How Much Attention is "All You Need"?

Stanford CS224N Custom Project

Ignacio Fernandez
Department of Computer Science
Stanford University
ifh919@stanford.edu

Duru Irmak Unsal
Department of Computer Science
Stanford University
duruu@stanford.edu

Abstract

We propose a novel approach to systematically analyze the attention requirements of transformer-based language models and identify the implications for designing these systems. We introduce the concept of *meaningful attention* as the smallest subset of attention weights whose sum exceeds a given threshold. By varying this threshold and evaluating the model's performance, we aim to determine the minimum number of tokens that need to be attended to while maintaining acceptable performance. Our experiment on the long document summarization task using the BookSum dataset demonstrate that the number of attended tokens can be significantly reduced without compromising performance. We provide insights into the sublinear relationship between the context length and the number of attended tokens, suggesting that the number tokens that attention mechanisms meaningfully attend to grows significantly more slowly than the context length. Our main contributions are: (1) a task-agnostic method for determining the optimal subset of attention weights, (2) an empirical analysis on the BookSum dataset showing the effectiveness of our approach, and (3) exploration into a possible attention patterns for transformer-based models.

1 Key Information

- Mentor: Yann Dubois

2 Introduction

The transformer architecture [1] has revolutionized the field of natural language processing, enabling the development of powerful language models that achieve state-of-the-art performance across a wide range of tasks. However, the quadratic complexity of self-attention with respect to sequence length presents a significant challenge when processing long sequences, such as those encountered in document-level tasks of summarization and question answering. This limitation then hinders the efficiency and accessibility of transformer-based models for a wide variety of real-world scenarios.

Existing approaches to mitigate this issue have focused on developing sparse attention mechanisms, which selectively attend to a subset of tokens to reduce computational overhead. While these methods have demonstrated promising results, they often rely on heuristics or predefined attention patterns that may not generalize well across different tasks and datasets. Moreover, the empirical limits on the trade-off between sparsity and performance have not been thoroughly explored for these mechanisms.

In this work, we propose a novel approach to systematically analyze the attention requirements of transformer-based models with respect to context length and use the findings to design an appropriate task-specific attention pattern. Our method introduces the concept of *meaningful attention*, defined as the smallest subset of attention weights whose sum reaches a given threshold of the total attention. By varying this threshold and evaluating the model's performance on a specific task, we then measure the minimum number of tokens that need to be attended to while maintaining acceptable performance.

Our contributions are threefold: (1) We present a task-agnostic method for determining the optimal subset of attention weights based on the notion of meaningful attention; (2) We demonstrate the effectiveness of our approach on the task of long document summarization using the BookSum dataset [2], showing that the number of attended tokens can, in theory, be significantly reduced without compromising performance; (3) We provide initial insights into the relationship between the

sparsity of attention and model performance, paving the way for the development of more efficient transformer-based models tailored to specific tasks and resource constraints.

3 Related Work

Sparse Attention Patterns. One approach to reduce the quadratic complexity of self-attention is to introduce sparsity in the attention matrix. The Sparse Transformer [3] employs a factorized sparse attention mechanism, where each token attends to a local window of tokens and a stride of global tokens. The Longformer [4] combines local sliding window attention with global attention on a few selected tokens, allowing the model to process longer sequences efficiently. BigBird [5] extends this idea by incorporating random attention connections to facilitate information flow across the sequence. While these methods have shown promising results, they often rely on predefined attention patterns that may not be optimal for all tasks.

Hierarchical Architectures. Another line of work focuses on developing hierarchical transformer architectures to handle long sequences. HAT [6] applies attention at multiple levels of granularity, such as sentence-level and document-level, to capture both local and global dependencies. These approaches have demonstrated improved performance on long document tasks, but they require modifications to the standard transformer architecture and may introduce computational overhead.

Attention Analysis. Understanding the behavior of attention in transformer models has garnered significant interest in the research community. Clark et al. [7] analyze the attention patterns in BERT and find that different attention heads specialize in capturing different linguistic phenomena. Kovaleva et al. [8] identify a set of common attention patterns across various transformer models and tasks. These studies provide useful insights into the inner workings of attention mechanisms but do not directly tackle the question of computational complexity on long contexts.

Our work builds upon these prior efforts by proposing a task-agnostic method for determining the optimal subset of attention weights based on their contribution to the overall attention distribution. By systematically analyzing the relationship between attention sparsity and model performance, we provide a more principled approach to optimizing transformer-based models for long sequence tasks.

4 Approach

Our approach aims to find the optimal threshold that preserves performance while minimizing the number of tokens meaningfully attended to. The key steps in our method are: (1) defining and identifying the subset of attention weights that are within a given threshold, (2) determining the optimal threshold for a specific task, and (3) using this threshold to understand the underlying attention pattern for generating a task-specific output.

4.1 Meaningful Attention

We define meaningful attention as the smallest subset of attention weights whose sum is at least equal to a given threshold τ. Formally, for the i-th query vector, this subset, S_i, is given by:

$$S_i = \underset{S \subseteq 1,\ldots,n}{\operatorname{argmin}} |S| \quad \text{subject to} \quad \sum_{j \in S} \alpha_{i,j} \geq \tau \tag{1}$$

where n is the total number of tokens in the context and $\alpha_{i,j}$ is the attention weight assigned by the i-th query vector to the j-th token. This formulation captures the idea that the most important tokens are those that are necessary to account for most of the attention distribution.

4.2 Optimal Threshold Selection

To determine the optimal threshold τ for a given task, we evaluate the model's performance at different threshold values. Let $P(\tau)$ be a performance metric of the model on a specific task when using a threshold τ. We define the optimal threshold τ^* as the maximum value of τ that satisfies:

$$P(\tau^*) \geq P_B + \gamma(P(1) - P_B) \tag{2}$$

where $\gamma \in (0, 1)$ is a hyperparameter that determines the fraction of the performance improvement from the baseline to the full attention (threshold $\tau = 1$) that we wish to retain.

4.3 Attention Masking for Content and Control Tokens

When applying our attention thresholding method to instruction tuned language models, we need to account for the presence of special tokens that are essential for the model's performance. These special tokens, which we refer to as *Control tokens*, include both delimiter tokens and task instruction tokens. To ensure that the model can effectively utilize the task instructions while selectively attending to the most relevant parts of the input and generated output, we modify our attention masking procedure.

Let C be the set of indices corresponding to the *Control tokens*, and X be the set of indices corresponding to the *Content tokens* (i.e., input and generated tokens). We define the smallest meaningful attention subset S_i as a subset of the *Content tokens* X, rather than the entire set of tokens. This allows the model to focus on the most important tokens from the content as well as essential Control tokens. Formally, S_i is given by:

$$S_i = \underset{S \subseteq X}{\operatorname{argmin}} |S| \quad \text{subject to} \quad \sum_{j \in S} \alpha_{i,j} \geq \tau \sum_{j \in X} \alpha_{i,j} \tag{3}$$

Here, S_i represents the smallest subset of *Content tokens* whose attention weights sum to at least a threshold proportion of the sum of all Content weights.

We then modify the attention masking procedure to keep the attention weights of the *Control tokens* intact, while masking out the attention weights of the *Content tokens* that are not part of the smallest meaningful attention subset S_i. The masked attention weights $\hat{\alpha}_{i,j}$ are given by:

$$\hat{\alpha}_{i,j} = \begin{cases} \alpha_{i,j} & \text{if } j \in S_i \cup C \\ 0 & \text{if } j \in X \setminus S_i \end{cases} \tag{4}$$

In other words, if a token is a *Control token* or part of the smallest meaningful attention subset S_i, its pre-normalization attention weight remains unchanged. If a token is a *Content token* but not part of S_i, its attention weight is set to zero. The masked attention weights are then re-normalized to ensure that they sum to 1:

$$\alpha'_{i,j} = \frac{\hat{\alpha}_{i,j}}{\sum_{k=1}^{n} \hat{\alpha}_{i,k}} \tag{5}$$

The attention thresholding and normalization process is detailed in Algorithm 1 and its effect is illustrated in Figure 1.

Algorithm 1 Attention Thresholding and Normalization

Input: Attention weights $\mathbf{a}$, threshold τ, *Control token* indices C, *Content token* indices X
Output: Normalized attention weights $\mathbf{a}''$
Compute *Content token* attention sum: $s \leftarrow \sum_{j \in X} a_j$
Zero out non-*Content token* attention weights: $\mathbf{x}[j] \leftarrow a_j \cdot \mathbf{1}_{j \in X}$
Sort the *Content token* attention weights: $\mathbf{s}, \text{indices} \leftarrow \text{sort}(\mathbf{x}, \text{descending})$
Compute cumulative sums: $\mathbf{c} \leftarrow \text{cumsum}(\mathbf{s})$
Determine threshold crossing index: $k \leftarrow \min\{i \mid c_i \geq \tau \cdot s\}$
Generate sorted mask for Content tokens: $\hat{\mathbf{m}}_X[i] \leftarrow \begin{cases} 1 & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$
Unsort mask for Content tokens: $\mathbf{m}_X \leftarrow \text{unsort}(\hat{\mathbf{m}}_X, \text{indices})$
Generate final mask: $\mathbf{m}[j] \leftarrow \begin{cases} 1 & \text{if } j \in C \\ \mathbf{m}_X[j] & \text{if } j \in X \end{cases}$
Apply final mask to attention weights: $\mathbf{a}' \leftarrow \mathbf{a} \odot \mathbf{m}$
Normalize: $\mathbf{a}'' \leftarrow \frac{\mathbf{a}'}{\sum \mathbf{a}' + \epsilon}$
return $\mathbf{a}''$

Our approach offers a flexible and adaptable method for analyzing the attention requirements of various transformer-based models across different tasks. By considering both input and generated tokens as Content tokens and keeping the attention weights of Control tokens intact, the method can be applied to models with different categorizations of input, such as instruction-based models or those with task-specific special tokens. Moreover, the threshold-based attention masking, which preserves

a proportion of the total attention weight rather than a fixed number of tokens, allows the method to adapt to the particular attention weight distribution of each model and task. This flexibility enables a more accurate and comprehensive analysis of the attention dynamics during the entire generation process, providing insights into what the most relevant parts of the content are for a given task.

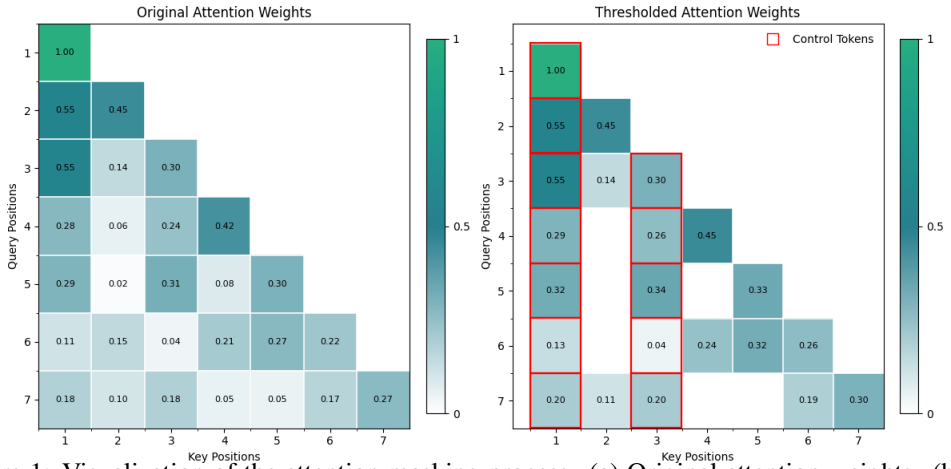

Figure 1: Visualization of the attention masking process. (a) Original attention weights. (b) Re-normalized attention weights after applying a threshold of 0.7 (with Control weights preserved).

The resulting attention pattern is expected to be sparser than the original one, with only the most important Content and the Control tokens being attended to. By applying this method to different context lengths, we can then analyze the relationship between the number of meaningfully attended tokens and the context length.

5 Experiments

5.1 Data

We use the BookSum dataset [2] for our experiments, which contains books with their corresponding chapter-level summaries. The task is to generate concise summaries for each chapter, making it an ideal scenario for analyzing attention patterns in the context of long-form narrative summarization. The chapter-level summaries in BookSum provide a natural way to evaluate the quality of the generated summaries and assess the impact of our attention thresholding approach on the model's performance. Moreover, the varying chapter lengths allow us to investigate the relationship between the number of meaningfully attended tokens and the context length.

5.2 Evaluation Method

We assess the impact of different attention thresholds on the language model's performance using the BERTScore F1 metric [9], which measures the semantic similarity between the generated and reference summaries by leveraging the contextual embeddings obtained from a pre-trained BERT model. BERTScore is more accurate under passagee paraphrases and reorderings, as well as under limited reference summaries, when compared to metrics like ROUGE [10] and BLEU [11], making it suitable for our purposes.

5.3 Experimental Details

We conduct our experiments using the LLaMA 3 8B Instruct model [12] quantized to 8-bit precision due to computational and memory constraints. The model's attention mechanism is modified to incorporate our proposed thresholding approach, as described in Section 4. We evaluate the model's performance at different attention thresholds, ranging from 0.05 to 1.0, with a step size of 0.05, and compute the BERTScore F1 metric between the generated and reference summaries for 199 chapters in the BookSum dataset.

The original code for the LLaMA attention module we modified can be found in the Hugging Face Transformers library [13].

To determine the optimal attention threshold, we compare the model's performance to a baseline where the model summarizes the chapter using only the book and chapter names, without access to the chapter's content. This baseline represents the performance achievable based solely on the model's prior knowledge and provides a more meaningful reference point than the lowest-performing threshold. We define the optimal threshold as the one that achieves at least 95% of the performance improvement from the baseline to full attention.

We acknowledge a potential limitation in our evaluation: the model may have memorized some of the chapters it is being evaluated on. However, the substantial difference in performance (approximately 0.1 on the BERTScore F1 metric) between the baseline (no chapter context) and the full attention model (with chapter) means we can still analyze significant performance variations to draw conclusions on the model's attention requirements. We also investigate this limitation with recent data that the model was not pre-trained on (see Appendix A.1).

Once the optimal attention threshold is determined, we analyze the relationship between the number of meaningfully attended tokens and the context length by applying our attention thresholding approach to the model while varying the input sequence length. For each sequence length across generations of summaries for 199 distinct chapters from the BookSum dataset, we compute and visualize the average number of meaningfully attended tokens.

5.4 Results

Optimal Threshold. Figure 2 illustrates the impact of different thresholds on the model's performance, as measured by BERTScore F1, across different chapter lengths and the overall average. The results demonstrate that a significant reduction in the threshold can be achieved without substantially compromising the model's performance. The similar trends observed across different chapter length buckets indicate that the relative pattern between performances holds for various lengths of summarization tasks.

Interestingly, we observe that the model's performance at a threshold of 0.95 slightly exceeds the performance at full attention (threshold of 1.0). This unexpected result may be attributed to the implicit regularization effect introduced by the attention thresholding process. By focusing on the most important tokens and reducing the influence of less relevant ones, the model may be able to better generalize for generating high-quality summaries. Further investigation is needed to fully understand the implications and replicability of this observation.

Based on the results presented in Figure 2, we identify an attention threshold of 0.75 as the optimal threshold that preserves at least 95% of the performance improvement from our baseline (no chapter text in context) to full attention (with chapter text). This threshold would strike a hypothetical balance between computational efficiency and summary quality, making it a suitable choice for our subsequent analyses. By setting the attention threshold to 0.75, we can then explore the relationship between the number of meaningfully attended tokens and the context length while ensuring that the model remains close to its potential performance.

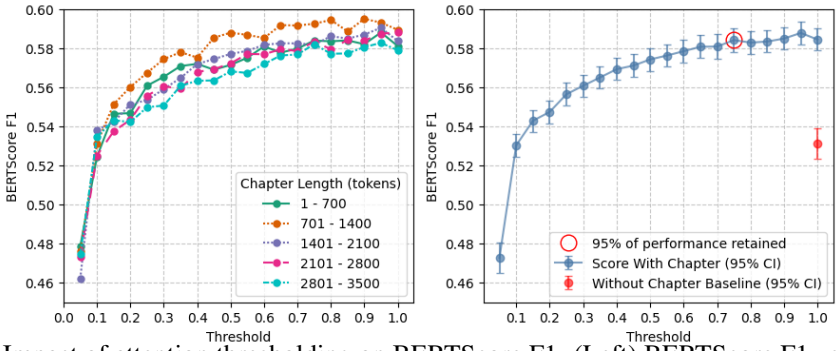

Figure 2: Impact of attention thresholding on BERTScore F1. (Left) BERTScore F1 vs. attention threshold for different chapter length buckets. (Right) Average BERTScore F1 across all chapter lengths vs. attention threshold. The attention threshold of 0.75 (circled) achieves at least 95% of the performance improvement from the baseline (without chapter text) to full attention (with chapter text). Error bars represent 95% confidence intervals. Number of Chapters evaluated: 199

Analyzing Attention. Figure 3 and Table 1 present the results of fitting different models to the relationship between context length and the average number of tokens attended at each generation step, using an attention threshold of 0.75. The scatter plot in Figure 3 shows the observed data points, while the lines represent the fitted square root, logarithmic, and linear models. The square root model (left) provides the best fit to the data, capturing the sublinear relationship between context length and the number of attended tokens. This is further supported by the model performance metrics in Table 1, where the square root model has the highest adjusted R-squared value of 0.612, the lowest Akaike Information Criterion (AIC) of 326300.46, and the lowest mean squared error (MSE) of 428.53 among the three models.

The sublinear relationship between context length and the number of attended tokens, as captured by the square root model, suggests that the number of attended tokens grows at a slower rate compared to the context length. In other words, as the input sequence becomes longer, the attention mechanism attends to a smaller fraction of the total tokens. This finding aligns with the hypothesis that the attention mechanism is able to efficiently identify and prioritize the most informative parts of the input, even as the sequence length increases.

The fitted models with their coefficients are as follows:

$$\text{Square Root Model: } \hat{y} = 9.6179 + 2.3644\sqrt{x} \tag{6}$$

$$\text{Logarithmic Model: } \hat{y} = -217.0985 + 44.1800\ln(x+1) \tag{7}$$

$$\text{Linear Model: } \hat{y} = 54.9845 + 0.0285x \tag{8}$$

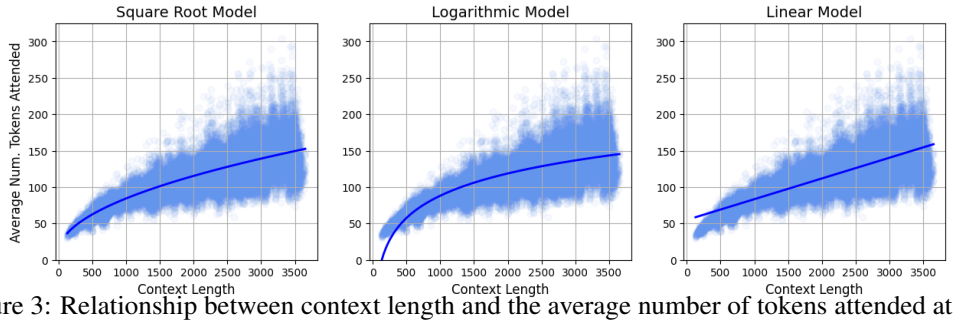

Figure 3: Relationship between context length and the average number of tokens attended at each generation step using an attention threshold of 0.75. The scatter plot represents the observed data points, while the lines show the fitted square root, logarithmic, and linear models. The square root model (left) provides the best fit to the data, capturing the sublinear relationship between context length and the number of attended tokens. Data from summary generations for 199 chapters.

Model	Adjusted R-squared	AIC	MSE
Square Root Model	0.612	326300.46	428.53
Logarithmic Model	0.598	327606.03	444.06
Linear Model	0.600	327439.85	442.06

Table 1: Comparison of model performance metrics for the relationship between context length and the average number of tokens attended at each generation step using a threshold of 0.75. The square root model provides the best fit based on the adjusted R-squared and AIC values, while having the lowest mean squared error (MSE).

While the square root model provides the best fit among the three models considered, it is important to note that this relationship may not necessarily hold true for all transformer-based models or across different tasks and domains. Further research is needed to investigate the generalizability of this finding and to explore other potential models that may better capture the relationship between context length and the number of attended tokens in various settings.

6 Analysis

We design a new attention mask based on the square root model for the relationship between context length and the number of tokens meaningfully attended to. The attention mask is kept the same across all layers as the relationship did not vary significantly across layers (see Appendix A.2). As shown in Figure 4, the square root attention mask includes a buffer of consecutive tokens followed by strides that include tokens at perfect square distances from the buffer. This pattern aims to capture both local and global context while maintaining a sparse attention distribution. In our experiments, we use a buffer size that is approximately 10 times larger than the model's intercept term suggests ($9.6179 \cdot 10 \approx 100$ tokens) and strides that are approximately 10 times more dense than the model suggests (by a factor of $2.3644 \cdot 10 \approx 25$ relative to perfect square distances). We tested the square root attention mask with this over-attending setup to obtain reasonable results without fine-tuning. When applied to the model, the square root mask achieves a BERTScore F1 of approximately 0.53, which is significantly lower than the performance of the model with full attention or the 0.7 threshold.

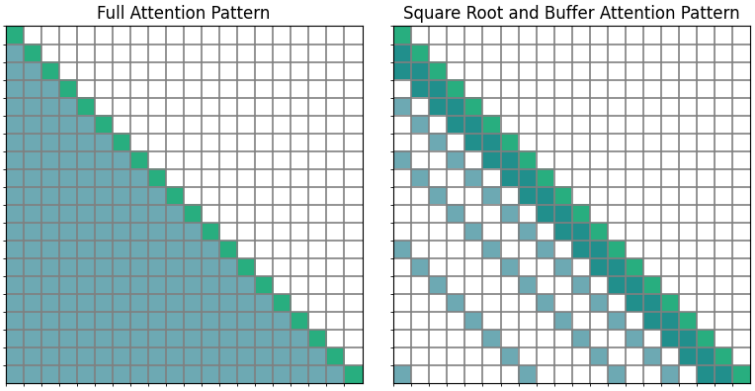

Figure 4: (Left) A visualization of the attention matrix for full attention. (Right) A visualization for an example attention mask with square root scaling. The mask includes a buffer of consecutive tokens followed by strides of tokens at perfect square distances from the buffer.

Generated Summaries

- **Full Attention (Threshold 1.0):** The text is a passage from Mary Shelley's novel "Frankenstein" and describes the early experiences of the monster, created by Victor Frankenstein, as he observes and learns from a family of cottagers. The monster is drawn to the family's gentle and kind nature, and he begins to learn their language and customs. He is fascinated by their relationships and emotions, and he longs to be accepted by them. Despite his physical deformity, the monster is intelligent and capable, and he begins to assist the family by performing tasks such as gathering firewood and clearing their path. As the seasons change, the monster becomes more confident and begins to form his own thoughts and desires, including a desire to be accepted by the family and to learn their language. The passage sets the stage for the monster's later interactions with the family and his own tragic fate.
- **Attention Threshold 0.7:** The text is an excerpt from the novel "Frankenstein" by Mary Shelley. It describes the narrator, the monster created by Victor Frankenstein, as he observes a family of cottagers living in a nearby village. The monster is fascinated by the family's gentle manners, beauty, and kindness towards each other, and he longs to join them and learn more about human society. He begins to learn their language by observing them and making connections between sounds and meanings. As the seasons change and the family's circumstances improve, the monster becomes more confident and begins to perform small acts of kindness for them, such as clearing their path from snow and collecting firewood. He hopes to eventually reveal himself to the family and earn their acceptance, but for now, he is content to observe and learn from them.
- **Square Root Attention Mask:** The text describes the early interactions of a monster with a family, specifically a couple and their young daughter. The monster, who is intelligent and capable, is initially drawn to the family's gentle and kind nature. He begins to assist them with tasks such as gathering firewood and clearing their path. As the seasons change, the monster becomes more confident and begins to form his own thoughts and desires, including a desire to be accepted by the family.

Figure 5: Generated summaries using different attention configurations.

Figure 5 shows the summaries generated for a passage from Mary Shelley's novel "Frankenstein" using full attention, an attention threshold of 0.7, and the square root attention mask. The full attention summary provides a comprehensive overview, while the threshold 0.7 summary maintains the core message with slightly less detail. The summary generated with the square root attention mask lacks depth and fails to capture some key details of the passage.

The relatively poor performance of the square root attention mask can be attributed to naively applying the mask without fine-tuning the model. Fine-tuning would allow the model to adapt its attention patterns to the specific task and dataset, likely resulting in improved summary quality. This highlights the importance of both carefully designing the attention mask as well as training models that maximize performance under these constraints.

While our initial results with the square root attention mask are not impressive, they serve as a proof-of-concept for the potential of using task-specific attention patterns derived from the relationship between context length and the number of attended tokens. Further research is needed to investigate more sophisticated methods for designing attention masks and to explore the impact of fine-tuning on the performance of models with said attention patterns.

7 Conclusion

In this paper, we present a novel approach to analyze the attention requirements of transformer-based language models, introducing the concept of meaningful attention and systematically exploring the relationship between context length and the number of attended tokens. Our findings demonstrate that a significant reduction in the number of attended tokens can be achieved without compromising the model's performance on the task of long document summarization, with the square root model capturing the sublinear relationship between context length and attended tokens.

These insights informed the design of a proof-of-concept square root attention mask, which, despite not outperforming the baseline without fine-tuning, showcases the potential for task-specific attention patterns that balance local and global context while maintaining sparsity. However, the limitations of our work, such as the focus on a single task and the reliance on attention weights as explanatory tools, highlight the need for further research to assess the generalizability of our findings and strengthen the foundations of our method.

Our approach offers a principled, data-driven framework for determining the appropriate level of sparsity in the attention mechanism, potentially leading to more efficient and interpretable attention patterns. The task-agnostic nature of our method makes it applicable to a wide range of domains, providing a foundation for future work on optimizing the attention requirements of transformer-based models across diverse tasks and settings.

8 Ethics Statement

The primary ethical challenges and societal risks of this project include the risk of overgeneralizing findings beyond the studied datasets and domains, which could lead to the development of models that prioritize efficiency over fairness, robustness, or responsible deployment. Without careful consideration of diverse requirements and characteristics across different user groups, linguistic backgrounds, and cultural contexts, the identified optimal attention patterns could inadvertently introduce biases or perpetuate existing inequalities. Additionally, by analyzing attention patterns and identifying potential weaknesses or biases in how language models process information, our method could inadvertently reveal vulnerabilities that bad actors could exploit to generate harmful, biased, or deceptive outputs. To mitigate these ethical risks, we emphasize the importance of further research evaluating the proposed approach across a wide range of datasets, tasks, and domains, with a particular emphasis on incorporating data from underrepresented and marginalized communities. This will help assess the generalizability of the findings, identify potential biases or limitations, and ensure that the developed models are inclusive and equitable.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[2] Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*, 2022. Version 2.

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

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

[5] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020.

[6] Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. An exploration of hierarchical attention transformers for efficient long document classification, 2022.

[7] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention, 2019.

[8] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert, 2019.

[9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[14] AI@Meta. Llama 3 model card. 2024.

[15] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers, 2022.

[16] Matteo Pagliardini, Daniele Paliotta, Martin Jaggi, and François Fleuret. Faster causal attention over large sequences through sparse flash attention, 2023.

[10] [14] [11] [9] [2] [12] [13] [15] [5] [16] [3] [4] [1]

A Appendix

A.1 Evaluating on Recent Data

We evaluated the model with a threshold of 0.75 on a recent news article from June 2024 reporting on the conviction of former U.S. President Donald Trump (link). The summary quality qualitatively shows that our identified threshold still performs well on texts not seen during training.

> Donald Trump, the 45th President of the United States, has been convicted of 34 felony counts related to a scheme to illegally influence the 2016 election through a hush money payment to Stormy Daniels, a porn actress who claimed to have had an affair with Trump. The verdict was delivered by a New York jury after more than nine hours of deliberation. Trump, who denied any wrongdoing, faces up to four years in prison on the charges, which carry a maximum sentence of 25 years. The conviction is a significant legal and political setback for Trump, who is seeking to reclaim the White House in the 2024 election.

A.2 Tokens Attended to Across Layers

We did not observe any meaningful pattern across the different layers.

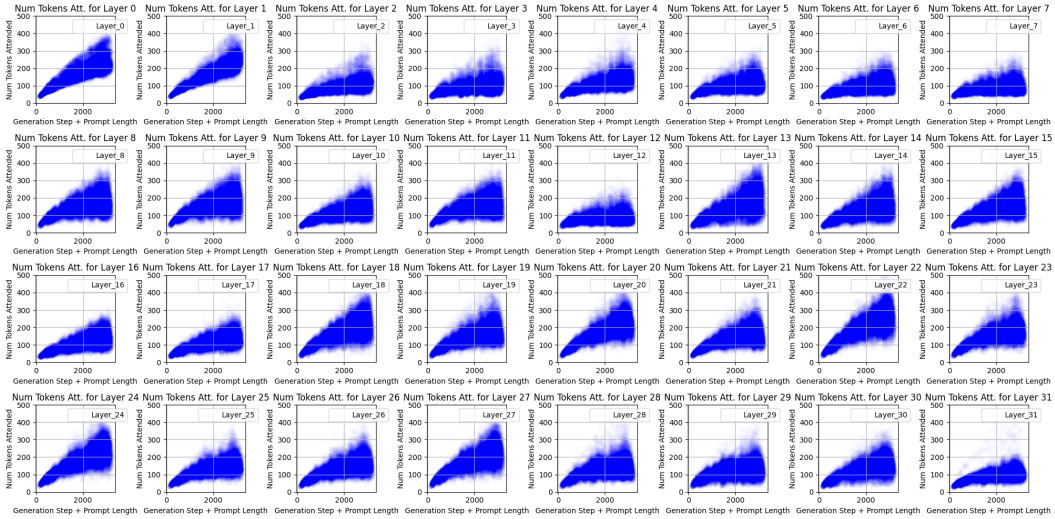

Figure 6: Number of Tokens Attended To Across Layers and Context Lengths

A.3 Tokens Attended to Across Heads

Similarly, we did not observe any meaningful pattern across the different heads.

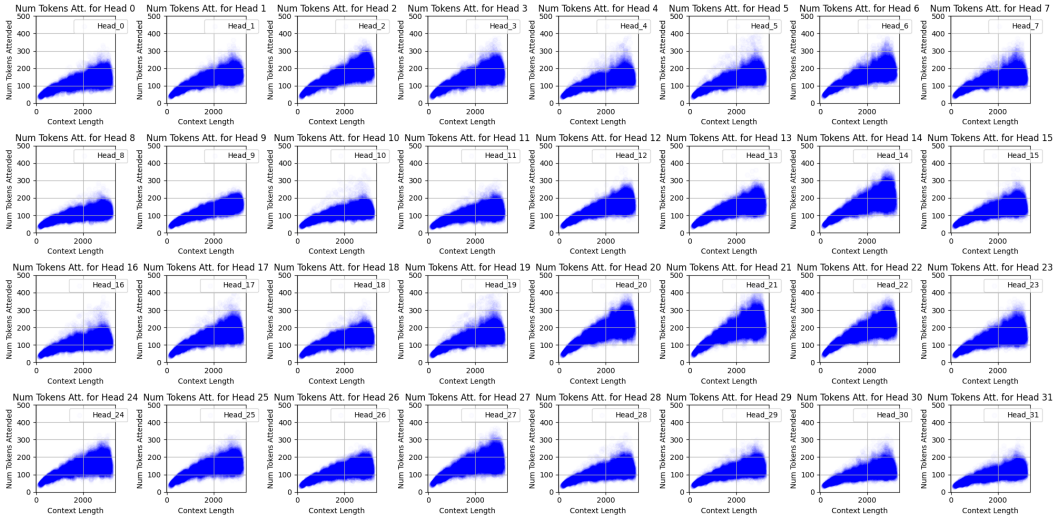

Figure 7: Number of Tokens Attended To Across Heads and Context Lengths