doi: 10.52202/085713-1396. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3babb6b453cb59d87cb58a1219ef914b-Paper-Conference.pdf.

Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya, Joseph Keshet, and Aviv Navon. Beyond transcription: Mechanistic interpretability in ASR. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37407–37416. Association for the Advancement of Artificial Intelligence, 2026. doi: 10.1609/aaai.v40i44.41073.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Ming-Hao Hsu, Liang-Hsuan Tseng, Hung-yi Lee, and Zhizheng Wu. TASLA: Text-aligned speech tokens with multiple layer-aggregation. *arXiv preprint arXiv:2510.14934*, 2025. URL https://arxiv.org/abs/2510.14934.

Ming-Hao Hsu, Yuxuan Hu, Shujie Liu, Jinyu Li, and Zhizheng Wu. Is text all you need? Text as a universal information bottleneck for speech LLMs. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=ux19efk62R.

Ming-Hao Hsu, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. Entity binding failures in speech LLM reasoning: Diagnosis and chain-of-thought intervention. In *Interspeech 2026*, pages 904–909, 2026b. doi: 10.21437/Interspeech.2026-640.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. The remarkable robustness of LLMs: Stages of inference? *arXiv preprint arXiv:2406.19384*, 2024. URL https://arxiv.org/abs/2406.19384.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Wenrui Liu, Qian Chen, Wen Wang, Guanrou Yang, Weiqin Li, Minghui Fang, Jialong Zuo, Xiaoda Yang, Tao Jin, Jin Xu, Zemin Liu, Yafeng Chen, Jionghao Bai, and Zhifang Guo. Speech token prediction via compressed-to-fine language modeling for speech generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10632–10641. ACM, 2025. doi: 10.1145/3746027.3755816.

Yujian Ma, Xikun Lu, Jinqiu Sang, Xianquan Jiang, and Ruizhe Li. Behind the scenes: Mechanistic interpretability of LoRA-adapted Whisper for speech emotion recognition. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5286–5290. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11464049.

Pooneh Mousavi, Yingzhi Wang, Mirco Ravanelli, and Cem Subakan. ALAS: An automatic latent alignment score for audio language models. *arXiv preprint arXiv:2505.19937*, 2025. URL https://arxiv.org/abs/2505.19937. Accepted to IEEE MLSP 2026.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed 2026-10-14.

OpenBMB. MiniCPM-o 2.6: A GPT-4o level MLLM for vision, speech and multimodal live streaming on your phone. Hugging Face model card, 2025. URL https://huggingface.co/openbmb/MiniCPM-o-2_6.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Ryan Soh-Eun Shim, Kwanghee Choi, Kalvin Chang, Ming-Hao Hsu, Florian Eichin, Zhizheng Wu, Alane Suhr, Michael A. Hedderich, David Harwath, David R. Mortensen, and Barbara Plank. Linear script representations in speech foundation models enable zero-shot transliteration. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9532–9544. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.464. URL https://aclanthology.org/2026.findings-acl.464/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, Heyu Chen, Jiawei Li, Lei Li, Menghang Zhu, Peidian Li, Qiying Wang, Sirui Deng, Weimin Xiong, Wenshan Huang, Wenyu Yang, Yilin Jiang, Yixin Yang, Yuanyuan Tian, Yue Ma, Yue Yu, Zihan Zhang, Zihao Yue, Bangjun Xiao, Bingquan Xia, Bofei Gao, Bowen Ye, Can Cai, Chang Liu, Chenhong He, Chunan Li, Dawei Zhu, Duo Zhang, Fengyuan Shi, Guoan Wang, Hailin Zhang, Hanglong Lv, Hanyu Li, Hao Tian, Heng Qu, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianguang Zuo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Linghao Zhang, Meng Chen, Nuo Chen, Peng Zhang, Qianli Chen, Qiantong Wang, Rang Li, Shaohui Liu, Shengfan Wang, Shicheng Li, Shihua Yu, Shijie Cao, Shimao Chen, Shuhao Gu, Weikun Wang, Wenhan Ma, Xiangwei Deng, Xing Yong, Xing Zhang, Xu Wang, Yifan Song, Yihao Zhao, Yingbo Zhao, Yizhao Gao, Yu Cheng, Yu Tu, Yudong Wang, Zhaojun Huang, Zhengju Tang, Zhenru Lin, Zhichao Song, Zhipeng Xu, Zhixian Zheng, and Zihan Jiang. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025. URL https://arxiv.org/abs/2512.23808.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18821–18835. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.857. URL https://aclanthology.org/2026.acl-long.857/.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, pages 31374–31410, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/34e278fbbd7d6d7d788c98065988e1a9-Paper-Conference.pdf.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.262. URL https://aclanthology.org/2025.emnlp-main.262/.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. URL https://arxiv.org/abs/2503.20215.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. When large language models meet speech: A survey on integration approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20298–20315. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1041. URL https://aclanthology.org/2025.findings-acl.1041/.

Kisung You. Semantics at an angle: When cosine similarity works until it doesn't. *arXiv preprint arXiv:2504.16318*, 2025. URL https://arxiv.org/abs/2504.16318.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Jialong Zuo, Guangyan Zhang, Minghui Fang, Shengpeng Ji, Xiaoqi Jiao, Jingyu Li, Yiwen Guo, and Zhou Zhao. Entropy-based coarse and compressed semantic speech representation learning. *arXiv preprint arXiv:2509.00503*, 2025. URL https://arxiv.org/abs/2509.00503.

Appendix

The appendix follows the order of Section 4. Appendices A to D support the alignment, answer-margin, attention, and behavioral results. Appendices E to G report the remaining models, and Appendix H gives selected cases. Appendices I to K give implementation details, broader impact, and asset licenses.

A Robustness of the Alignment Analysis

We examine how the alignment diagnostics vary with the DTW base layer, distance metric, and band constraint, and whether the middle-layer pattern persists without pooling.

A.1 DTW Sensitivity and Band Statistics

The cross-layer alignment analysis in Section 3 uses DTW to align speech-frame and text-token positions at a fixed base layer. This token correspondence is held fixed when comparing different layer pairs. A cross-layer best-match curve records the matched text-layer index at each speech layer.

Setup. The DTW base layer ℓ_0 is swept over 0, 5, 10, and 20, with cosine versus L_2 cost computed after feature-wise standardization within each route, and with or without a width-64 Sakoe–Chiba band. The sweep uses Qwen2.5-Omni with BF16 weights on 100 items per benchmark. Let p_ℓ denote the matched text-layer index at speech layer ℓ, for $\ell = 0, \ldots, L-1$. We report the R^2 of a linear fit $p_\ell = a\ell + b$ and the stall fraction $\frac{1}{L-1}\sum_{\ell=1}^{L-1} \mathbf{1}[p_\ell = p_{\ell-1}]$, which counts adjacent speech-layer pairs mapped to the same text layer. The fit measures linearity rather than distance to the identity diagonal.

Results. The sweep in Figure 7 shows greater sensitivity at base layer 0 than at the higher bases. Path linearity is $R^2 \leq 0.75$ at base layer 0 against $R^2 \geq 0.80$ at every higher base in all four panels, and the stall fraction is highest at base layer 0 in three of the four panels. The width-64 band leaves both statistics unchanged in every setting. We retain bases 10 and 20 for SpeechMMLU and BBH, respectively. The native-length, position-permutation controls in Appendix A.2 supply the quantitative test of whether the middle-layer lift and band survive without pooling. All layer-pair comparisons use the same estimated token correspondence.

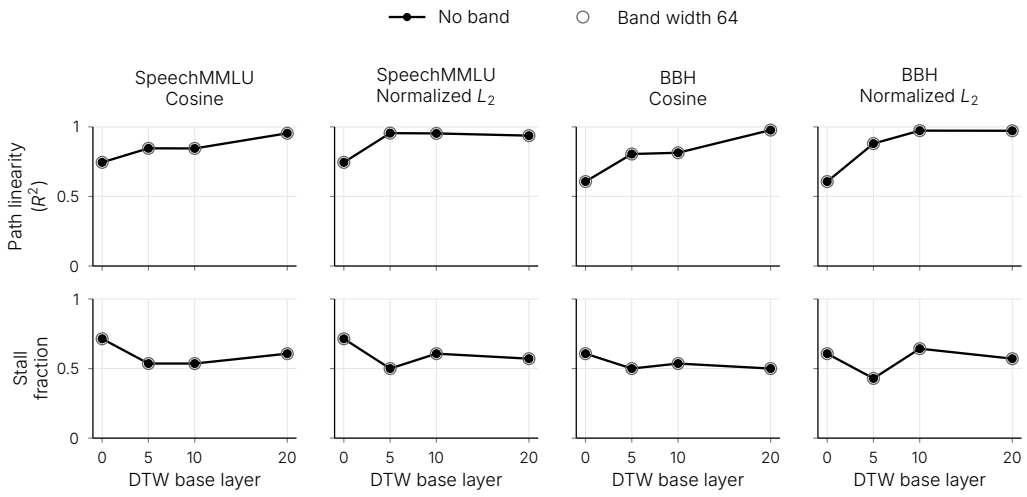


Figure 7: **Cross-layer best-match sensitivity to DTW settings.** Qwen2.5-Omni, 100 items per benchmark. The top row reports the R^2 of a linear fit to matched text-layer indices across speech layers, and the bottom row the fraction of adjacent speech-layer pairs mapped to the same text layer. Open rings mark the width-64 band and coincide with the unbanded values. Pooling-free, null-adjusted results appear in Appendix A.2.

A.2 Native-Length Alignment without Interpolation or Averaging

We test whether the Phase-II CKA lift and broad near-peak band persist without the interpolation and temporal averaging used in the original heatmaps. We freeze 100 items per benchmark and use the same IDs for Qwen2.5-Omni and MiniCPM-o, yielding four model–benchmark cells with no failed items. Both models use BF16 weights without quantization, processing one item at a time. The BBH sample contains 25 items from each of its four categories, and the SpeechMMLU sample is balanced across subjects. For each item, we isolate the final user content span, excluding padding, instructions outside that span, and assistant positions. We compute one cosine-distance DTW path on native hidden states at layer 10 for SpeechMMLU or layer 20 for BBH and reuse it at every layer pair. For each text token, we select the lower median of the unique speech-frame indices assigned by that path. Thus each CKA row contains one native speech state and one native text state, with no interpolation, truncation, or within-group averaging. Linear CKA uses float64 centered-Gram arithmetic after conversion of the hidden states to float32.

For each item, let A be the Phase-II mean of the row-wise CKA maxima minus the corresponding Phase-I mean. Let $W_{.95}$ be the Phase-II mean number of text layers reaching at least 95% of each speech layer's maximum CKA. We retain the original phase windows: 0–9, 10–21, and 22–28 for Qwen2.5-Omni on both benchmarks, 0–3, 4–20, and 21–28 for MiniCPM-o on SpeechMMLU, and 0–12, 13–23, and 24–28 for MiniCPM-o on BBH. We compare each endpoint with its within-item mean over 20 fixed non-identity text-position permutations, reusing each permutation at every layer pair. We bootstrap paired items 10,000 times with seed 20260724, stratifying BBH by category, and report percentile 95% intervals. Items are the resampling unit, with permutations and layer pairs summarized within each item.

Table 5: **Native-length CKA controls.** The first two endpoints subtract the within-item mean position-permutation endpoint and report item means with 95% bootstrap intervals. The last column gives the alignment-free matched-minus-wrong-item differences in A and $W_{.95}$, respectively, without an inferential interval.

Model	**Benchmark**	ΔA	$\Delta W_{.95}$	**Cross-item** $(\Delta A, \Delta W_{.95})$
Qwen2.5-Omni	SpeechMMLU	.0952 [.0867, .1034]	7.56 [6.88, 8.22]	(.0299, 1.37)
Qwen2.5-Omni	BBH	.0981 [.0925, .1038]	5.64 [5.07, 6.21]	(.0081, –0.72)
MiniCPM-o	SpeechMMLU	.1343 [.1250, .1438]	5.17 [4.52, 5.80]	(.0521, –0.46)
MiniCPM-o	BBH	.0916 [.0840, .0993]	5.47 [4.98, 5.95]	(–.0031, –3.30)

In all four model–benchmark combinations (Table 5), the CKA lift and band width exceed their position-permutation means, with positive 95% intervals. The lower confidence bound for the observed $W_{.95}$ also exceeds one. We separately remove token alignment by computing CKA across the 100 paired final-content states, one state per item and layer. We compare matched items with 20 wrong-item derangements within category or subject. Text layer 0 is undefined because the fixed final text token has zero cross-item variance, so matched and deranged matrices both use speech layers 0–28 and text layers 1–28. Only Qwen2.5-Omni on SpeechMMLU satisfies all three directional criteria of positive lift above the derangement mean, width above one, and width above the derangement mean. This analysis changes the CKA sample axis from token positions to items and provides mixed corroboration of alignment-free geometry. The native-length result therefore supports independence from interpolation and mean pooling while retaining the assumption of a monotonic token correspondence.

We also compare mean pooling, center-frame selection, a random frame averaged over 20 selection seeds, the full repeated DTW path, and repeated group means after resampling both sequences to 256 positions. The no-averaging conditions retain the lift and broad-band criteria in all four cells, including width thresholds of 90% and 99%. These fixed-length controls isolate aggregation and path weighting, whereas the native-length control above additionally removes resampling.

A.3 Best-Match Paths and Phase Windows

Figure 8 redraws the main CKA heatmaps as best-match paths together with the three phase windows.

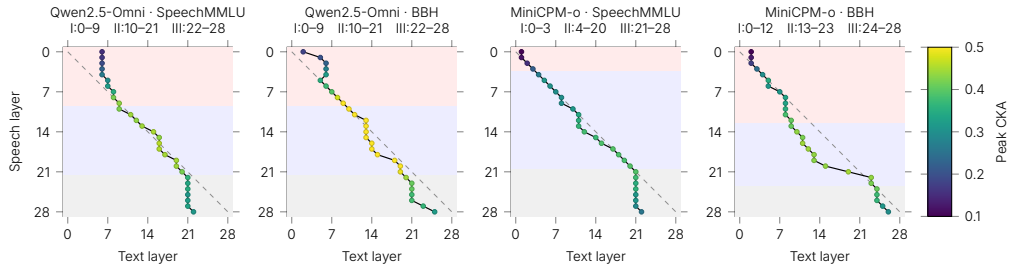


Figure 8: **Best-match paths and phase windows from the CKA heatmaps.** Black lines connect each speech layer to its best-matching text layer, point colors give the peak CKA in that speech-layer row, and the dashed diagonal marks equal layer indices. Background shading marks the three phase windows, whose speech-layer ranges appear above each panel and are listed in Appendix A.2. The late plateau describes repeated best-matching text layers, with functional behavior tested separately by answer readouts.

B Answer-Margin Readouts

B.1 Paired Decision Readouts and Margin Definitions

We use all 1,000 BBH items and define correctness groups from the shared-instruction free-generation evaluation in Appendix D.3. The group sizes are 508 for both routes correct, 233 for only T2T correct, 107 for only S2T correct, and 152 for both routes wrong. For the paired readout, we append two newline characters, `The answer is`, and a trailing space to each route's category instruction inside the user message. We then apply the chat template with its assistant generation prefix and read the final valid prompt position. Thus the probe predicts the next generated token from the assistant-prefix state, rather than continuing an assistant-side answer phrase. This controlled prompt provides a common decision position and is distinct from the prompt used to assign free-generation correctness groups. We retain all 29 hidden-state coordinates, comprising the embedding output and 28 Transformer outputs. At intermediate coordinates we apply the model's final RMSNorm once, followed by the complete vocabulary projection. At the final coordinate we index the model's native full-vocabulary logits directly, avoiding both a second normalization and rounding changes from a candidate-only projection. Candidate scores are promoted to float32 before subtraction and softmax.

Hyperbaton uses the candidates A and B, with bare and leading-space uppercase forms. The remaining categories use yes and no, each with lowercase and initial-capital forms, with and without a leading space. We retain only single-token forms, and no token belongs to both candidates. For candidate c, we define $z_c = \max_{v \in V_c} z_v$, where V_c contains its retained surface tokens. The signed logit margin is $z_y - \max_{c \neq y} z_c$, and the primary plotted probability margin is $q_y - \max_{c \neq y} q_c$, where $q = \mathrm{softmax}(z)$ at temperature one over the task's two candidates. We also compute a full-vocabulary probability margin by summing the full-vocabulary softmax mass over V_c for each candidate before taking the gold-minus-best-alternative difference. These probability views have explicit denominators but remain sensitive to logit scale and are not calibrated confidence estimates.

For each margin, we average the paired T2T-minus-S2T difference over layers 23–28 within each item. The crossover contrast subtracts the mean difference in the only-S2T-correct group from that in the only-T2T-correct group. We bootstrap route-paired items independently within each category-by-correctness cell 10,000 times with seed 20260725, preserving the observed cell sizes. Table 2 reports item-weighted group means and percentile intervals. The candidate-probability crossover is positive in all four categories, whereas the full-vocabulary crossover is slightly negative for web of lies (–0.00024). Grouping on observed correctness makes this a localization of an outcome-conditioned difference, not a causal estimate or an independent correctness test. Figure 9 gives the signed-logit counterpart of the paired probability analysis.

B.2 Projected-Logit and Probe Diagnostics

The entropy, candidate-readout and margin plots in this subsection and in Appendices E to G read the final prompt position, apply the model's final RMSNorm once at intermediate coordinates, and index

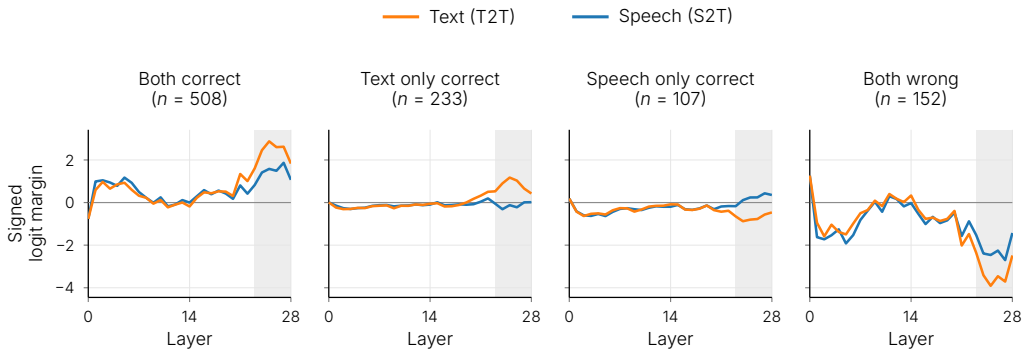


Figure 9: **Signed-logit counterpart of the probability replication.** The items, readout and groups match Figure 4. Curves show mean correct-minus-competing logits, with the fixed late window shaded.

the native logits at the last coordinate, as in the paired analysis of Appendix B.1. They use different prompts, the 300 aggregate items per benchmark of the figure reruns, and BF16 weights, so their absolute scales are not directly comparable with the paired readout. Margins are candidate logits after an assistant answer cue, the correct option minus the best other option. Figure 10 shows the margins for Qwen2.5-Omni, and Figures 11 to 13 show the entropy, probe, and update diagnostics, respectively.

Correctness groups. For each model, benchmark and route, we decode the 300 items greedily with at most 256 new tokens, truncate at the first conversation-role marker, and parse the answer with the deterministic parser of Appendix D.3, counting unparsed answers as wrong. BBH uses the category instruction of that section, and SpeechMMLU a zero-shot instruction that ends with `The answer is (A).` to `(D).`; the text route receives the question followed by the same instruction. Qwen2.5-Omni on BBH keeps the groups of the 1,000-item analysis, restricted to these items. The groups depend on the free-generation answer alone, so they can disagree with the sign of a readout margin. Table 6 lists accuracies, unparsed answers and group sizes. MiniCPM-o, Qwen2-Audio on BBH, and the speech route of LLaMA-Omni break the answer format more often, so their groups partly reflect formatting.

Table 6: **Free-generation correctness groups of the 300 items per benchmark.** Accuracy counts unparsed answers as wrong. Group sizes are both correct, text only correct, speech only correct, and both wrong. Qwen2.5-Omni on BBH uses the groups of Appendix D.3, whose unparsed counts are reported there.

Model	Benchmark	S2T acc.	T2T acc.	Unparsed S2T / T2T	Group sizes
Qwen2.5-Omni	SpeechMMLU	66.7	72.3	0 / 4	183 / 34 / 17 / 66
Qwen2.5-Omni	BBH	62.0	69.7	–	147 / 62 / 39 / 52
MiniCPM-o	SpeechMMLU	50.0	57.3	64 / 73	101 / 71 / 49 / 79
MiniCPM-o	BBH	56.7	50.0	51 / 98	96 / 54 / 74 / 76
Qwen2-Audio	SpeechMMLU	33.3	55.3	10 / 8	70 / 96 / 30 / 104
Qwen2-Audio	BBH	49.0	42.3	26 / 81	78 / 49 / 69 / 104
LLaMA-Omni	SpeechMMLU	16.3	67.0	114 / 16	33 / 168 / 16 / 83
LLaMA-Omni	BBH	29.3	57.3	114 / 54	50 / 122 / 38 / 90

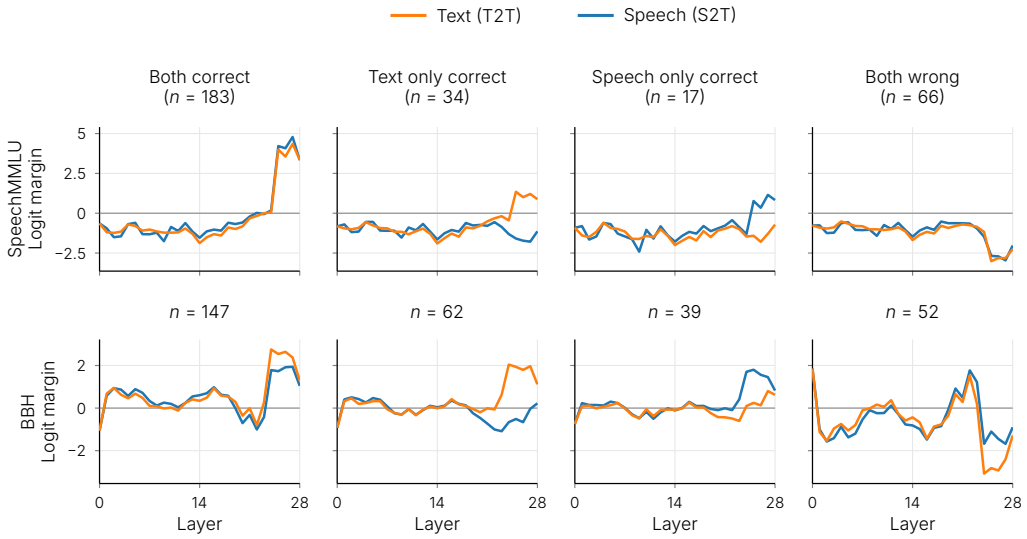


Figure 10: **Qwen2.5-Omni logit margin by correctness group.** Rows are SpeechMMLU and BBH, and columns are the free-generation groups with their sizes. Curves show the mean candidate-readout margin of the group. The BBH groups are those of the 1,000-item analysis in Figure 4 restricted to 300 items. In the text-only-correct group the late-layer margin (mean of the last six coordinates) is +0.63 for text and −1.40 for speech on SpeechMMLU, and +1.58 and −0.46 on BBH.

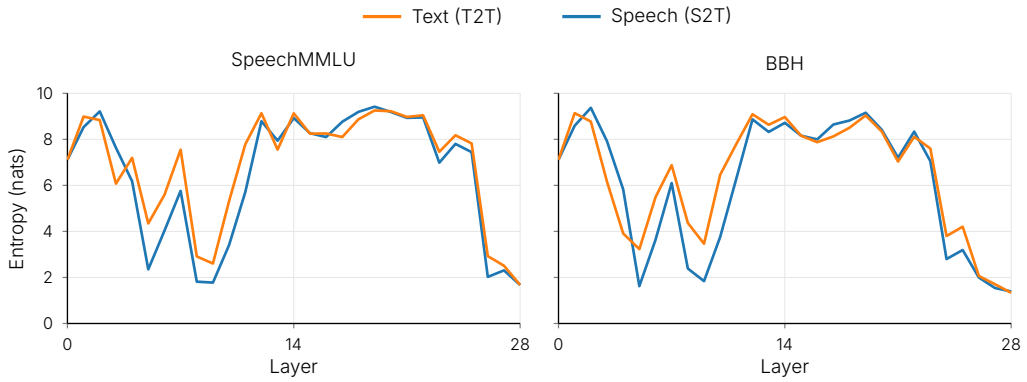


Figure 11: **Layer-wise projected-logit entropy of Qwen2.5-Omni.** Curves show the mean over 300 items per benchmark. Speech and text have similar entropy trajectories, with mean absolute differences of 0.67 nats on SpeechMMLU and 0.76 nats on BBH.

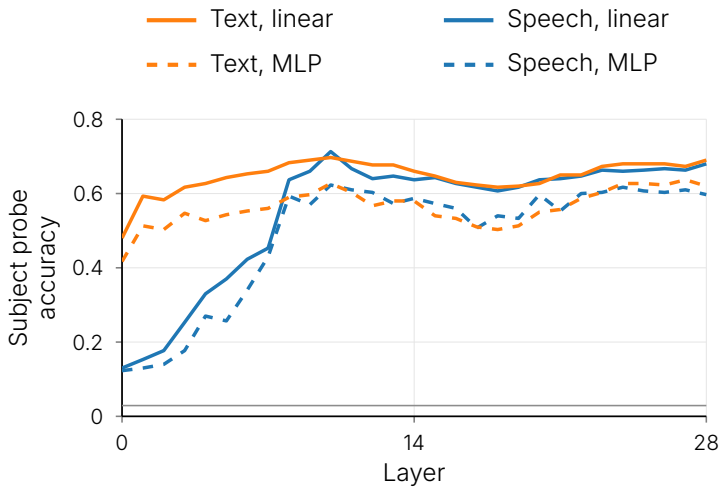


Figure 12: **Layer-wise subject probe of Qwen2.5-Omni.** Logistic-regression (linear) and one-hidden-layer MLP probes on the mean hidden state of the evidence tokens predict the SpeechMMLU subject of the 300 items (34 subjects, chance 2.9%), with 5-fold cross-validated accuracy. Linear speech accuracy rises from 0.13 at the embedding output to 0.71 at layer 10 and text accuracy from 0.48 to 0.70, and both stay between 0.6 and 0.7 afterwards; fold standard deviations reach 0.10. These curves describe probe performance rather than generated-answer correctness.

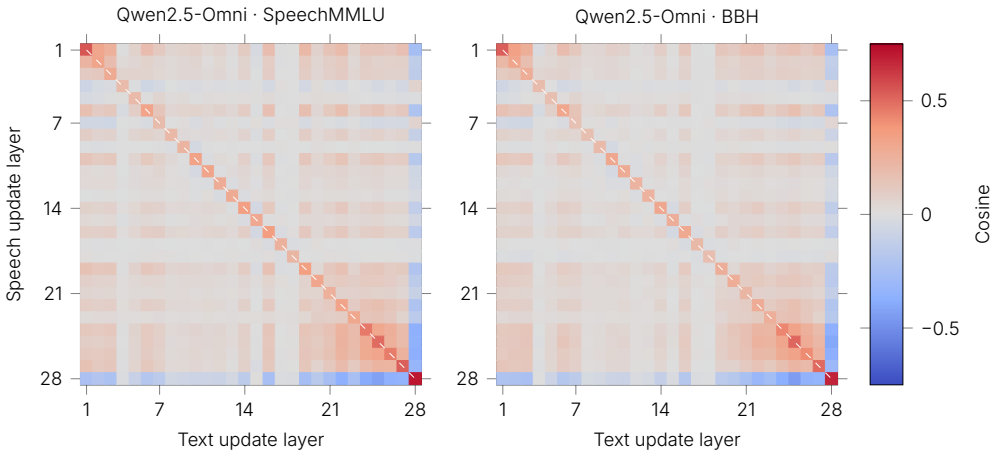


Figure 13: **Cosine similarity between speech and text layer-wise updates in Qwen2.5-Omni.** An update is the difference between consecutive hidden states after the fixed base-layer-0 token alignment, averaged over 100 items per benchmark. For every speech layer the largest cosine lies at the text layer with the same index. The mean diagonal cosine is 0.33 on SpeechMMLU and 0.30 on BBH, against 0.03 off the diagonal.

C Attention Analyses

C.1 All-Head Attention Coverage

Table 7 gives the all-head attention coverage summary cited in Section 4.4, together with normalized entropy.

Table 7: **All-head attention coverage audit for Qwen2.5-Omni on SpeechMMLU.** We summarize all head–layer pairs within each item, then report the across-item medians of the within-item medians and 90th percentiles (Q90). The positive-pair fraction is averaged over items. Δ denotes S2T minus T2T, and Ent. is attention entropy divided by the log of valid prompt length. Attention is normalized over all valid prompt keys at the final prompt query. These windows differ from the CKA phase windows. Speech usually needs more tokens to cover 90% of the attention mass, while normalized entropy differences vary across heads.

Layers	Items	Median cov. ratio	Median $\Delta\text{Cov}_{.90}$	Q90 $\Delta\text{Cov}_{.90}$	Pairs $\Delta\text{Cov} > 0$	Q90 ΔEnt.
7–21	100	2.20×	+16.0	+83.5	78.4%	+0.037
23–28	100	3.74×	+48.0	+102.5	86.1%	+0.073

C.2 Selected-Head Attention Summary

Table 8 reports a selected-head attention diagnostic, with examples in Figure 5. The speech peak mass is 0.105 versus 0.638 for text, top-10 mass is 0.483 versus 0.933, and 90% coverage requires 101 versus 8 tokens. These selected summaries emphasize large cross-modal differences, and coverage counts depend on input length.

Table 8: **Selected-head attention summary for Qwen2.5-Omni on SpeechMMLU.** Δ is S2T minus T2T, computed from the displayed values. $\text{Entropy}_{\text{norm}}$ is normalized attention entropy, $p_{\max}$ is the largest attention coefficient, top-10 mass sums the ten largest coefficients, and $\text{Cov}_{.90}$ counts the tokens needed to cover 90% of attention.

Metric	T2T	S2T	Δ
$\text{Entropy}_{\text{norm}}$	0.361	0.657	+0.296
$p_{\max}$	0.638	0.105	–0.533
Top-10 mass	0.933	0.483	–0.450
$\text{Cov}_{.90}$ count	8	101	+93

C.3 Fixed-Input Attention Interventions and Their Interpretation

We intervene on the final prompt query of every head in Qwen2.5-Omni, using evaluation mode, eager attention, and batch size one. Within each baseline–treatment pair, the input tensors, token count, positions, and model weights are identical. For the target content span, we normalize the post-softmax attention coefficients to a conditional distribution p_j, replace it by $p'_j = p_j^{1/\tau} / \sum_k p_k^{1/\tau}$, and restore the original per-head span mass. All coefficients outside the span and the current value vectors remain unchanged by the local edit. We apply the resulting attention-output difference and let later computation proceed normally, without matching the attention-output norm. Speech sharpening uses τ = .6 at layers 23–28. We compare it with sharpening at the six middle layers 12–17, a within-span permutation of the sharpened late coefficients using seed 1234, and reciprocal text diffusion at layers 23–28 with τ = 1.6666667. The permutation is a location control whose intervention magnitude is not matched to sharpening.

We cross each condition with two answer-label mappings and two rule orders on the same 1,000 items. Semantic class 0 denotes A for hyperbaton and no otherwise, and class 1 denotes B or yes. The identity mapping assigns classes 0 and 1 to X and Y, and the swapped mapping reverses this assignment. Rule order states either the class-0 or the class-1 rule first, followed by `Output ONLY a single character: X or Y.` For this intervention assay, each candidate score is the

log-sum-exp of the native logits for its bare and leading-space forms, unlike the maximum-over-forms readout in Appendix B.1. The prespecified direction-free diagnostic is the per-item change in $|z_X - z_Y|$ within each mapping/order cell. We report the 16 cell estimates in Table 9, using category-stratified paired-item bootstrap intervals with 10,000 resamples and seed 20260724. The four cells reuse the same items and are not independent replications, and the intervals are marginal rather than simultaneous.

Table 9: **Change in absolute answer-code separation.** Each entry is the mean treatment-minus-baseline change with its 95% interval on the same 1,000 items. Positive values indicate greater separation regardless of which semantic answer is favored. Order 01 states the class-0 rule first, and order 10 states the class-1 rule first.

Condition	Order	Identity mapping	Swapped mapping
S2T late sharpening	01	+.0291 [.0237, .0346]	+.0076 [.0026, .0127]
	10	+.0332 [.0274, .0389]	+.0129 [.0076, .0184]
S2T middle sharpening	01	−.0039 [−.0091, .0012]	+.0056 [.0019, .0093]
	10	+.0006 [−.0048, .0059]	−.0029 [−.0069, .0013]
S2T late permutation	01	−.0314 [−.0410, −.0219]	−.0004 [−.0084, .0077]
	10	−.0344 [−.0436, −.0253]	+.0054 [−.0032, .0138]
T2T late diffusion	01	−.0177 [−.0251, −.0103]	−.0051 [−.0097, −.0006]
	10	−.0201 [−.0268, −.0133]	−.0301 [−.0351, −.0250]

Late speech sharpening increases separation in all four cells, whereas middle sharpening has an interval entirely above zero in one cell and the permutation control in none. Reciprocal text diffusion decreases separation in all four cells. Our criterion for interpreting a change as semantic improvement requires at least 75% agreement between baseline semantic predictions under identity and swapped mappings for each modality and rule order. Agreement is 73.7% and 59.9% for S2T, and 84.9% and 69.7% for T2T, so three of the four comparisons fall below the criterion. The prespecified mapping-averaged correct-log-odds change is +.00381 [.00117, .00652], but label-dependent semantic predictions make this change ambiguous as a measure of answer improvement. The intervention therefore establishes a change in absolute answer-code separation at fixed input. The largest measured span-mass error is 5.96×10^{-8}, the outside-span coefficient change is zero, and the $\tau = 1$ check reproduces native logits exactly on its four-item check set.

We separately test whether matched text-query guidance moves the full speech next-token distribution toward the matched text distribution. At the same late layers, we score current speech keys with the matched text query, mix the resulting speech-conditional attention distribution with the current one at weight .25, and preserve current values and total span mass. On all 1,000 items, the reduction in full-vocabulary Jensen–Shannon divergence is +.000033 nats [−.000202, .000273]. Its advantage over a category- and semantic-label-matched wrong-item query is +.000208 [−.000015, .000438]. Both intervals include zero, leaving uncertain whether text-query guidance brings the speech output distribution closer to the text route or improves on wrong-item guidance.

Within-speech correctness comparison. On the 750 BBH items held out from the 250-item intervention pilot, we average normalized speech-span entropy across heads and layers 23–28 at the final prompt query. The assay's baseline semantic readout gives 439 correct and 311 incorrect items, with mean entropies .73205 and .72795. The equally category-weighted correct-minus-incorrect contrast is .00109 [−.00032, .00251], using 10,000 category-by-correctness bootstrap resamples with seed 20260725. The interval spans zero, providing no clear difference in mean entropy between correct and incorrect items under this readout. The comparison uses the intervention assay's baseline readout, not the separate free-generation correctness groups.

D Behavioral Protocols and Repair

D.1 Behavioral Protocols

Table 10 distinguishes the behavioral settings used in the paper. The category breakdown and 178-item repair subset use a 1,024-token diagnostic evaluation, with archived scoring that includes fallback parsing. The shared-instruction comparison instead uses a 256-token limit and deterministic

parsing, with unparsed answers counted as wrong. Its 233-item subset, paired readouts, and complete transition counts therefore form a separate comparison.

Table 10: **Behavioral settings and their baseline accuracies.** All values are percentages for Qwen2.5-Omni on BBH. A dash denotes a route not used for that comparison. Baseline differences identify separate settings, so their accuracy changes and correctness groups are interpreted within each row.

Setting	T2T	S2T	Use in the paper
Benchmark overview	72.7	59.9	Aggregate model comparison and overview repair changes
1,024-token diagnostic	71.9	59.7	Category breakdown, duration groups, 178-item gap subset
Calibration	—	62.3	Input- and last-layer moment matching
Word repetition	62.4	—	$r = 1$ reference for the text stress test
Shared instructions, 256 tokens	74.1	61.5	233-item gap subset, paired readouts, complete behavioral transitions

D.2 Speech-Token Merging Control

The token-merging intervention in Table 4 tests compression after speech projection. It is an inference-time mid-layer token-merging intervention applied only to the speech path of Qwen2.5-Omni on BBH, with no finetuning and no changes to the text path. Compression starts after the early projection phase, where speech representations are already closer to text representations, and is applied across zero-based Thinker modules 7–21, corresponding to paper layers 8–22.

For each selected layer, we isolate the audio-token span, split audio hidden states into an even-index destination set and an odd-index source set, L2-normalize both sets, and assign each source token to its highest-cosine destination token. The highest-scoring source–destination pairs are merged by weighted averaging until the configured removal budget is reached, after which the source tokens are removed and the attention mask and KV cache are rebuilt for the shortened sequence. Non-audio tokens are never merged. The control uses target keep ratio 0.5, minimum 64 audio tokens, maximum 512 audio tokens, zero cosine threshold, and a back-loaded merge schedule over these modules. The intervention changes retained state content as well as token count, so its effects combine compression and representation changes.

D.3 Shared-Instruction Evaluation and Complete Transitions

This evaluation applies shared instructions to all behavioral conditions and reports both repaired and newly introduced errors. This protocol yields 233 text-correct/speech-wrong items, compared with 178 in the 1,024-token diagnostic setting (Table 10). We evaluate Qwen2.5-Omni-7B on the same 1,000 VoiceBench BBH IDs, with 250 items in each of hyperbaton, navigate, sports understanding, and web of lies. We use the official general system prompt and the identical modality-neutral category instruction for speech, gold text, and ASR transcripts. Hyperbaton is instructed to end with `The answer is (A).` or `The answer is (B).`, and the remaining categories use the corresponding yes/no endings. All routes allow reasoning and use greedy decoding with at most 256 new tokens. Before the common deterministic parser, we truncate generated text at the first new conversation-role marker: `\nHuman:`, `\nUser:`, `\nAssistant:`, `\nsystem\n`, `\nuser\n`, or `\nassistant\n`. This format guard applies no additional output-token cap or prefix restriction. An unparsed answer counts as wrong and remains in the denominator. No item is skipped and no inference failure is recorded in any of the six conditions. Gold-text accuracy is 74.1% and direct-speech accuracy is 61.5%, with 3 and 19 unparsed answers, respectively. The runs use BF16 model weights without quantization, PyTorch 2.6.0 with CUDA 12.4, and Transformers 4.52.1.

We generate the tempo condition with SoX 14.4.2, using the pitch-preserving `tempo -s 1.2` effect. Silence trimming uses 20-ms RMS windows with 10-ms hops, a threshold 40 dB below peak RMS, and 50-ms padding around retained regions. Both transformations operate on 48-kHz mono PCM-32 audio, with stereo sources converted by channel averaging. Silence trimming uses no normalization, crossfade, or resampling. The aggregate retained waveform-frame fractions are .8333 for tempo and .8283 for silence, and neither transformation introduces additional clipped

samples. The merging condition follows Appendix D.2 and records an aggregate full-sequence retention fraction of .5612.

Table 11: **Complete paired transitions relative to direct speech.** Repairs are wrong-to-correct transitions, harms are correct-to-wrong transitions, and retained C/W denotes unchanged correct/wrong outcomes across all 1,000 items. Gap repairs use the 233 items correct only under gold text as their denominator. Accuracy changes are percentage points with category-stratified paired-bootstrap 95% intervals.

Condition	Acc. (%)	Unparsed	Repairs	Harms	Retained C/W	Gap repairs	ΔAcc. [95% CI]
Tempo 1.2×	63.7	20	151	129	486/234	113/233	+2.2 [−1.1, 5.4]
Silence trimming	62.8	22	139	126	489/246	104/233	+1.3 [−1.9, 4.5]
Token merging	60.0	21	139	154	461/246	99/233	−1.5 [−4.8, 1.8]
ASR cascade	71.5	14	219	119	496/166	181/233	+10.0 [6.7, 13.4]

We use 10,000 category-stratified paired bootstrap resamples with base seed 20260725 and contrast-specific deterministic seeds, together with exact two-sided McNemar tests on discordant pairs. The gap repair rates are 48.50% [42.06, 54.94] for tempo, 44.64% [38.20, 51.07] for silence, and 42.49% [36.48, 48.50] for merging. The paired tempo-minus-merging repair-rate difference is +6.01 points [−2.15, 14.59], and the silence-minus-merging difference is +2.15 points [−6.44, 10.31]. Thus the point estimates favor waveform compaction, while neither comparison establishes a repair advantage and the net-accuracy intervals for all three interventions include zero.

The cascade transcribes the original audio with Whisper large-v2 in English at temperature zero and supplies that transcript to the same text route. We apply Unicode NFC and whitespace normalization without restoring gold-text content. Across all 1,000 items, micro-averaged WER is 5.27%, mean item WER is 5.03%, and mean option-marker recall is 99.2%. The cascade gains 10.0 accuracy points over direct speech, but its difference from gold text is −2.6 points [−5.4, .1], with exact McNemar $p = .0783$. The comparison establishes a benefit of the cascade over direct speech in this protocol while leaving the residual cascade–text difference uncertain. The archived records include post-execution hashes of the inputs, transformed audio, transcripts, predictions, model files, code, and launch settings. These identify the evaluated artifacts but do not guarantee bitwise reproducibility across kernels or hardware.

D.4 Transcript Audit

Table 12 reports the Whisper-large-v2 transcript audit behind the word error rates in Section 4.1. Format-marker recall is 99.5% on BBH and 98.4% on SpeechMMLU.

Table 12: **ASR transcript audit.** We transcribe 200 randomly sampled items per benchmark with Whisper-large-v2. Word error rate (WER), substitution (Sub.), insertion (Ins.), and deletion (Del.) rates are averaged over items. Med. WER is the median item WER. Marker recall measures recognition of "the answer is" on BBH and that phrase plus "question" and "option A/B/C/D" on SpeechMMLU.

Dataset / slice	N	WER	Med. WER	Marker	Sub.	Ins.	Del.
SpeechMMLU	200	3.28%	0.00%	98.4%	1.89%	0.66%	0.72%
BBH overall	200	4.65%	3.45%	99.5%	4.09%	0.26%	0.29%
Hyperbaton	45	2.55%	0.00%	100.0%	1.82%	0.60%	0.13%
Navigate	51	3.09%	0.00%	100.0%	2.76%	0.06%	0.28%
Sports	54	3.22%	2.44%	100.0%	2.48%	0.31%	0.43%
Web-of-lies	50	9.66%	9.84%	98.0%	9.24%	0.11%	0.31%

D.5 Audio Duration

In the 1,024-token diagnostic, text-correct/speech-wrong items have longer audio than both-correct items on average, 23.58 versus 22.41 seconds, with a 95% bootstrap interval of [0.44, 1.89] seconds for the difference (Table 13). This duration comparison is pooled across categories.

Table 13: **BBH temporal-load metadata by correctness group.** Durations use the 1,000-item, 1,024-token diagnostic setting in Table 10. Confidence intervals are 95% bootstrap intervals over items, pooled across categories.

Group	Count	Duration (s)	Seconds per word
Both correct	541	22.41 [22.04, 22.77]	0.44 [0.43, 0.44]
Only T2T correct	178	23.58 [22.97, 24.18]	0.45 [0.44, 0.46]
Only S2T correct	56	21.23 [20.32, 22.17]	0.45 [0.43, 0.46]
Both wrong	225	21.54 [20.97, 22.11]	0.42 [0.42, 0.43]

D.6 Text-Redundancy Stress Test

Table 14 reports a text-side proxy stress test. Repeating each word reduces T2T accuracy relative to the unrepeated input in both benchmark-specific stress tests. The repeated text is out of distribution and changes both length and linguistic form.

Table 14: **Text-redundancy stress test for Qwen2.5-Omni.** Each word of the text prompt is repeated r times, and entries are T2T accuracy in percent.

r	SpeechMMLU	BBH
1	73.9	62.4
2	67.4	56.2
4	57.3	55.4
8	51.2	53.5

D.7 Speech-Native Boundary on MMSU

Table 15 reports selected speech-native MMSU tasks. For MMSU, Qwen2.5-Omni has 1 versus 257 unparsed speech/text answers among 2,420 items, and MiniCPM-o has 1 versus 4. Excluding seven shared invalid or ambiguous source labels preserves the speech advantage, so the boundary result combines acoustic cues with answer-formatting behavior.

Table 15: **MMSU results on selected speech-native tasks.** Acoustic cues can be answer-relevant on MMSU. Scores use deterministic parsing, counting unparsed answers as wrong. These selected tasks have large speech advantages, with Δ denoting S2T minus text accuracy in percentage points.

Task	N	Qwen2.5-Omni			MiniCPM-o		
		S2T	Text	Δ	S2T	Text	Δ
Continuation writing	110	100.0	22.7	+77.3	99.1	22.7	+76.4
Homophone reasoning	101	97.0	26.7	+70.3	95.0	35.6	+59.4
Intent detection	115	88.7	4.3	+84.3	87.0	25.2	+61.7
Polysemy reasoning	98	99.0	43.9	+55.1	96.9	38.8	+58.2
Deixis resolution	103	94.2	24.3	+69.9	83.5	51.5	+32.0
Dialogue reasoning	114	83.3	32.5	+50.9	86.8	41.2	+45.6
Code-switch QA	111	90.1	44.1	+45.9	93.7	54.1	+39.6
Speech translation	119	95.0	58.8	+36.1	90.8	61.3	+29.4

E Additional Results for MiniCPM-o

We also report projected-logit entropy, layer-wise decision margins, and norm controls for **MiniCPM-o** on SpeechMMLU and VoiceBench BBH using the 300 aggregate items per dataset with BF16 weights and no quantization.

E.1 Projected-Logit Entropy

Across both datasets, speech and text entropy both decrease in the last layers and end below 2 nats at the final coordinate. The two curves differ more than in Qwen2.5-Omni, with mean absolute differences of 1.34 nats on SpeechMMLU and 0.85 nats on BBH. Beyond layer 10, entropy first falls below 3 nats at layer 22 for speech and layer 25 for text on SpeechMMLU, and at layer 22 for both routes on BBH. Option-level margins complement this distribution-wide summary by measuring preference for the correct candidate. Figure 14 provides the corresponding layer-wise entropy trajectories.

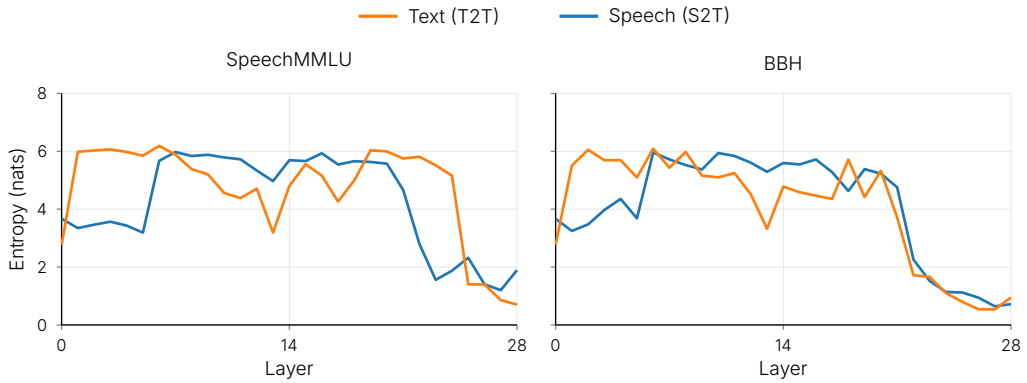


Figure 14: **MiniCPM-o layer-wise projected-logit entropy.** Curves show the mean over 300 items per benchmark. Speech and text both show a late-layer decrease, with larger differences between the routes than in Qwen2.5-Omni.

E.2 Layer-wise Margins by Correctness Group

In the text-only-correct group the text margin is clearly higher than the speech margin in the late layers (mean of the last six coordinates 4.65 versus 1.56 on SpeechMMLU, and 1.92 versus 0.45 on BBH), but the speech margin stays positive on both benchmarks rather than turning negative. Speech margins are also smaller when both routes are correct (3.67 versus 6.17, and 2.07 versus 3.82). Between 17% and 33% of the answers are unparsed (Table 6), so the groups partly reflect formatting. Figure 15 shows the full groupwise margin curves.

E.3 Pre-LN and Post-LN Norms

For each item and block, we take the median L2 norm over the evidence tokens of the block input before and after the block's input layer norm, and average over the 300 items. Pre-LN norms reach 535 to 553 for speech and 675 to 718 for text at the last block, whereas post-LN norms are 71 to 77 and differ between modalities by a factor of at most 1.6. These curves describe activation magnitude, which does not determine the direction of the normalized states. Figure 16 reports the norm-control results.

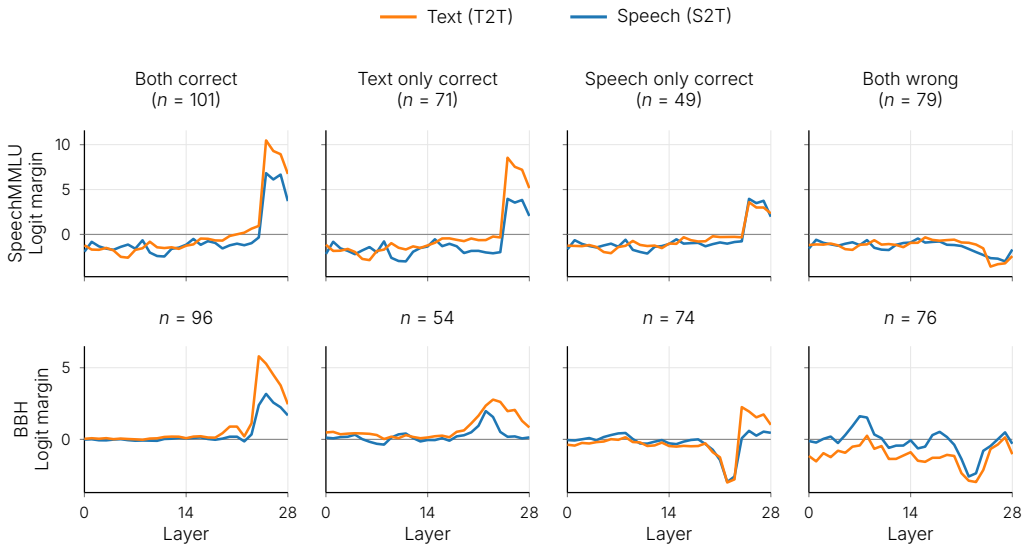


Figure 15: **MiniCPM-o layer-wise logit margin by correctness group.** Rows are SpeechMMLU and BBH. In the text-only-correct group the text margin exceeds the speech margin in late layers, and speech margins are smaller than text margins when both routes are correct.

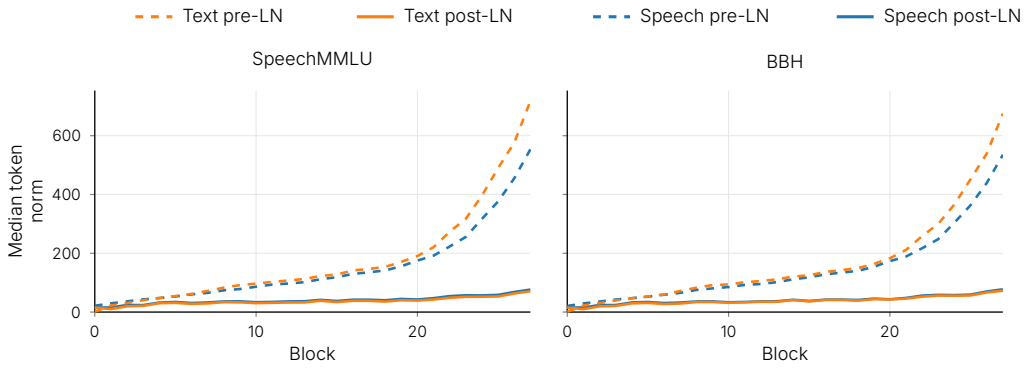


Figure 16: **MiniCPM-o pre-LN and post-LN norms.** Left is SpeechMMLU and right is BBH. Dashed curves are the block input before the input layer norm and solid curves after it. Post-LN norms are comparatively stable across modalities despite late pre-LN divergence.

F Additional Results for LLaMA-Omni

We also examine **LLaMA-Omni**, using 33 hidden-state coordinates including the embedding output, with hidden dimension 4096. We run Llama-3.1-8B-Omni in BF16 on the same 300 aggregate items with its own Llama-3 chat template and system prompt. Its speech route always receives the 300 projected positions of a 30-second padded input, and we analyze the positions that cover audio.

F.1 Cross-Layer CKA

Similarity is high across most layer pairs: CKA ranges from 0.40 to 0.66 on SpeechMMLU and from 0.51 to 0.82 on BBH, and only the first two text layers are clearly lower. The best-match path maps every speech layer to text layer 15 on SpeechMMLU and to three text layers on BBH, so, compared to Qwen2.5-Omni, the high-similarity region extends across more text layers and the correspondence is less layer-specific. Figure 17 shows the cross-layer CKA heatmaps.

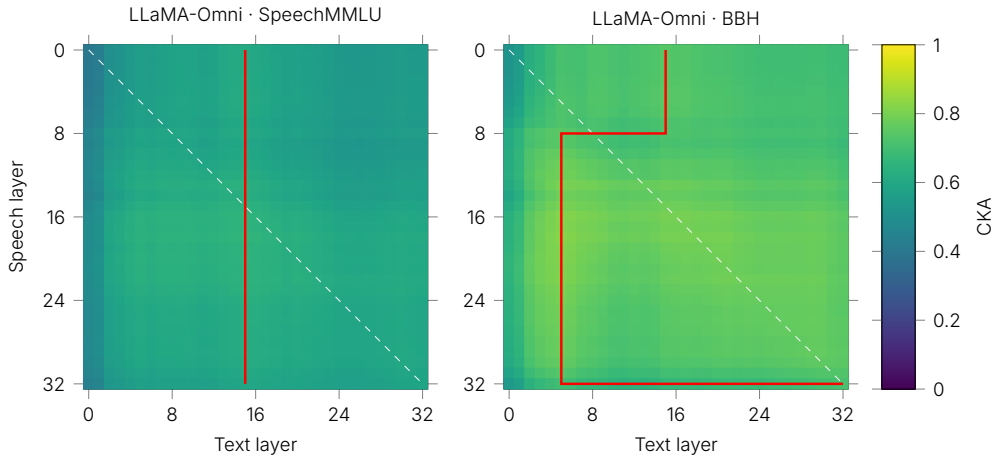


Figure 17: **LLaMA-Omni cross-layer CKA heatmaps.** Left is SpeechMMLU with DTW base layer 10 and right is BBH with base layer 20. Red lines mark the best-matching text layer of each speech layer and dashed lines the diagonal. Similarity is high across most layer pairs and lowest at the first two text layers.

F.2 Projected-Logit Entropy

Entropy trends are broadly aligned between modalities, with mean absolute differences of 1.31 nats on SpeechMMLU and 1.19 nats on BBH. Beyond layer 10, entropy first falls below 3 nats at layer 23 for speech and layer 19 for text on SpeechMMLU, and at layer 22 for both routes on BBH. We compare option-level margins below to examine which answer each route favors. Figure 18 provides the entropy curves.

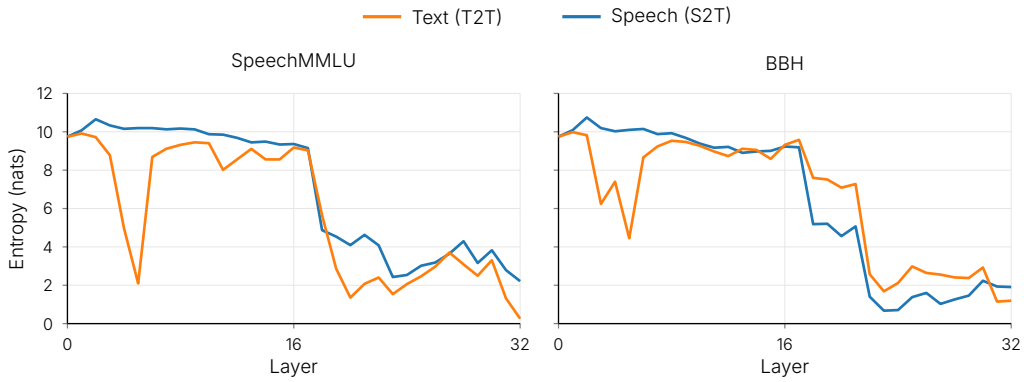


Figure 18: **LLaMA-Omni layer-wise projected-logit entropy.** Curves show the mean over 300 items per benchmark. Entropy curves are broadly comparable across modalities.

F.3 Layer-wise Margins by Correctness Group

As with Qwen2.5-Omni, the text-only-correct group has a negative speech margin and a positive text margin in the late layers (mean of the last six coordinates −1.50 and +3.04 on SpeechMMLU, −0.44 and +1.02 on BBH). The effect is smaller on BBH than on SpeechMMLU. Speech margins are also negative when both routes are correct on SpeechMMLU (−0.57), and 114 of the 300 speech answers on each benchmark are unparsed, so text-only-correct items dominate the groups (168 and 122). Figure 19 shows the groupwise margin trajectories.

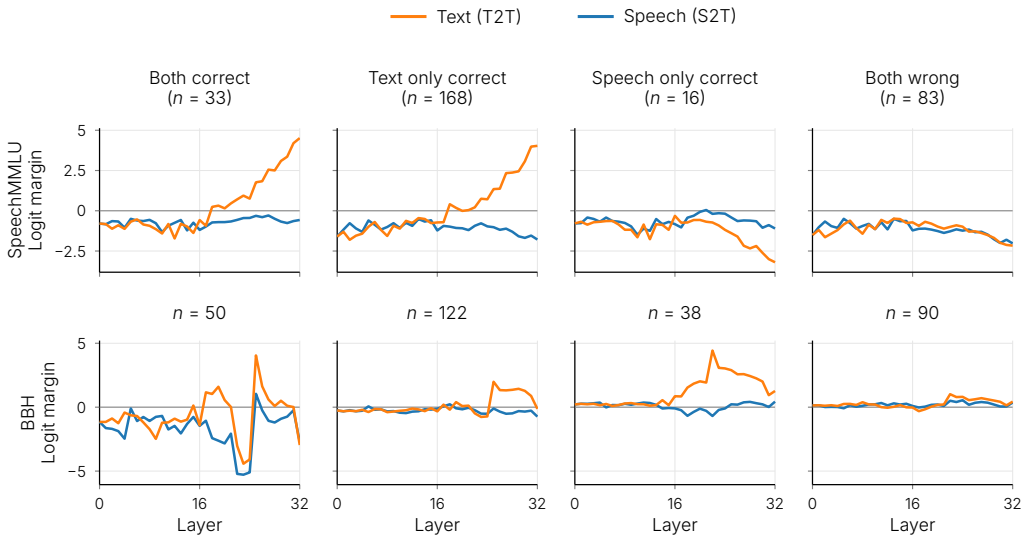


Figure 19: **LLaMA-Omni layer-wise logit margin by correctness group.** Rows are SpeechMMLU and BBH. The text-only-correct group exhibits negative speech margins and positive text margins in late layers.

F.4 Pre-LN and Post-LN Norms

Pre-LN norms rise from about 5 at block 8 to 40 to 45 at the last block, whereas post-LN norms stay at about 20 to 35 after the first blocks. The speech-to-text ratio of the post-LN norm ranges from 0.64 to 1.08 and ends at 1.04 to 1.05, so the cross-modal difference is small after normalization. Figure 20 provides the norm-control visualizations.

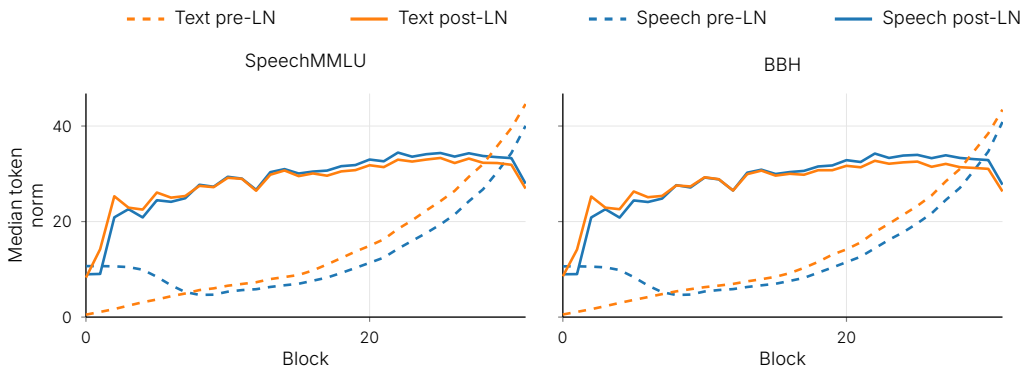


Figure 20: **LLaMA-Omni pre-LN and post-LN norms.** Left is SpeechMMLU and right is BBH. Post-LN norms are comparatively stable across modalities while pre-LN norms grow with depth.

G Additional Results for Qwen2-Audio

We examine **Qwen2-Audio**, whose speech accuracy on SpeechMMLU is below the 25% uniform-choice baseline and whose modality gap is smaller on BBH. Its CKA heatmaps show low cross-modal similarity through the middle layers, rising only in the last speech layers. We use *Representation Failure* to describe this weak-alignment pattern, in contrast to the middle-layer alignment followed by late decision instability seen in Qwen2.5-Omni and MiniCPM-o.

G.1 Cross-Layer CKA

Unlike Qwen2.5-Omni and MiniCPM-o, Qwen2-Audio does not exhibit a clear diagonal or broad middle-layer band in its CKA heatmap. Instead, the best-match path collapses toward a near-vertical trajectory that maps many speech layers to the same text layer. On SpeechMMLU, 21 of the 33 speech layers share text layer 19, and speech layers 9–32 use only text layers 18–20. On BBH, speech layers 9–32 map to text layers 21, 23, 25, and 26, as shown in Figure 21.

Low CKA through the middle layers. The key diagnostic observation is that Qwen2-Audio shows low CKA values not just in early layers, as in Phase I of the other models, but through the middle of the network: the row maxima of speech layers 10–21 stay at or below 0.07 on SpeechMMLU and 0.13 on BBH. In Qwen2.5-Omni and MiniCPM-o, CKA heatmaps exhibit a pronounced high-similarity band in middle layers where speech and text representations converge. Qwen2-Audio lacks this band. Similarity rises only in the last speech layers, reaching 0.45 on SpeechMMLU and 0.66 on BBH at the final layer. The measured speech and text trajectories therefore show weak geometric correspondence through the middle of the network.

Task-dependent impact. Weak geometric correspondence accompanies a 43.07-point gap on SpeechMMLU and a 2.2-point gap on BBH (Table 3). The task difference motivates separating geometric alignment from task-specific answer formation, as in the readouts below.

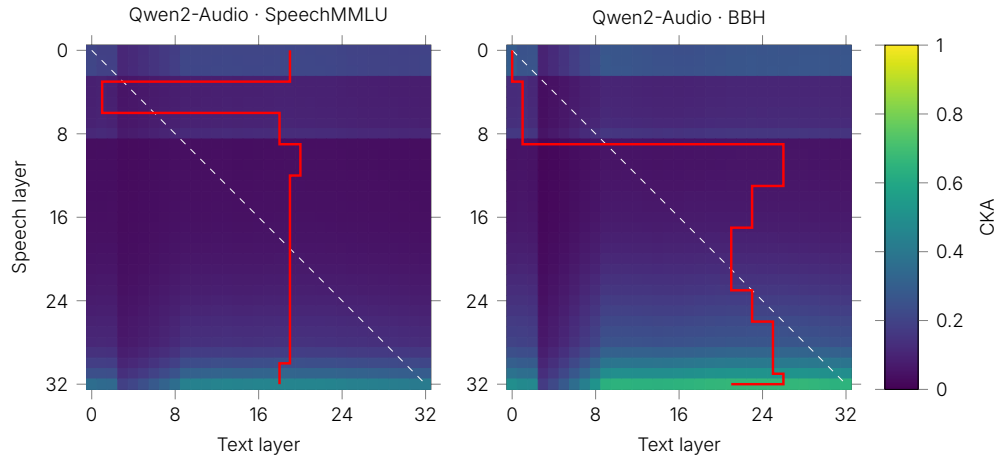


Figure 21: **Qwen2-Audio cross-layer CKA heatmaps.** Left is SpeechMMLU with DTW base layer 10 and right is BBH with base layer 20. Red lines mark the best-matching text layer of each speech layer and dashed lines the diagonal. Unlike Qwen2.5-Omni and MiniCPM-o, Qwen2-Audio lacks a clear high-similarity middle-layer band.

G.2 Candidate-Only Readout

On SpeechMMLU, text begins to concentrate probability mass onto the correct option in late layers, reaching 0.45 at the last layer against a chance level of 0.25, whereas speech stays within 0.24–0.27 and does not elevate the correct option above competitors. On BBH both routes stay between 0.45 and 0.55 at every layer, around the two-option chance level of 0.5. Figure 22 provides the candidate-only readout curves.

G.3 Projected-Logit Entropy

Entropy is comparable across modalities, with mean absolute differences of 0.75 nats on Speech-MMLU and 0.71 nats on BBH. Beyond layer 10, it first falls below 3 nats at layer 19 on SpeechMMLU and at layer 23 on BBH for both routes, well before the final layers. On SpeechMMLU, speech entropy exceeds text entropy by more than 1 nat at layers 22–26, so its late decrease is delayed, and

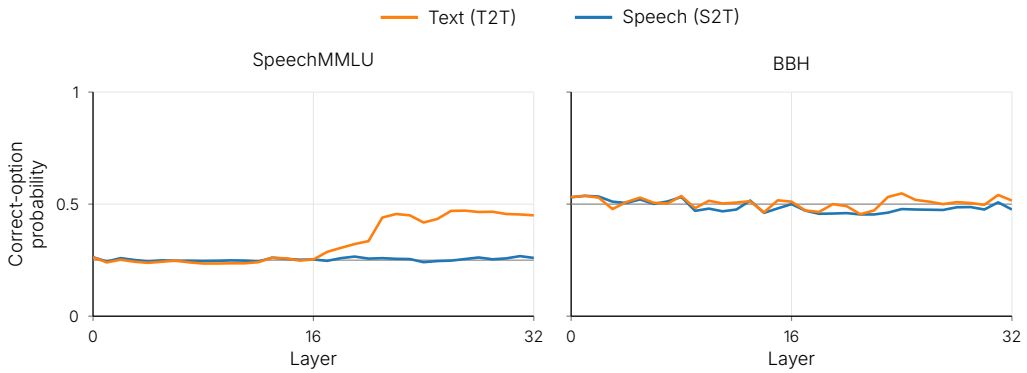


Figure 22: **Qwen2-Audio candidate-only readout.** Per-layer probability of the correct answer option among candidates for S2T and T2T at the decision position after an assistant answer prefix, averaged over 300 items per benchmark. Thin lines mark chance, 0.25 on SpeechMMLU and 0.5 on BBH. Speech does not raise the correct option, especially on SpeechMMLU.

the gap is below 1 nat from layer 27 on. Figure 23 shows the layer-wise entropy trajectories for Qwen2-Audio.

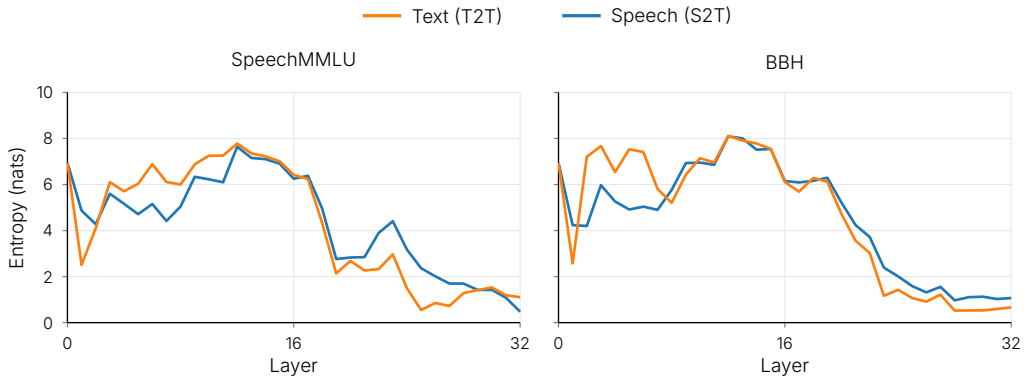


Figure 23: **Qwen2-Audio layer-wise projected-logit entropy.** Curves show the mean over 300 items per benchmark. Speech entropy is not uniformly higher than text entropy across layers.

G.4 Layer-wise Margins by Correctness Group

On SpeechMMLU, text margins are positive in late layers in both groups where text is correct (mean of the last six coordinates +3.15 and +1.49), while speech margins stay negative in every group (−0.99 to −2.03). On BBH the margins are small, and the text margin is negative in the text-only-correct group (−0.44, against −1.11 for speech). Speech-only-correct cases number 30 on SpeechMMLU and 69 on BBH, and the SpeechMMLU group shows no positive speech margin (−1.50). Figure 24 provides the groupwise margin curves.

G.5 Pre-LN and Post-LN Norms

Pre-LN norms differ between modalities by up to a factor of 45 across blocks (speech-to-text ratio 0.75 to 45) and reach 170 to 180 for speech and 226 to 227 for text at the last block, whereas post-LN norms, i.e., the normalized stream entering attention, differ by at most 25% (ratio 0.97 to 1.25). Normalization reduces the cross-modal difference in activation magnitude. Figure 25 shows the pre-LN and post-LN norm controls for Qwen2-Audio.



Figure 24: **Qwen2-Audio layer-wise logit margin by correctness group.** Rows are SpeechMMLU and BBH. Margins favor the correct option less strongly under speech than under text on SpeechMMLU.

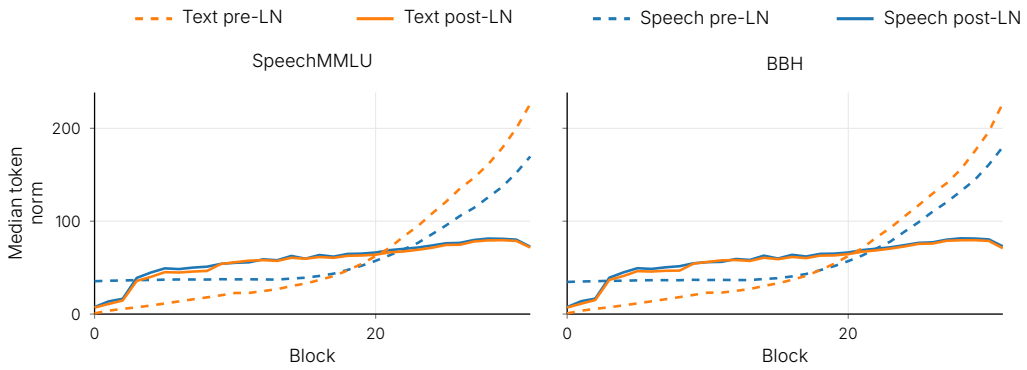


Figure 25: **Qwen2-Audio pre-LN and post-LN norms.** Left is SpeechMMLU and right is BBH. Post-LN norms remain comparatively stable across modalities.

H Qualitative Cases

We inspect four examples from an earlier comparison in which the model answered correctly under text input and incorrectly under speech input.

Setup. These instances have matched prompts where T2T was correct and S2T was incorrect in that comparison, and this selection is not re-measured here. We re-execute each item with Qwen2.5-Omni in BF16 and plot, for each layer, the probability assigned to the correct option, normalized over the answer candidates after an assistant answer prefix. This view isolates whether the correct answer becomes a plausible competitor during intermediate computation, or whether it never gains probability mass under speech.

Results. Only `hyperbaton_105` shows a text advantage at the answer layers: the text route reaches 0.90 at the last layer while the speech route stays at 0.22. In `hyperbaton_111` the order reverses, with the speech route at 0.90 and the text route at 0.35 at the last layer. In the two SpeechMMLU cases the routes end at similar probabilities after fluctuating through the middle and late layers (speech and text 0.50 and 0.46 for `anatomy_1`, 0.48 and 0.46 for `anatomy_11`). This readout is a candidate probability at a single position and does not reproduce the free-generation

outcome that selected these items, so the traces show layer-to-layer variability rather than a general speech failure. Table 16 and Figure 26 summarize the selected prompts and the corresponding per-layer traces. In `web_of_lies_74`, baseline S2T answers "No" while gold-text T2T answers "Yes" in the 1,024-token diagnostic, and tempo 1.2× flips S2T to "Yes" without changing lexical content. Under the shared-instruction evaluation, both routes answer this item correctly.

Table 16: **Qualitative cases (text correct, speech wrong).** We summarize selected questions where T2T answered correctly but S2T failed in the earlier comparison that selected them. The accompanying figure plots the per-layer probability of the correct option, normalized over answer candidates.

Dataset	Item	Question summary
SpeechMMLU	`anatomy_1`	Which craniofacial condition is associated with a "dished face" profile?
SpeechMMLU	`anatomy_11`	Which signs follow damage to the third cranial nerve after an orbital injury?
BBH	`bbh_hyperbaton_105`	Choose the correct ordering of adjectives describing a repulsive, normal-size, green, old-fashioned cardboard surfboard.
BBH	`bbh_hyperbaton_111`	Choose between "little black driving ship" and "driving little black ship."
BBH	`bbh_web_of_lies_74`	Jamey tells the truth. Raymond says Jamey lies. Audrie says Raymond lies. Michaela says Audrie lies. Rashida says Michaela lies. Does Rashida tell the truth?

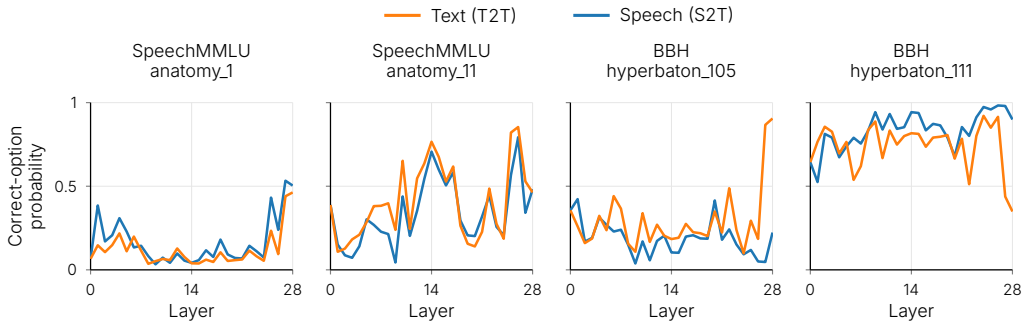


Figure 26: **Candidate-probability traces for four instances selected as text-correct and speech-wrong.** Curves show, for each fixed item, the probability of the correct option after normalization over answer candidates in Qwen2.5-Omni. Only `hyperbaton_105` has a higher text probability at the last layer.

I Implementation Details

I.1 Hyperparameters and Setup

Key implementation choices that affect all analyses are reported below. Table 17 lists the shared settings used for DTW and CKA analyses in this paper.

I.2 Compute Resources

The confirmatory alignment, paired-readout, and intervention runs used NVIDIA A800 GPUs with 80 GB memory, with independent conditions or shards distributed over four GPUs. Each native-center alignment cell used one GPU and 100 items. Its recorded processing interval was 178 seconds for Qwen2.5-Omni on BBH, 342 seconds on SpeechMMLU, 175 seconds for MiniCPM-o on BBH, and 327 seconds on SpeechMMLU. These intervals run from metadata initialization before model loading through completion. The four cells therefore used about 0.284 GPU-hours within those intervals.

Table 17: **Summary of implementation hyperparameters.**

Parameter	Value or Method
Datasets and Audio	
SpeechMMLU	Dataset-provided paired `question_text`, answer options, `question_audio`, and speaker identifier from MiMo-Audio-Eval
VoiceBench BBH	Dataset-provided paired `prompt`, reference answer, and audio clip for `hyperbaton`, `navigate`, `sports_understanding`, and `web_of_lies`
T2T Prompt	The paired text field corresponding to the audio item. No ASR transcript is used for gold-text T2T
Hidden State Extraction	
Extraction Window	Last user segment, i.e., question turn only
Valid Tokens	Positions where `attention_mask=1`
DTW Alignment	
Base Layer ℓ_0	Layer 10 for SpeechMMLU, Layer 20 for BBH
Metric	Cosine similarity with per-row normalization
Step Penalty λ	0, i.e., standard DTW
Band Constraint	None, i.e., unconstrained
Many-to-one Aggregation	Speech frames aligned to the same text token are mean-pooled before CKA, standardized L2, and norm computation
Cross-Layer Aggregation	
Item Selection	Fixed set of 100 items with seed 0
CKA Method	Per-example linear CKA with double-centering, averaged across the selected items
ASR-to-LLM Control	
Transcript Prompting	Same downstream LLM prompted with ASR transcript. Explicit question/option markers are preserved or restored when present

The study also includes behavioral inference, other diagnostics, and discarded pilots, for which a complete total compute accounting is unavailable.

J Broader Impact

Separating internal similarity from answer reliability can help developers evaluate speech interfaces before using them for consequential tasks. Our results also show that a change in an internal diagnostic need not improve accuracy, which can prevent misplaced confidence in a repair. Suppressing acoustic variation indiscriminately could remove information needed for emotion, intent, or accessibility-related communication. The benchmark setting does not measure performance disparities across accents or disabilities, so the reported diagnostics should not be treated as evidence of equitable deployment performance.

K Existing Assets and Licenses

We use existing pretrained models and benchmark items without training new model weights or collecting new human-subject data. Table 18 records the licenses stated by the linked providers, checked in October 2026. The MiMo-Audio-Eval toolkit license does not by itself specify the license of SpeechMMLU's synthesized audio. OpenAI's original Whisper release states MIT for code and weights, while the Hugging Face model card labels the checkpoint Apache-2.0.

Table 18: **Provider-stated licenses of existing assets.** Links point to the official model cards, dataset cards, or repositories. These are source declarations rather than a complete audit of every underlying asset.

Asset and source	Stated license or condition
Qwen2.5-Omni-7B	Apache-2.0
MiniCPM-o 2.6	Apache-2.0 for code and weights
Qwen2-Audio-7B-Instruct	Apache-2.0
LLaMA-Omni	Code: Apache-2.0. Model: academic, noncommercial use with citation
Whisper-large-v2	MIT in the OpenAI release, with differing checkpoint-card metadata
MiMo-Audio-Eval toolkit	Apache-2.0
SpeechMMLU	Audio license not stated in the dataset card
MMLU	MIT
VoiceBench / BBH	Apache-2.0 / MIT
MMSU	MIT

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction distinguish geometric correspondence, correctness-conditioned readouts, and the fixed-input causal effect on separation magnitude.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 6 and the affected results discuss benchmark scope, model coverage, answer-label sensitivity, uncertain route convergence, and uncertainty in repair gains.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper is empirical and diagnostic. It does not present theoretical results or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No].

Justification: The controlled alignment, paired-readout, attention-intervention, and shared-instruction evaluations are specified in Appendices A.2, B.1, C.3, and D.3, respectively. Complete settings for some additional diagnostics, including the probe training, text-repetition entropy, ASR-correct subset selection, and DTW path summaries, are unavailable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The study uses open-weight models and public benchmark sources. A complete public release of the diagnostic code is not provided in this version.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No].

   Justification: The methodology and appendix give model choices, alignment settings, and controlled intervention protocols. Some additional diagnostic settings remain incomplete, as described in the reproducibility answer above.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The alignment controls, paired readout contrasts, attention edits, and behavioral comparisons report item-bootstrap confidence intervals, with the resampling unit, strata, and number of resamples specified. These quantify item uncertainty rather than training-seed variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: Appendix I.2 reports accelerator type, memory, and measured processing intervals for the native alignment controls. Complete per-run and total compute accounting for the entire study is unavailable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work analyzes existing public models and benchmark items, does not collect private data, and is intended to improve transparent understanding of speech LLM failures.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: Appendix J discusses improved reliability evaluation, misplaced confidence in internal metrics, loss of useful acoustic information, and unmeasured deployment disparities.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a new high-risk model, dataset, or scraped corpus.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No].

Justification: Appendix K credits the assets and lists provider-stated licenses. The Speech-MMLU dataset card does not specify a license for its synthesized audio, and a complete audit of underlying source assets is unavailable.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release a new dataset, model checkpoint, or benchmark asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The research does not involve crowdsourcing experiments or new human-subject data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

Justification: The research does not involve new human-subject experiments requiring IRB review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: The studied systems are LLM-based speech models, and their use is central to the empirical subject of the paper rather than an incidental writing aid.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.