ON THE EXISTENCE AND BEHAVIOR OF SECONDARY ATTENTION SINKS

Jeffrey T.H. Wong*[1], **Cheng Zhang**[1], **Louis Mahon**[2], **Wayne Luk**[1], **Anton Isopoussu**[2], **Yiren Zhao**[1]
[1]Imperial College London [2]UnlikelyAI
{tsz.wong20,cheng.zhang122,w.luk,a.zhao}@imperial.ac.uk
{louis,anton}@unlikely.ai

ABSTRACT

Attention sinks are tokens, often the beginning-of-sequence (BOS) token, that receive disproportionately high attention despite limited semantic relevance. In this work, we identify a class of attention sinks, which we term secondary sinks, that differ fundamentally from the sinks studied in prior works, which we term primary sinks. While prior works have identified that tokens other than BOS can sometimes become sinks, they were found to exhibit properties analogous to the BOS token. Specifically, they emerge at the same layer, persist throughout the network and draw a large amount of attention mass. Whereas, we find the existence of secondary sinks that arise primarily in middle layers and can persist for a variable number of layers, and draw a smaller, but still significant, amount of attention mass. Through extensive experiments across 11 model families, we analyze where these secondary sinks appear, their properties, how they are formed, and their impact on the attention mechanism. Specifically, we show that: (1) these sinks are formed by specific middle-layer MLP modules; these MLPs map token representations to vectors that align with the direction of the primary sink of that layer, (2) the ℓ_2-norm of these vectors determines the sink score of the secondary sink, and also the number of layers it lasts for, thereby leading to different impacts on the attention mechanisms accordingly, (3) the primary sink weakens in the middle layers, coinciding with the emergence of secondary sinks. We observe that in larger-scale models, the location and lifetime of the sinks, together referred to as sink levels, appear in a more deterministic and frequent manner. Specifically, we identify three sink levels in QwQ-32B and six levels in Qwen3-14B. We open-sourced our findings at github.com/JeffreyWong20/Secondary-Attention-Sinks.

1 INTRODUCTION

Attention sinks were first identified by Xiao et al. (2023), where the BOS token was observed to receive anomalously high attention weights. This phenomenon has since shown broad practical implications, including LLM quantization (Son et al., 2024; Liu et al., 2024), KV-cache optimization (Cai et al., 2025; 2024), efficient LLM serving (Xiao et al., 2023), and model enhancement (Yu et al., 2024).

Many recent studies investigated the formation and functional role of the BOS sink. Cancedda (2024) analyzes attention sinks from a spectral subspace perspective, while Gu et al. (2024) interprets them as a form of positional bias that mitigates over-mixing. Building on this line of work, Queipo-de Llano et al. (2025) further examines their role in the depth-wise organization of the model, proposing that attention sinks serve as a mechanism for information compression along the depth dimension.

Sun et al. (2024) and Yu et al. (2024) show that attention sinks are not limited to the BOS token; instead, multiple tokens can function as attention sinks. Ruscio et al. (2025) further analyzes this phenomenon from a geometric perspective, showing that the emergence of multiple attention sinks is closely tied to the model's positional embedding scheme. In this view, attention sinks act as reference points in a high-dimensional representation space, enabling the model to establish a stable internal coordinate system. However, the multiple sinks identified in previous work are fundamentally the

*Part of this work was carried out during an internship at Unlikely AI.

same as the BOS sink: they emerge at the same layers and persist throughout the network. In contrast, we identify a new type of sink that differs in both its layer of emergence and its lifetime.

In this work, we show that all of the multiple tokens that act as attention sinks can be organized into distinct *sink levels*. The primary level normally corresponds to the BOS sink: it emerges at the same layer as the BOS token and persists throughout the network. Additional sink levels arise in the middle layers and persist for a variable number of layers; we refer to these as `Secondary Sinks`.

We first identify `Secondary Sinks`, especially their *sink levels*, the token sets and positions in which they frequently occur, through characterizing the similarities and differences between `Secondary Sinks` and the BOS sink across a range of models. (Section 3). We then quantify the contributions of different layers to their formation via an empirical analysis of their emergence across network depth (Section 4). Finally, we examine their impact on attention scores throughout the network.(Section 5) . We make the following conclusions:

- Unlike the primary sink, which emerges in early layers and persists throughout the entire network, `Secondary Sinks` arise primarily in middle layers and persist only for a couple of layers. They can be found at any position in the generation sequence and semantically uninformative tokens.
- `Secondary Sinks` shares a similar direction with the primary sink. This direction is encoded in specific **middle-layer l_{start} MLP modules**, which convert multiple orthogonal directions to the same sink direction. After l_{start}, a set of semantically uninformative tokens are transformed into attention sinks. The layers preceding l_{start} play a key role in constructing this set, distinguishing these tokens from other semantically uninformative tokens.
- Different levels of `Secondary Sinks` exhibit distinct lifetimes and attention sink strength. Both are strongly correlated with the ℓ_2-norm of l_{start} output. Larger models show clearer differentiation between sink levels, and models going through extensive post-training on reasoning data show stronger `Secondary Sinks` phenomenon.
- `Secondary Sinks` show a compensating effect relative to the BOS sink. The BOS gradually decays and reaches its weakest strength in the middle layers, coinciding with the emergence of `Secondary Sinks` phenomenon.

2 Background and Preliminary

Let f_θ be a decoder-only transformer with L layers and hidden size h. At each decoder layer l, the decoder receives a hidden sequence of length t, $\boldsymbol{H}^l \in \mathbb{R}^{t\times h} = \{\boldsymbol{h}_0^l, \boldsymbol{h}_1^l, \ldots, \boldsymbol{h}_t^l\}^T$, where $\boldsymbol{h}_i$ is the hidden representation at position i in layer l.

Decoder blocks Each decoder layer l consists of a multi-head self-attention (MHSA) module and a multi-layer perceptron (MLP) , both of which operate on the decoder input hidden state, also referred to as the residual stream $\boldsymbol{H}^l \in \mathbb{R}^{t\times h}$. The MHSA produces an output $\boldsymbol{O}^l \in \mathbb{R}^{t\times h}$, and the MLP produces $\boldsymbol{F}^l \in \mathbb{R}^{t\times h}$; in both cases, the outputs are added back to the residual stream. Decoders may use either a pre-norm or post-norm architecture for which the normalization is applied either before or after each module. The majority of modern models employ pre-norm as shown in Figure 3.

Position Embedding Position embeddings P provide tokens with positional information in the attention mechanism. Common P include absolute positional embeddings (Vaswani et al., 2017), rotary positional embeddings (RoPE) (Su et al., 2024), and NTK[1]-aware scaled RoPE (Peng et al., 2023; Xiong et al., 2024). RoPE is used in models such as LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2023), while NTK-aware scaled RoPE is adopted by Qwen (Team et al., 2024) and CodeLlama (Roziere et al., 2023). Compared to RoPE, NTK-aware scaled RoPE increases the rotary base resulting in less rotation with position (Xiong et al., 2024).

Attention Sink Attention sink refers to the phenomenon where semantically uninformative tokens are assigned disproportionately high attention weights across diverse model architectures and scales, resulting in vertical patterns in the attention weight matrix (Xiao et al., 2023).

[1]NTK stands for Neural Tangent Kernel, this refers to changing the base frequency in RoPE based on the network capability, allowing improved performance on long-context sequences.

Recent work Ruscio et al. (2025) suggests that reducing the rotary frequency weakens the dot-product advantage of the first token. As a result, standard RoPE often produces a strong attention sink at the BOS token, whereas NTK-aware scaled RoPE can give rise to additional attention sinks at other positions in the sequence. In this work, we evaluate 11 model families that use distinct rotary base frequencies (10K-500K). Our findings on QwQ and the Qwen2/2.5/3 model families aligns with Ruscio et al. (2025). However, we observe that CodeLlama and several other model families, despite using very large rotary bases, do not exhibit such secondary sinks. The underlying cause of this discrepancy remains an open question.

In this work, we show that while multiple tokens can act as attention sinks, they can be organised into discrete levels. The primary level corresponds to the BOS sink: these sinks emerge in the same layer as the BOS token and persist throughout the network. Additional levels appear in later layers and persist for varying depths. Building on this, we study the properties of additional sink levels, the variation of this phenomenon across different model sizes and post-training methods, then examine their formation across network depth, and finally analyze their resulting impact on attention.

3 Properties of Secondary Sinks

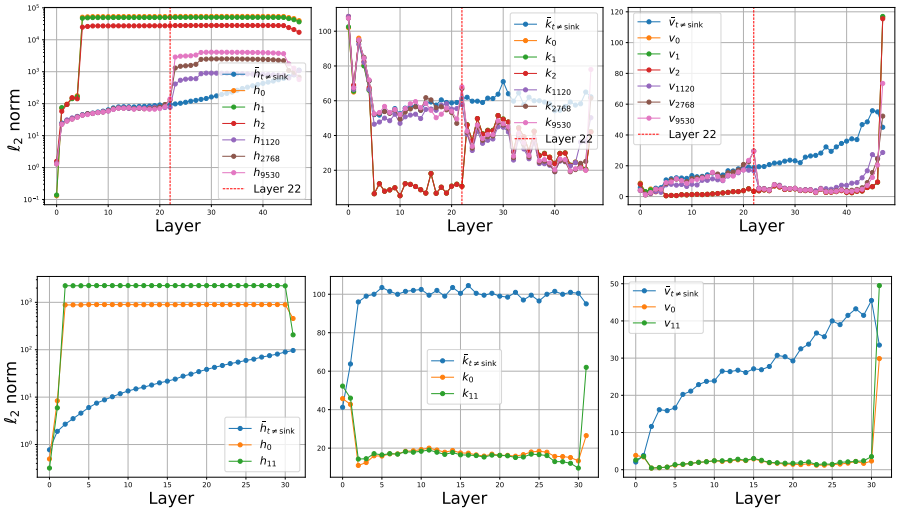

Figure 1: **(Bottom)** In LLaMA2-7B-Chat, the 1st, 11th tokens act as primary sinks: they exhibit significantly larger l_2 norms in their hidden states, while having much smaller key and value norms compared to other tokens. **(Top)** DeepSeek-14B also shows primary sinks (tokens 0, 1, and 2). In addition, it exhibits other attention sinks that emerge in middle layers (e.g., tokens 1120, 2768).

We perform extensive empirical experiments to study the properties of `Secondary Sinks` by comparing them with the BOS sink. To observe the `Secondary Sinks`, we generate reasoning traces from LLMs (DeepSeek-14B, DeepSeek-32B, Qwen-3-32B) on datasets including AIME24 (MAA, 2024), Math (Hendrycks et al., 2021), and then threshold the attention weights to locate sink tokens across different layers. We find that these attention sinks consistently exhibit at least an order of magnitude higher ℓ_2-norms than average token[2] and align along the similar the BOS sink direction.

Based on this observation, we developed an efficient method to identify attention sinks by computing the pairwise cosine similarity between each hidden state $\boldsymbol{h}^l_{t!=0}$ and the BOS token $\boldsymbol{h}^l_0$ at layer l, and then thresholding with cosine similarity > 0.95. We also observe that this phenomenon can be triggered by naturally distributed inputs rather than model-generated inputs. Accordingly, we use reasoning traces generated by DeepSeek-14B on the two datasets and feed them into 11 model families: Deepseek-distill, Qwen2, Qwen2-Math, Qwen2.5, Qwen2.5-Math, QwQ, Qwen3, LlaMA-3.1, Phi-4, Mathtral, CodeLlama. In the main body of the paper, we focus on results from DeepSeek-14B on AIME24; we observe the same behaviour in other models that exhibit secondary sinks (See Appendix Section A).

[2]Tokens excluding attention sinks.

Secondary Sinks Figure 1 (top) shows the ℓ_2-norms of the hidden states, keys, and values of attention sinks, compared with those of an average token in DeepSeek-14B denoted by $\bar{\boldsymbol{h}}_{t\neq\text{sink}}$. Token $\boldsymbol{h}_0$ corresponds to the BOS token, while tokens $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are tokens from the chat template. We classify all of these sinks as *primary sinks*, as they emerge simultaneously and persist for the same number of layers as the BOS sink. In addition to these BOS sinks, we observe *secondary sinks* at positions 1120, 2768, and 9530. These tokens behave like normal tokens up to layer 22, after which they are converted into sinks. Before layer 22, their hidden-state, key, and value norms follow the average-token trend; afterward, they closely track the BOS sink, whose hidden-state norm is several orders of magnitude higher than that of normal tokens, while its key and value norms are lower.

Not all models exhibit secondary sinks. For example, although multiple sinks also appear in LLaMA-7B-Chat Figure 1 (Bottom), they are exclusively primary sinks. Among the 11 model families we analyzed, secondary sinks were found in Qwen2, Qwen2-Math, Qwen2.5, Qwen2.5-Math, Qwen3, and QwQ. Full detection results on all 11 model families can be found in Appendix A. These sinks typically occur in semantically uninformative tokens Table 3 and can appear at any position during the generation Figure 5.

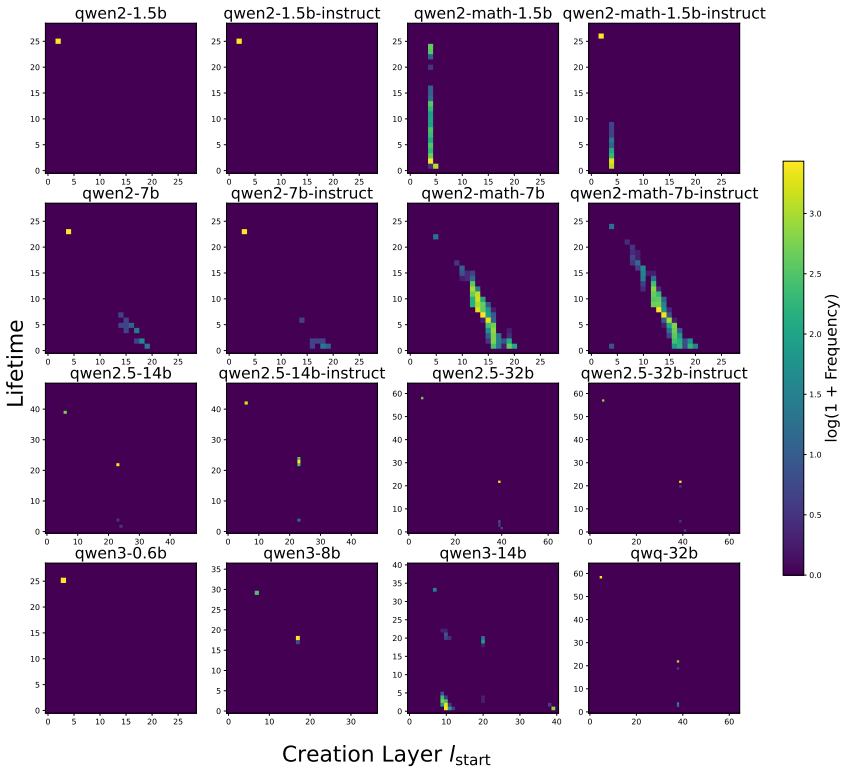

Figure 2: Sink levels (l_{start}, *lifetime*) at different model scales.

Sink Levels As well as differentiating primary from secondary sinks, we can, to some extent, go further and group secondary sinks into multiple 'levels', based on the layer they begin in and how many layers they persist. We associate each sink with an attribute pair (l_{start}, *lifetime*), and refer to sinks with different attribute pairs as belonging to different sink levels. We quantify the distribution of sink levels across models of different sizes and families, as shown in Figure 2.

Secondary sinks are absent or very weak in small base models. They begin to emerge more prominently after mid-training on large amounts of mathematical data, as evidenced by the transition from Qwen2-1.5B/7B to Qwen2-Math-1.5B/7B. This suggests that secondary sinks arise as a mechanism to support enhanced reasoning capability. As model scale increases further, we observe that in the majority of models the number of distinct sink levels decreases: sink lifetimes and creation layers become less dispersed and increasingly concentrated at a small number of characteristic levels across the network.

4 Causal Formation Of Secondary Sinks

The secondary sinks exhibit an intriguing formation patterns across model layers. As shown in Figure 1, they only become apparent as sinks after layer 22; before this layer, their hidden state ℓ_2-norms are similar to those of normal tokens. To investigate their origins, we track tokens that eventually become secondary sinks, and compute their cosine similarity with the BOS sink token as they pass through the different components of layer 22. Here we focus on presenting results on Deepseek-14B on AIME24. Results on other models and datasets can be found in Section A. Our analysis reveals that these sinks primarily emerge from specific MLP modules in the network. Additionally, we perform clustering analysis on early layers and conduct token swapping experiments to trace the formation these sinks back to very early layers of the model.

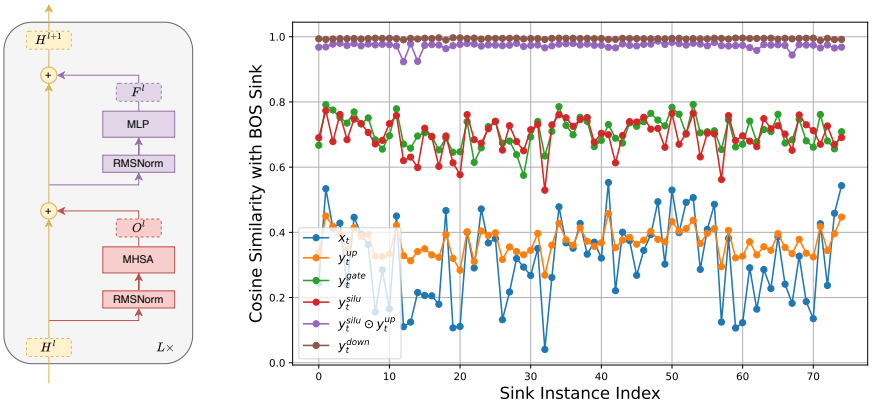

Figure 3: **(Left)** Decoder overview **(Right)** Cosine Similarity Between the Different Secondary Sinks and BOS Sink Token as they go through the Layer 22 MLP in Deepseek-14B.

Cosine Similarity & PCA Analysis For each thinking trace, we compute the cosine similarity between the hidden vector that eventually becomes a secondary sink and that of the BOS sink token at each layer. For simplicity, we name the tokens that eventually become secondary sinks as "future secondary sinks".

Figure 3 (Right) shows the cosine similarity of 75 future secondary sink tokens from AIME24 in DeepSeek-14B (layer 22). The MLP inputs $\boldsymbol{x}_t$ initially have low similarity to the BOS direction, but this similarity increases through the MLP; by the output $\boldsymbol{f}_t$, the vectors are nearly fully aligned with the BOS sink.

This suggests that the sink direction is encoded in the MLP and that future sink tokens contain enough information to recover it. Following the linear hypothesis (Park et al., 2023), we hypothesize that these tokens share a common key that retrieves the sink direction. We test this by performing PCA on the concatenated MLP inputs $\boldsymbol{X}^l$ from 250 sink tokens. The spectrum is strongly low-rank. Feeding the top individual principal components through the MLP shows that these PCs are mapped to the same sink direction. Figure 7 plots the MLP output norm and cosine similarity with the BOS sink when feeding $\pm\alpha\mathrm{PC}_i$. We find that the MLP amplifies components aligned with the BOS sink and suppresses those that are misaligned, producing the large hidden-state norms.

Clustering After identifying the layers at which secondary sinks begin to emerge, we conduct a clustering analysis on the hidden states $\boldsymbol{h}_t^l$, attention outputs $\boldsymbol{o}_t^l$, and MLP outputs $\boldsymbol{f}_t^l$. These vectors correspond to two categories: (1) normal uninformative tokens and (2) future secondary sinks.

As shown in Figure 6, in early layers, normal uninformative tokens and future sinks do not form clearly separable clusters. As the network depth increases, these two categories gradually become distinguishable, forming two increasingly distinct clusters. Both the attention and MLP modules contribute to representations that separate the two types of tokens. Notably, we observe the emergence of two well-defined clusters as early as layer 19. This shows that, although the *effect* of the sink

doesn't emerge until layer 22, the *decision to create the sink* begins in earlier layers and is mostly complete before layer 22 itself.

To better isolate the decision-making contributions from attention, MLP, and the hidden state, we substitute their outputs at future sink tokens with the average semantically uninformative token. See Appendix C for further details.

5 Secondary Sinks Effect on Attention

In this section, we first examine the difference in attention scores between secondary sinks and primary sinks. We then show that secondary sinks exhibit varying sink-score and persist for different numbers of layers. Both the sink-score and the duration across layers are correlated with the norm of the MLP output from which the sinks are formed.

Attention Sink Score We follow Queipo-de Llano et al. (2025) to define the attention sink-score for token position k at layer l as:

$$\text{sink-score}_k^{(l,h)} = \frac{1}{T-k}\sum_{t=k}^{T-1} a_{tk}^{(l,h)}, \qquad \text{sink-score}_k^{l} = \frac{1}{H}\sum_{h=0}^{H} \text{sink-score}_k^{(l,h)} \tag{1}$$

As shown in Figure 4 (Left) , the sink-score for the BOS sink shows a valley shape across the model depth, and it is when it reaches its lowest sink score that the secondary sinks started emerging. This suggests the secondary sinks may serve to compensate for the decay in the BOS sink.

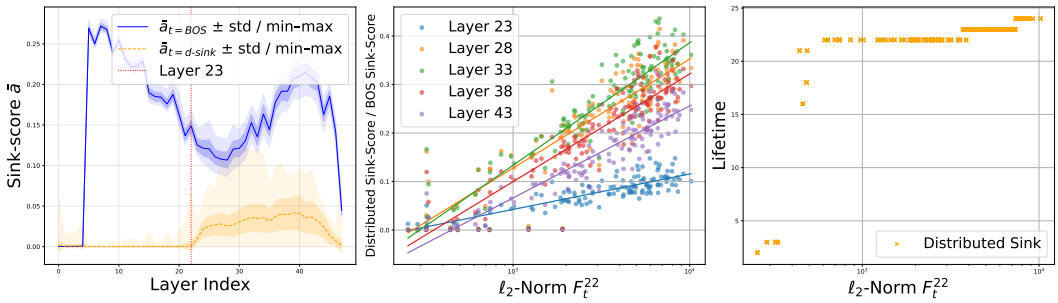

Figure 4: **(Left)** Sink-score comparison between BOS and `Secondary Sinks` across whole model depth. **(Mid)** `Secondary Sinks` Sink-score grows log-linearly with ℓ_2-norm l_{start}. **(Right)** Sink Lifetime growths with ℓ_2-norm l_{start}.

Attention Sink Lifetime We observe that secondary sinks persist for varying depths, ranging from as few as 2 layers to as many as 22 layers, approximately half of the network. As shown in Figure 4 (right), both the sink lifetime and the sink score of the secondary sink are strongly correlated with the ℓ_2-norm of the MLP output $\boldsymbol{f}_t^{22}$ at l_{start}. The ratio between the sink scores of the secondary sink and the BOS sink exhibits a log-linear relationship with $\boldsymbol{f}_t^{22}$, with the largest slope observed at layer 33. In contrast, the sink lifetime increases monotonically with the log-norm, exhibiting distinct linear regimes separated by plateaus.

6 Conclusion

This work identifies a class of multi-token attention sinks that in certain LLMs, that differ from those described in previous works. We empirically study their properties, formation, and influence on attention patterns. We have presented evidence that secondary sinks are fundamentally different from BOS sinks. Investigating the root cause of secondary sinks emergence during pre-training, as well as why post-training processes amplify their effects, is a promising direction for future research. Moreover, its effect on the text generation and downstream performance remains an open question.

REFERENCES

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Zefan Cai, Wen Xiao, Hanshi Sun, Cheng Luo, Yikai Zhang, Ke Wan, Yucheng Li, Yeyang Zhou, Li-Wen Chang, Jiuxiang Gu, et al. R-kv: Redundancy-aware kv cache compression for training-free reasoning models acceleration. *arXiv preprint arXiv:2505.24133*, 2025.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. *arXiv preprint arXiv:2402.09221*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengzhuo Xu, Lu Hou, Jun Yao, and Chun Yuan. Intactkv: Improving large language model quantization by keeping pivot tokens intact. *arXiv preprint arXiv:2403.01241*, 2024.

MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2024*, February 2024. URL `https://maa.org/math-competitions/american-invitational-mathematics-examination-aime`.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. *arXiv preprint arXiv:2508.02546*, 2025.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. *arXiv preprint arXiv:2406.12016*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

A Additional Detection Results

Model Family	Model Name	RoPE Base θ	Primary Sink Layer	Primary Sink Count	Top 3 Secondary Sink Layers	Secondary Sink Count
DeepSeek	deepseek-ai/DeepSeek-R1-Distill-Llama-8B	500000	1	30	X	X
	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	10000	2	60	X	X
	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	10000	3	60	11: 82.55% 12: 9.19%	1284
	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	1000000	4	60	22: 100%	75
	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	1000000	4	90	38: 98.08%, 39: 1.92%	52
Qwen3	Qwen/Qwen3-0.6B-Base	1000000	2	30	X	X
	Qwen/Qwen3-0.6B	1000000	2	30	X	X
	Qwen/Qwen3-1.7B-Base	1000000	2	30	X	X
	Qwen/Qwen3-1.7B	1000000	2	30	X	X
	Qwen/Qwen3-4B-Base	1000000	6	30	16: 100%	86
	Qwen/Qwen3-4B	1000000	6	30	16: 100%	105
	Qwen/Qwen3-4B-Thinking-2507	5000000	6	30	16: 100%	81
	Qwen/Qwen3-4B-Instruct-2507	5000000	6	30	16: 100%	80
	Qwen/Qwen3-8B-Base	1000000	6	30	16: 100%	104
	Qwen/Qwen3-8B	1000000	6	30	16: 100%	134
	Qwen/Qwen3-14B-Base	1000000	6	30	9: 80.51%, 8: 14.07%, 10: 3.13%	6803
	Qwen/Qwen3-14B	1000000	6	30	9: 74.0%, 8: 13.85%, 38: 9.4%	7662
	Qwen/Qwen3-32B	1000000	6	30	9: 38.36%, 8: 36.67%, 7: 23.06%	33727
Qwen2.5	Qwen/Qwen2.5-1.5B	1000000	2	30	X	X
	Qwen/Qwen2.5-1.5B-Instruct	1000000	2	30	X	X
	Qwen/Qwen2.5-7B	1000000	3	30	11: 50%, 12: 50%	32
	Qwen/Qwen2.5-7B-Instruct	1000000	3	30	12: 56.76%, 11: 43.24%	37
	Qwen/Qwen2.5-14B	1000000	5	30	22: 100%	87
	Qwen/Qwen2.5-14B-Instruct	1000000	5	30	22: 100%	94
	Qwen/Qwen2.5-32B	1000000	5	30	38: 98.61%, 40: 1.39%	72
	Qwen/Qwen2.5-32B-Instruct	1000000	5	30	38: 100%	59
Qwen2.5-Math	Qwen/Qwen2.5-Math-1.5B	10000	2	30	4: 100%	2777
	Qwen/Qwen2-Math-1.5B-Instruct	10000	1	30	4: 100%	1622
	Qwen/Qwen2.5-Math-7B	10000	3	30	11: 61.91%, 12: 26.34%, 13: 6.23%	95854
	Qwen/Qwen2-Math-7B-Instruct	10000	3	30	12: 35.68%, 11: 29.37%, 13: 17.78%	13423
Qwen2	Qwen/Qwen2-1.5B	1000000	1	30	X	X
	Qwen/Qwen2-1.5B-Instruct	1000000	1	30	X	X
	Qwen/Qwen2-7B	1000000	3	30	17: 42.86%, 19: 14.29%, 14: 14.29%	7
	Qwen/Qwen2-7B-Instruct	1000000	3	30	18: 40%, 16: 20%, 19: 20%	10
Qwen2-Math	Qwen/Qwen2-Math-1.5B	10000	3	30	3: 87.64%, 4: 12.36%	356
	Qwen/Qwen2-Math-1.5B-Instruct	10000	1	30	3: 100%	103
	Qwen/Qwen2-Math-7B	10000	4	30	12: 33.16%, 13: 21.8%, 14: 14.94%	68709
	Qwen/Qwen2-Math-7B-Instruct	10000	3	30	12: 35.53%, 13: 22.88%, 11: 15.3%	49682
Llama-3.1	meta-llama/Llama-3.1-8B	500000	1	30	X	X
	meta-llama/Llama-3.1-8B-Instruct	500000	1	30	X	X
Phi-4	microsoft/phi-4	250000	2	30	4: 100%	30
	microsoft/Phi-4-reasoning	500000	2	30	X	
	microsoft/phi-4-mini-reasoning	10000	3	30	X	
	microsoft/Phi-4-mini-instruct	10000	3	30	X	
Mathstral	mistralai/Mathstral-7B-v0.1	1000000	1	30	X	X
Code-LLaMA	meta-llama/CodeLlama-7b-hf	1000000	1	30	X	X
	codellama/CodeLlama-7b-Instruct-hf	1000000	1	60	X	X
	meta-llama/CodeLlama-13b-hf	1000000	2	30	X	X
	codellama/CodeLlama-13b-Instruct-hf	1000000	2	30	X	X
QwQ	Qwen/QwQ-32B	1000000	1	30	38: 97.37%, 40: 2.63%	38

Table 1: Model configurations and secondary sink detection results based on 30 traces generated by DeepSeek-14B on AIME24. X indicates that no secondary sinks were detected. *RoPE Base θ* denotes the parameter used in computing the rotary frequency; larger θ means lower frequency. *Primary Sink Layer* indicates the layer at which primary sinks appear. *Top 3 Secondary Sink Layers* denotes the top 3 creation layer l_{start} at which tokens are converted into secondary sinks, with $N\%$ indicating the proportion of total sinks created by that layer. *Secondary Sink Count* reports the total number of detected secondary sink tokens.

Model Family	Model Name	RoPE Base θ	Primary Sink Layer	Primary Sink Count	Top 3 Secondary Sink Layers	Secondary Sink Count
DeepSeek	deepseek-ai/DeepSeek-R1-Distill-Llama-8B	500000	1	500	X	X
	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	10000	2	1000	X	X
	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	10000	3	1000	11: 69.14%, 4: 22.4%, 12: 6.39%	5807
	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	1000000	4	1500	22: 99.53%, 23: 0.47%	214
	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	1000000	4	1000	38: 98.55%, 39: 1.21%; 40: 0.24%	413
Qwen3	Qwen/Qwen3-0.6B-Base	1000000	2	500	X	X
	Qwen/Qwen3-0.6B	1000000	2	500	X	X
	Qwen/Qwen3-1.7B-Base	1000000	2	500	X	X
	Qwen/Qwen3-1.7B	1000000	2	500	X	X
	Qwen/Qwen3-4B-Base	1000000	6	500	16: 100%	504
	Qwen/Qwen3-4B	1000000	6	500	16: 100%	692
	Qwen/Qwen3-4B-Thinking-2507	5000000	6	500	16: 100%	645
	Qwen/Qwen3-4B-Instruct-2507	5000000	6	500	16: 100%	623
	Qwen/Qwen3-8B-Base	1000000	6	500	16: 100%	684
	Qwen/Qwen3-8B	1000000	6	500	16: 100%	829
	Qwen/Qwen3-14B-Base	1000000	6	500	9: 84.96%, 8: 9.37%, 10: 3.5%	36533
	Qwen/Qwen3-14B	1000000	6	500	9: 81.97%, 8: 9.07%, 38: 5.56%	37349
	Qwen/Qwen3-32B	1000000	6	500	8: 37.0%, 9: 34.92%, 7: 26.68%	203789
Qwen2.5	Qwen/Qwen2.5-1.5B	1000000	2	500	X	X
	Qwen/Qwen2.5-1.5B-Instruct	1000000	2	500	X	X
	Qwen/Qwen2.5-7B	1000000	3	500	11: 48.44%, 12: 48.44%, 13: 1.56%	64
	Qwen/Qwen2.5-7B-Instruct	1000000	3	500	12: 54.17%, 11: 41.67%, 15: 4.17%	72
	Qwen/Qwen2.5-14B	1000000	5	500	22: 99.61%, 23: 0.39%	258
	Qwen/Qwen2.5-14B-Instruct	1000000	5	500	22: 100%	222
	Qwen/Qwen2.5-32B	1000000	5	500	38: 98.39%, 39: 1.41%, 40: 0.2%	497
	Qwen/Qwen2.5-32B-Instruct	1000000	5	500	38: 99.35%, 39: 0.65%	309
Qwen2.5-Math	Qwen/Qwen2.5-Math-1.5B	10000	2	500	4: 100%	6667
	Qwen/Qwen2-Math-1.5B-Instruct	10000	1	500	4: 98.74%, 1: 0.87%, 2: 0.38%	4700
	Qwen/Qwen2.5-Math-7B	10000	3	500	11: 55.88%, 12: 28.87%, 13: 7.91%	149922
	Qwen/Qwen2-Math-7B-Instruct	10000	3	500	12: 34.07%, 11: 29.45%, 13: 19.03%	24575
Qwen2	Qwen/Qwen2-1.5B	1000000	1	500	X	X
	Qwen/Qwen2-1.5B-Instruct	1000000	1	500	X	X
	Qwen/Qwen2-7B	1000000	3	500	17: 30.0%, 18: 30.0%, 16: 15.0%	20
	Qwen/Qwen2-7B-Instruct	1000000	3	500	17: 30.0%, 18: 30.0%, 15: 15.0%	20
Qwen2-Math	Qwen/Qwen2-Math-1.5B	10000	3	500	3: 93.04%, 4: 6.9%, 6: 0.06%	1739
	Qwen/Qwen2-Math-1.5B-Instruct	10000	1	500	3: 99.92%, 4: 0.08%	1238
	Qwen/Qwen2-Math-7B	10000	4	500	12: 30.88%, 13: 23.25%, 14: 16.26%	114579
	Qwen/Qwen2-Math-7B-Instruct	10000	3	500	12: 33.96%, 13: 22.37%, 11: 17.11%	82327
Llama-3.1	meta-llama/Llama-3.1-8B	500000	1	500	X	X
	meta-llama/Llama-3.1-8B-Instruct	500000	1	500	X	X
Phi-4	microsoft/phi-4	250000	2	500	4: 100%	500
	microsoft/Phi-4-reasoning	500000	2	500	X	
	microsoft/phi-4-mini-reasoning	10000	3	500	X	
	microsoft/Phi-4-mini-instruct	10000	3	500	X	
Mathstral	mistralai/Mathstral-7B-v0.1	1000000	1	500	X	X
Code-LLaMA	meta-llama/CodeLlama-7b-hf	1000000	1	1000	X	X
	codellama/CodeLlama-7b-Instruct-hf	1000000	1	1000	X	X
	meta-llama/CodeLlama-13b-hf	1000000	2	500	X	X
	codellama/CodeLlama-13b-Instruct-hf	1000000	2	500	X	X
QwQ	Qwen/QwQ-32B	1000000	1	500	38: 95.69%, 40: 2.59%, 39: 0.86%	116

Table 2: Model configurations and secondary sink detection results based on 500 traces generated by DeepSeek-14B on AIME-500. X indicates that no secondary sinks were detected. *RoPE Base* θ denotes the parameter used in computing the rotary frequency; larger θ means lower frequency. *Primary Sink Layer* indicates the layer at which primary sinks appear. *Top 3 Secondary Sink Layers* denotes the top 3 creation layer l_{start} at which tokens are converted into secondary sinks, with $N\%$ indicating the proportion of total sinks created by that layer. *Secondary Sink Count* reports the total number of detected secondary sink tokens.

B Properties of Secondary Sinks

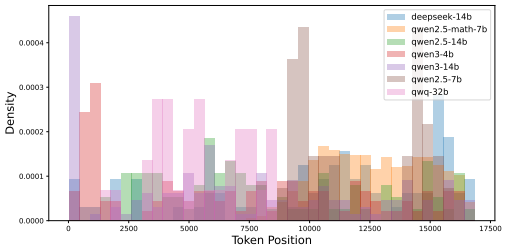

Figure 5: Distribution of sinks across token index.

Model	Secondary Sinks Token Distribution						
Deepseek-14b	" " 56.41%	"1" 17.95%	"^" 5.13%	"2" 5.13%	"," 2.56%	"(" 2.56%	" I" 2.56%
Qwen3-14b	"\n" 19.87%	" " 19.87%	"1" 13.91%	"0" 9.93%	"9" 4.64%	"2" 3.97%	"," 3.97%
Qwen2-7B	"," 6.13%	"1" 5.28%	"2" 3.62%	" " 3.56%	".\n" 3.50%	"0" 3.48%	"4" 2.98%
Qwen2.5-14B	" " 74.71%	"0" 9.20%	"1" 5.75%	"8" 2.30%	"=" 2.30%	"_" 1.15%	"t" 1.15%

Table 3: Frequency of secondary sink tokens.

C CAUSAL FORMATION OF SECONDARY SINKS

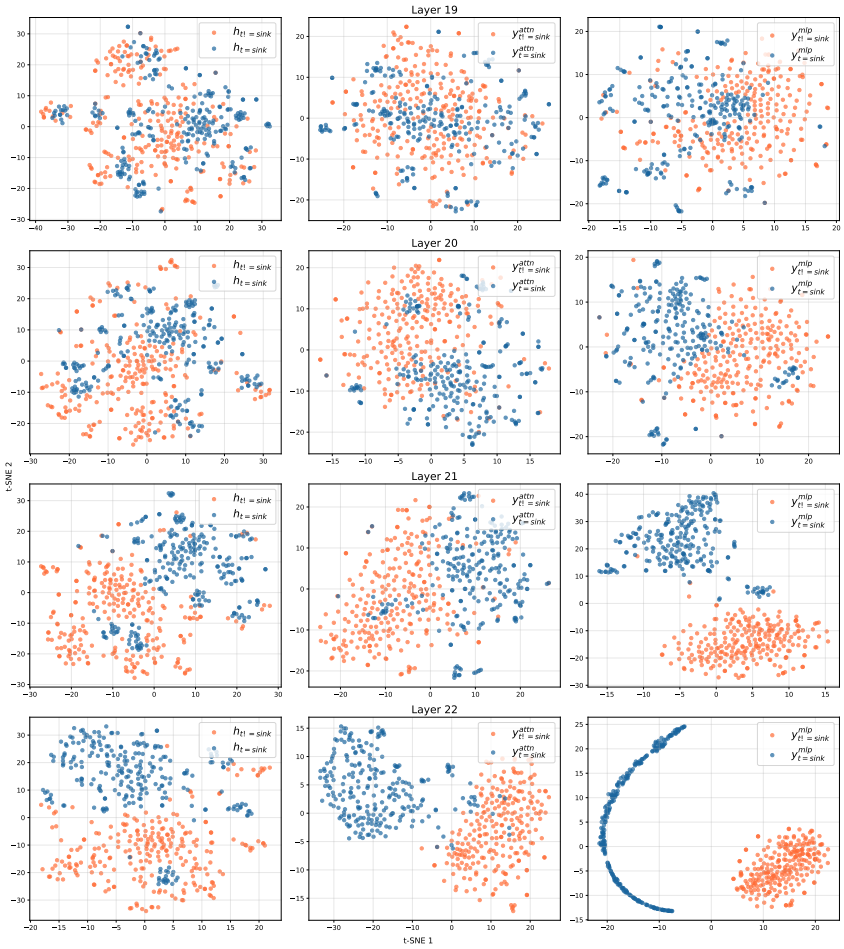

Figure 6: Hidden State, Attention Output and MLP Output t-sne Clustering between normal semantic uninformative tokens and future sinks tokens.

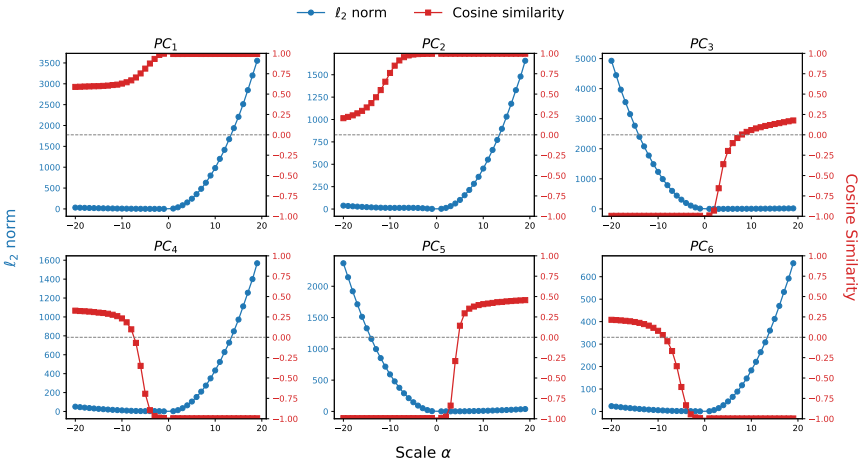

Figure 7: The ℓ_2-norm and cosine similarity of the MLP output relative to BOS sink with $\pm\alpha \mathrm{PC_i}$ as input.

Token Swapping To further verify the formation of secondary sinks in early layers, we conduct token swapping experiments between future secondary sinks and average tokens[3]. By swapping the hidden vectors of these tokens, MLP outputs, and attention outputs at early layers, we observe that we can often suppress the sink from emerging in later layers. The later we swap out the activations, the more effective it is at suppressing the future sink. Figure 8 shows the percentage of the time the future sink is suppressed, as a function of the layer in which it is swapped.

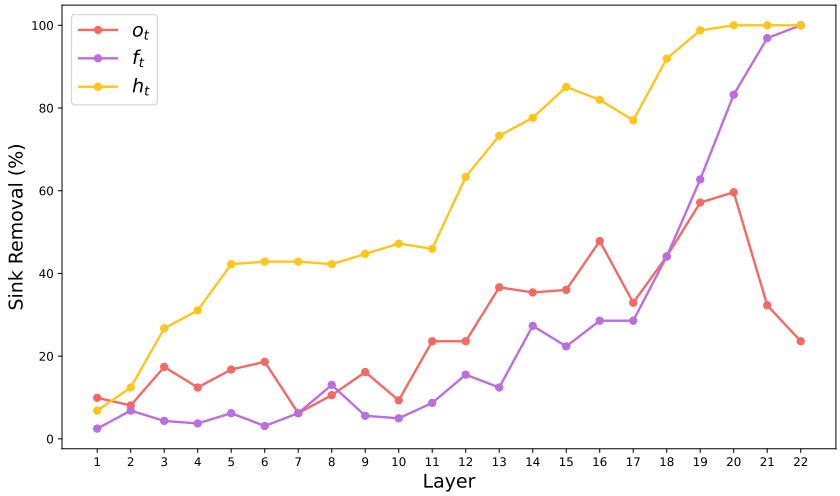

Figure 8: Swapping the activation of the secondary sink at hidden state $\boldsymbol{h}_t^l$, attention output $\boldsymbol{o}_t^l$, and MLP output $\boldsymbol{f}_t^l$ with that of an average uninformative token.

[3]We average the hidden vectors of all non-sink tokens at a given layer to obtain the average token representation.