# Interleaved Speech Language Models Latently Work In Text

**Talia Sternberg** **Gallil Maimon** **Yossi Adi**
The Hebrew University of Jerusalem
talia.sternberg@mail.huji.ac.il

arXiv:2606.22473v2 [cs.CL] 6 Sep 2026

## Abstract

Speech language models (SLMs) increasingly combine speech and text, often by interleaving their tokens within a single sequence. Yet how these two modalities interact in the model's latent space remains unclear. In this work, we analyze interleaved speech–text LMs from different model families and training configurations using three complementary methods. We reveal that these models pass through an *implicit latent transcription* phase in which the text token matching the spoken word becomes decodable in intermediate layers, *despite not being trained for speech recognition*. This phenomenon occurs in diverse, natural speech, and intermediate representations also encode likely text continuations. We further show that implicit transcription emerges most when combining text-LM pretraining and speech–text interleaving, and that its prevalence is positively associated with spoken factual-knowledge retrieval. Our analysis sheds light on the internal interaction between speech and text modalities in interleaved SLMs.

## 1 Introduction

Speech language models (SLMs) are gaining popularity as the basis for universal speech processing systems as well as dialogue models (Arora et al., 2026). Such models hold potential to reason about speech (Yosha et al., 2026; Sakshi et al., 2025), as well as answer spoken questions or instructions (Nachmani et al., 2024; Chen et al., 2024). However, several recent works have indicated lacking semantic and knowledge capabilities in speech-only SLMs (Cuervo and Marxer, 2024).

More recently Speech LMs started integrating text data and pre-trained text LMs into speech LMs (Défossez et al., 2024; Hassid et al., 2023; Xie and Wu, 2024). One such method is speech-text interleaving, in which models are trained over "code switching" data which contains both speech and text tokens as a single stream (Nguyen et al., 2025; Zeng et al., 2025; Manakul et al., 2026). Maimon et al. (2025b) demonstrated that this interleaving method improves scaling dynamics of the semantic abilities of SLMs relative to the speech-only paradigm, even when considering speech-in speech-out performance. While these methods improve speech-in speech-out capabilities, the mechanism underlying this behavior, and the internal interplay between speech and text modalities, still remain unclear.

In this work, we analyze latent dynamics between modalities using three complementary decoding methods. We use the logit lens (nostalgebraist, 2020), which projects intermediate hidden states into the token vocabulary, and complement it with a latent-space retrieval method (Krojer et al., 2026) and supervised linear probes (Alain and Bengio, 2018; Belinkov, 2021). This analysis reveals a clear pattern in SLMs' internal representations (Figure 1): interleaved SLMs operate in a meaningful textual latent space within the hidden layers of the transformer. Specifically, speech-derived representations become decodable as the corresponding text transcription and as hypotheses about the next word, before being projected back into the speech-token space. This occurs despite the models not being explicitly trained for speech transcription. We refer to this phenomenon as *implicit latent transcription*.

We further show in Section 4.4 that these findings occur across diverse speakers, domains, and recording conditions - from controlled synthesized prompts to natural speech.

In Section 4.5, we analyze which training decisions impact this phenomenon. Notably, we highlight that implicit latent transcription emerges most strongly when combining initialization from a trained text LM and speech-text interleaving. Conversely, models without interleaving or without text-LM initialization exhibit substantially weaker transcription signals.

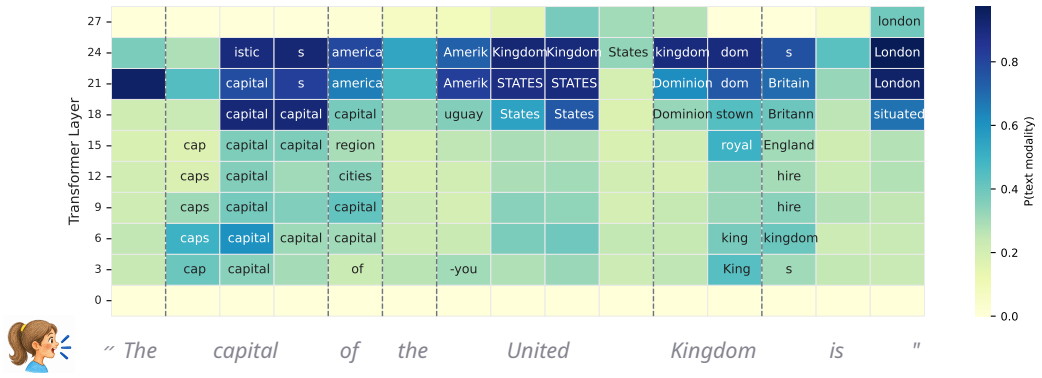



Figure 1: **Implicit transcription emerges without speech-recognition supervision.** Logit-lens analysis of intermediate states for the spoken prompt "The capital of the United Kingdom is...". Cells show textual tokens probability, from light yellow for zero to dark blue for high probability. The labels show the most probable relevant textual token at each position; notably, the model predicts "London" although it was not spoken in the prompt.

Finally, we analyze how this phenomenon correlates with basic spoken factual knowledge retrieval. Our results, presented in Section 4.6, suggest that implicit transcription is positively associated with factual knowledge abilities. We also qualitatively analyze examples and find that transcriptions often build gradually over the course of a spoken word and sometimes contain acoustic errors (Section 4.7).

Our main contributions are: (i) Showing that interleaved Speech LMs latently transcribe the current spoken word and encode hypotheses about textual continuations, while not being trained for speech recognition, with consistent evidence across three complementary methods. (ii) Highlighting the importance of both text LM initialization, and interleaving data in eliciting this behavior. (iii) Analyzing the relationship between implicit transcription and spoken knowledge abilities. The dataset, audio samples, and additional implicit-transcription visualizations are available on our project page.[1]

## 2 Background

**Interleaved Speech Language Models** are trained on both discrete speech units and text tokens. These models commonly include three components: a speech-to-unit module, a joint speech-text LM, and a unit-to-speech module (Arora et al., 2026). The speech-to-unit module converts raw audio into discrete units, typically using a self-supervised model such as HuBERT (Hsu et al., 2021) and quantizing these representations. For more on speech tokenization see Mousavi et al. (2025a). The joint Speech LM then models sequences containing both speech units and text tokens, often initialized from a pretrained text LM (Hassid et al., 2023). Finally, generated speech units can be converted back into waveform audio using a unit-to-speech decoder (Lakhotia et al., 2021).

We specifically focus on interleaved speech-text SLMs, where speech units and text tokens are mixed within the same sequence (Nguyen et al., 2025). Such models show promising results with elegant modeling of a single stream of tokens (Nguyen et al., 2025; Maimon et al., 2025b; Zeng et al., 2025). For model training, given time-aligned transcriptions of speech samples, each word is assigned either the speech or text modality. Consecutive words with the same modality are grouped into spans: text spans are tokenized as text, while speech spans are replaced by their discrete units. This yields sequences such as `[TEXT] the monkey climbed [SPEECH] Hu14 Hu62 Hu9 ... Hu31 [TEXT] tree.`

Such a modeling approach trains the SLM to generate cross-modal continuations within the same sentence, under the assumption that semantic information is transferred across modalities.

## 3 Approach

### 3.1 Interpreting Latent Embeddings

Let $h_i^{(\ell)}$ denote the hidden state at speech-token position $i$ and layer $\ell$. We aim to understand what information is encoded in speech LM latent representations, and how this information evolves across the layers of the model. To do so, we decode these

[1] https://pages.cs.huji.ac.il/adiyoss-lab/slm_work_in_text/

representations using three complementary methods.

**Logit Lens.** The logit lens (nostalgebraist, 2020) applies the model's final normalization and output projection to intermediate hidden states:

$$p^{(\ell)}_{\text{logit}}(x \mid i) = \text{softmax}\left(W_{\text{out}}\,\text{Norm}_f\left(h_i^{(\ell)}\right)\right)_x,$$

where $\text{Norm}_f$ is the model's final RMSNorm and $W_{\text{out}}$ is its learned vocabulary projection.

Although this projection is normally applied only at the final layer, applying it throughout the network yields a layer-wise distribution over the vocabulary and reveals which token predictions are already linearly decodable at each depth. Since transformer predictions are formed gradually through the shared residual stream (Geva et al., 2022), this provides a direct view of how vocabulary-aligned information emerges and changes across layers. Logit-lens analyses have been widely used to study the evolution of intermediate predictions (Dar et al., 2023; Wendler et al., 2024; Yang et al., 2024; Halawi et al., 2024; Neo et al., 2025).

**Latent Lens.** The logit lens only reveals information aligned with the model's output projection. We therefore complement it with Latent Lens (Krojer et al., 2026), which tests whether speech hidden states are geometrically close to text representations produced by the same model.

We collect contextualized occurrences of the evaluation words and random distractors from held-out Red Pajama text. At each layer, we average the representations of all occurrences of each word to obtain one model-specific text representation per word. We then rank the words by their cosine similarity to each speech hidden state and treat the top-$k$ words as the decoded candidates.

**Linear Probes.** We additionally use simple linear probes (Alain and Bengio, 2018; Belinkov, 2021) to test whether the current and next words are linearly recoverable from speech hidden states. Each probe is a single affine classifier followed by a softmax. For each model and layer, we train separate current- and next-word probes while keeping the SLM frozen. Each probe predicts over the 10,000 most frequent tokenizer-specific RedPajama tokens , together with all evaluation targets, and is trained on $5M$ aligned speech-token positions from People's Speech dataset (Galvez et al., 2021). The probe measures whether lexical information is linearly recoverable, even when it is not aligned with the model's output head or the text representations used by the latent lens.

Across all decoding methods, we lowercase text and strip punctuation to ensure exact string matching, then evaluate on the first word-start token.

### 3.2 Word-Level Analysis of Speech Latents

A spoken word spans a variable number of speech tokens, and no fixed position within the span necessarily contains the strongest information about its identity. Let $S(w)$ denote the speech-token positions aligned with word $w$.

**Modality Mass** is calculated using the logit lens, as it is the only method that produces a probability distribution over the joint speech-text vocabulary. Let

$$p_{\text{logit}}(i)^{(\ell)} = \text{softmax}\Big(W_{\text{out}}\,\text{Norm}_f\Big(h_i^{(\ell)}\Big)\Big)$$

denote the vocabulary distribution at position $i$ and layer $\ell$. For modality $m \in \{\text{text}, \text{speech}\}$, we define the word-level modality score as

$$s_m^{(\ell)}(w) = \max_{i \in S(w)} \sum_{v \in \mathcal{V}_m} p_{\text{logit}}(i)^{(\ell)}(v),$$

where $\mathcal{V}_m$ is the corresponding modality vocabulary.

**Word Recovery.** For decoding method $j \in \{\text{logit}, \text{latent}, \text{probe}\}$, let $p_j(i)^{(\ell)}$ denote its output scores at position $i$ and layer $\ell$. These are vocabulary probabilities for the logit lens and probe, and cosine-similarity scores over bank entries for the latent lens.

A word is recovered at layer $\ell$ if its target label appears among the top-$k$ candidates at any position in $S(w)$. Because different words become decodable at different depths, we report cumulative Recall@$k$ through layer $L$:

$$\frac{1}{|\mathcal{D}|}\sum_{w \in \mathcal{D}} \mathbf{1}\Big[\exists\, \ell \le L,\ i \in S(w) \\ \text{s.t. } y_w \in \text{TopK}\big(p_j(i)^{(\ell)}, k\big)\Big].$$

where $\mathcal{D}$ is the set of evaluated word instances and $y_w$ is the target label. Thus, a word is counted once it becomes decodable at any layer up to $L$.

We use max aggregation because the signal is sparse and its position varies across words. Robustness to this choice, along with localization, layer-wise recall, and persistence, is reported in Appendices A.4.1 and A.4.3.

**Matched-frequency baseline.** To control for word-frequency effects, we compare each gold word with 20 distractor words sampled from the same RedPajama frequency band and score them using the same Recall@$k$ protocol. Full sampling details are provided in Appendix A.2.

### 3.3 Evaluation Data

We test whether speech–text interleaving enables models to access factual knowledge from speech. Because the factual content of large speech corpora is unknown, we compare text-pretrained interleaved models against speech-only and randomly initialized baselines. Better spoken fact completion by the interleaved models would suggest that joint training supports cross-modal acquisition or retrieval of factual knowledge.

Existing speech benchmarks do not offer controlled, single-word fact-completion tasks, while standard question-answering datasets often allow multiple valid answers. We therefore construct a diagnostic dataset of 282 short prompts with a single unambiguous completion, such as *"The capital of France is ..."*. The examples span colors, days and months, object functions, capital cities, arithmetic, and other common-sense categories. We synthesize the prompts with Kokoro-82M (Hexgrad, 2025) and obtain word-level alignments using Whisper large-v3 (Radford et al., 2023).This controlled format enables precise analysis of factual retrieval. Statistics and examples appear in Table 5, and the dataset is publicly available. We additionally validate our main findings on natural speech corpora spanning diverse speakers, domains, and recording conditions, see 4.4.

**Knowledge Evaluation in Speech LMs.** We then use the above dataset for factual knowledge evaluation. Each true fact is paired with a matched counterfactual example by replacing the correct answer with an incorrect one, chosen randomly from the same category, i.e *'The color of a banana is yellow'* Vs. *'The color of a banana is red'*. We score a model as correct when it assigns higher likelihood to the true fact than to the wrong fact.

$$\log p(\text{fact}) > \log p(\text{counterfactual fact}).$$

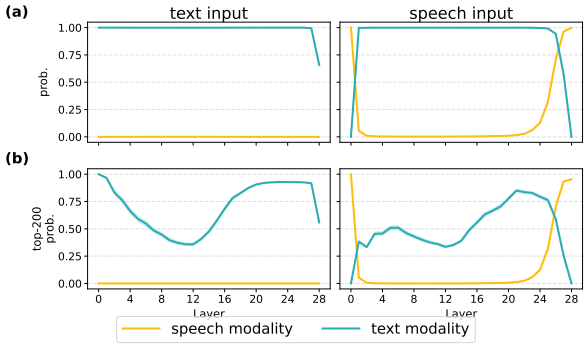



Figure 2: **Speech LMs operate in text**. Modality distribution of inner state Logit Lens. (a) Sum of probabilities over all speech tokens and text tokens respectively. (b) Same but only considering top 200 tokens.

We report the percentage of examples for which this condition holds. Such likelihood based evaluation is commonly used for Speech LMs (Hassid et al., 2023; Maimon et al., 2025c).

### 3.4 Experimental Setup

We analyze interleaved Speech LMs based on Llama 3.2-3B (Dubey et al., 2024) and Qwen 2.5 (Qwen et al., 2025), trained on mixtures of speech, text, and speech–text interleaved data (Maimon et al., 2025b). Model and training details are provided in Appendix A.3.

To isolate the factors underlying implicit transcription, we train controlled Llama 3.2-3B models while varying initialization and data composition. Models are initialized either from a pretrained text LM (P) or randomly (R), and trained on speech only (S), balanced speech and text (ST), or speech, text, and interleaved data. In the latter setting, interleaved examples constitute $1/3$, $2/3$, or $5/6$ of the training tokens, denoted I-1/3, I-2/3, and I-5/6, respectively. Importantly, all models are trained only with next-token prediction and receive no ASR or transcription loss. Time-aligned transcripts are used only to construct the interleaved sequences; the model is never trained to map a speech span directly to its corresponding transcript.

## 4 Results

### 4.1 Speech LMs Operate In Text Latent Domain

We first examine whether speech representations remain in the speech-token domain throughout the network. Using the modality-mass measure defined in Section 3.2, we compare the logit-lens probability assigned to speech and text tokens across layers.

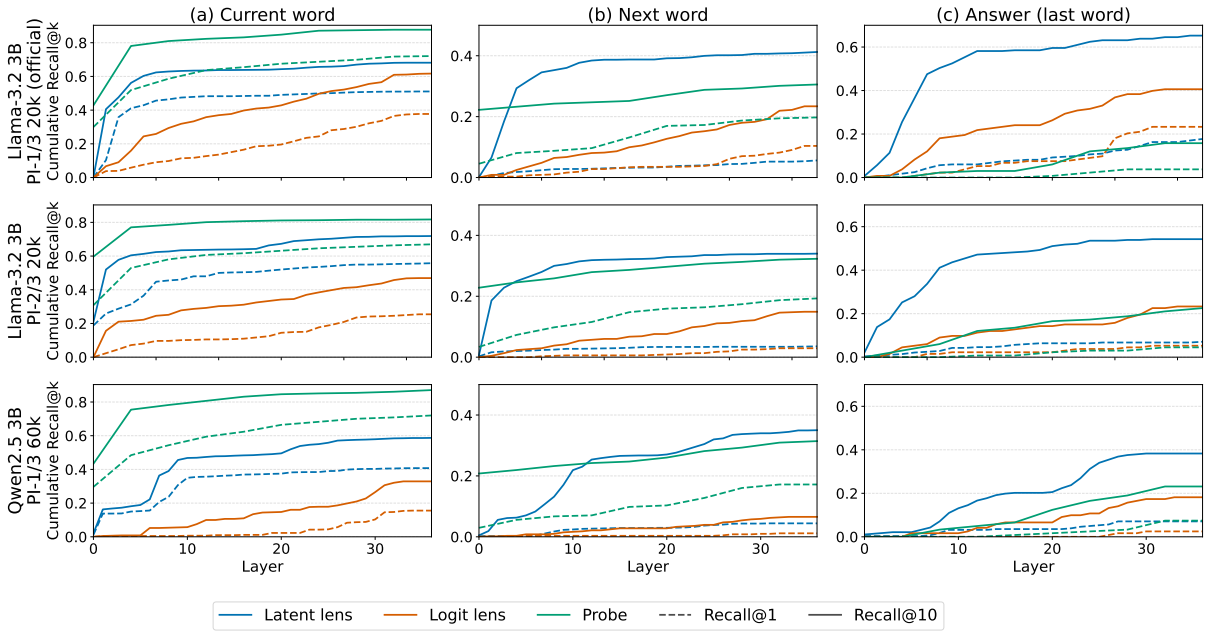



Figure 3: **Implicit transcription and textual continuation are recoverable with complementary decoding methods.** Cumulative $Recall@1$ (dashed) and $Recall@10$ (solid) across layers for the current word, next word, and final answer, using the logit lens, latent lens, and linear probes. Current-word information is recovered most strongly, while next-word and answer information are weaker but remain decodable across models and methods. Full logit lens Recall@k curves and matched-frequency baselines are shown in Figure 4; corresponding results for the other methods are reported in Appendix A.4.2. The baselines are substantially lower and are omitted here for clarity.

Figure 2a shows a clear three-stage pattern for SIMS-Llama3.2-3B. Early layers are dominated by speech tokens (layers 0–2), intermediate layers shift strongly toward text tokens (2–25), and the final layers return to the speech-token domain (26–28). Thus, speech inputs pass through a text-like intermediate regime before being mapped back to speech for generation.

For text inputs, the distribution remains concentrated on text tokens throughout the network, indicating that this transition is specific to speech. Because the text vocabulary is substantially larger, we repeat the analysis using only the top-200 tokens (Figure 2b). The same pattern persists, showing that the intermediate shift toward text is not explained solely by vocabulary size.

## 4.2 Speech LMs Implicitly Transcribe Words

A shift toward the text vocabulary does not by itself establish that speech representations encode the spoken content in text. We therefore test whether the transcription of each spoken word can be recovered from its aligned speech-token hidden states using the logit lens, latent lens, and linear probes.

Figure 3a shows that current-word transcription is strongly recoverable across all three models and decoding methods. Recovery is substantial even at Recall@1 and increases further at Recall@10. For Llama-3.2 3B PI-1/3, cumulative Recall@10 reaches approximately 62% with the logit lens, 69% with the latent lens, and over 80% with the linear probe. The effect also generalizes to Llama-3.2 3B PI-2/3 and Qwen2.5 3B PI-1/3, although with lower overall recall. Thus, the result is neither model-specific nor an artifact of a particular decoder: speech hidden states align with the correct vocabulary token, lie close to contextualized text representations of the same word, and make its identity linearly recoverable.

Recovery accumulates primarily through the intermediate layers, consistent with the text-modality shift observed in Section 4.1. Non-cumulative layer-wise analyses further show that this pattern is not an artifact of cumulative aggregation: across decoding methods, transcription recall is concentrated in the intermediate layers and decreases toward the output (Appendix A.4.3, Figures 10 - 12).

Additional analyses further characterize this transcription signal. Using the logit lens, we find that recovery remains substantial under stricter within-span selectors and is typically localized to only a few speech-token positions (Appendix A.4.1).

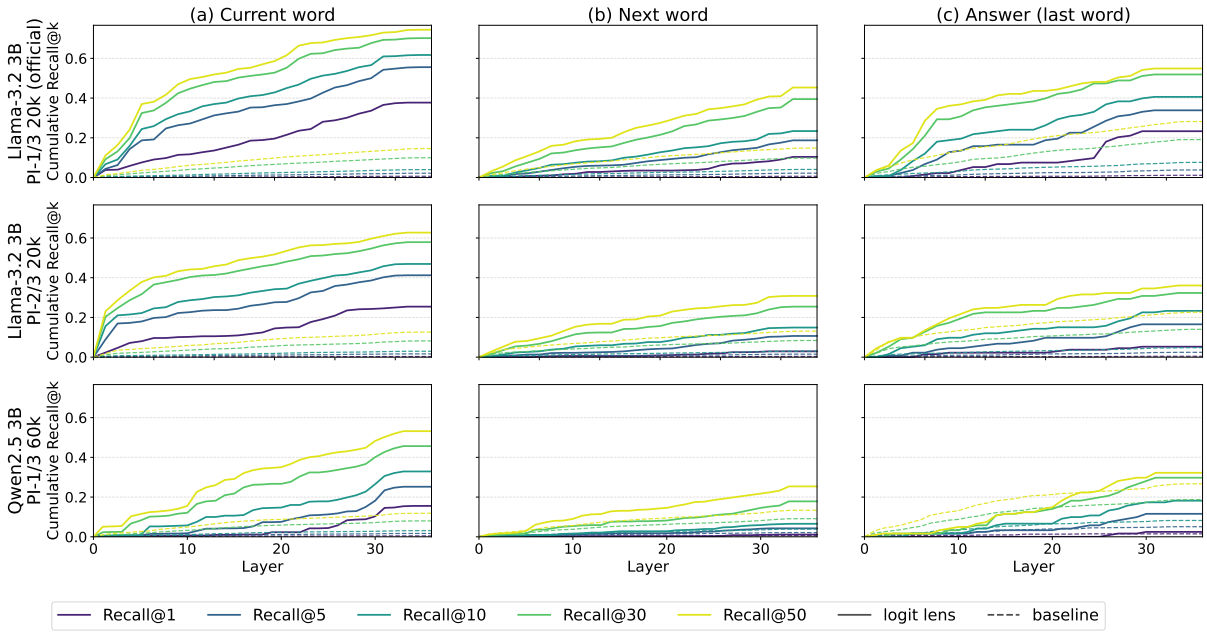



Figure 4: **Full cumulative recovery using the logit lens.** Cumulative Recall@$k$ across layers for the current word, next word, and final answer. Solid curves show recovery of the gold target, while dashed curves show the matched-frequency control. Current-word transcription is strongly recoverable across interleaved models, while next-word and answer information are weaker but remain above the frequency-matched baseline.

Moreover, among words recovered at least once, a substantial subset remains recoverable across multiple consecutive layers (Appendix A.4.3).

Matched-frequency controls are substantially lower for all three decoding methods. Figure 4 shows the full logit lens Recall@k curves, demonstrating that the same pattern holds across different values of $k$. Corresponding full Recall@k curves for the latent lens and linear probes are reported in Appendix A.4.2. Because the models are trained only with next-token prediction and receive no ASR or transcription objective, these results show that their intermediate speech representations implicitly encode the corresponding textual transcription. We refer to this as *implicit latent transcription*.

### 4.3 Models Predict the Continuation in Text

We next ask whether the intermediate text-like representations encode only the current spoken word or also its expected textual continuation. Using the same protocol, we replace the current-word target with either the next word in the sentence or, at the final prompt position, the correct factual answer.

Figure 3b shows that next-word information is recoverable across models and decoding methods, although less strongly than the current transcription. This weaker recovery is partly expected because the next word is not always uniquely determined. For example, after "United," both "Kingdom" and "States" may be plausible continuations. Lower recall may therefore reflect target ambiguity as well as a weaker predictive signal.

We consequently evaluate the final prompt word against the factual answer, which provides a less ambiguous continuation target. As shown in Figure 3c, answer information is recoverable across all three methods, particularly for Llama-3.2 3B PI-1/3 20k (official). Figure 1 illustrates one such example, where "London" becomes decodable although it was not spoken in the prompt. Additional examples are provided in Appendix A.4.4.

Matched-frequency controls are substantially lower for both continuation targets, with the same pattern holding across $k$ values for the logit lens (Figure 4); corresponding results for the other methods are reported in Appendix A.4.2.

Together, these results show that intermediate speech representations encode not only the current transcription, but also likely textual continuations. However, we do not observe a consistent transcription-then-prediction ordering across the dataset, suggesting that the two signals may overlap rather than form a fixed sequential process.

| Model | Data | Logit lens | | Latent lens | | Linear probe | |
|---|---|---|---|---|---|---|---|
| | | Cur | Next | Cur | Next | Cur | Next |
| Llama-3.2 3B PI-1/3 20k (official) | Natural | 71.7 | 27.1 | 40.3 | 20.0 | 89.4 | 57.4 |
| | Controlled | 61.9 | 23.9 | 68.1 | 41.2 | 87.7 | 30.6 |
| Llama-3.2 3B PI-2/3 20k | Natural | 57.7 | 17.4 | 51.7 | 28.2 | 87.7 | 58.5 |
| | Controlled | 49.0 | 15.7 | 71.9 | 34.0 | 81.7 | 32.3 |
| Qwen2.5 3B PI-1/3 60k | Natural | 40.8 | 16.5 | 43.6 | 28.9 | 88.0 | 56.7 |
| | Controlled | 32.9 | 6.5 | 58.7 | 35.0 | 87.1 | 31.4 |

Table 1: **Implicit transcription and textual continuation generalize to natural speech.** Cumulative Recall@10 on natural speech and the controlled common-sense dataset using the logit lens, latent lens, and linear probes. Natural-speech results are aggregated across seven corpora. Cur and Next denote recovery of the current and next words, respectively. Corpus-level results and matched-frequency controls are reported in Appendix A.4.6.

| Model | Text pre-trained | Text | Inter. | Inter. frac. | Cur | Next | Ans |
|---|---|---|---|---|---|---|---|
| SIMS Llama-3.2 PI-1/3 (official) | ✓ | ✓ | ✓ | 1/3 | **61.88** | **23.91** | **41.60** |
| Llama-3.2 PI-1/3 (ours) | ✓ | ✓ | ✓ | 1/3 | **48.75** | **15.31** | **26.40** |
| Llama-3.2 PI-2/3 | ✓ | ✓ | ✓ | 2/3 | **49.06** | **15.78** | 24.00 |
| Llama-3.2 PI-5/6 | ✓ | ✓ | ✓ | 5/6 | 2.66 | 1.72 | 4.00 |
| Llama-3.2 PST | ✓ | ✓ | x | – | 3.75 | 3.28 | 0.00 |
| Llama-3.2 PS | ✓ | x | x | – | 0.16 | 0.00 | 0.00 |
| Llama-3.2 RST | x | ✓ | x | – | 5.78 | 6.72 | 4.80 |
| Llama-3.2 RI-1/3 | x | ✓ | ✓ | 1/3 | 2.19 | 5.78 | 7.20 |

Table 2: Recall@10 for different Speech LMs, i.e the percentage of examples in which the correct current word (Cur), next word (Next), or answer word (Ans) appears among the top-10 predicted tokens across the relevant spoken word in any transformer layer.

### 4.4 Validation on Natural Speech

We test whether implicit transcription generalizes to natural speech using $2,100$ utterances sampled evenly from seven corpora spanning diverse speaking styles and recording conditions.

Table 1 reports cumulative Recall@$k$ alongside results on the controlled common-sense dataset; corpus-level results are provided in Appendix A.4.6. The same qualitative patterns are observed across settings and corpora. Thus, both effects extend beyond synthesized prompts.

### 4.5 What Impacts Implicit Transcription Prevalence?

We next ask whether implicit transcription depends on text-LM initialization, speech–text interleaving, or their combination. We compare Llama-3.2-3B variants spanning pretrained and random initialization, with and without interleaving, and different interleaving fractions.

Because the three decoding methods show consistent qualitative patterns on the representative models, we use the more computationally efficient logit lens for the full model sweep in this section and Section 4.6. We report final-layer cumulative Recall@10: the percentage of targets recovered at any position in the aligned word span and at any layer up to the model output.

As shown in Table 2, for the setting of $k = 10$, implicit transcription is strongest in text-pretrained models trained with interleaved speech-text data. Both the official and our Llama-3.2-PI-1/3 variants show clear transcription signals when interleaved data fraction was at at most $2/3$. In contrast, models trained without interleaved data show substantially weaker transcription, even when text is included in the training mixture. Similarly, randomly initialized models show only weak transcription signals, including when trained with interleaved data. These results suggest that, in this setting, neither exposure to text tokens nor interleaved data alone is sufficient; implicit transcription emerges most clearly from their combination with a pretrained textual prior. Results for different values of $k$ show the same pattern (Tables 6, 7, 8 in Appendix A.4.5).

The fraction of interleaved data also affects the strength of implicit transcription. Models trained with lower or intermediate interleaved-token fractions, such as $1/3$ and $2/3$, show strong implicit transcription, whereas the model trained with a much higher interleaved-token fraction, $5/6$, shows

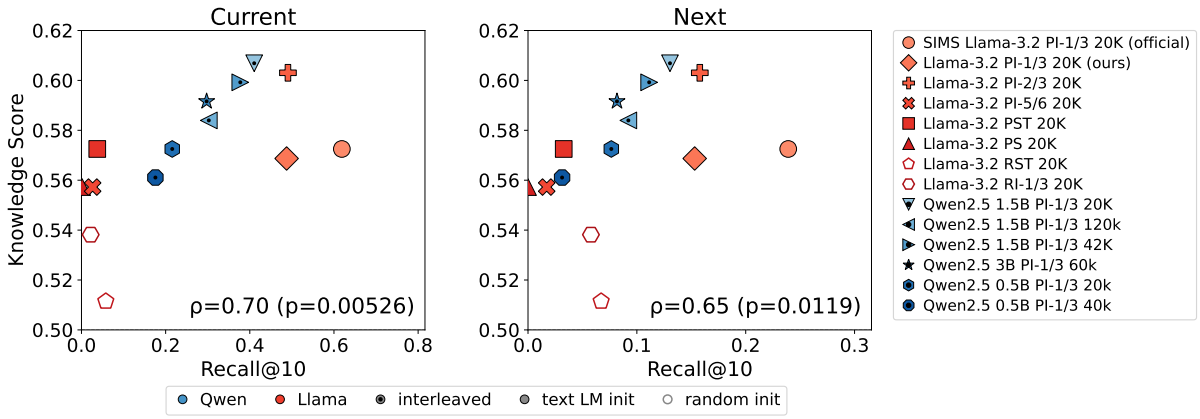



Figure 5: **Implicit transcription correlates with factual knowledge retrieval.** Across model variants, knowledge accuracy increases with current-word (left) and next-word (right) transcription Recall@10, aggregated across aligned speech-token positions and layers.

a substantially weaker signal. This suggests that the effect is not simply monotonic in the amount of interleaved data: excessive interleaving may shift the training distribution or reduce exposure to pure modality-specific contexts.

Overall, these ablations suggest that implicit transcription is not merely a byproduct of using a unified vocabulary or exposing the model to text tokens. Rather, it appears most clearly when a text-pretrained model receives sufficient interleaved speech-text supervision to align spoken units with textual representations.

### 4.6 Does Implicit Transcription Correlate With Factual Knowledge?

We test whether stronger implicit transcription is associated with better factual-knowledge retrieval across the official SIMS models and our variants from Section 4.5. These models span Llama 3.2 and Qwen 2.5 architectures, model sizes, initializations, training budgets, and data mixtures. Following Section 4.5, we use the same final-layer cumulative logit-lens Recall@10 for the cross-model correlation analysis. For each model, we compare current- and next-word recovery with accuracy on our common-sense knowledge benchmark.

We test the monotonic association between transcription recall and factual accuracy using two-sided Spearman rank correlations. Both measures are positively correlated with factual accuracy: current-word recall yields $\rho = 0.70$ ($p = 0.0053$), while next-word recall yields $\rho = 0.65$ ($p = 0.0119$). Both correlations remain significant under the Bonferroni-adjusted threshold of $\alpha = 0.025$.

However, this across-model correlation does not establish sufficiency or causality. Randomly initialized models, for example, show non-zero transcription scores despite low knowledge accuracy, and aggregation may obscure differences across layers and positions. We therefore interpret this as an association rather than a mechanistic explanation.

### 4.7 Qualitative Analysis

As part of a qualitative analysis of implicit transcription, we observed recurring phenomena that may provide broader context for this behavior. We present these findings as exploratory observations, and leave a broader and more systematic investigation of their scope and implications to future work. For this analysis, we also consider single word utterances, synthesized using the same approach as our evaluation data, to provide a simpler setting for interpretation.

**Implicit Transcription Builds Gradually.** Implicit transcription sometimes appears to emerge incrementally over the course of a spoken word. For example, the logit lens first decodes *white* as "why" (Figure 14), *lime* as "lie" (Figure 13), *kingdom* as "king" (Figure 17), and *Pakistan* (Figure 15) as "pack", before later converging to the full word. These examples suggest that the model may construct textual representations from partial acoustic evidence, updating them as more of the word becomes available. More examples can be found in Appendix A.4.4. As a result, intermediate representations can temporarily correspond to acoustically plausible but semantically unrelated words. Such transient mismatches may contribute to the broader speech - text representation gap, and we leave a systematic investigation of this effect to future work.

**Implicit Transcription Has Errors.** We also observe incorrect but acoustically similar implicit transcriptions. For example, *lime* is decoded as "line" (Figure 13, Appendix A.4.4). Unlike the gradual prefix effect above, these cases suggest that the model occasionally settles on a text interpretation that is acoustically close to the input but is incorrect. Such errors highlight that the implicit transcription process may introduce ambiguity or noise into the speech-derived text representation.

## 5 Related Work

**Interpreting Latent Computation in LMs.** Several works suggest that multilingual LLMs share information across languages through English-centric latent representations; Wendler et al. (2024) show this in Llama-2, and related English-mediated processing or steering effects are reported by Zhao et al. (2024); Mahmoud et al. (2025). Related multi-modal work proposes a similar semantic-hub view, where representations from different modalities align in intermediate layers (Wu et al., 2025). Closest to our analysis, Neo et al. (2025) apply logit lens to visual-token hidden states in LLaVA, showing that image representations can become decodable in the language vocabulary despite not corresponding to discrete text tokens. Complementarily, Nikankin et al. (2026) further show that vision and text may rely on distinct circuits, with alignment emerging only in late layers. Together, these works suggest that cross-modal reasoning requires non-textual inputs to enter a shared workspace early enough. However, they focus on multilingual text or vision-language models, whereas we study whether such a text workspace emerges during speech processing.

**Analyzing Speech-Text LMs.** Several studies have tried to interpret speech-text LMs by analyzing the sources of the modality gap: the mismatch between speech and equivalent text inputs. Recent work suggests that this gap is partly caused by the different structure of the two modalities; for example, redundant speech tokenization can lead to diffuse attention and weaker decision sharpening compared to text (Hsu et al., 2026). In SpiRit-LM (Nguyen et al., 2025), layer-wise analyses are used to explain why interleaving speech and text is beneficial: the two modalities become increasingly geometrically aligned across layers, suggesting that interleaving helps induce a shared latent structure that may reduce the speech-text gap. Related studies further measure speech-text alignment across layers (Mousavi et al., 2025b; Xiang et al., 2025). Other work finds that audio LMs often rely on transcript-like information, behaving similarly to ASR→LLM cascades on text-sufficient tasks (Chen et al., 2025; Billa, 2026). Recent mechanistic studies of ASR models further analyze how transcript tokens emerge across layers, using decoder logit-lens analyses, probing, and activation patching to trace acoustic and semantic information during transcription (Lioubashevski et al., 2024; Glazer et al., 2025). Nevertheless, these works focus on audio-to-text models or static speech-text alignment. In contrast, we study a generative speech-text LM that both consumes and produces speech tokens, and ask whether speech processing internally passes through a textual workspace despite the model not being explicitly trained for speech-to-text transcription.

## 6 Discussion & Conclusion

Our work, clearly indicates that Speech LMs implicitly move to a text latent space in which the corresponding text transcription and hypotheses about the next word become decodable before being projected back to speech tokens. We observe this behavior consistently across three complementary decoding methods and show that it generalizes from controlled synthesized prompts to natural speech. We also demonstrate that this positively correlates with spoken fact knowledge abilities.

This work immediately opens up future research question, most excitingly can this behavior be optimized directly leading to overall better performance. However, acting on these insights explicitly could also have negative sides. Specifically, we left for future work the analysis of the effect of implicit transcription on acoustic abilities. As acoustic abilities are a key motivation for modeling speech directly and avoiding *explicit* transcription, this poses a key question.

Another interesting question that arises from this study is "what limits the spoken abilities relative to text and leads to a modality gap?". Given that SLMs latently work in text, one could expect a small modality gap and yet this is not the case. Future work could further analyze if this has to do with transcription error, compute "wasted" on transcription, or lack of temporal compression making language reasoning more challenging.

## Limitations

Our work shows that implicit transcription emerges in intermediate layers, but does not identify the precise mechanism that computes it, such as the specific heads, layers, or pathways involved. More generally, our decoding analyses establish that information is accessible in intermediate representations, but not that it is causally used by the model. The logit lens is limited to information aligned with the model's output projection, while linear probes demonstrate linear decodability through a learned classifier but do not establish that the model itself uses this information during generation. Although agreement across complementary methods strengthens the evidence for a text-like intermediate representation, it does not establish its causal role. Moreover, although transcription-like signals are positively associated with factual knowledge retrieval from speech, the correlation reaches only up to $\rho = 0.70$ and is not sufficient to explain all variation across models. We do not causally test this relationship by intervening on models to increase implicit transcription and measuring whether knowledge retrieval improves, and leave this analysis for future work.

**Acknowledgments.** This research work was supported by Israel Science Foundation (ISF), grant number $2049/22$.

## References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2026. On the landscape of spoken language models: A comprehensive survey. *Preprint*, arXiv:2504.08528.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Jayadev Billa. 2026. The cascade equivalence hypothesis: When do speech llms behave like asr→llm pipelines? *Preprint*, arXiv:2602.17598.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2025. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *Preprint*, arXiv:2510.10444.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Santiago Cuervo and Ricard Marxer. 2024. Scaling properties of speech language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *Preprint*, arXiv:2111.09344.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45.

Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya, Joseph Keshet, and Aviv Navon. 2025. Beyond transcription: Mechanistic interpretability in asr. *Preprint*, arXiv:2508.15882.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations*, volume 2024, pages 42749–42787.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. *Preprint*, arXiv:1511.03690.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, and 1 others. 2023. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36:63483–63501.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus*

*Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing.

Hexgrad. 2025. Kokoro-82m (revision d8b4fc7).

Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. 2026. Anatomy of the modality gap: Dissecting the internal states of end-to-end speech llms. *Preprint*, arXiv:2603.01502.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 7669–7673. IEEE.

Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the spoken wikipedia for speech data and beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. 2026. Latentlens: Revealing highly interpretable visual tokens in llms. *Preprint*, arXiv:2602.00462.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Daria Lioubashevski, Tomer Schlank, Gabriel Stanovsky, and Ariel Goldstein. 2024. Looking beyond the top-1: Transformers determine top tokens in order. *Preprint*, arXiv:2410.20210.

Omar Mahmoud, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. 2025. Improving multilingual language models by aligning representations through steering. *Preprint*, arXiv:2505.12584.

Gallil Maimon, Avishai Elmakies, and Yossi Adi. 2025a. Slamming: Training a speech language model on one GPU in a day. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12201–12216, Vienna, Austria. Association for Computational Linguistics.

Gallil Maimon, Michael Hassid, Amit Roth, and Yossi Adi. 2025b. Scaling analysis of interleaved speech-text language models. In *Second Conference on Language Modeling*.

Gallil Maimon, Amit Roth, and Yossi Adi. 2025c. Salmon: A suite for acoustic language model evaluation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Potsawee Manakul, Woody Haosheng Gan, Martijn Bartelds, Guangzhi Sun, William Held, and Diyi Yang. 2026. Scaling open discrete audio foundation models with interleaved semantic, acoustic, and text tokens. *arXiv preprint arXiv:2602.16687*.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, Bhuvana Ramabhadran, Benjamin Elizalde, Loren Lugosch, Jinyu Li, Cem Subakan, Phil Woodland, Minje Kim, Hung yi Lee, Shinji Watanabe, and 2 others. 2025a. Discrete audio tokens: More than a survey! *Transactions on Machine Learning Research*.

Pooneh Mousavi, Yingzhi Wang, Mirco Ravanelli, and Cem Subakan. 2025b. Alas: Measuring latent speech-text alignment for spoken language understanding in multimodal llms. *Preprint*, arXiv:2505.19937.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michele Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. In *International Conference on Learning Representations*, volume 2024, pages 51883–51898.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 57172–57189.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2026. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

nostalgebraist. 2020. Interpreting gpt: The logit lens. https://www.

lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong blog post.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, pages 28492–28518.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pages 84929–84964.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Preprint*, arXiv:2101.00390.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *Preprint*, arXiv:2411.04986.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. *Preprint*, arXiv:2510.12116.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

Iddo Yosha, Gallil Maimon, and Yossi Adi. 2026. Stresstest: Can your speech lm handle the stress? *Preprint*, arXiv:2505.22765.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Yuxiao Dong, Jie Tang, and 1 others. 2025. Scaling speech-text pre-training with synthetic interleaved data. In *International Conference on Learning Representations*, volume 2025, pages 49396–49419.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Preprint*, arXiv:2402.18815.

# A Appendix

## A.1 Common Sense Dataset

Table 5 reports statistics for the different subsets of our evaluation dataset, along with one representative example from each subset.

## A.2 Matched-Frequency Baseline

For each gold word, we sample 20 control words from the same frequency band in a fixed RedPajama word-frequency table based on the Llama-3.2 tokenizer. The same reference is used for all models. Controls are restricted to words in the latent-lens bank and exclude the gold word and all evaluation words. If a band contains too few candidates, we expand the pool to adjacent bands and then to all alphabetic bank words.

Controls are scored using the same span-level and cumulative Recall@$k$ procedure as the gold words. We report the mean recall across the 20 samples with 95% confidence intervals.

## A.3 Experimental Setup

All Llama-3.2 variants are trained with the same architecture and optimization setup, enabling controlled comparison across initialization and data composition. Matching the original SIMS paper, all models are trained with a next-token prediction objective, sequence length 2048, an effective batch size of 32 samples, and 20K training steps. Training is performed with SLAMKIT (Maimon et al., 2025a).

**Interleaving.** We follow Zeng et al. (2025), sampling speech-segment lengths from a Poisson distribution with $\lambda = 10$ until speech spans cover $\eta = 0.3$ of the words. Time-aligned transcripts are used only to identify word boundaries and assign each span to either speech or text. Thus, each word appears in only one modality within an interleaved sequence, rather than as a parallel speech–transcript pair. The models are trained solely with next-token prediction, without an ASR loss or direct speech-to-text alignment objective.

**Training Data** For speech training data, we follow Maimon et al. (2025b) and use the same English speech mixture: LibriSpeech (Panayotov et al., 2015), LibriLight (Kahn et al., 2020), VoxPopuli (Wang et al., 2021), TED-LIUM (Hernandez et al., 2018), People's Speech (Galvez et al., 2021), SWC (Köhn et al., 2016), and synthetic sTinyStories (Maimon et al., 2025a). Text-only data is taken from RedPajama (Weber et al., 2024), filtered with Gopher rules (Rae et al., 2021). Speech is represented with discrete HuBERT-style units (Hsu et al., 2021): mHuBERT features are quantized using a 500-unit k-means codebook. Text is tokenized with the Llama 3.2 tokenizer. In interleaved configurations, speech and text spans are mixed using Whisper large-v3-turbo alignments (Radford et al., 2023).

## A.4 Additional Results

### A.4.1 Span Aggregation and Localization Robustness

A spoken word corresponds to a variable-length span of speech tokens, and the position at which its transcription becomes decodable may differ across words. Our main analysis therefore uses any-position aggregation: a word is considered recovered if its target text token appears among the top-$k$ predictions at any position within its aligned speech span. Here, we test whether our findings depend on this choice and quantify how localized the transcription signal is within each span.

Let $S(w) = (i_1, \ldots, i_m)$ denote the ordered speech-token positions aligned with word $w$, and let $p_i^{(\ell)}$ denote the logit-lens distribution at position $i$ and layer $\ell$. We evaluate the following span selectors:

- **Max over position:** the word is recovered if its target token appears among the top-$k$ predictions at any position in $S(w)$.
- **Final 25%:** recovery is evaluated over only the final $\lceil m/4 \rceil$ positions in the span.
- **Final token:** recovery is evaluated only at the final position $i_m$.
- **First token:** recovery is evaluated only at the first position $i_1$.
- **Mean positional recall:** we compute the fraction of positions in the span at which the target token appears among the top-$k$ predictions. For example, if the target appears at two of five positions, the score is $2/5$, whereas any-position aggregation assigns a score of one.

These variants differ only in how recovery is aggregated across positions within a word span.

As shown in Table 3, max over position aggregation yields the highest recall, but the effect remains substantial under stricter selectors. For the official

| Span selector | Llama PI-1/3 | Llama PI-2/3 | Qwen PI-1/3 |
|---|---|---|---|
| Max over positions | 61.7 | 46.9 | 32.9 |
| Final 25% | 38.6 | 27.8 | 14.7 |
| Final token | 33.9 | 24.6 | 9.2 |
| First token | 9.2 | 5.8 | 1.8 |
| Mean over positions | 22.9 | 16.6 | 12.1 |

Table 3: Cumulative Recall@10 under alternative within-span selectors. Although any-position aggregation is the most sensitive, implicit transcription remains detectable when the analysis is restricted to the final quarter or final token of each word span.

Llama PI-1/3 model, Recall@10 decreases from 61.7% using any position to 38.6% using only the final quarter of the span and 33.9% using only the final token. The low recall at the first token is consistent with the model requiring additional acoustic context before the word becomes decodable, although imperfect boundaries and brief silences may also contribute. Mean positional Recall@10 for the official Llama PI-1/3 model is 22.9%. The gap between the two metrics indicates that transcription is often decodable somewhere within a word span, but is not uniformly present across all of its positions.

To quantify this localization directly, we define the number of successful positions for word $w$ at layer $\ell$ as

$$n_{\text{hit}}(w, \ell, k) = \sum_{i \in S(w)} \mathbf{1}\left[y_w \in \text{TopK}\left(p_i^{(\ell)}, k\right)\right]$$

where $S(w)$ is the set of speech-token positions aligned with word $w$, $p_i^{(\ell)}$ is the logit-lens distribution at position $i$ and layer $\ell$, and $y_w$ is the target text token.

We evaluate this statistic over word–layer pairs containing at least one top-$k$ hit. At $k = 10$, the target token appears in at most two positions in 82.2% of successful word–layer pairs, while the median aligned word span contains 6 speech tokens. Thus, the transcription signal is typically concentrated in a small subset of the aligned span rather than distributed uniformly across it.

Overall, max-over-position aggregation increases sensitivity to localized signals whose position varies across words, but does not create the implicit-transcription effect. The signal remains clearly detectable under stricter fixed-position selectors, while the mean positional and localization analyses confirm that it is generally concentrated at only a few positions within each speech span.

| | ≥5 layers | ≥8 layers | ≥10 layers | ≥15 layers |
|---|---|---|---|---|
| Recall@1 | 24.8% | 12.4% | 7.4% | 5.4% |
| Recall@5 | 31.8% | 18.7% | 12.1% | 7.9% |
| Recall@10 | 37.7% | 24.4% | 17.3% | 11.1% |

Table 4: Persistence of implicit transcription in Llama-3.2 3B PI-1/3 20k (official). Values report the fraction of word instances recovered at least once that remain recoverable for at least the specified number of layers.

### A.4.2 Cumulative Recall Results Over Models and Decoding Methods

Figure 4 reports the full cumulative Recall@$k$ curves for the logit lens in the main text, while Figures 7 and 11 report the corresponding results for the latent lens and linear probes, respectively, for $k \in \{1, 5, 10, 30, 50\}$. Across methods and models, the same overall pattern persists across values of $k$. Each figure also reports matched-frequency controls, evaluated with the same span-level and cumulative protocol. These baselines remain substantially below gold-word recovery across methods and targets, showing that the effect is not explained by lexical frequency alone.

### A.4.3 Layer-Wise Recall and Persistence

Our main analysis reports cumulative recall, which measures whether a word has become decodable by a given model depth. To identify the layers at which the transcription signal is actively present, we additionally compute non-cumulative, layer-wise recall.

Figures 10 - 12 present layer-wise recall. Across models, recall generally increases through the intermediate layers, peaks in the late-intermediate layers, and drops sharply near the output. Thus, the transcription signal is not uniformly distributed across model depth, but is most strongly decodable from intermediate hidden states.

We next test whether recovery is restricted to isolated layers or persists across model depth. For each word recovered at least once, we identify its first recovery layer and measure the number of consecutive layers for which the target remains among the top-$k$ predictions at some position in the aligned speech span.

The persistence results are shown in Table 4. At $k = 10$, 37.7% of the recovered word instances remain decodable for at least 5 layers, 24.4% for at least 8 layers, and 17.3% for at least 10 layers. Even under the stricter Recall@1 criterion, 24.8% of recovered words persist for at least 5 layers.

These results show that implicit transcription is not limited to isolated layer-wise hits. Although

persistence varies across words, a substantial subset remains decodable across several consecutive intermediate layers, supporting the interpretation of an extended intermediate transcription regime rather than a purely transient effect.

#### A.4.4 Transcription Examples

Figures 13-28 show logit lens analyses across different models and spoken inputs. Overall, the trends are similar to those presented in the main paper: several models exhibit clear implicit transcription and next word prediction. However, the strength and extent of this phenomenon vary across models.

Figures 13-16 show examples of gradual transcription and possible transcription errors for single-word inputs. Figures 17-25 show examples of transcription for different inputs across different models, while Figure 26 shows a case in which transcription does not occur at all, for a randomly initialized model. This is consistent with the low transcription capacity observed for this model in Table 2 and Section 4. Natural Speech implicit transcription examples can be found in figures Figure 27 - 28, and Figure 29 contains an example for implicit transcription given by Latent Lens. In all figures, cells are colored by textual-token probability, dark blue indicates high probability and light yellow is zero textual probability. Transcriptions are written for textual token according to the most probable token given by the decoding method.

#### A.4.5 Transcription Recall For Different Top Ks Across Different Models

We analyze $\mathrm{Recall@}k$ for the Llama3.2-3B variants with $k \in \{1, 5, 30\}$. The results show consistent trends across different values of $k$ and are reported in Tables 6, 7, and 8. The cumulative recall patterns across all values of $k$ and all models are shown in Figure 9.

#### A.4.6 Natural-Speech Results by Corpus

We evaluate 300 utterances from each of seven natural-speech corpora: People's Speech, LibriSpeech, VoxPopuli, TED-LIUM, the Spoken Wikipedia Corpus (SWC), VCTK, and Flickr8k (Harwath and Glass, 2015). These corpora span different speakers, recording conditions, and speaking styles.

Tables 9–10 report corpus-level cumulative Recall for the logit lens, latent lens, and linear probes, respectively. Each entry reports gold-word recovery, with the matched-frequency baseline in parentheses. Across corpora, methods, and models, current-word recovery is consistently stronger than next-word recovery, and both generally remain substantially above their frequency-matched controls. Absolute performance varies across corpora, indicating sensitivity to their acoustic and linguistic properties, but the transcription signal is not restricted to a particular speech domain.

Natural-speech implicit transcription examples are shown in Figures 27–28.

### A.5 Model Evaluation

For reference, and as a proxy for the general semantic abilities of the trained models we also report standard likelihood speech metrics in Figure 6. We note that indeed all models perform as expected with models randomly initialized and without interleaving under-performing those with. We also see that our version of SIMS performs comparably to the official version once again showing that our training setup works as expected.

### A.6 AI Tools Usage

AI tools have been used to assist in fixing grammar mistakes and sentence paraphrasing. Additionally, AI tools have been partially used to enhance code implementations. However, the authors carefully reviewed all content, ensuring these tools were only used as supportive aids and in a responsible manner.

| Category | # Examples | Example |
|---|---|---|
| Colors | 16 | *The color of grass is → green* |
| Days of the week | 7 | *The day that comes after Sunday is → Monday* |
| Months of the year | 14 | *The month that comes after January is → February* |
| Object functions | 21 | *People use eyes to → see* |
| Common-sense facts | 13 | *Water freezes into → ice* |
| Country languages | 10 | *The official language of France is → French* |
| Family relations | 18 | *The mother of mother is called → grandmother* |
| Numerical facts | 27 | *The number of seconds in a minute is → 60* |
| Opposites | 35 | *the opposite of hot is → cold* |
| Baby animals and professions | 10 | *a baby dog is called a → puppy* |
| Capital cities | 30 | *The capital of France is → Paris* |
| Simple arithmetic | 58 | *one plus one equals → 2* |
| Number sequences | 23 | *the number after one is → two* |
| **Total** | **282** | – |

Table 5: Statistics and representative examples from the factual-completion dataset. Each example consists of a short prompt with the final answer omitted and a single target completion.

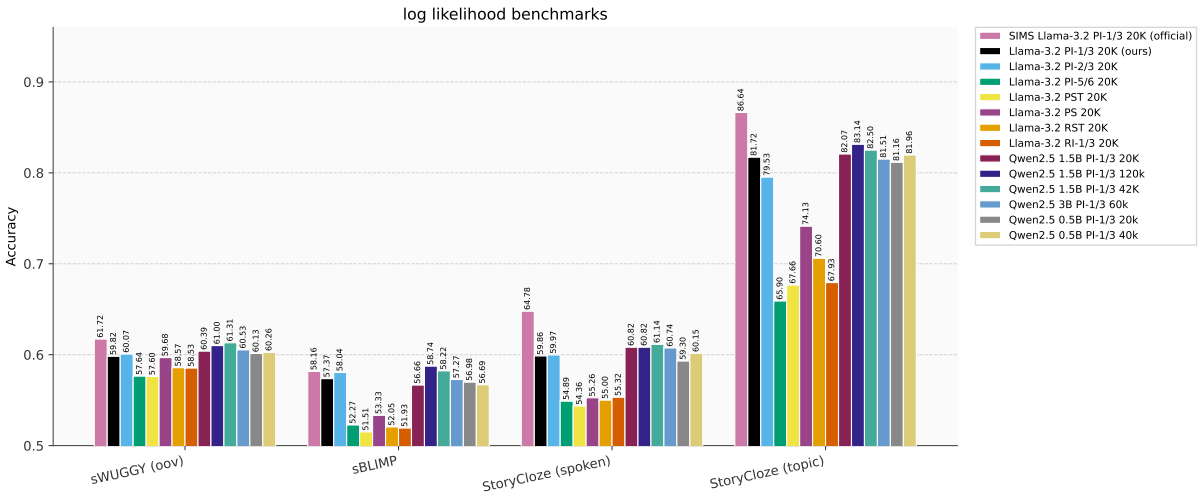



Figure 6: Log Likelihood based evaluations for all models we used

| Model | Text pre-trained | Text | Inter. | Inter. frac. | Cur | Next | Ans |
|---|---|---|---|---|---|---|---|
| SIMS Llama-3.2 PI-1/3 (official) | ✓ | ✓ | ✓ | 1/3 | **37.19** | **10.31** | **23.20** |
| Llama-3.2 PI-1/3 (ours) | ✓ | ✓ | ✓ | 1/3 | **32.66** | **4.84** | **12.80** |
| Llama-3.2 PI-2/3 | ✓ | ✓ | ✓ | 2/3 | **27.34** | **3.59** | **5.60** |
| Llama-3.2 PI-5/6 | ✓ | ✓ | ✓ | 5/6 | 0.00 | 0.00 | 0.00 |
| Llama-3.2 PST | ✓ | ✓ | x | – | 0.16 | 0.31 | 0.00 |
| Llama-3.2 PS | ✓ | x | x | – | 0.00 | 0.00 | 0.00 |
| Llama-3.2 RST | x | ✓ | x | – | 0.00 | 0.00 | 0.00 |
| Llama-3.2 RI-1/3 | x | ✓ | ✓ | 1/3 | 0.00 | 0.00 | 0.00 |

Table 6: Recall@1 for different Speech LMs, i.e., the percentage of examples in which the correct current word (Cur), next word (Next), or answer word (Ans) appears as the top predicted token across the relevant spoken word in any transformer layer.

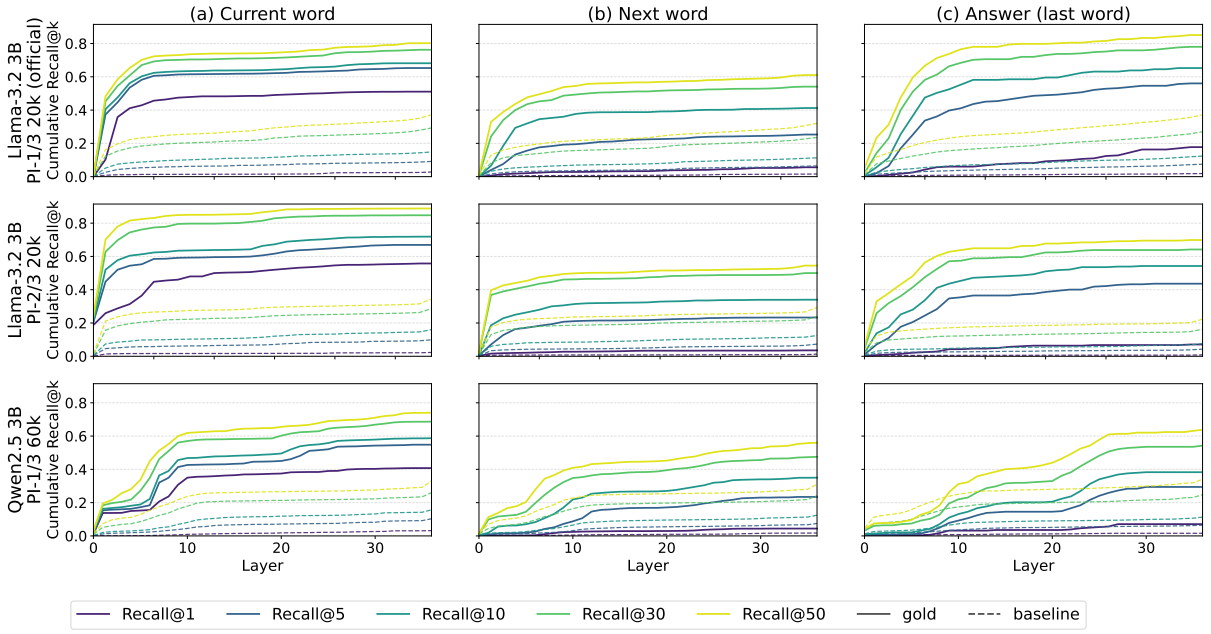



Figure 7: **Full cumulative recovery using the latent lens.** Cumulative Recall@$k$ across layers for the current word, next word, and final answer, obtained by retrieving the nearest contextualized text representations. Solid curves show recovery of the gold target, while dashed curves show the matched-frequency control. Speech hidden states are strongly aligned with the corresponding textual representations, particularly for the current spoken word.

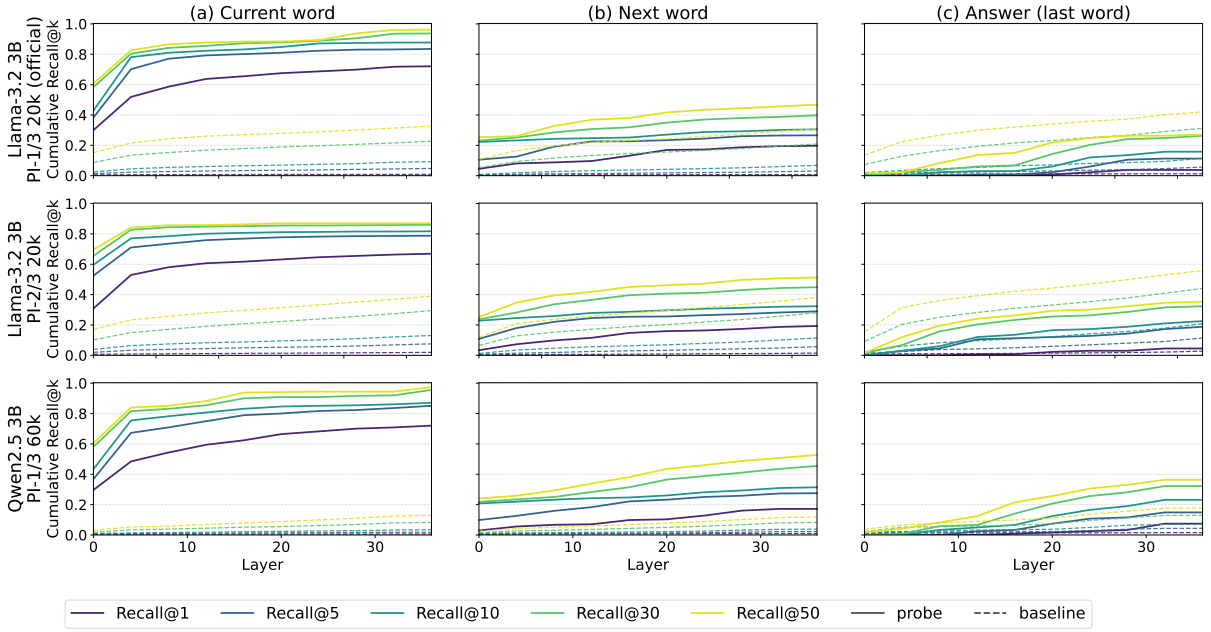



Figure 8: **Full cumulative recovery using linear probes.** Cumulative Recall@$k$ across layers for the current word, next word, and final answer. Separate supervised linear probes are trained for each model, layer, and prediction target while the Speech LM remains frozen. Solid curves show recovery of the gold target, while dashed curves show the matched-frequency control. The high recall demonstrates that lexical and continuation information is strongly linearly recoverable from intermediate speech hidden states.

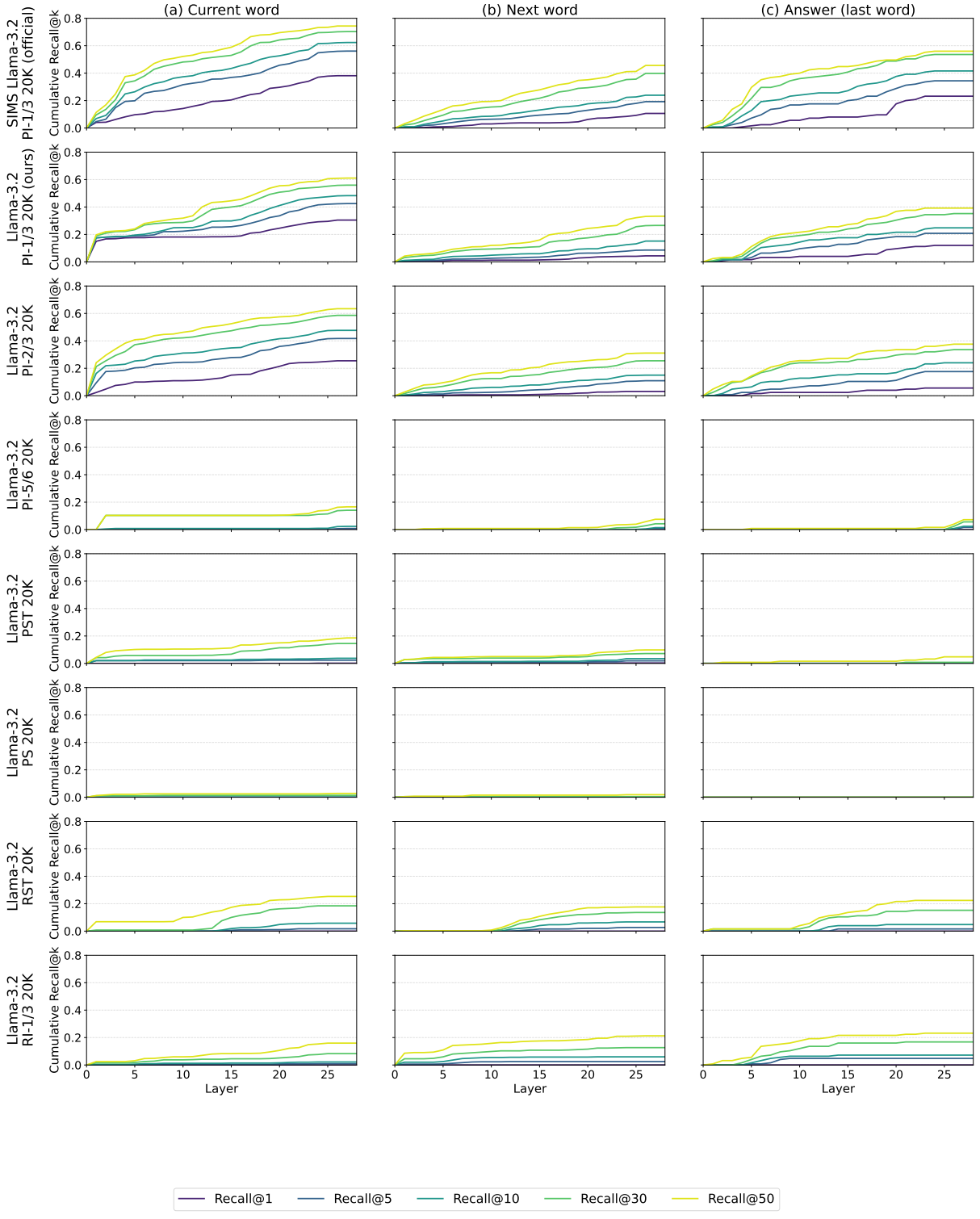



Figure 9: **Implicit transcription and textual continuation emerge in speech hidden states for text pre-trained with interleaving.** We apply the logit lens to speech-token hidden states and report Recall@k up to a given layer, for the current transcription word, the next word, and the final answer. Although the models are not explicitly trained for transcription, current-word transcription emerges reliably in intermediate layers across models, while next-word and answer-word predictions are weaker but still decodable.

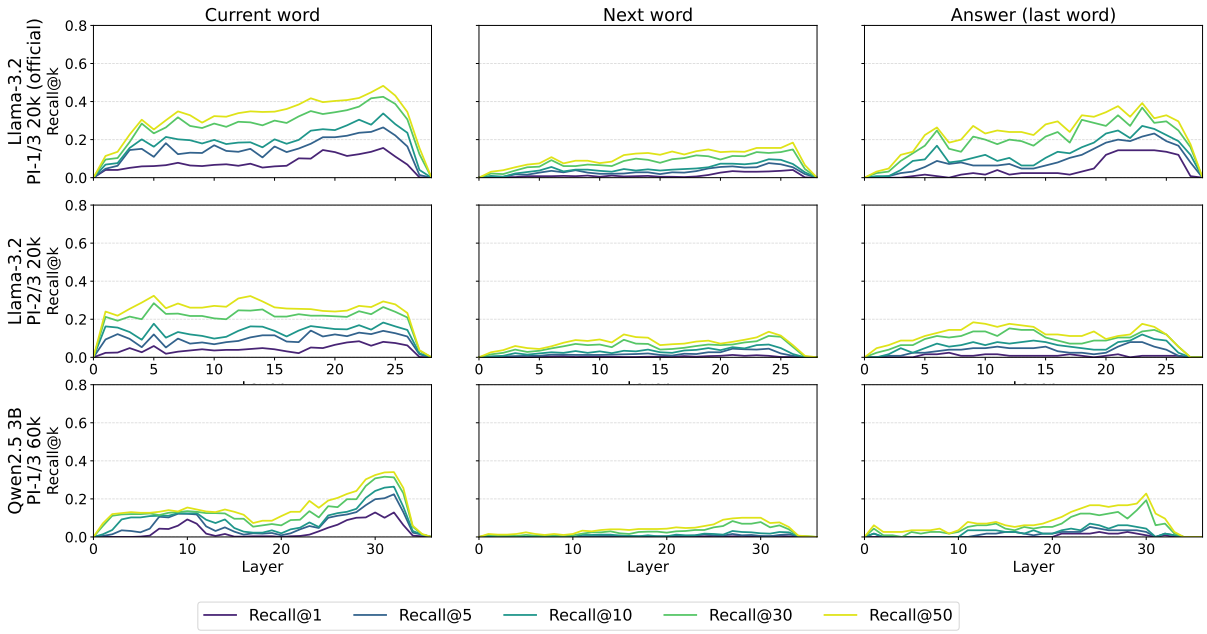



Figure 10: **Layer-wise recovery using the logit lens.** Layer-wise Recall@(k), computed using the logit lens on speech-token hidden states, is shown for the current transcription word, the next word, and the final answer. Although the models are not explicitly trained for transcription, the current word is reliably recoverable from intermediate layers across models, while next-word and answer-word predictions are weaker but remain decodable. Transcription emerges primarily in the middle layers.

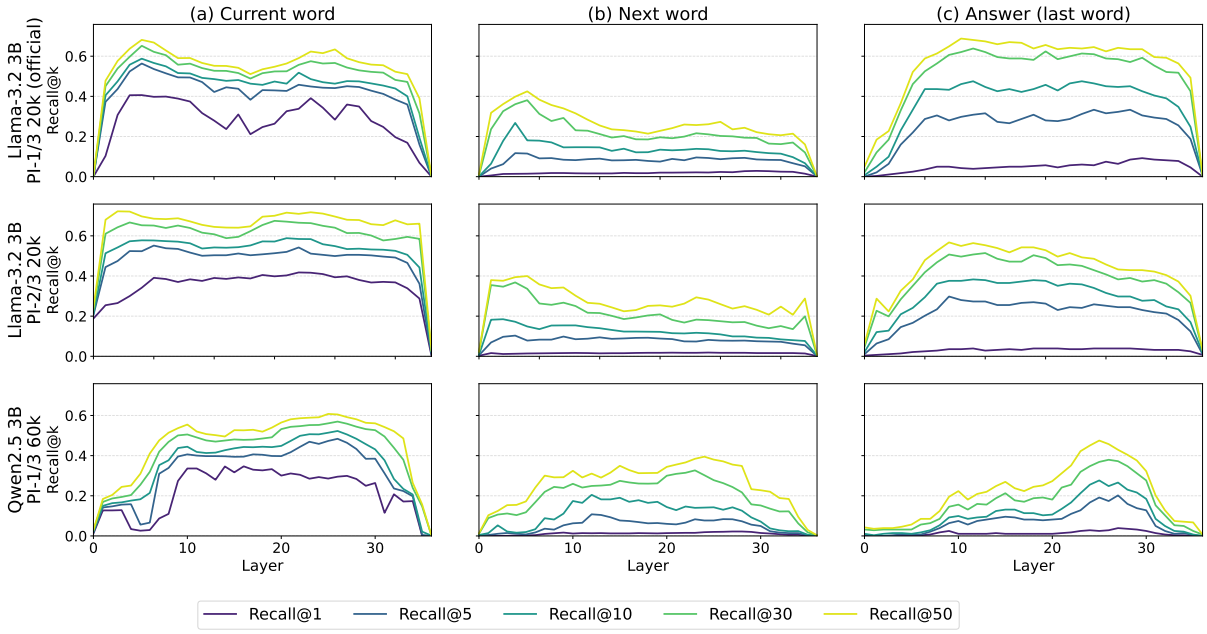



Figure 11: **Layer-wise recovery using the latent lens.** Layer-wise Recall@(k), computed using the latent lens on speech-token hidden states, is shown for the current transcription word, the next word, and the final answer. Although the models are not explicitly trained for transcription, the current word is reliably recoverable from intermediate layers across models, while next-word and answer-word predictions are weaker but remain decodable. Transcription emerges primarily in the middle layers.

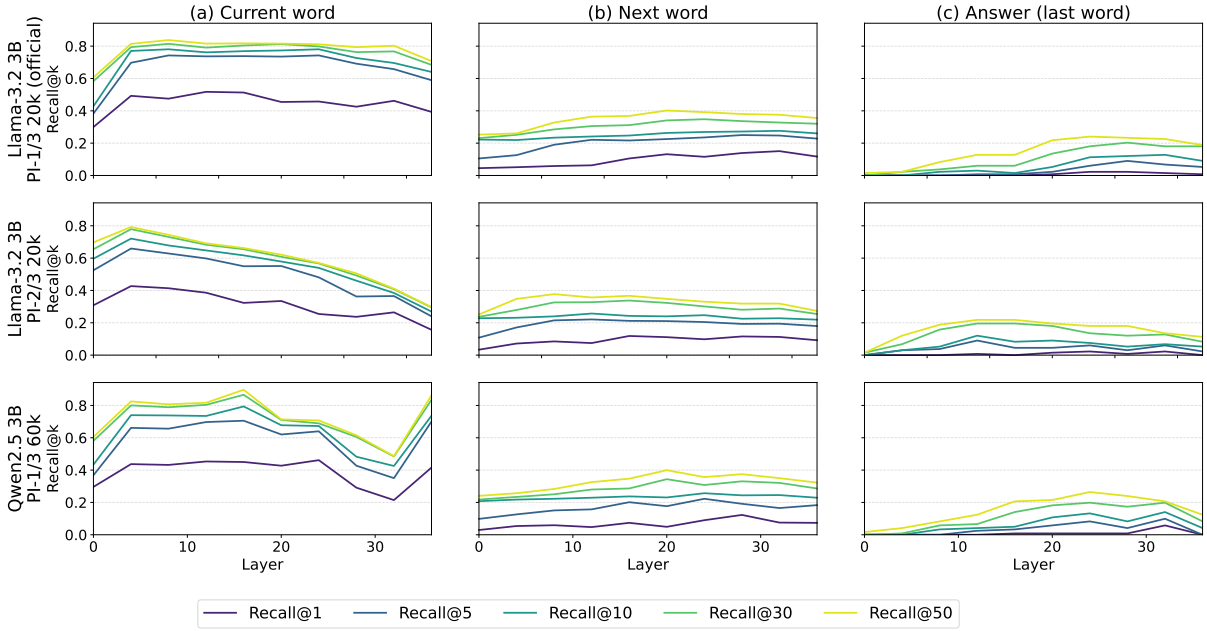



Figure 12: **Layer-wise recovery using the linear probe.** Layer-wise Recall@(k), computed using the linear probe on speech-token hidden states, is shown for the current transcription word, the next word, and the final answer. Although the models are not explicitly trained for transcription, the current word is reliably recoverable from intermediate layers across models, while next-word and answer-word predictions are weaker but remain decodable. Transcription emerges primarily in the middle layers.

| Model | Text pre-trained | Text | Inter. | Inter. frac. | Cur | Next | Ans |
|---|---|---|---|---|---|---|---|
| SIMS Llama-3.2 PI-1/3 (official) | ✓ | ✓ | ✓ | 1/3 | **55.16** | **19.22** | **34.40** |
| Llama-3.2 PI-1/3 (ours) | ✓ | ✓ | ✓ | 1/3 | **42.81** | **8.91** | **21.60** |
| Llama-3.2 PI-2/3 | ✓ | ✓ | ✓ | 2/3 | **43.75** | **11.87** | 18.40 |
| Llama-3.2 PI-5/6 | ✓ | ✓ | ✓ | 5/6 | 1.09 | 1.09 | 3.20 |
| Llama-3.2 PST | ✓ | ✓ | x | – | 2.34 | 1.87 | 0.00 |
| Llama-3.2 PS | ✓ | x | x | – | 0.00 | 0.00 | 0.00 |
| Llama-3.2 RST | x | ✓ | x | – | 1.72 | 2.66 | 1.60 |
| Llama-3.2 RI-1/3 | x | ✓ | ✓ | 1/3 | 0.94 | 2.34 | 4.80 |

Table 7: Recall@5 for different Speech LMs, i.e., the percentage of examples in which the correct current word (Cur), next word (Next), or answer word (Ans) appears among the top-5 predicted tokens across the relevant spoken word in any transformer layer, using the logit lens.

| Model | Text pre-trained | Text | Inter. | Inter. frac. | Cur | Next | Ans |
|---|---|---|---|---|---|---|---|
| SIMS Llama-3.2 PI-1/3 (official) | ✓ | ✓ | ✓ | 1/3 | **73.75** | **45.78** | **56.80** |
| Llama-3.2 PI-1/3 (ours) | ✓ | ✓ | ✓ | 1/3 | **61.56** | **33.59** | **40.00** |
| Llama-3.2 PI-2/3 | ✓ | ✓ | ✓ | 2/3 | **65.00** | **32.34** | 38.40 |
| Llama-3.2 PI-5/6 | ✓ | ✓ | ✓ | 5/6 | 17.03 | 7.50 | 8.80 |
| Llama-3.2 PST | ✓ | ✓ | x | – | 18.91 | 9.38 | 4.80 |
| Llama-3.2 PS | ✓ | x | x | – | 2.81 | 1.87 | 0.00 |
| Llama-3.2 RST | x | ✓ | x | – | 25.31 | 17.50 | 22.40 |
| Llama-3.2 RI-1/3 | x | ✓ | ✓ | 1/3 | 15.62 | 21.41 | 24.00 |

Table 8: Recall@30 for different Speech LMs, i.e., the percentage of examples in which the correct current word (Cur), next word (Next), or answer word (Ans) appears among the top-30 predicted tokens across the relevant spoken word in any transformer layer, using the logit lens.

| Dataset | Current word | | | Next word | | |
|---|---|---|---|---|---|---|
| | Llama PI-1/3 | Llama PI-2/3 | Qwen PI-1/3 | Llama PI-1/3 | Llama PI-2/3 | Qwen PI-1/3 |
| Controlled | 68.1 (14.8) | 71.9 (16.0) | 58.7 (15.7) | 41.2 (11.3) | 34.0 (12.5) | 35.0 (12.5) |
| People's Speech | 56.8 (5.4) | 64.0 (5.7) | 55.4 (4.7) | 23.3 (5.5) | 31.6 (5.9) | 32.0 (5.0) |
| LibriSpeech | 47.1 (7.0) | 58.4 (7.4) | 46.8 (5.8) | 23.2 (7.0) | 31.4 (7.5) | 32.6 (6.2) |
| VoxPopuli | 26.3 (6.4) | 38.5 (6.6) | 33.8 (4.8) | 16.1 (6.4) | 27.1 (6.8) | 27.3 (5.1) |
| TED-LIUM | 64.1 (8.6) | 72.1 (8.6) | 58.4 (7.4) | 31.0 (8.5) | 34.8 (8.9) | 35.8 (7.6) |
| SWC | 24.5 (6.2) | 32.8 (6.4) | 27.7 (5.5) | 15.1 (6.2) | 21.7 (6.6) | 22.2 (5.8) |
| VCTK | 76.3 (10.6) | 77.7 (9.4) | 67.5 (10.0) | 36.4 (9.6) | 28.1 (9.4) | 33.6 (9.0) |
| Flickr8k | 68.3 (10.8) | 76.6 (10.1) | 61.5 (9.7) | 25.5 (8.8) | 30.7 (9.4) | 32.0 (9.3) |
| Natural speech | 40.3 (6.3) | 51.7 (6.8) | 43.6 (5.7) | 20.0 (6.4) | 28.2 (6.9) | 28.9 (5.9) |

Table 9: **Natural-speech recovery using the latent lens.** Cumulative Recall@10 for the current and next words. Parentheses report matched-frequency baselines. Llama PI-1/3 denotes Llama-3.2 3B PI-1/3 20k (official); Llama PI-2/3 denotes Llama-3.2 3B PI-2/3 20k; and Qwen PI-1/3 denotes Qwen2.5 3B PI-1/3 60k.

| Dataset | Current word | | | Next word | | |
|---|---|---|---|---|---|---|
| | Llama PI-1/3 | Llama PI-2/3 | Qwen PI-1/3 | Llama PI-1/3 | Llama PI-2/3 | Qwen PI-1/3 |
| Controlled | 87.7 (9.3) | 81.7 (13.1) | 87.1 (3.5) | 30.6 (6.8) | 32.3 (11.4) | 31.4 (3.8) |
| People's Speech | 94.6 (8.0) | 94.8 (10.7) | 94.8 (2.3) | 70.9 (6.3) | 81.5 (9.6) | 76.0 (1.7) |
| LibriSpeech | 85.6 (8.5) | 85.4 (11.1) | 86.0 (3.1) | 49.4 (7.4) | 52.2 (11.0) | 51.9 (2.6) |
| VoxPopuli | 91.5 (9.1) | 91.9 (12.1) | 92.5 (3.3) | 55.5 (7.4) | 57.4 (11.4) | 58.3 (2.3) |
| TED-LIUM | 87.9 (8.4) | 87.9 (11.3) | 88.1 (2.7) | 52.7 (7.2) | 56.1 (10.7) | 54.9 (1.9) |
| SWC | 82.1 (7.3) | 81.8 (10.0) | 82.3 (3.5) | 43.9 (6.2) | 47.2 (9.9) | 46.6 (3.0) |
| VCTK | 97.4 (9.7) | 91.7 (11.5) | 93.2 (2.6) | 57.6 (8.0) | 53.5 (10.9) | 54.7 (2.1) |
| Flickr8k | 87.5 (6.8) | 81.7 (9.2) | 82.0 (3.2) | 39.4 (4.7) | 42.8 (8.4) | 41.7 (2.5) |
| Natural speech | 89.4 (8.2) | 87.7 (10.6) | 88.0 (2.9) | 57.4 (6.7) | 58.5 (9.9) | 56.7 (2.3) |

Table 10: **Natural-speech recovery using linear probes.** Cumulative Recall@10 for the current and next words. Parentheses report matched-frequency baselines.

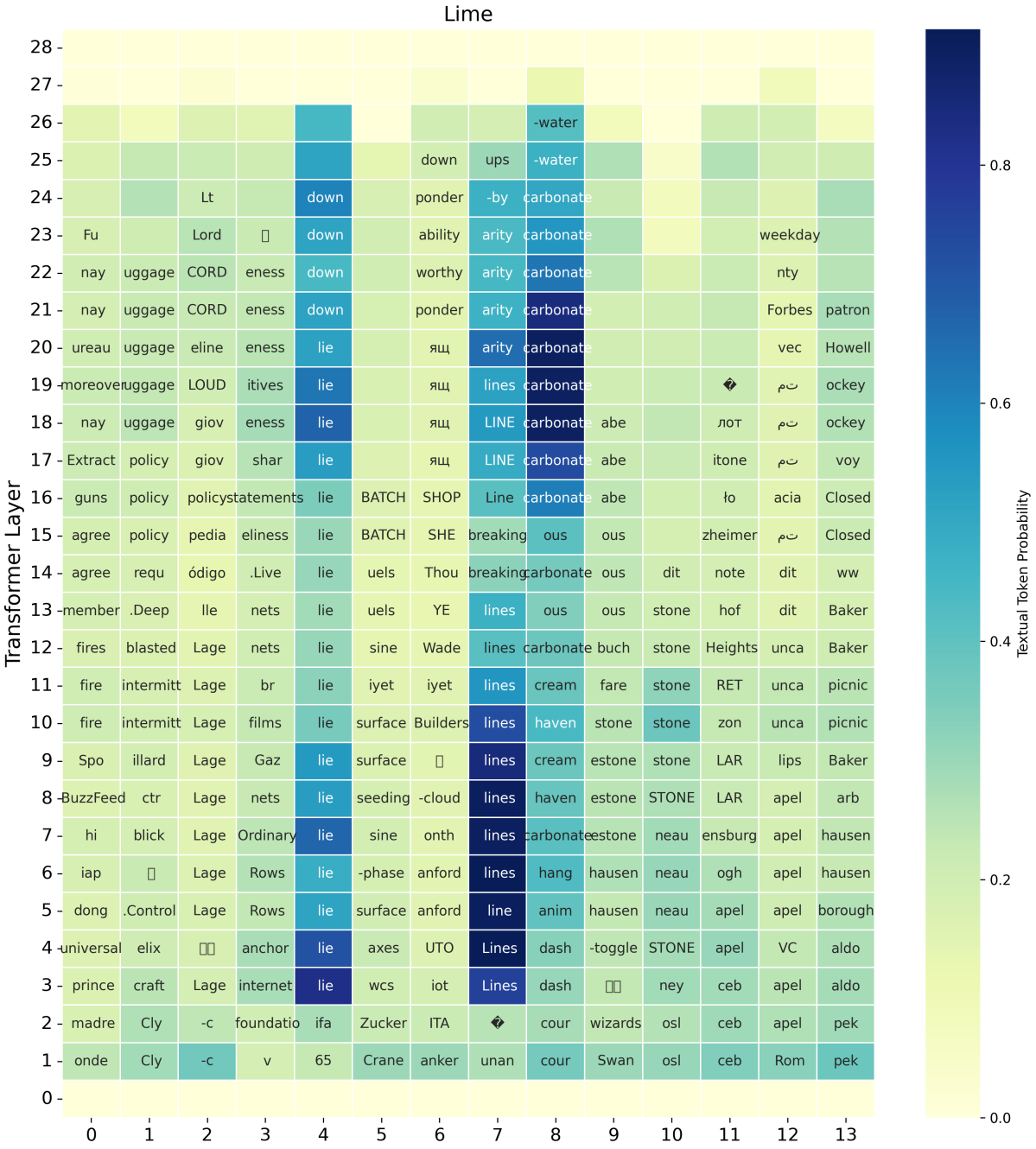



Figure 13: Logit lens of intermediate states for the spoken input "lime", using the Llama-3.2 PI-1/3 (official) model. The example shows both gradual transcription, with an early prediction of "lie", and a transcription error, where the model predicts "line" instead of "lime".

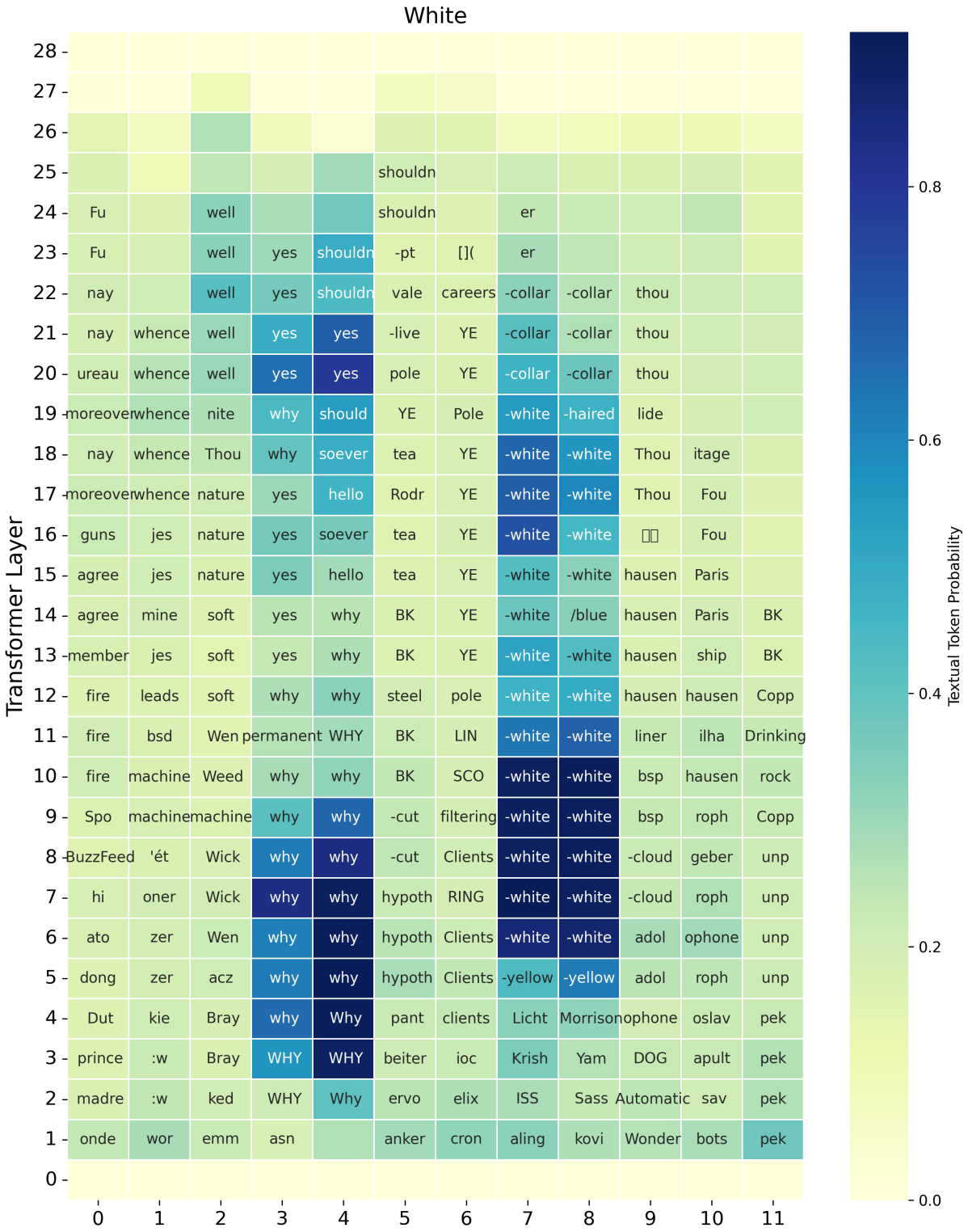



Figure 14: Logit lens of intermediate states for the spoken input "white", using the Llama-3.2 PI-1/3 (official) model. The example shows gradual transcription, where the model first predicts the partial form "why" before later converging to the full word "white".

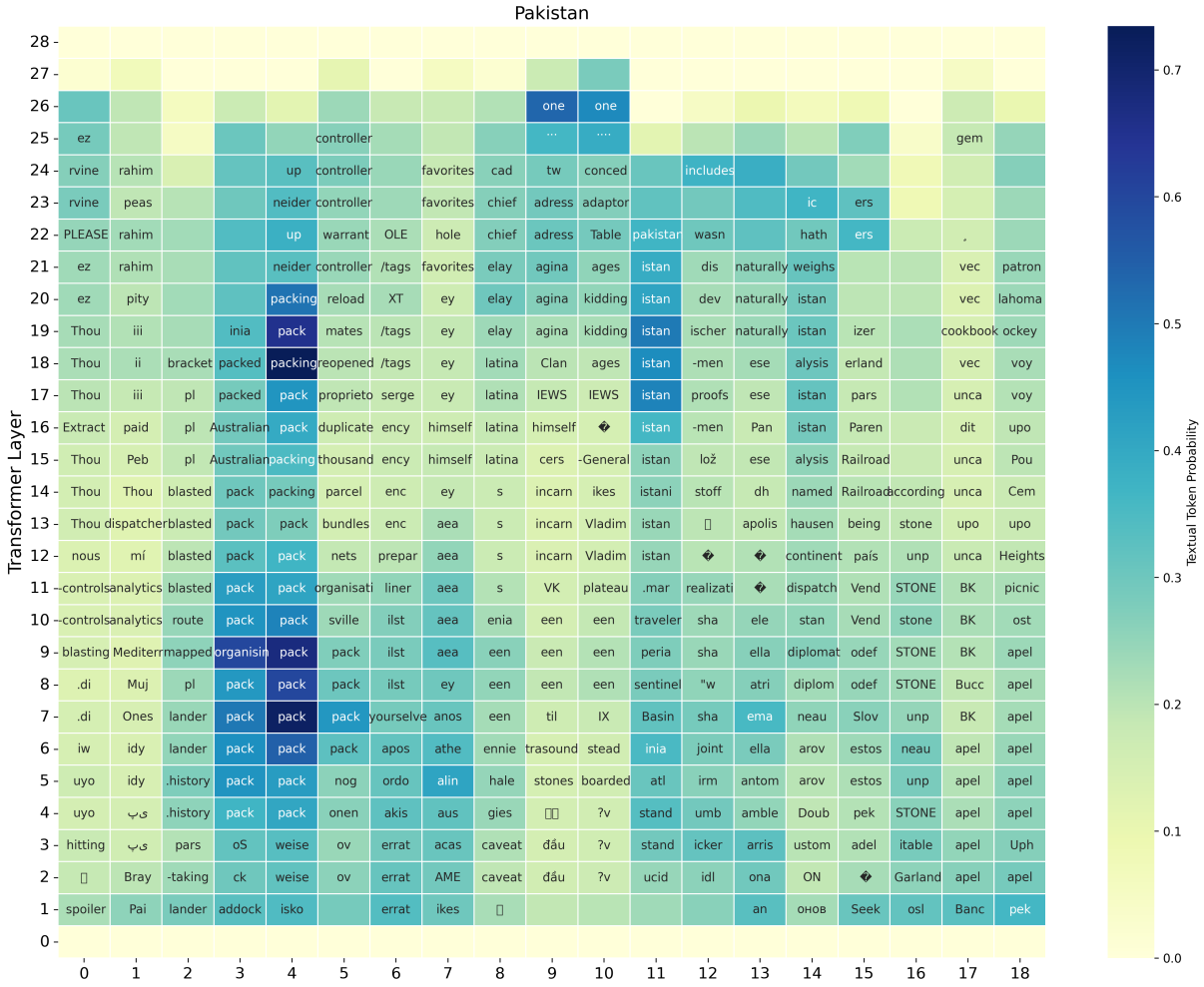
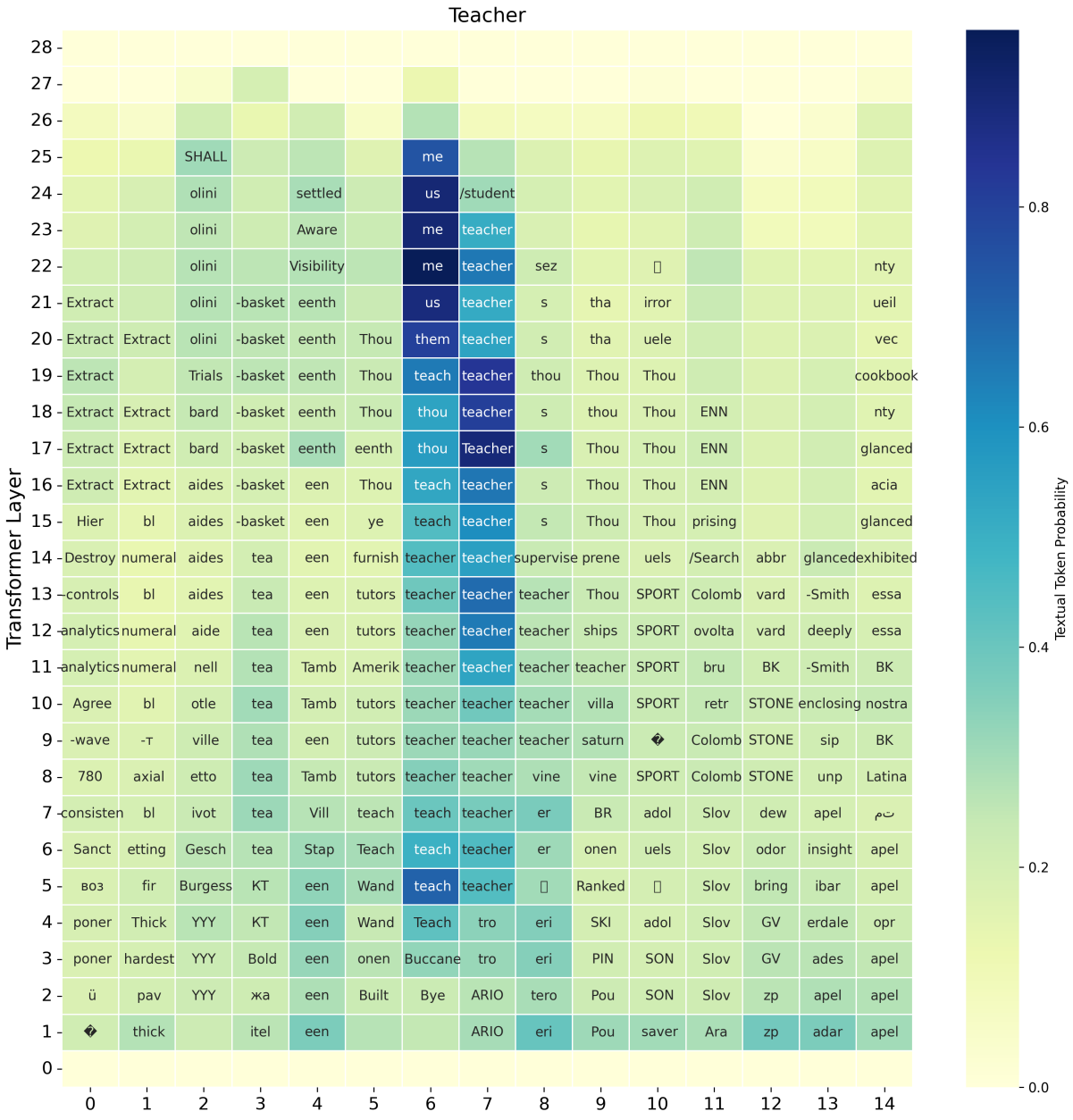



Figure 15: Logit lens of intermediate states for the spoken input "Pakistan", using the Llama-3.2 PI-1/3 (official) model. The example shows gradual transcription, where the model first predicts the partial form "pack" and later updates toward later components of the word, such as "istan", before converging to "Pakistan".


Teacher
Transformer Layer
Textual Token Probability
0.0
0.2
0.4
0.6
0.8
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
28 Extract Extract Extract Extract Extract Extract Hier Destroy controls analytics analytics Agree -wave 780 consisten Sanct воз poner poner ü
27 Extract Extract Extract Extract Extract Extract bl numeral bl numeral numeral bl -т axial bl etting fir Thick hardest pav
26 olini olini olini Trials bard bard aides aides aide nell otle ville etto ivot Gesch YYY YYY YYY
25 SHALL olini olini olini -basket -basket -basket

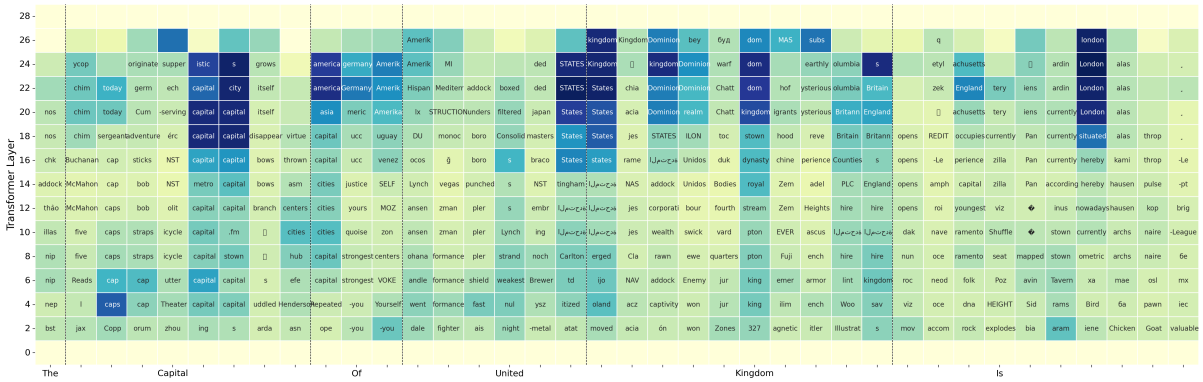



Figure 17: Logit Lens of inner states of the spoken input: "The capital of United Kingdom is...", using Llama-3.2 PI-1/3 (official).This version contains less filters on the time and layer domain and thus show higher resolution in the inner mechanism by logit lens perspective

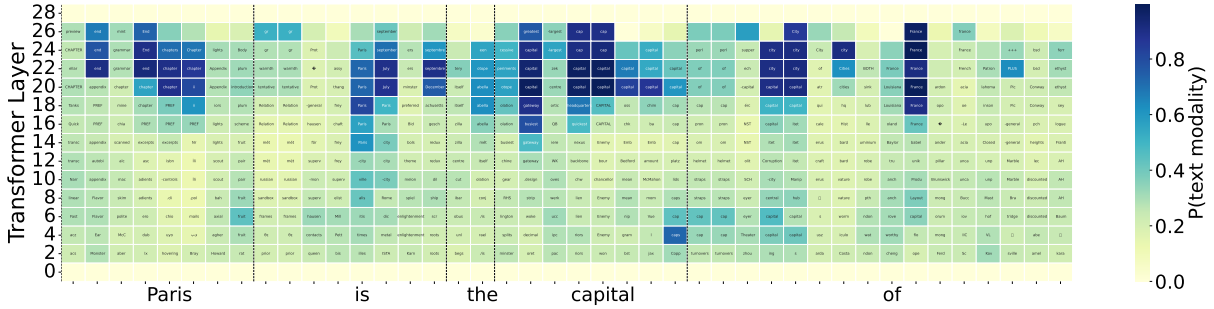



Figure 18: Logit lens of inner states of the spoken input: "Paris is the capital of ...", using Llama-3.2 PI-1/3 (official).

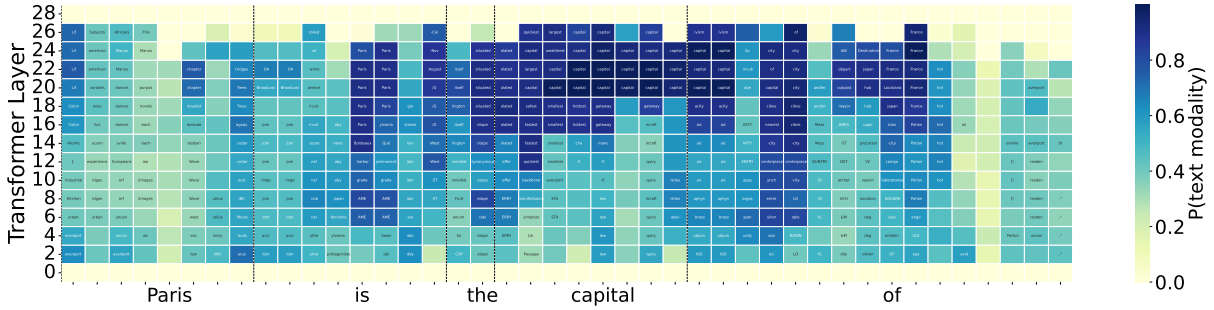



Figure 19: Logit Lens of inner states of the spoken input: "Paris is the capital of ...", using Llama-3.2 PI-1/3 (ours).

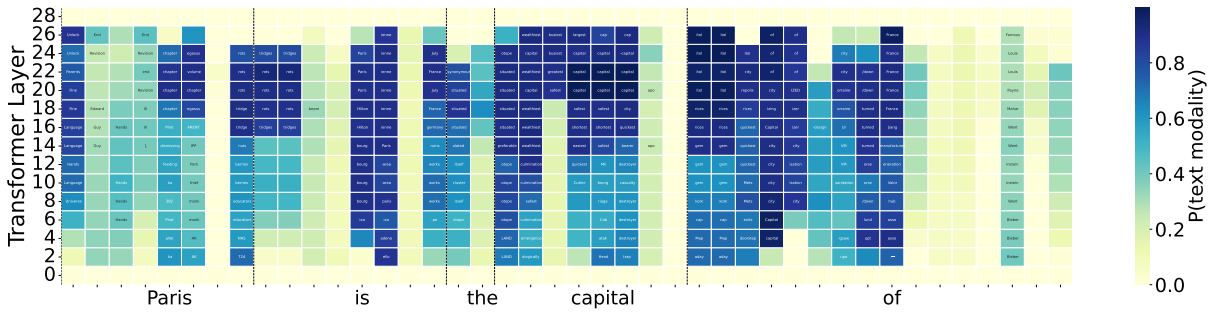



Figure 20: Logit Lens of inner states of the spoken input: "Paris is the capital of ...", using Llama-3.2 PI-2/3.

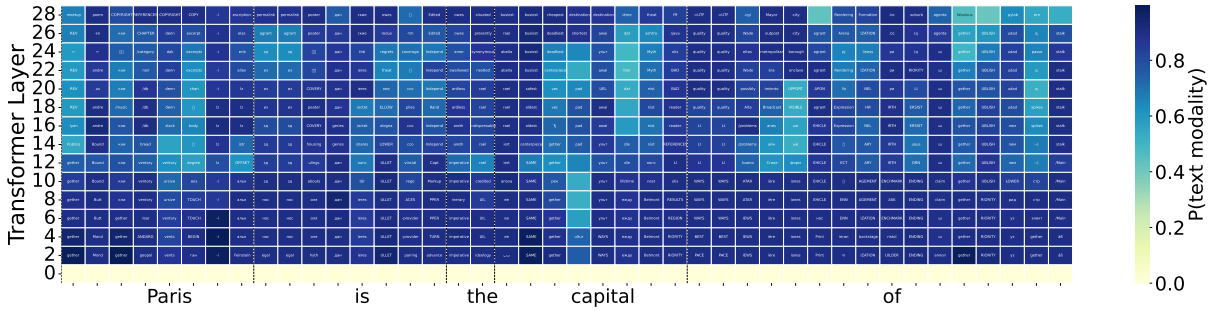



Figure 21: Logit Lens of inner states of the spoken input: "Paris is the capital of...", using Qwen2.5-3B PI-1/3 60k

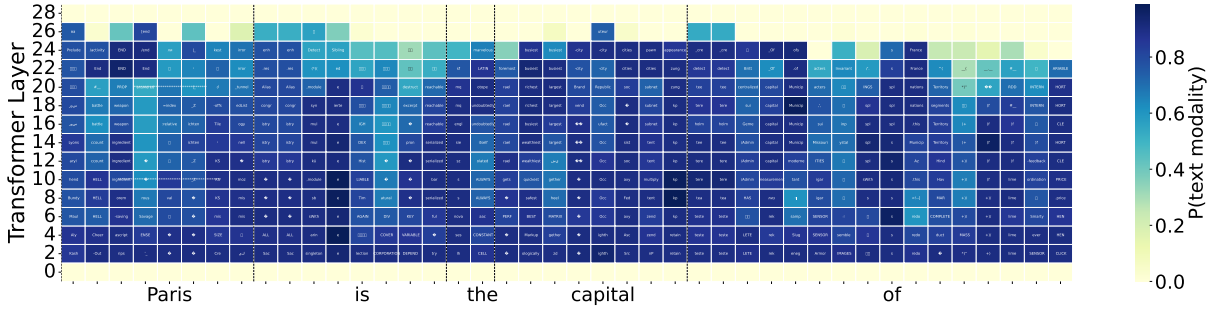



Figure 22: Logit Lens of inner states of the spoken input: "Paris is the capital of...", using Qwen2.5-1.5B PI-1/3 42K

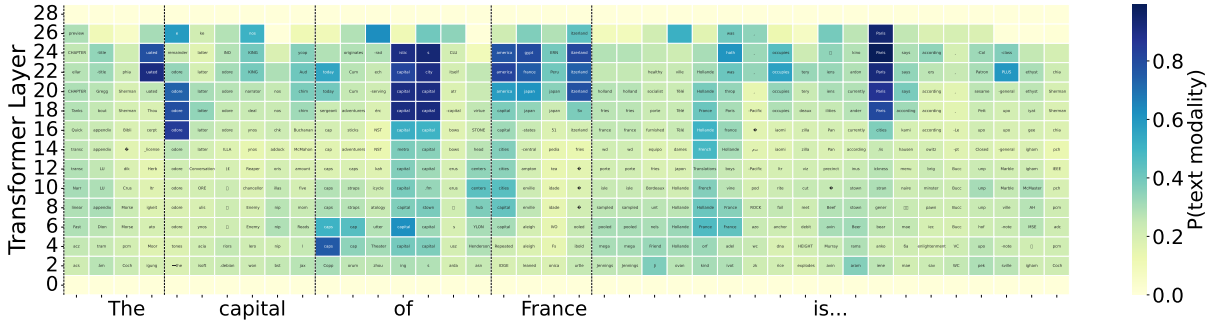



Figure 23: Logit Lens of inner states of the spoken input: "The capital of France is...", using Llama-3.2 PI-1/3 (official).

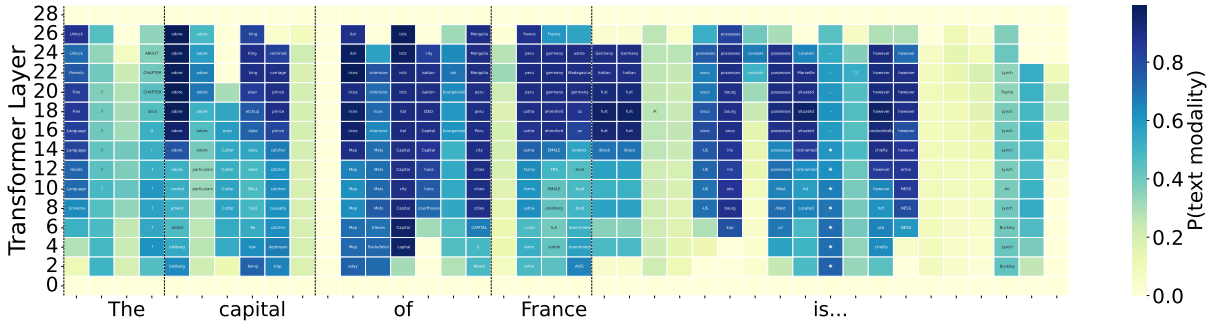



Figure 24: Logit Lens of inner states of the spoken input: "The capital of France is...", using Llama-3.2 PI-2/3

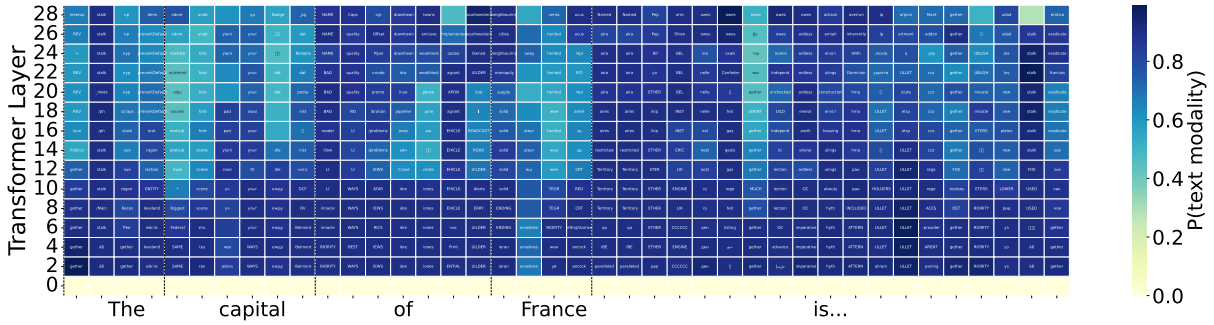



Figure 25: Logit Lens of inner states of the spoken input: "The capital of France is...", using Qwen2.5-3B PI-1/3 60k.

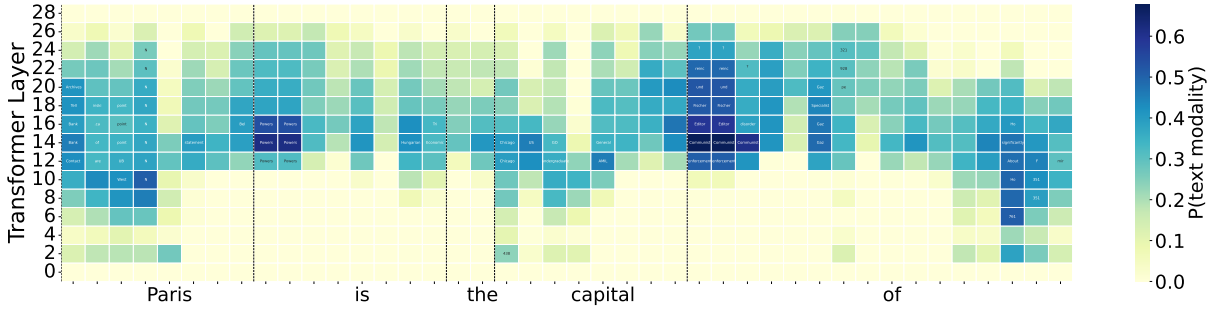



Figure 26: Logit Lens of inner states of the spoken input: "Paris is the capital of ...", using Llama3.2-3B RST

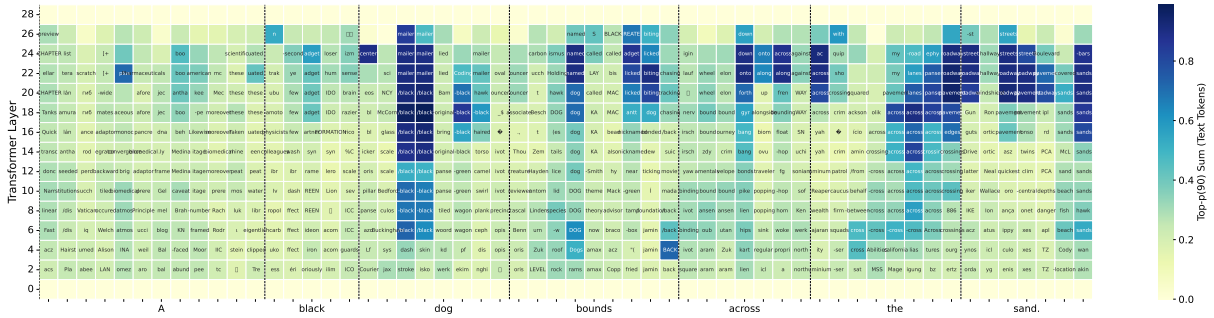



Figure 27: Logit Lens of inner states of the natural spoken input from Flickr8k dataset: "A black dog bounds across the sannd.", using using Llama-3.2 PI-1/3 (official)

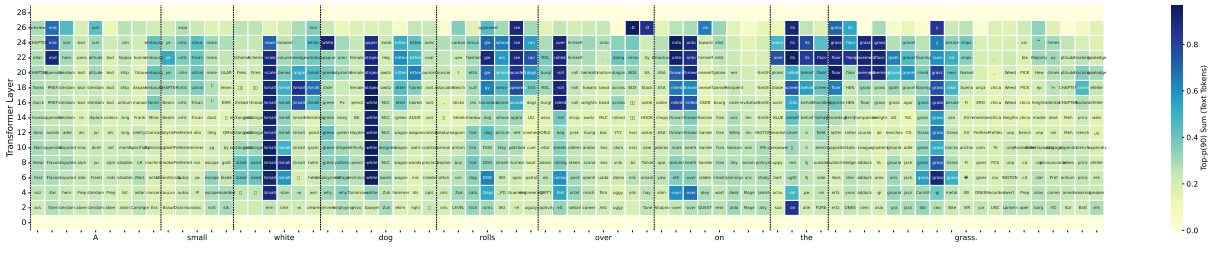



Figure 28: Logit Lens of inner states of the natural spoken input from Flickr8k dataset: "A small white dog rolls over on the grass.", using using Llama-3.2 PI-1/3 (official)

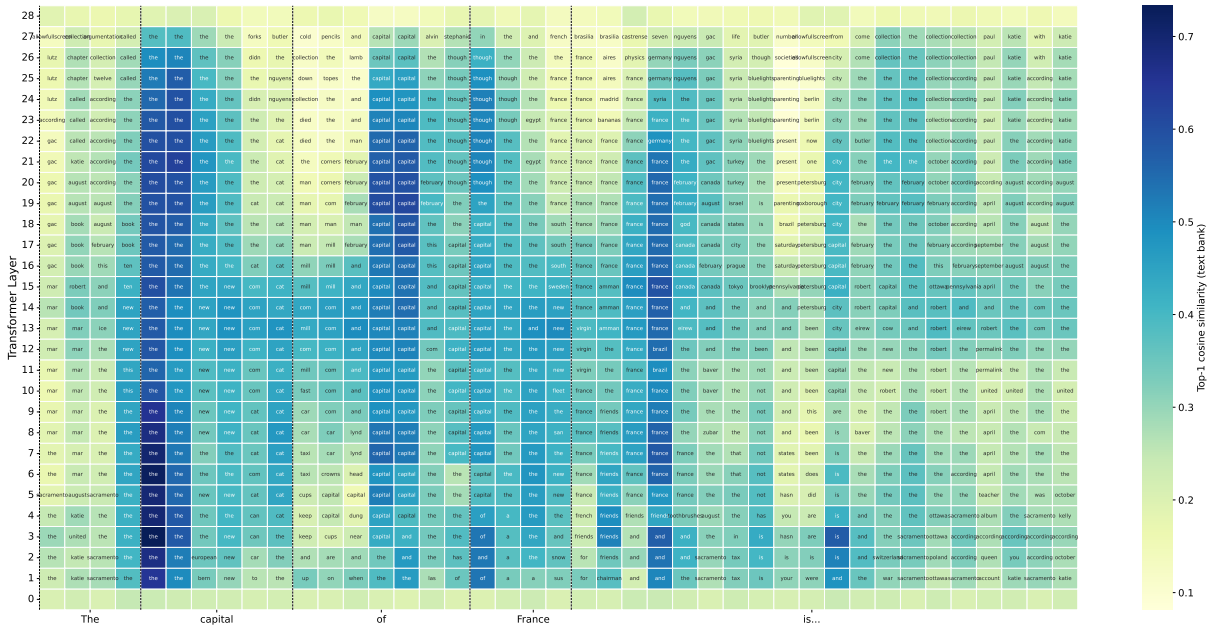



Figure 29: Latent Lens of inner states of the spoken input: "The capital of France is...", using Llama-3.2 PI-1/3 (official).