Analysis of Hidden State Representations in Llama 3 8B Instruct for a Word Reversal Task

Zelong Xu

May 22, 2025

Abstract

This report investigates the internal hidden state representations of the Llama 3 8B Instruct model when performing a 5-word reversal task using 3-shot prompting. We analyze hidden states extracted from embedding, middle, and final layers using Principal Component Analysis (PCA), linear probing, and cosine similarity. Our findings reveal a progression from vocabulary-centric encoding in early layers to a clear differentiation of input/output roles and more abstract, task-oriented representations in deeper layers.

Contents

1 Project Overview

This project aims to analyze the characteristics of hidden state representations within the Llama 3 8B Instruct model as it processes a structured sequence-to-sequence task: word reversal. The core methodology involves:

- Prompting the model with few-shot examples to perform a 5-word reversal.
- Extracting hidden states from the embedding layer (Layer 0), a middle layer (Layer 16), and the final transformer block's output (Layer -1).
- Applying analytical techniques: Principal Component Analysis (PCA) for visualization, linear probing to assess the decodability of specific information, and cosine similarity to measure representational similarity.

The task was configured for 5-word sequences, using a 3-shot prompting strategy with the Llama 3 Instruct chat template. The end-of-turn token `<|eot_id|> (ID: 128009)` was used as the primary stop signal for generation. Word-token alignment was performed based on character offsets, and samples with a word extraction ratio below 80% (non-strict mode) were excluded from hidden state analysis.

2 Experimental Results and Analysis

2.1 Task Performance Evaluation

The Llama 3 8B Instruct model was evaluated on 100 generated samples for the 5-word reversal task. The performance metrics are summarized in Table 1.

Table 1: Word Reversal Task Performance (Llama 3 8B Instruct, 5-words, 3-shot)

Metric	Score
Token Exact Match Accuracy	0.500
Decoded Word Sequence Match Accuracy	0.730
Character-level Accuracy (no marker/spaces)	0.835

Note: Token exact match compares generated token IDs (including the end marker) against the ground truth token IDs. Word sequence match compares space-separated words after decoding and stripping special tokens. Character-level accuracy ignores spaces and end markers.

The model achieved a decoded word sequence accuracy of 73%, indicating a good capability to perform the reversal task at the semantic (word) level. The lower token exact match accuracy (50%) suggests discrepancies often arise from minor tokenization differences or subtle formatting variations (e.g., leading newlines from chat templates) rather than complete failure on the task logic for many of these cases.

2.2 Information Encoding in Hidden States (Linear Probing)

Linear probes were trained to decode specific types of information from the hidden states of selected layers. Key findings are presented in Table 2.

Table 2: Linear Probe Mean Accuracies Across Layers

Probing Task	Layer 0	Layer 16	Layer -1 (Final)
Input Word Identity (from input states)	1.000	1.000	0.998
Output Word Identity (from output states)	0.985	0.976	0.689
Reversed Input Word (from output states)	0.985	0.976	0.689
Token Role (Input vs. Output Token)	0.409	1.000	1.000

Note: Scores are mean accuracies from 2-fold or 5-fold (for position type) cross-validation, depending on class sample sizes. Word identity probes predict the specific word; Token Role probe classifies if a token belongs to the input prompt or the generated output.

Key Observations from Linear Probing:

- **Word Identity**: The identity of both input and output words is highly decodable from Layer 0 and Layer 16 states. The decomposition of the output word identity slightly decreases at the final layer (Layer -1), potentially as representations become more focused on next-token prediction logits.
- **Token Role (Input vs. Output Context)**: A significant finding is the evolution of role encoding. At Layer 0 (post-embedding), the model's hidden states show poor linear separability between input and output tokens (accuracy ≈ 0.41, close to the random chance for a potentially imbalanced binary classification). However, by Layer 16, this separability becomes perfect (accuracy 1.0) and is maintained at Layer -1. This suggests that the model learns to clearly distinguish the input context from the generation phase in its middle to late layers.

2.3 Evolution of Representational Similarity (Cosine Similarity)

Cosine similarity was used to measure the similarity between hidden state representations of:

(a) An input word at position i (Input[Word i]) and its corresponding word in the reversed output sequence (Output[corr. pos]).

(b) An input word at position i (Input[Word i]) and a different input word at position j (Input[Word j]) as a control.

Results are presented for states derived from the average of a word's tokens (avgWord State). Findings for last-token states were very similar.

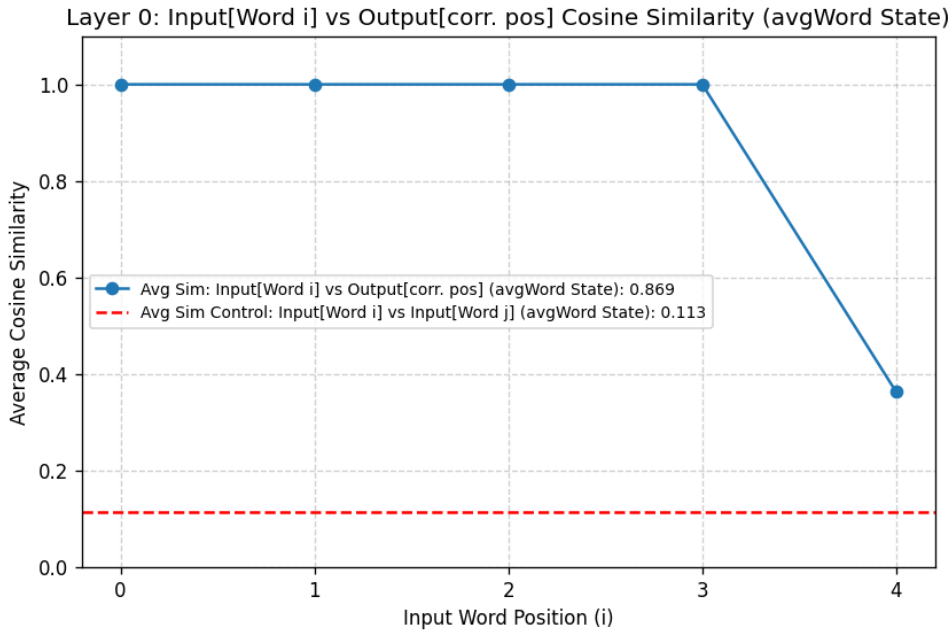

Figure 1: Layer 0: Cosine Similarity (avgWord State). Input[Word i] vs Output[corr. pos] shows high similarity (overall 0.869), especially for initial positions (1.0), while control similarity (Input[Word i] vs Input[Word j]) is low (0.113). The drop at Input Position 4 for corresponding similarity is notable.

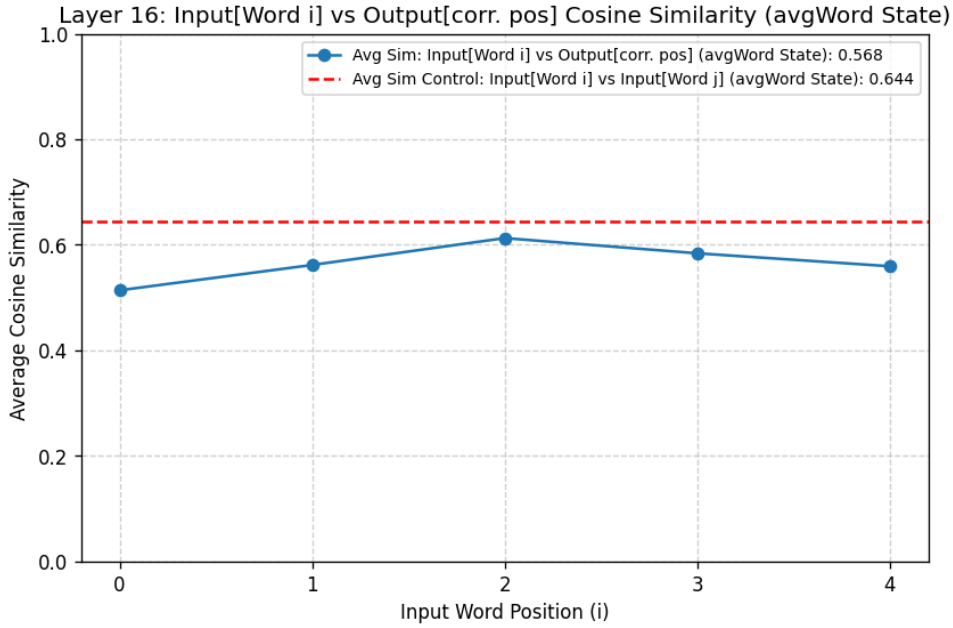

Figure 2: Layer 16: Cosine Similarity (avgWord State). Similarity between corresponding input/output words drops (overall 0.568), while control similarity among different input words increases (0.644), even exceeding the corresponding similarity.

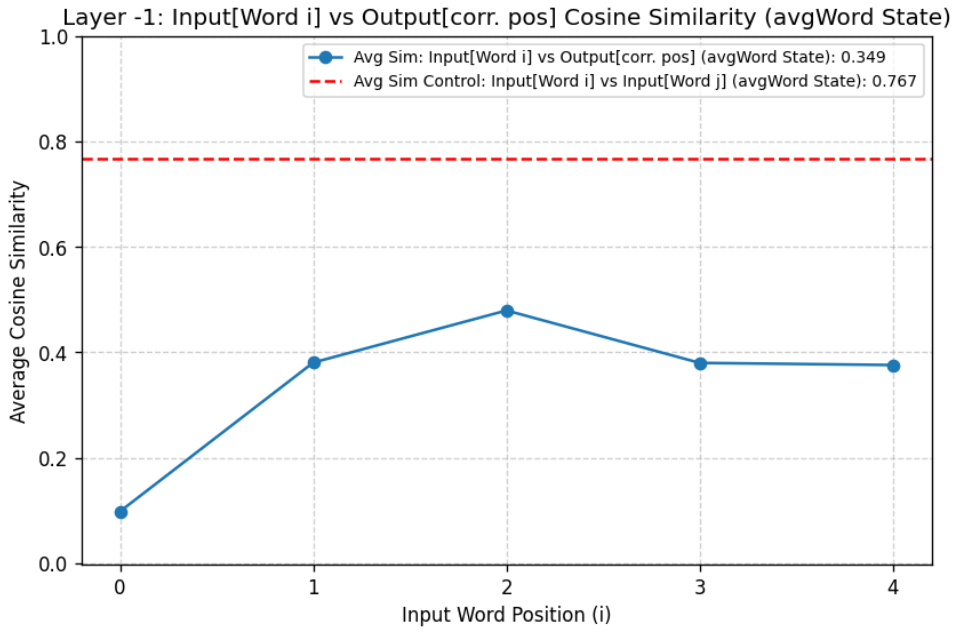

Figure 3: Layer -1 (Final): Cosine Similarity (avgWord State). Similarity between corresponding input/output words is lowest (overall 0.349). Control similarity among different input words is highest (0.767).

Key Observations from Cosine Similarity:

- **Layer 0**: Representations are vocabulary-centric. The same word in the input and its corresponding (reversed) position in the output have very high state similarity, suggesting position/role has minimal impact on the initial embedding. The high similarity (1.0) for input positions 0-3 vs. their output counterparts suggests identical words at these corresponding positions often result in nearly identical average hidden states. The drop at input position 4 warrants further investigation.
- **Layer 16**: Contextual information begins to dominate. The direct similarity between an input word and its target output word decreases. Conversely, the similarity among different input words increases, and is now higher than the input-to-corresponding-output similarity. This indicates that representations are becoming more contextualized, perhaps forming a more holistic representation of the input sequence or an intermediate state of the reversal process.
- **Layer -1 (Final Layer)**: Representations are highly task-specific and abstracted from the initial embeddings. The similarity between an input word and its corresponding output word is at its lowest. The high similarity among different input words persists, suggesting that at this stage, the "input context" as a whole might be represented more cohesively than individual input-output mappings based on original word embeddings.

2.4 PCA Visualization of Hidden States

PCA was applied to visualize the 2D projection of hidden states. Figures 4 and 5 show states colored by their role (Input Word vs. Output Word) and styled by their word position within the sequence (Pos0 to Pos4). We present results for avgWord states; lastToken states showed similar patterns.

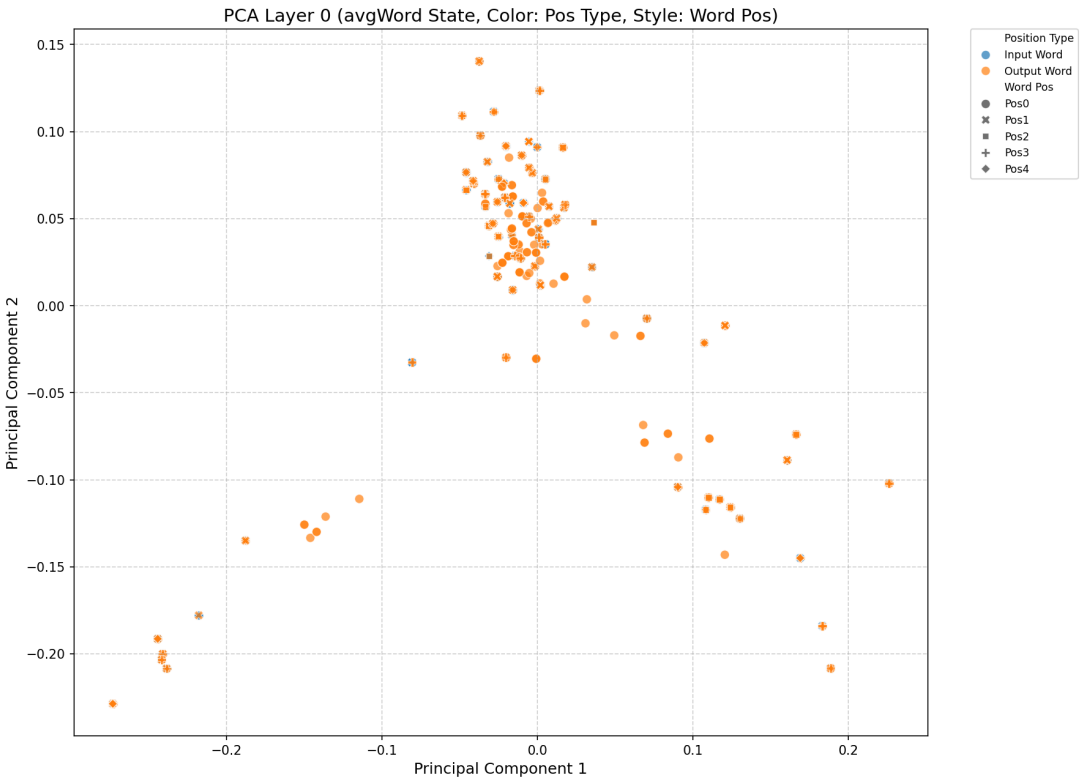

Figure 4: PCA of Layer 0 Hidden States (avgWord State), Colored by Position Type (Input/Output Word), Styled by Word Position. Input and Output word states are highly intermingled.

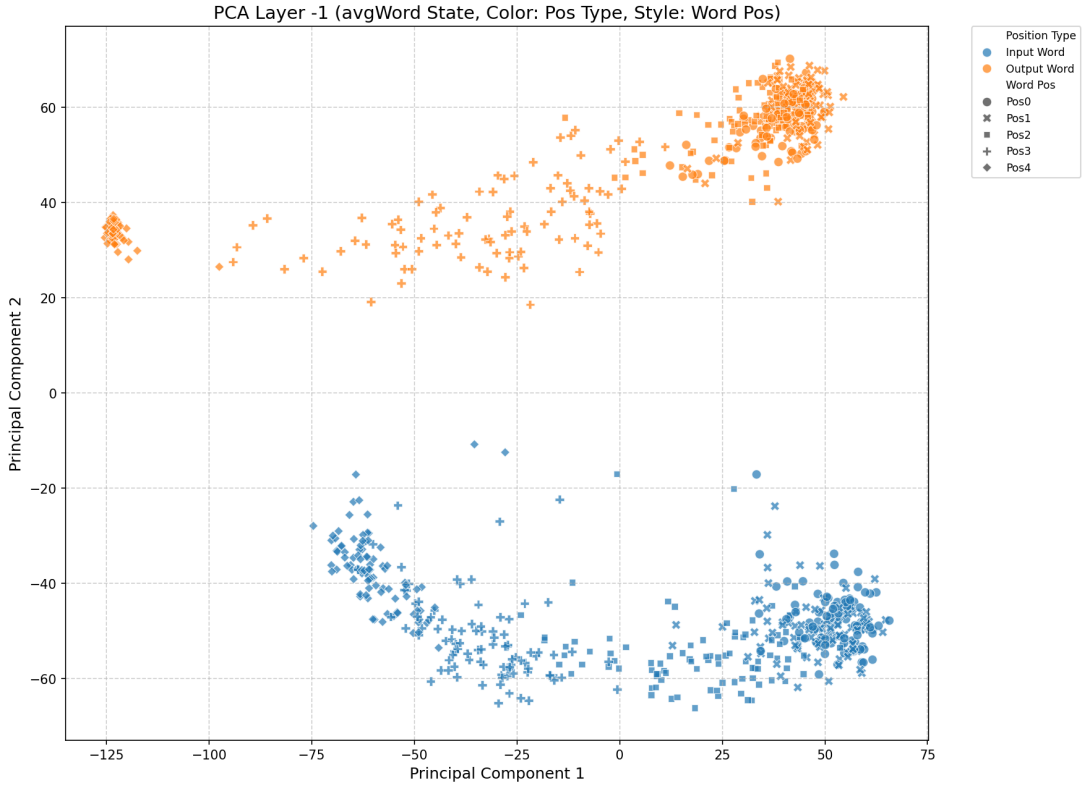

Figure 5: PCA of Layer -1 (Final) Hidden States (avgWord State), Colored by Position Type, Styled by Word Position. Input and Output word states form clearly distinct clusters.

Key Observations from PCA:

- **Layer 0**: As seen in Figure 4, the hidden states of input words and output words are largely undifferentiated in the PCA projection, consistent with the linear probe findings for token role at this layer.
- **Layer 16 (not shown here but similar to Layer -1's trend)** and **Layer -1 (Final Layer)**: Figure 5 (and similarly for Layer 16) clearly shows that the hidden states for input words and output words form distinct clusters. This visual evidence strongly supports the linear probe result that the model learns to differentiate these roles in its deeper layers. Within these clusters, different word positions (styles) also show some structuring, though less pronounced than the input/output separation.

Visualizing states colored by individual words (not shown for brevity due to the large vocabulary) at Layer 0 typically shows tight clusters for identical words, regardless of their input/output role, further supporting the vocabulary-centric nature of early representations. This clustering becomes less word-specific and more context/role-specific in deeper layers.

3 Conclusion

The Llama 3 8B Instruct model demonstrates a hierarchical approach to solving the word reversal task. Initial layers (Layer 0) focus on vocabulary–level encoding with minimal distinction between input and output roles. Middle (Layer 16) and final layers (Layer -1) progressively develop a clear, linearly separable representation of token roles (input vs. output) and transform word representations in a context-dependent manner, leading to lower direct similarity with original input embeddings but likely encoding the necessary information for the reversal. PCA visualizations corroborate these findings, showing a transition from mixed to clearly separated

input/output state clusters in deeper layers.