What are you sinking? A geometric approach on attention sink

Valeria Ruscio, Umberto Nanni, Fabrizio Silvestri

Sapienza University of Rome

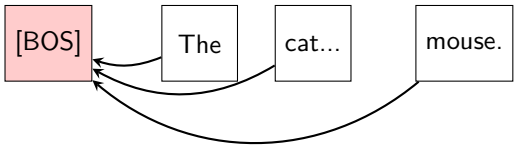
SAPIENZA
UNIVERSITÀ DI ROMA

The Puzzle of the Attention Sink

What is it? A consistent pattern where specific tokens—often the first token—receive a huge amount of attention (e.g., 30-40%) from many other tokens, regardless of semantic content.

The core questions this raises:

- Why would the word 'mouse' pay so much attention to a generic '[BOS]' token?
- Why does model performance *collapse* if you try to prevent this behavior?
- Is it a bug, an artifact, or a feature?



All tokens attend to the sink.

Why Sinks Emerge: A Tale of Pressure and Bias

Our research shows that attention sinks are not an accident, but an inevitable consequence of two fundamental forces working together inside the model.

1. The Pressure to be Sparse

- The softmax function forces attention weights to live on a **probability simplex** ($\sum \alpha_i = 1$).
- This creates pressure towards sparse, low-entropy solutions. It's more efficient to focus a limited "attention budget" on a few points.

2. The Bias for a Stable Target

- This focused attention needs a stable, easy-to-find target.
- The **Positional Encoding** provides this via mathematical asymmetry. In standard RoPE, the first token receives no rotation ($\mathbf{R}_0 = \mathbf{I}$), making it a unique and computationally convenient anchor.

Models Build Geometric "Reference Frames"

We discovered that models don't just process sequences; they build an internal, stable coordinate system. They do this by creating **attention sinks**.

Attention Sinks: Specific tokens (like '[BOS]' or commas) that consistently attract a large portion of the attention, regardless of semantic content.

These sinks aren't noise; they are Reference Frames

They act like an origin point $(0, 0)$ on a map, providing a stable geometric anchor that all other tokens use to orient themselves and establish consistent relationships.

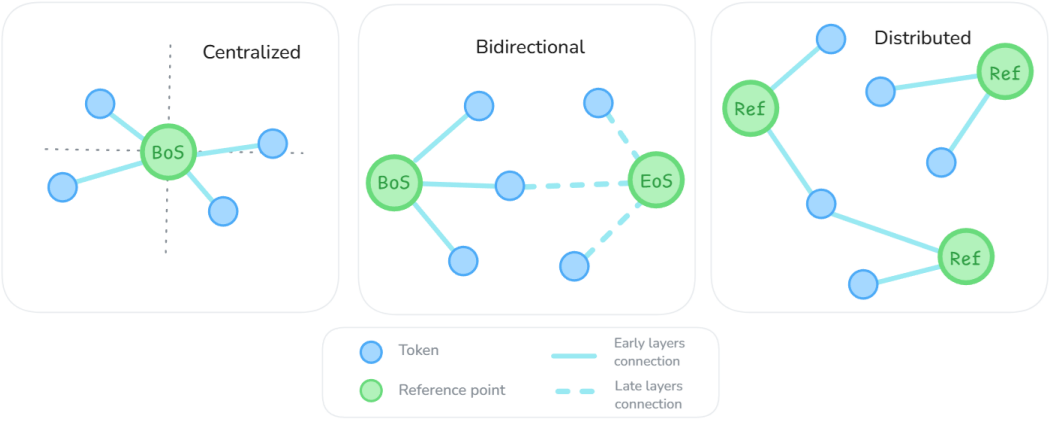


Figure: Geometric interpretation of the three frame types.

How We Proved It: A Multi-Faceted Analysis

To uncover these geometric structures, we employed a combination of advanced techniques to analyze the models from different perspectives.

1. Topology

We used **Persistent Homology** to compute Betti numbers (β_0, β_1), revealing the stable connectivity and cyclic structures in the attention graph.

2. Spectral Analysis

We analyzed the **Laplacian matrix** of the attention graph, using the Fiedler value to measure connectivity and other metrics to find "star-like" patterns.

3. Vector Geometry

We measured **Directional Influence** and **Geometric-Semantic Alignment** to see how reference points shape the transformations in the value space.

Evidence of Early Emergence: A Random Matrix Theory View

The Question: Are reference frames a high-level semantic feature learned late in training, or something more fundamental?

The Method: Tracking Non-Randomness

- We used **Random Matrix Theory (RMT)** to track when attention matrices stop behaving like random noise.
- We measured the deviation from the theoretical **Marchenko-Pastur distribution**, which describes the eigenvalues of random matrices.
- Key metrics included the **Spectral Gap** (λ_1/λ_2) and **Participation Ratio**.

The Finding: Structure is Immediate

- Structure emerges almost instantly. We see significant deviation from random within the **first 8 training steps**.
- This emergence is **non-monotonic with model size**. Mid-sized models (Pythia-6.9B) establish frames most efficiently.
- The largest models (12B+) show a **phase transition**, concentrating attention into fewer, stronger dimensions.

Conclusion

Reference frames are not learned on top of language; they are the geometric scaffolding the model builds at the very beginning to make learning possible.

Finding 6: Frames Dictate Value Space Transformations

The reference frame isn't just an attention pattern; it defines the model's strategy for transforming token representations.

- **Centralized (LLaMA):** Prioritizes **geometry over semantics**. The consistently negative "Geometric-Semantic Alignment" (-0.29) shows it enforces its coordinate system, even if it conflicts with word meanings.
- **Distributed (Qwen):** Balances geometry and semantics. The alignment score stays near zero, indicating a flexible strategy that adapts to local context.
- **Bidirectional (XLM-R):** Performs a **two-phase computation**. It starts by using semantic relationships (positive alignment, $+0.20$) and then transitions to a rigid geometric structure in deeper layers (strong negative alignment, -0.39).

Thank you!

ruscio@diag.uniroma1.it