The Singular Anchor: First Token Dominance in Large Language Model Attention Sinks

Khurram Khalil
Department of Electrical Engineering and Computer Science
University of Missouri-Columbia, USA
khurram.khalil@missouri.edu

Abstract

Large Language Models rely on "attention sinks"—initial sequence tokens that accumulate disproportionate attention—for efficient context management. However, the precise formation and positional dominance of these natural sinks remain under-characterized. We present the first systematic empirical study investigating attention sink patterns across three LLM families (GPT-2, Llama, Mistral) and five text categories. Our analysis reveals that the absolute first token (P1) overwhelmingly serves as the dominant natural attention sink, attracting significantly more attention ($p < 0.001$, Cohen's $d > 6.0$) than subsequent initial tokens across all architectures. While P1 dominance is universal, its strength varies by model family—Mistral exhibits the strongest P1 reliance—and is significantly modulated by input characteristics, with short texts eliciting maximal P1 attention and code texts minimal. These findings challenge assumptions about distributed sink importance and provide foundational insights for designing efficient long-context models.

1 Introduction

The Transformer architecture's self-attention mechanism enables Large Language Models (LLMs) to process long-range dependencies, but its quadratic complexity limits context window scaling. Recent work has identified "attention sinks"—initial sequence tokens that attract disproportionate attention—as a solution for efficient long-context processing [1]. While engineered sink preservation has proven effective in streaming applications, the natural formation and characteristics of these sinks across different architectures remain poorly understood.

Several fundamental questions persist: How consistently do natural attention sinks manifest across diverse LLM families? Do multiple initial tokens contribute equally as sinks, or does a specific position dominate? How do architectural choices and input characteristics influence sink utilization? Understanding these mechanisms is crucial for optimizing KV cache management, designing efficient attention mechanisms, and developing interpretable long-context models.

This paper presents the first large-scale empirical investigation into natural attention sink dynamics across three major LLM families (GPT-2, Llama, Mistral) and diverse text types. Our key finding reveals that the absolute first token (P1) serves as a dominant "singular anchor," consistently attracting significantly more attention than other initial positions across all tested architectures.

2 Methodology

2.1 Experimental Setup

We analyzed five publicly available decoder-only LLMs spanning three architectural families: `gpt2` (124M), `gpt2-medium` (355M), `microsoft/DialoGPT-medium` (355M),

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

`meta-llama/Llama-2-7b-hf` (7B), and `mistralai/Mistral-7B-v0.1` (7B). Models were accessed via Hugging Face Transformers and run in evaluation mode with appropriate precision settings.

To assess content-dependent effects, we curated 25 text samples across five categories: *narrative*, *technical*, *dialogue*, *code*, and *short* texts (5 samples per category). This design enables statistical analysis of both architectural and content influences.

2.2 Attention Analysis Protocol

For each model-text pair, we performed forward passes with `output_attentions=True` to extract attention weights. We defined the first $N = 4$ tokens as the potential sink region and computed our primary metric—P1 Attention Strength—as the average attention directed to the first token from all subsequent tokens, averaged across layers:

$$\text{P1\_Attn} = \frac{1}{L_{model}} \sum_{l=1}^{L_{model}} \left(\frac{1}{L_{actual} - N} \sum_{j=N+1}^{L_{actual}} \bar{A}_{j,1}^{(l)} \right) \tag{1}$$

where $\bar{A}_{j,1}^{(l)}$ represents the head-averaged attention from token j to position 1 in layer l.

We also computed attention to positions P2-P4 and the P1 Dominance Ratio:

$$\text{P1\_Dom\_Ratio} = \frac{\text{P1\_Attn}}{\sum_{k=2}^{N} \text{Pk\_Attn} + \epsilon} \tag{2}$$

2.3 Statistical Analysis

We employed rigorous statistical testing ($\alpha = 0.05$) including paired-sample t-tests for positional comparisons, one-way ANOVA for group differences, and Tukey HSD for post-hoc analysis. Effect sizes were computed using Cohen's d for practical significance assessment.

3 Results

3.1 Universal P1 Dominance

Our most striking finding is the overwhelming dominance of the first token (P1) as the primary attention sink across all tested architectures and text types. Figure 1 shows P1 receives an average attention score of 0.495 (SD=0.109) compared to P2 (0.013, SD=0.011), P3 (0.011, SD=0.010), and P4 (0.015, SD=0.013).

Paired-sample t-tests confirmed P1's statistical dominance: P1 vs P2 ($t(124) = 36.82, p < .001, d_z = 6.61$), P1 vs P3 ($t(124) = 36.73, p < .001, d_z = 6.59$), and P1 vs P4 ($t(124) = 35.15, p < .001, d_z = 6.31$). These exceptionally large effect sizes indicate fundamental differences in attention allocation patterns.

3.2 Architectural Variations

While P1 dominance is universal, its absolute strength varies significantly across model families. Figure 2 shows a one-way ANOVA revealed significant family effects ($F(2, 122) = 9.49, p < .001, \eta^2 = 0.135$). Post-hoc analysis showed: Post-hoc analysis showed that **Mistral models** exhibited the highest P1 attention (0.567, SD=0.101), while **Llama models** showed intermediate levels (0.497, SD=0.125), and the **GPT-2 family** demonstrated the lowest but still substantial P1 attention (0.471, SD=0.083). Notably, Mistral's Sliding Window Attention architecture showed significantly stronger P1 reliance than both GPT-2 ($p < .001$, Cohen's $d = 1.09$) and Llama ($p = .031$, $d = 0.61$), suggesting that local attention constraints may amplify reliance on global anchors.

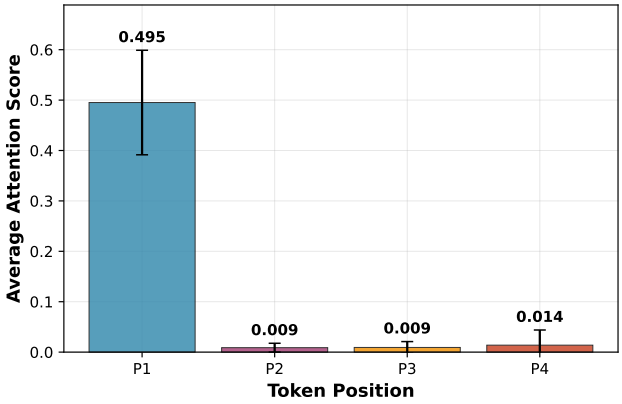

Figure 1: Average attention score received by each of the first four token positions (P1-P4) from subsequent tokens, aggregated across all models and text samples.

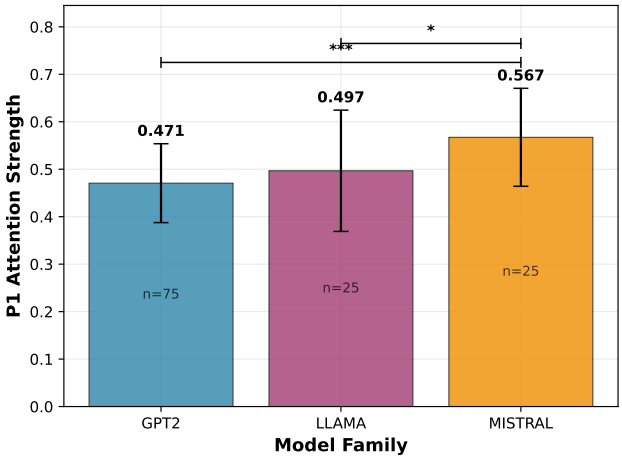

Figure 2: Mean P1 attention strength by model family. Error bars represent standard deviation. Significance indicators: *** $p < .001$, * $p < .05$.

3.3 Content-Dependent Modulation

Input text characteristics significantly modulate P1 attention strength ($F(4, 120) = 100.30, p < .001, \eta^2 = 0.770$). Figure 3 shows the clear hierarchy. Post-hoc analysis revealed a clear hierarchy where **short texts** showed the highest P1 attention (0.665, SD=0.079), followed by **narrative texts** (0.489, SD=0.037) and **technical texts** (0.479, SD=0.032) which demonstrated similar levels of high P1 attention. **Dialogue texts** exhibited moderate P1 attention (0.434, SD=0.045), while **code texts** showed the lowest P1 reliance (0.407, SD=0.045). This hierarchy was consistent across model families, suggesting that P1 reliance inversely correlates with local structural complexity. Short texts lacking internal structure maximize global anchor reliance, while structured code with explicit syntactic patterns shows minimal P1 dependence.

4 Discussion and Implications

4.1 The Singular Anchor Phenomenon

Our findings reveal P1 as a "singular anchor" rather than part of a distributed multi-token sink system. This suggests an emergent optimization where models utilize the most stable position—the first token—as the primary conduit for global contextual information. Unlike sparse attention patterns that reduce computational complexity, P1 sinks provide a complementary strategy for maintaining global coherence.

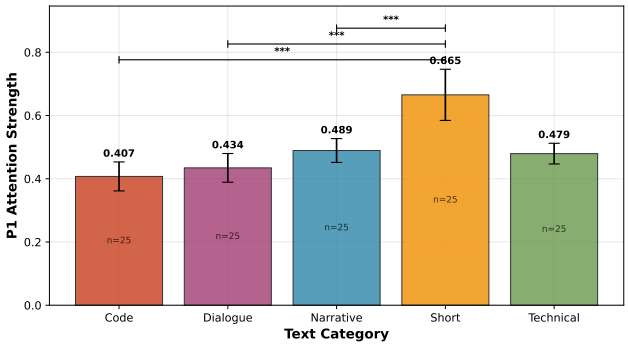

Figure 3: Mean P1 attention strength by text category. Error bars represent standard deviation. Significance indicators: *** $p < .001$.

4.2 Architectural Insights

Mistral's pronounced P1 reliance, despite its Sliding Window Attention design, suggests that local attention constraints may necessitate stronger global anchors. This finding has immediate implications for efficient long-context architectures: hybrid approaches combining sparse local attention with enhanced global anchors merit exploration.

4.3 Practical Applications

These results inform several practical applications:

KV Cache Optimization: Understanding P1's dominant role enables more efficient cache allocation strategies. Instead of treating multiple initial tokens equally, systems could allocate enhanced capacity specifically to P1.

Content-Adaptive Mechanisms: The robust content-dependent patterns suggest that attention mechanisms could dynamically adjust sink allocation based on text type classification, optimizing both efficiency and performance.

Interpretability: P1's consistent role makes it a high-value target for probing global context representation and understanding how models maintain long-range coherence.

4.4 Limitations and Future Work

Our study focuses on decoder-only architectures with specific attention mechanisms. Future work should investigate encoder-decoder models, alternative positional encodings, and causal intervention studies to establish functional necessity rather than mere correlation. Additionally, understanding what specific information accumulates in P1 representations across different contexts remains an open question.

5 Conclusion

This work provides the first systematic empirical characterization of natural attention sinks across diverse LLM architectures. Our key finding—that the absolute first token serves as a dominant singular anchor—challenges assumptions about distributed sink importance and reveals a fundamental self-organizing mechanism in transformer-based context management. While this P1 dominance is universal, its strength varies significantly with both architectural choices and input characteristics. These insights provide crucial foundations for designing more efficient long-context models, optimizing attention mechanisms, and advancing our understanding of how LLMs process and maintain global contextual information.

The identification of P1 as a singular anchor opens new avenues for targeted architectural improvements, content-adaptive attention mechanisms, and interpretability research. Future work should focus on causal validation of P1's functional necessity and development of optimization strategies that leverage these natural patterns for enhanced efficiency and performance.

References

References

[1] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[3] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[5] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[6] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[8] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[9] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.