Token Identity Drift in Transformer Language Models

David McQueen
Independent Researcher

December 22, 2025

Abstract

I investigate how individual token representations in transformer language models evolve across layers. Using cosine similarity between hidden states and original token embeddings, I measure how rapidly lexical identity decays and how strongly contextual influences shape representations. The results show that token identity collapses within the first few layers, that semantic context induces mid-layer divergence between senses of the same token, and that late layers reconverge representations toward task-relevant abstractions. These findings support the view that transformer representations are not stable semantic units but dynamic, context-dependent computations.

1 Introduction

Transformer language models begin from discrete token embeddings. However, my observations suggest that these embeddings function only as initial conditions for deeper contextualization. Understanding how token identity interacts with contextual representations is crucial for interpretability, polysemy analysis, tokenization design, and representational geometry.

In this paper, I address two core questions:

- How rapidly does a token's representation drift from its original embedding?
- How do identical tokens in different contexts diverge and reconverge across layers?

I answer these questions through controlled experiments on GPT-2 family models, using cosine similarity to quantify representational drift and convergence.

2 Method

For a target token t and a prompt P containing t, I extract the hidden state $h_\ell(t)$ at each transformer layer ℓ. Let $e(t)$ be the original token embedding from the model's embedding table. I compute:

$$\text{Identity}(\ell) = \text{cosine}(h_\ell(t), e(t)) \tag{1}$$

to measure how similar the layer's representation is to the original embedding.

For pairs of contexts P_a and P_b containing the same token, I compute:

$$\text{Context}(\ell) = \text{cosine}(h_\ell(t_a), h_\ell(t_b)) \tag{2}$$

to quantify the similarity between representations in different semantic contexts.

I define:

- **Half-life layer:** the first layer where identity similarity falls below a threshold.
- **Context dominance layer (CDL):** the first layer where context similarity becomes less than identity similarity.

3 Experiments

I define a set of semantic contrasts on tokens with multiple senses:

- `Ġlead`: verb vs metal
- `Ġbank`: finance vs river

Prompts are designed to isolate these contexts while keeping sentence structure similar.

Models tested include GPT-2 (117M), DistilGPT-2, and GPT-2-Medium.

4 Results

Identity similarity curves show that tokens rapidly lose alignment with their original embeddings within the first few layers.

Context similarity curves reveal a characteristic pattern:

- **High early:** shared lexical origin
- **Mid layers:** semantic divergence
- **Late layers:** partial reconvergence

Figure 1 shows the identity and context similarity curves for `Ġlead` in GPT-2-Medium.

Across models, I observe:

- Identity similarity collapses within 2–4 layers.
- Semantic divergence peaks in mid layers and often recedes in later layers.
- Context similarity remains well above identity similarity in deeper layers for many tokens.

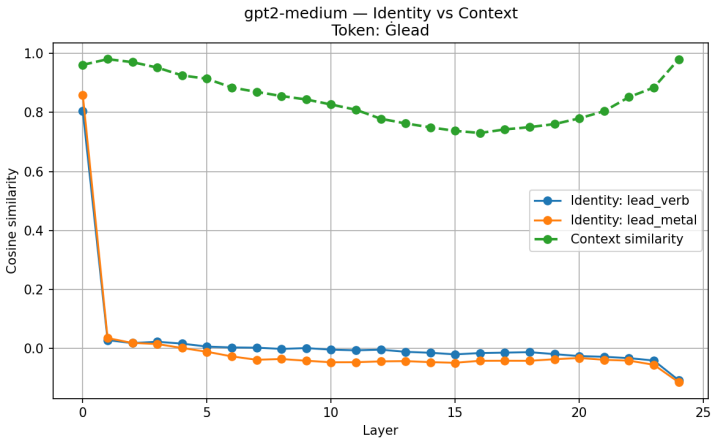

Figure 1: Identity vs Context similarity for `Ġlead` in GPT-2-Medium.

5 Discussion

My findings suggest that transformer layers serve not to preserve lexical identity but to construct *contextual meaning trajectories*. Identity collapse precedes semantic differentiation, indicating that token-level meaning is not encoded in embedding alignment but in relational structure formed through attention.

Reconvergence of context similarity in late layers reflects the model's shift from fine semantic representation to shared predictive structure. Negative identity similarity in deep layers indicates a full basis rotation rather than noise.

6 Conclusion

I have quantified how token identity drifts and how contextual representations evolve across transformer layers. The results show that semantic meaning in transformers is not anchored in static embeddings but emerges dynamically through depth.

Future work could explore:

- Larger models and different architectures (e.g., rotary positional embeddings)
- Subword vs word comparisons
- Attention pattern decomposition

Acknowledgements

I thank the open-source community for tools such as HuggingFace Transformers and GPT-2 models, and all readers and collaborators who help improve interpretability research.