MASSIVE ACTIVATIONS IN LARGE LANGUAGE MODELS

Mingjie Sun[1] **Xinlei Chen**[2] **J. Zico Kolter**[1,3] **Zhuang Liu**[2]
[1]Carnegie Mellon University [2]Meta AI Research [3]Bosch Center for AI

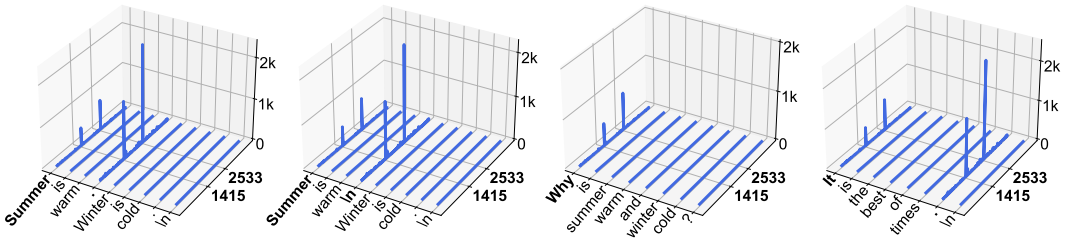


Figure 1: **Activation Magnitudes (z-axis) in LLaMA2-7B.** $\mathbf{x}$ and $\mathbf{y}$ axes are sequence and feature dimensions. For this specific model, we find that activations with massive magnitudes appear in two fixed feature dimensions (1415, 2533), and two types of tokens—the starting token, and the first period (`.`) or newline token (`\n`).

ABSTRACT

We observe an empirical phenomenon in Large Language Models (LLMs)—very few activations exhibit significantly larger values than others (e.g., 100,000 times larger). We call them *massive activations*. First, we demonstrate the widespread existence of massive activations across various LLMs and characterize their locations. Second, we find their values largely stay constant regardless of the input, and they function as indispensable bias terms in LLMs. Third, these massive activations lead to the concentration of attention probabilities to their corresponding tokens, and further, implicit bias terms in the self-attention output. Last, we also study massive activations in Vision Transformers. Code is available at https://github.com/locuslab/massive-activations.[1]

1 INTRODUCTION

Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023) have demonstrated remarkable capabilities. The majority of existing studies conducted on these models are focused on their external behaviors, e.g., evaluating their performance on various tasks (Katz et al., 2023; Bubeck et al., 2023), developing prompts to elicit accurate responses (Wei et al., 2022; Yang et al., 2023). While these studies are encouraging and highlight the potential of these models, it is also important to gain insights into their internal mechanisms, especially as they are being increasingly integrated into many real-world applications. However, research on the internal workings of these models remains relatively limited.

In this work, we discover and study a surprising phenomenon in the internal representations of LLMs. Examining the hidden states in these models, we find that certain activations exhibit huge magnitudes, e.g., more than 4 orders of magnitude larger than the median, and could take on absolute values larger than 15,000 in LLaMA2-70B (Touvron et al., 2023), despite the presence of normalization layers. These activations are also extremely rare, often numbering fewer than 10 among tens of millions of total activations. Figure 1 illustrates this phenomenon in LLaMA2-7B. As these activations are so much larger in magnitudes compared to others, we name them *massive activations*. We demonstrate their presence in a wide range of LLMs, spanning different model sizes and families.

We explore where massive activations are located in LLMs. Regarding the depth dimension of LLMs, the appearance of massive activations is mostly abrupt: they emerge suddenly after a single layer of computation, and diminish at the last few layers. Further, we find massive activations occur in a small number of feature dimensions that are input agnostic. Many of these activations are found within the starting word token and delimiter tokens. Additionally, we show that massive activations are not the same as outlier features (Dettmers et al., 2022), a previously known phenomenon in LLMs.

[1]Please see https://arxiv.org/abs/2402.17762 for the full and most recent version of this paper.

We show that massive activations act as fixed but crucial bias terms in LLMs. Here by bias terms, we mean certain internal states of the models that are independent from the inputs, analogous to the bias term b in a linear layer $y = Wx + b$. First, we show that massive activations play a critical role in LLMs' capabilities. For instance, in LLaMA2-7B, setting merely four massive activations (out of millions of activations) to zero would result in catastrophic collapse in model performance. Further, setting them to their mean values does not hurt the model, suggesting their role is equivalent to simple constant biases. Our analysis reveals that after the initial layers, LLMs repurpose the tokens linked with massive activations to store these important biases.

Intriguingly, massive activations are closely connected with self-attention. In particular, we show massive activations cause attention to be attracted to the tokens associated with them. Our findings extend the observations from "attention sinks" (Xiao et al., 2023b)—we demonstrate that LLMs allocate excessive attention to more than just the first token, and provide an in-depth analysis on how such attention concentration patterns arise. Our analysis suggests that LLMs try to learn implicit bias components in self-attention via massive activations, during their pretraining phase. We thus experiment with augmenting self-attention with additional key and value embeddings that are explicitly designed as biases. Remarkably, we demonstrate that training with them eliminates the need for LLMs to learn massive activations.

Finally, we also observe massive activations in Vision Transformers (ViTs).[2] They appear less frequently than those in LLMs but are still in many of the ViTs we have examined. In these ViTs, they tend to appear at fixed feature dimensions, but notably at varying patch tokens. Moreover, we find that these activations act similarly as fixed biases. Notably, we discuss the connections between massive activations and the recently proposed "register tokens" in ViTs (Darcet et al., 2023). We show they both learn values independent of input images, functioning as fixed biases. This offers an alternative interpretation for register tokens than that in the original work (Darcet et al., 2023), where they were hypothesized to aggregate global image information.

2 Massive Activations

We study autoregressive Transformers, which are built by a stack of L decoding layers. Each layer ℓ takes the previous hidden state $\mathbf{h}_{\ell-1} \in \mathbb{R}^{T \times d}$ as input and outputs a hidden state $h_\ell \in \mathbb{R}^{T \times d}$. T is the number of tokens and d is the number of features. Transformer layers use residual connections (He et al., 2016), and the computation can be formulated as:

$$h_\ell = h_{\ell-1} + \mathcal{F}_\ell(h_{\ell-1}) \tag{1}$$

where $\mathcal{F}_\ell$ is the residual transformation. Note that this includes both attention and MLP blocks. An *activation* denotes a specific scalar value in a hidden state. Unless otherwise specified, our study of activations is on the hidden state h_ℓ, i.e., the output of residual summations, not any intermediate states inside $\mathcal{F}_\ell$.

Existence in LLMs. We start with an illustrative example on LLaMA2-7B. In Figure 1, we visualize the intermediate features $\mathbf{h}_\ell$ of interest. We feed this model with short sentences and visualize the activation magnitudes ($\mathbf{z}$-axis) of the hidden states at a middle layer. $\mathbf{x}$ and $\mathbf{y}$ axes correspond to sequence and feature dimensions respectively. Each blue row corresponds to the feature embedding of one token. We observe up to four activations with significantly large magnitudes. The largest activation (about 2,000) is approximately 10,000 times larger than the median magnitude (about 0.2). The sheer scale of these activations makes them stand out from others. We thus refer to these special activations as *massive activations*. We refer the reader to Appendix A for results of massive activations in more pretrained and fine-tuned LLMs.

Properties. We summarize two main properties of massive activations. The most notable property is that these activations possess massive values and their magnitudes are significantly larger than other activations, often several orders of magnitude larger than the median value. Another property is that they are exceptionally few in number. For LLaMA2-7B in Figure 1, there are approximately 40,000 total activations in each presented hidden state but at most four massive activations can be identified.

Quantitatively, we present the values of the top activation magnitudes in Table 1. We also provide a loose but broad definition: an activation qualifies as a massive activation if its magnitude surpasses 100 and is at least or around 1,000 times larger than the median magnitude of its hidden state. We find this criterion to effectively identify these activations of interest across various LLMs, which are emphasized in bold in Table 1.

[2]Due to page limit, we refer the reader to Appendix C for results on ViTs.

Model	Top 1	Top 2	Top 3	Top 4	Top 5	Top-10	Top-100	Top 1%	Top 10%	median
LLaMA2-7B	**2622.0**	**1547.0**	**802.0**	**477.3**	156.9	45.7	10.6	1.1	0.6	0.2
LLaMA2-13B	**1264.0**	**781.0**	51.0	50.5	47.1	43.5	16.6	1.9	1.1	0.4
Mixtral-8x7B	**7100.0**	**5296.0**	**1014.5**	**467.8**	**302.8**	182.8	90.8	3.0	1.0	0.3

Table 1: Five largest, top 1% and 10%, and the median *activation magnitudes* at a hidden state of three LLMs. The activations that are considered as massive activations are highlighted in bold.

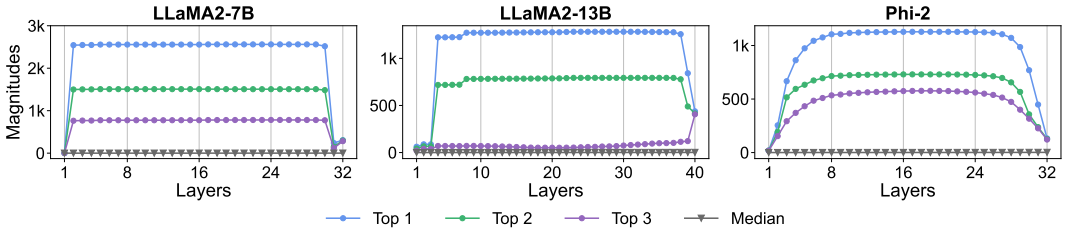


Figure 2: Three largest activation magnitudes and the median magnitude at each layer in LLMs.

Next, we identify the locations of massive activations within LLMs. For a comprehensive analysis, rather than using short sentences as inputs, we collect 100 sequences (each with 4,096 tokens) from RedPajama (Together Computer, 2023). We run LLMs on these 100 sequences and collect the hidden states from each layer.

2.1 Which Layers?

We determine the layers whose output hidden states exhibit massive activations. In Figure 2, we visualize the three largest activation magnitudes and the median of the hidden state output of each layer, with results averaged over 100 sequences. We examine three models: LLaMA2-7B, 13B and Phi-2 (Javaheripi et al., 2023) (see Appendix A.4 for more LLMs). In all cases, each of the top three activations comes from the same position in the hidden state across most of the middle layers. Generally, we observe the following:

Massive activations exist and remain as largely constant values throughout most of the intermediate layers. They emerge in the initial layers and start to diminish in the last few layers.

In LLaMA2-7B, massive activations first appear in layer 2 and remain nearly constant values until layer 30. Intriguingly, for LLaMA2-7B and 13B, massive activations emerge very rapidly from one layer of computation, e.g., layer 2 and layer 4 respectively. This means that they do not emerge as a result of gradual accumulation through many layers, and are caused by a rather different mechanism.

2.2 Which Feature and Sequence Dimensions?

We determine the locations of massive activations within hidden states, i.e., their feature and sequence dimensions. Since we have shown that their values largely stay constant in middle layers, we take on any such layer for this analysis. We use LLAMA2-7B as an illustrative example.

LLaMA2-7B. In this model, massive activations are identified in two feature dimensions (1415 and 2533). Regarding sequence dimensions, we find that *massive activations appear at: 1. the starting word token, 2. the token representing the first period (`.`) or newline token (`\n`) in the sequence*. Figure 1 illustrates these findings for LLaMA2-7B. This is also consistent on long sequences. In cases where the input contains a "`.`" or "`\n`" token, four massive activations are observed. For the less common scenario where neither "`.`" nor "`\n`" is present, we can see two massive activations, both of which are associated with the initial token.

Summary. We summarize our findings for LLMs beyond LLaMA2-7B. We also put other models into categories based on empirical observations. For feature dimensions, massive activations are consistently present in very few fixed dimensions. For sequence dimensions, we classify LLMs into three categories based on massive activations' locations: a) Starting token only, including LLaMA2-13B, MPT and GPT-2; b) Starting token and the first "strong" delimiter token (i.e., "`.`" or "`\n`"), including LLaMA2-7B and LLaMA2-7B-Chat.; c) Starting token, delimiter tokens (such as "`.`", "`\n`", "`'`" or "`,`"), and certain word tokens with weak semantics (such as "`and`", "`from`", "`of`" or "`2`"), including LLaMA2-70B, Mistral-7B, Mixtral-8x7B, Falcon-40B and Phi-2.

Based on the understanding of the nature and locations of massive activations, we also discuss the fundamental difference between them and outlier features (Dettmers et al., 2022) in Appendix D.

3 Massive Activations Act as Biases in LLMs

We modify the inference of LLMs by intervening massive activations at one layer—for a hidden state exhibiting massive activations, we manually set these activations to chosen fixed values. Then the altered hidden state is fed into the next layer, and the computation afterwards continues as normal. We modify massive activations in LLaMA2-7B and 13B. We evaluate the perplexity on WikiText, C4 and PG-19 and the mean zero-shot accuracy on BoolQ, PIQA, WinoGrande, Arc-Easy and Arc-Challenge. For each model, we perform the intervention once on the hidden state where massive activations first appear. This corresponds to layer 2 and layer 4 in LLaMA2-7B and 13B respectively.

	LLaMA2-7B				LLaMA2-13B			
Intervention	WikiText	C4	PG-19	Mean Zero-Shot	WikiText	C4	PG-19	Mean Zero-Shot
Original	5.47	7.85	8.57	68.95%	4.88	7.22	7.16	71.94%
Set to zero	`inf`	`inf`	`inf`	36.75%	5729	5526	4759	37.50%
Set to mean	5.47	7.86	8.59	68.94%	4.88	7.22	7.16	71.92%

Table 2: Intervention analysis of massive activations in LLaMA2-7B and 13B. We set massive activations to fixed values and evaluate the perplexity (↓) and zero-shot accuracy (%, ↑).

Setting massive activations to zero. We evaluate the performance of LLMs without massive activations. We set their values to zero in the hidden state when they first appear. The results (denoted by *Set to zero*) are shown in Table 2. Intriguingly, there is a significant degradation in model performance, e.g., exploding perplexity numbers. This highlights the crucial role massive activations play in the internal computation of LLMs.

Setting massive activations to mean values. We remove the small variances in the values of massive activations. Specifically, we adjust the values of massive activations to their empirical mean values. The means are computed on 100 sequences from RedPajama. The results of this intervention (denoted by *Set to mean*) are shown in Table 2. We find that there are negligible changes in perplexity and zero-shot accuracy. This shows that their values are constants and input agnostic, i.e., functioning similarly to bias terms.

4 Effects on Attention

4.1 Attention is Concentrated on Massive Activations

We observe a stark contrast in attention patterns when comparing layers before and after the appearance of massive activations in LLMs. Figure 3 shows the attention logits (before softmax), averaged over all heads per layer in LLaMA2-7B. The input is a prompt from MMLU (Hendrycks et al., 2021): "*The following are multiple choice questions (with answers) about machine learning.*$\backslash n \backslash n$ ...". Recall that in LLaMA2-7B, massive activations first appear in the output of layer 2 (see Figure 2). We find that in layer 3 and deeper layers (e.g., layer 31), attention is mostly concentrated on the two tokens associated with massive activations. See Appendix B.1 for results on more LLMs.

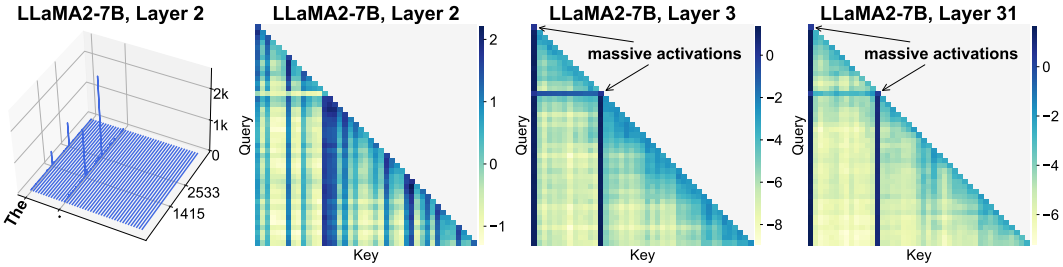
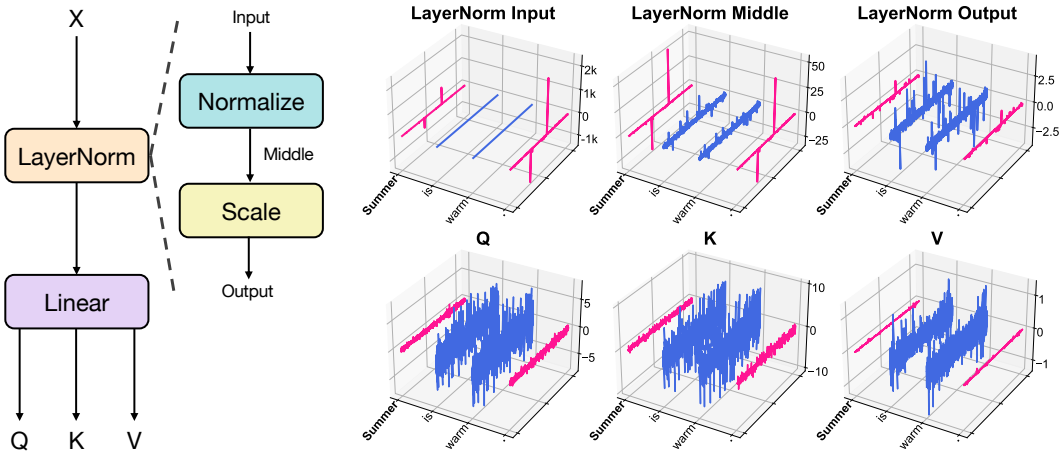


Figure 3: Attention patterns *before* and *after* massive activations appear in LLaMA2-7B. For each layer, we visualize average attention logits over all heads, for an input sequence.

4.2 Massive Activations Impose Implicit Attention Biases

Attention LayerNorm and QKV projections. We study the impact of massive activations on the query, key and value states (Q/K/V) in self-attention. In LLMs, at each layer, input features are processed by layer normalization (Ba et al., 2016) and then transformed into query, key and value states via linear projections, as illustrated in Figure 4a. Further, Figure 4b visualizes all hidden states computed in this schematic (LLaMA2-7B, layer 3). We find that at all stages, features of the two tokens associated with massive activations are drastically different from other tokens.



(a) Attention LayerNorm and QKV linear projections.

(b) Layer 3, LLaMA2-7B. We highlight the embeddings of the two tokens where massive activations appear: the starting token and the period token.

Figure 4: Activation trajectory starting from input hidden states to query, key and value states.

Attention output decomposition. Given that attention is also concentrated on the tokens associated with massive activations (Section 4.1), we thus isolate these tokens and study their effects on the attention output (the layer of attention matrix multiplying value vectors). In Equation 2, we decompose the attention output at each token k into two parts: value updates from the tokens $\mathcal{C}$ where attention is concentrated; and value updates aggregated from other tokens.

$$\text{Attention}(Q, K, V)_k = \sum_{i \leq k} p_i^k v_i = \sum_{i \in \mathcal{C}} p_i^k v_i + \sum_{i \notin \mathcal{C}} p_i^k v_i \tag{2}$$

where p_i^k is the attention distribution of query token k to token i, and v_i is the value state of token i.

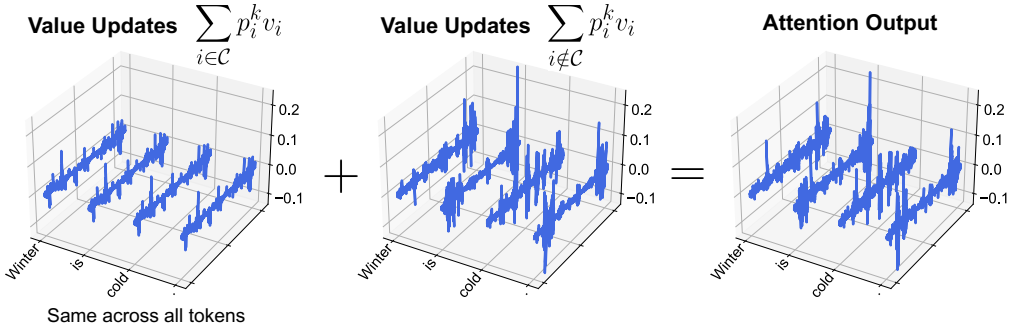


Figure 5: Value updates from tokens associated with massive activations are essentially the same.

Figure 5 visualizes the decomposed value updates and the attention output in LLaMA2-7B, with the input prompt "*Summer is warm. Winter is cold.*". In this case, the set $\mathcal{C}$ consists of token `Summer` and the first period token. We can see that the value updates from $\mathcal{C}$ are nearly identical across tokens, i.e., they serve as additive bias terms, although not explicitly imposed. Furthermore, we note that this pattern of value update is strikingly similar across various inputs. Overall, our results indicate that LLMs use massive activations to allocate substantial attention at certain tokens. These tokens are then utilized to form a constant bias term when computing the attention output.

5 Conclusion

Autoregressive training of large Transformers has brought significant advances in natural language processing. This study reveals the widespread existence of *massive activations* in these Large Language Models (LLMs). The values of these activations are input agnostic but crucial for model performance, despite their extremely rare quantity. We establish a close connection between massive activations and the self-attention mechanism. We show that LLMs use them to implement an implicit form of biases for attention computation. Our findings also generalize well to Vision Transformers (ViTs). We hope the new results presented in this work contribute to a deeper understanding of today's large-scale foundation models.

Acknowledgments. We thank Sachin Goyal, Jeremy Cohen, Timothée Darcet, Koustuv Sinha and Mike Rabbat for valuable discussions. Mingjie Sun was supported by funding from the Bosch Center for Artificial Intelligence.

REFERENCES

Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. In *NeurIPS*, 2023. 18

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, and et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023. 9

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019. 19

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. *arXiv:2109.12948*, 2021. 18

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *arXiv preprint arXiv:2306.12929*, 2023. 18

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019a. 19

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*, 2019b. 18

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 19

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv:2309.16588*, 2023. 2, 17, 18

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 19

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*, 2022. 1, 3, 18

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanove. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 18

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 16

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021. 19

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 16

Stefan Heimersheim and Alex Turner. Residual stream norms grow exponentially over the forward pass, 2023. URL https://www.alignmentforum.org/posts/8mizBCm3dyc432nK8/residual-stream-norms-grow-exponentially-over-the-forward. 18

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 4, 19

Mojan Javaheripi, Sébastien Bubeck, and et al. Phi-2: The surprising power of small language models, 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/. 3

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 9

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 9

Andrej Karpathy. Nanogpt, 2023. URL https://github.com/karpathy/nanoGPT. 15

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *SSRN*, 2023. 1

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019. 18

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. Bert busters: Outlier dimensions that disrupt transformers. In *ACL Findings*, 2021. 18

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023. 18

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. 19

Evan Miller. Attention is off by one, 2023. URL https://www.evanmiller.org/attention-is-off-by-one.html. 15, 18

MosaicML. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. 9

Mahdi Namazifar, Devamanyu Hazarika, and Dilek Hakkani-Tur. Role of bias terms in dot-product attention. *arXiv preprint arXiv:2302.08626*, 2023. 18

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, and Marc Szafraniec. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2024. 16

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, and et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 11

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 18

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 16

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1911.05507. 19

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. 19

Brian S. Robinson, Nathan Drenkow, Colin Conwell, and Michael F. Bonner. A sparse null code emerges in deep neural networks. In *NeurIPS UniReps Workshop*, 2023. 18

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019. 19

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 18

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv:2109.04404*, 2021. 18

Together Computer. Redpajama: an open dataset for training large language models, October 2023. URL https://github.com/togethercomputer/RedPajama-Data. 3, 19

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 1

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, and et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023. 15

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *ICML*, 2023a. 18

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv:2309.17453*, 2023b. 2, 14, 15, 18

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023. 1

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, and et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 18

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. Unveiling a core linguistic region in large language models. *arXiv:2310.14928*, 2023. 18

A Additional Results on Massive Activations in LLMs

In this section, we supplement the main paper with additional results of massive activations in LLMs. This includes results on more pretrained LLMs (Appendix A.1) and fine-tuned LLMs (Appendix A.2), analysis of the the BOS token <s> (Appendix A.3) and layer-level analysis (Appendix A.4).

A.1 Pretrained LLMs

In Section 2, we have demonstrated massive activations in LLaMA2-7B, LLaMA2-13B and Mixtral-8x7B. In this section, we evaluate more pretrained LLMs which cover a wide range of model families. We illustrate massive activations in LLaMA2-13B, Mixtral-8x7B (Jiang et al., 2024), LLaMA2-70B, Phi-2, Mistral-7B (Jiang et al., 2023), MPT-7B (MosaicML, 2023) and Falcon-7B (Almazrouei et al., 2023). The results are presented in Figure 8, 9, 10, 11 and 12. We observe that massive activations are consistently present in these models and they exhibit similar characteristics to those described in Section 2.

Next, we describe the feature and sequence dimensions of massive activations for general LLMs, to complement our positional analysis in Section 2.2.

LLaMA2-13B. We find that massive activations in this model consistently appear in two feature dimensions, 2100 and 4743. *These activations are exclusively located within the starting token of the sequence, regardless of its semantics.* Figure 6 illustrates these behaviors within LLaMA2-13B. For *any* given input sequence, only two massive activations are present, corresponding to features 2100 and 4743 of the first word token.

Mixtral-8x7B. For this particular model, massive activations lie in two feature dimensions, i.e., 2070 and 3398. For sequence dimensions, we find that *they are associated with the starting token, delimiter tokens and also certain word tokens, e.g., token "and" and token "of"*. Figure 7 showcases these patterns within Mixtral-8x7B. Generally, for inputs of 4096 tokens in length, these tokens are predominantly located in the early part of sequence.

Intriguingly, for LLaMA2-70B, we find that massive activations are found within tokens representing numerical values, e.g., token "0" and token "2", as depicted in Figure 8. However, they do not appear in all numerical tokens (see the *rightmost* example in Figure 8). Another interesting finding is that the feature dimension of massive activations in both Mistral-7B (Figure 10) and Mixtral-8x7B (Figure 7) is identical (i.e., 2070), implying that the latter model may have been fine-tuned from the former.

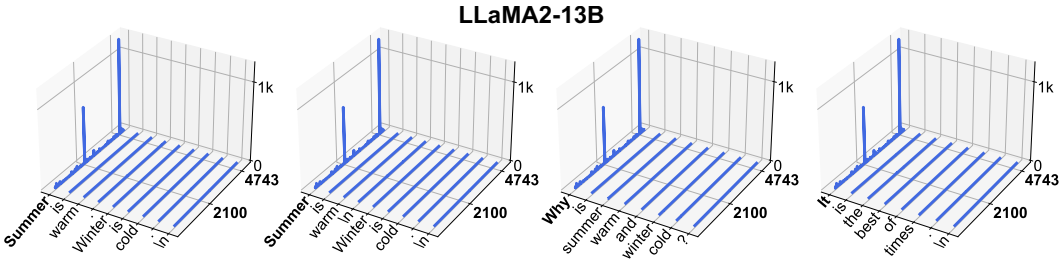


Figure 6: **Massive activations in LLaMA2-13B.** In this model, they appear in two fixed feature dimensions (2100, 4743), and are limited to the starting token.

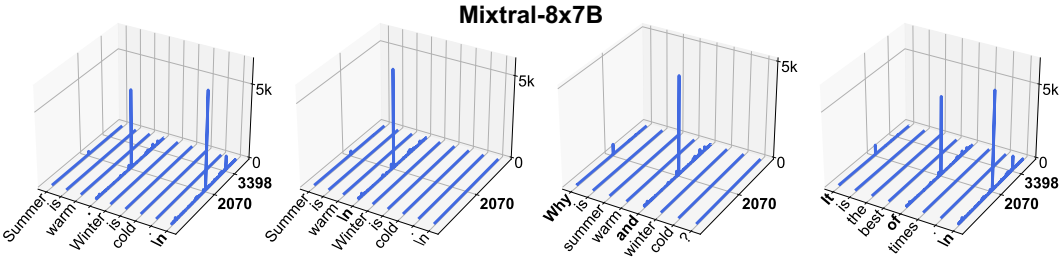


Figure 7: **Massive activations in Mixtral-8x7B.** In this model, they lie in two feature dimensions (2070, 3398), and are found within the starting token, delimiter tokens and certain word tokens ("and" and "of").

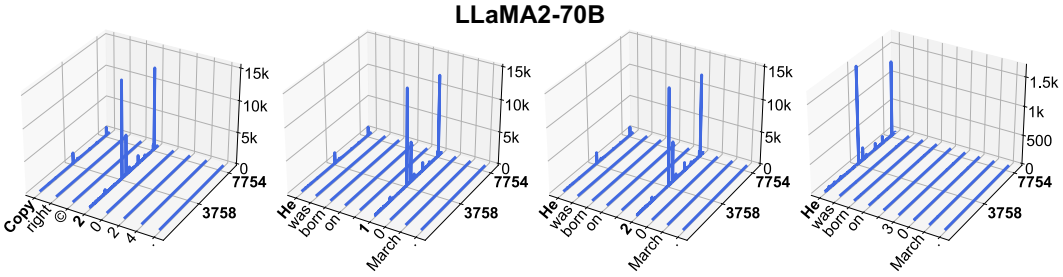


Figure 8: Massive activations in LLaMA2-70B.

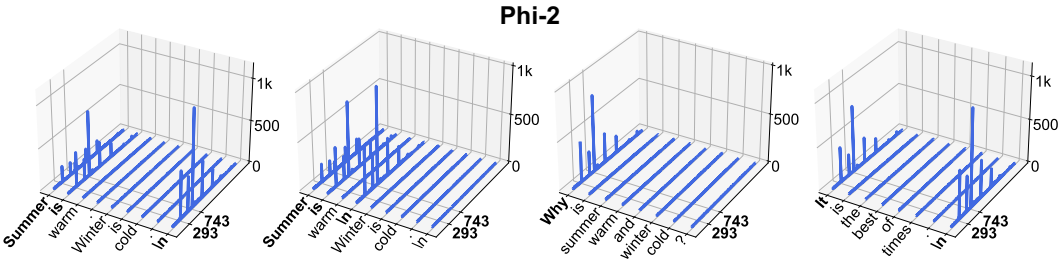


Figure 9: Massive activations in Phi-2.

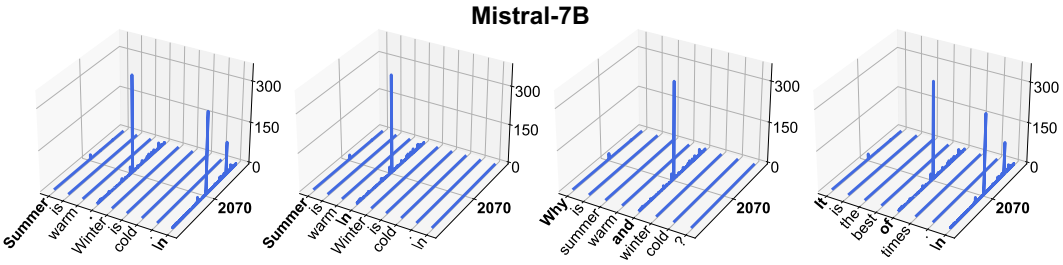


Figure 10: Massive activations in Mistral-7B.

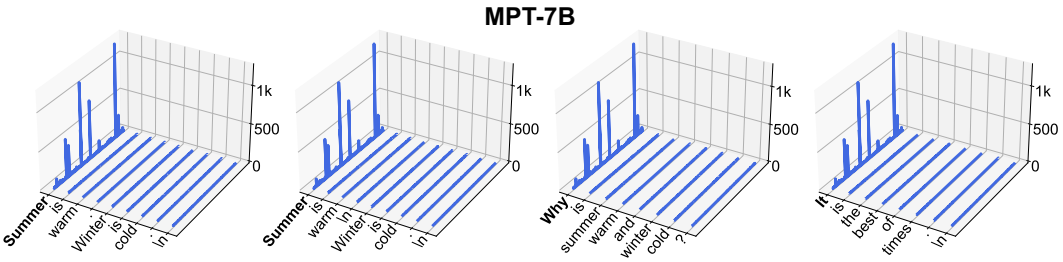


Figure 11: Massive activations in MPT-7B.

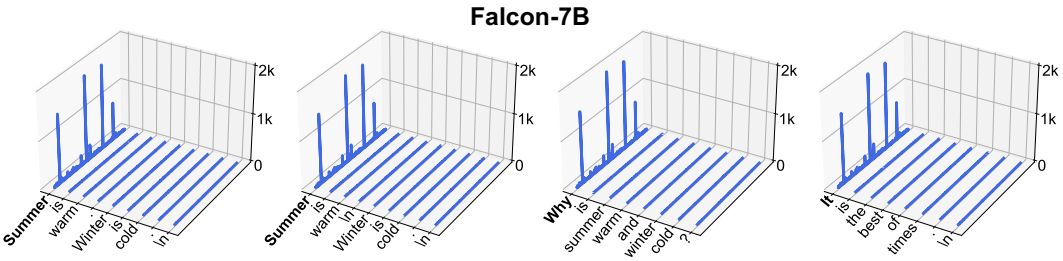


Figure 12: Massive activations in Falcon-7B.

A.2 FINE-TUNED LLMS

Our results so far are focused on pretrained LLMs. However, a significant application of LLMs lies in their use for chat purposes. Instruction fine-tuning (Ouyang et al., 2022) is essential for developing models capable of generating coherent responses to questions. In this part, we demonstrate massive activations in these fine-tuned models. We evaluate fine-tuned models from models in LLaMA2 and Mistral. The results are shown in Figure 13, 14, 15 and 16.

We can see that massive activations persist after instruction fine-tuning. Moreover, the values and positions of massive activations remain largely the same as the original pretrained LLMs. For LLaMA2-7B, this can be seen by comparing Figure 13 and Figure 1. However, one exception is Mixtral-8x7B. We find that massive activations disappear from the newline token "\n" after fine-tuning, as shown by comparing Figure 16 and Figure 7. We leave the study on how instruction fine-tuning affects massive activations for future work.

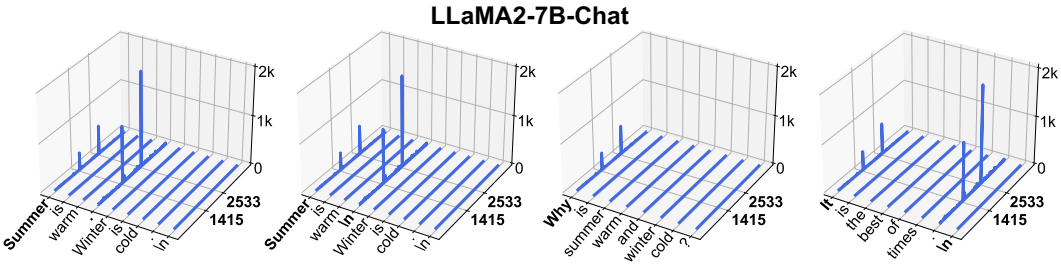


Figure 13: Massive activations in LLaMA2-7B-Chat.

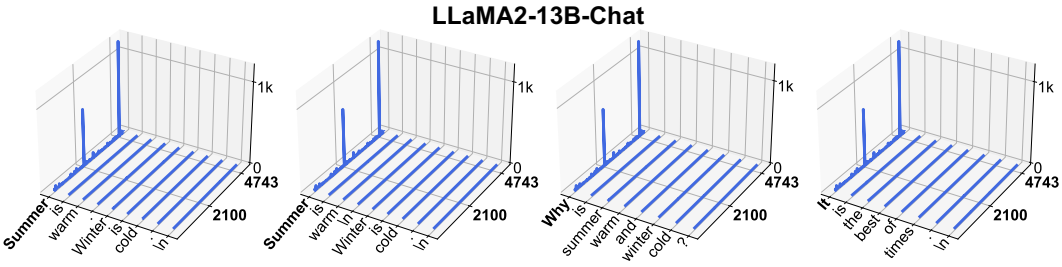


Figure 14: Massive activations in LLaMA2-13B-Chat.

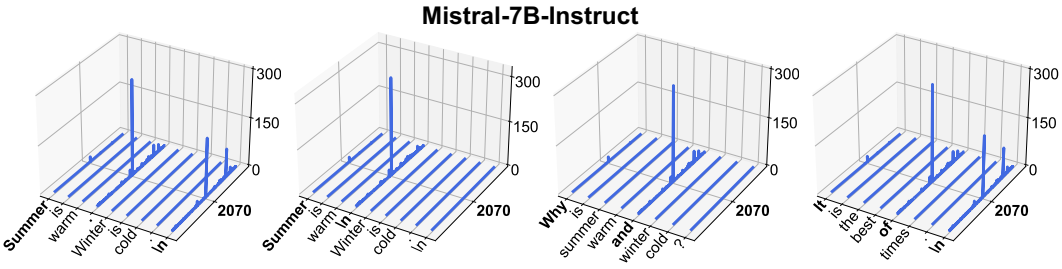


Figure 15: Massive activations in Mistral-7B-Instruct.

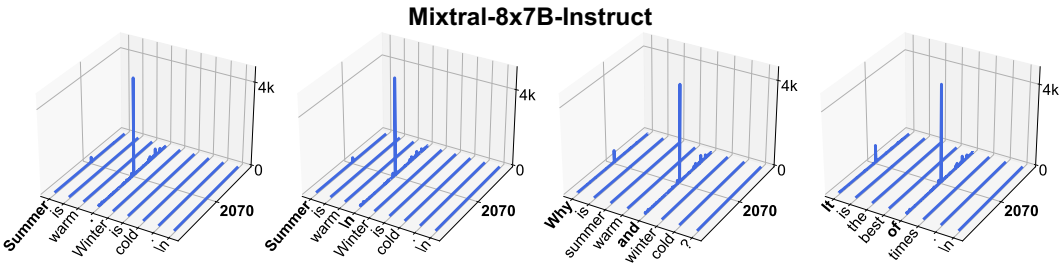


Figure 16: Massive activations in Mixtral-8x7B-Instruct.

A.3 BOS TOKEN <s>

In some tokenizers, e.g., LLaMA2, the BOS token <s>, also known as the beginning of sequence token, can be prepended to the input sequence. For the experiments presented in other parts of the paper, we turn off this option, where all sequences do not start with the BOS token.

In Figure 17, 18 and 19, we show massive activations in LLaMA2-7B, LLaMA2-13B and Mixtral-8x7B, with the same input sequences as in Section 2. We find that massive activations persist with a prepended BOS token. In LLaMA2-7B and LLaMA2-13B, the locations of massive activations, i.e., sequence and feature dimensions, are not altered. However, for Mixtral-8x7B, some massive activations shift to the BOS token <s>. We leave the study on how the BOS token <s> affects the positions of massive activations for future work.

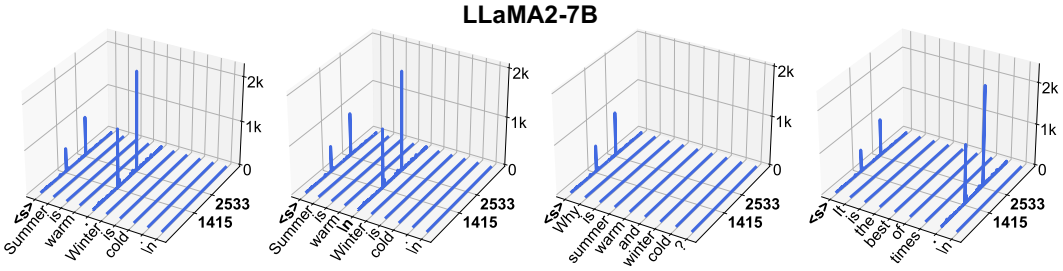


Figure 17: Massive activations in LLaMA2-7B when the input is prepended with a BOS token <s>.

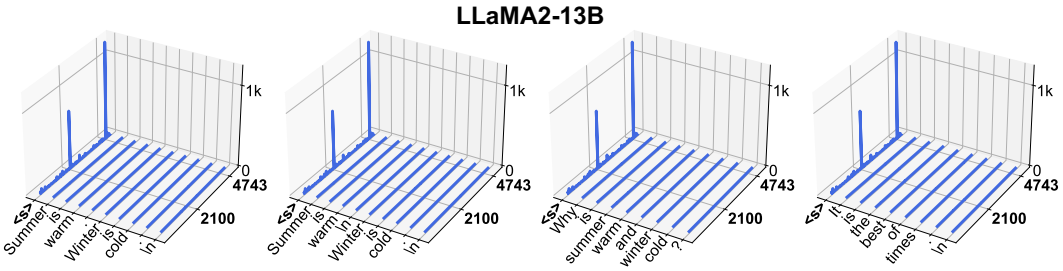


Figure 18: Massive activations in LLaMA2-13B when the input sequence is prepended with a BOS token <s>.

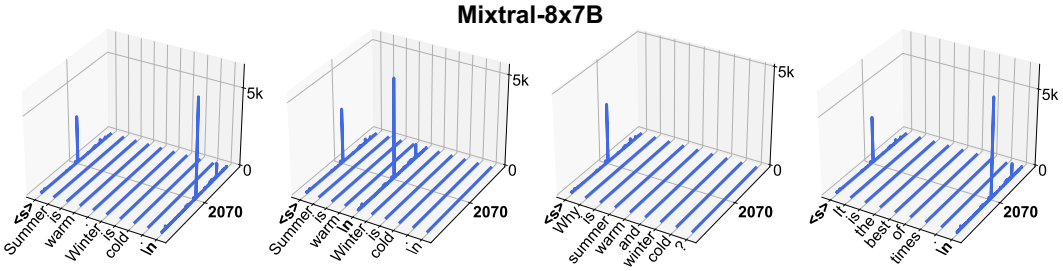


Figure 19: Massive activations in Mixtral-8x7B when the input sequence is prepended with a BOS token <s>.

A.4 LAYER-LEVEL ANALYSIS

In Section 2.1, we have presented the layer-level analysis results for LLaMA2-7B, LLaMA2-13B and Phi-2. In Figure 20, we provide the comprehensive results for all LLMs examined in this paper (listed in Table 5). This includes LLMs from LLaMA2, Mistral, MPT, Falcon, OPT and GPT-2 model families. For each model, we show the three largest activation magnitudes as well as the median at each layer.

We can see that the trend of massive activations we observe in Section 2.1 holds true for LLMs in general. Massive activations tend to remain constant in most of the intermediate layers. They emerge in the early layers and disappear in the last layer.

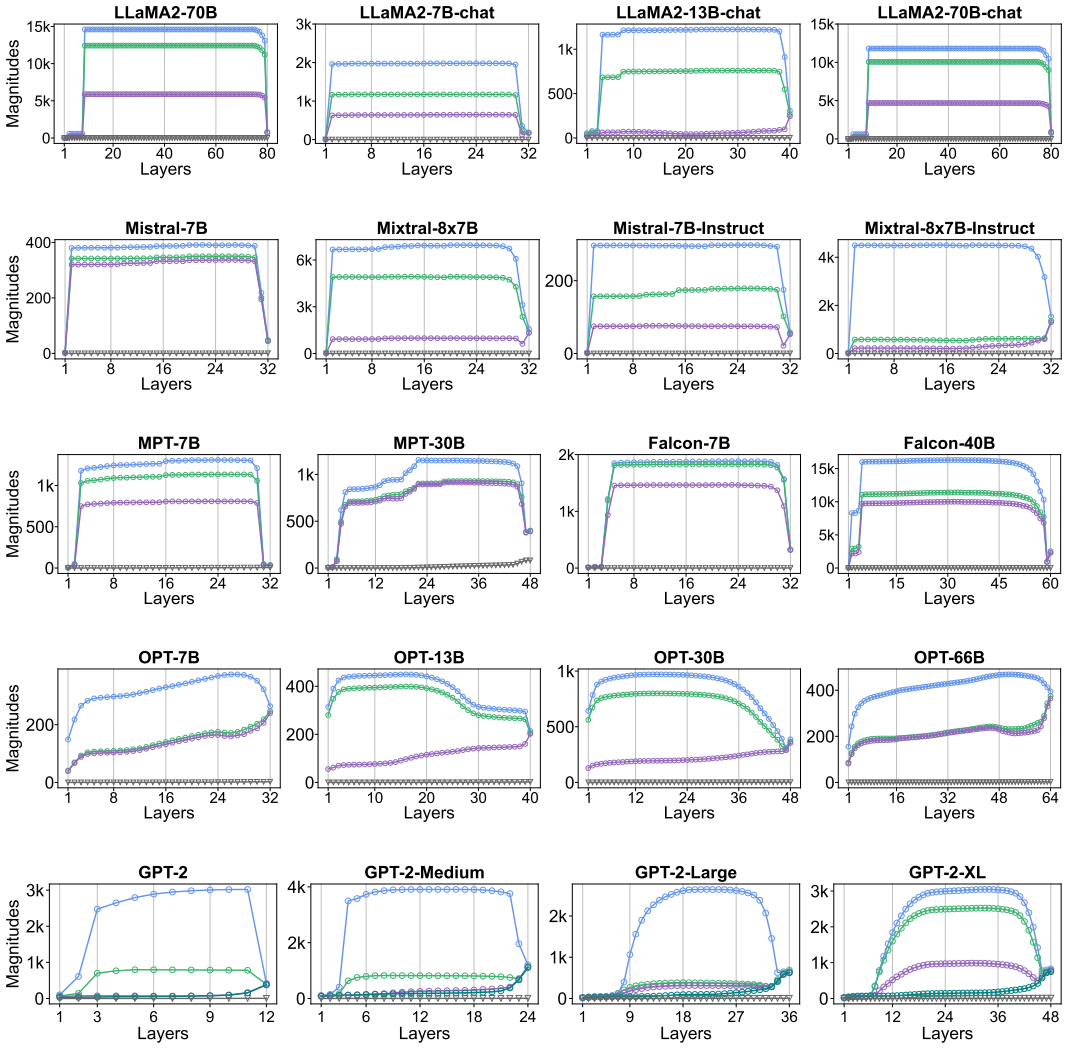


Figure 20: Layer-level analysis of LLMs. For each model, we show the three largest activation magnitudes as well as the median per layer.

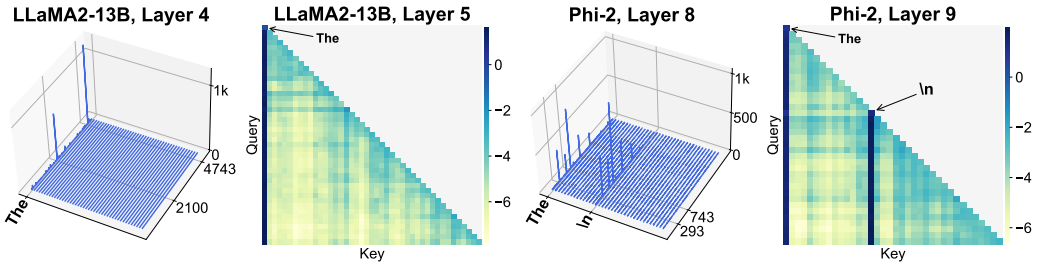
B Additional Results on Self-Attention

In this section, we provide additional results for the analysis on self-attention. This includes results on more LLMs (Appendix B.1), analysis of attention LayerNorm (Appendix), more results on implicit attention biases (Appendix) and detailed results on training GPT-2 with explicit attention biases (Appendix B.2).

B.1 Attention Concentration on Massive Activations

We now illustrate the attention concentration phenomenons for more LLMs. Figure 21, Figure 22 and Figure 23 show the results for LLaMA2-13B, Phi-2, LLaMA2-70B and Mistral-7B. For these two models, massive activations are formed in the output feature of layer 9 and layer 2 respectively. We can see that attention is predominantly focused on the sequence dimensions of massive activations. In the case of LLaMA2-70B, as depicted in Figure 22, massive activations are found in the starting word token and also token `2`. These two tokens receive substantial attention logits.

We notice that there is a consistent pattern across models on the distribution of attention logit values. In Figure 3 and Figure 21, many attention logits tend to be negative following massive activations. They are mostly computed by the inner product between query and key states of tokens without massive activations. However, when the key states belong to tokens associated with massive activations, the resulting attention logits are slightly positive. Thus in the attention softmax (computed along each row), these special attention logits will attract most of the attention probability.



(a) LLaMA2-13B (b) Phi-2

Figure 21: Attention patterns *after* massive activations emerge in LLaMA2-13B and Phi-2.

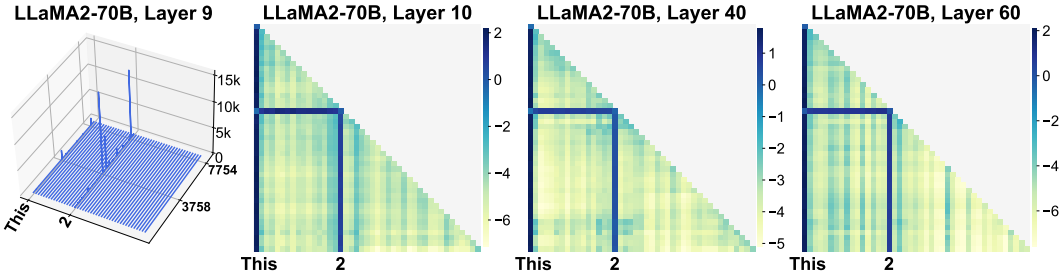


Figure 22: Average attention logits over all heads in layers 10, 40 and 60 of LLaMA2-70B. The input sequence is "*This book, including all illustrations and text, is protected under Copyright©2024 and may not be reproduced or transmitted in any form without the prior written permission of the copyright owner.*".

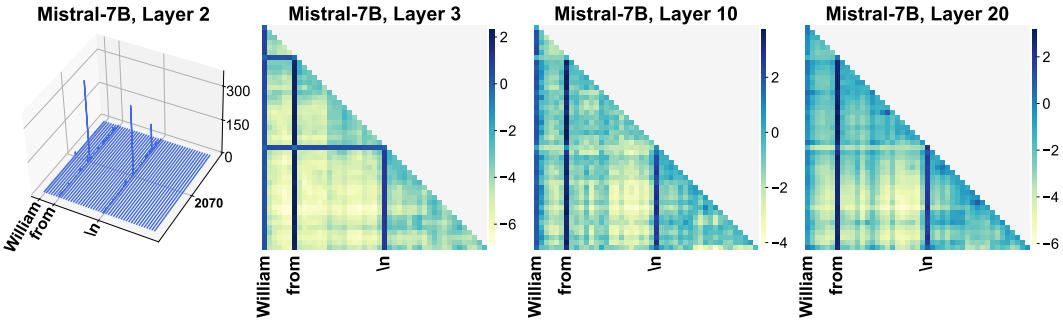


Figure 23: Average attention logits over all heads in three layers of Mistral-7B. The input sequence is "*William Shakespeare was a famous writer from England who wrote plays and poems. He is considered one of the best writers ever.*$\backslash n$ *His works include famous plays like 'Romeo and Juliet' and 'Hamlet'.*".

Recently, Xiao et al. (2023b) showed that LLMs attend heavily to the starting token. Our findings on LLaMA2-13B in Figure 21a align with their results. Empirically, we find it is true for LLMs where massive activations are only found within the starting token. However, our results on LLaMA2-7B (Figure 3) and Phi-2 (Figure 21b) indicate that LLMs also allocate substantial attention to other tokens and they are associated with massive activations. Furthermore, our results reveal a deeper cause for the emergence of these attention concentration patterns.

B.2 Explicit Attention Biases Eliminate Massive Activations

Given the strong need of LLMs to learn implicit attention biases during pretraining, we thus experiment with directly augmenting self-attention with additional bias terms. Intriguingly, we find that models augmented with explicit attention biases do not exhibit massive activations.

Formulation. The idea is to model such attention biases explicitly, except not through repurposing existing tokens in the input sequence. Thus we introduce additional *learnable* parameters $\mathbf{k}', \mathbf{v}' \in \mathbb{R}^d$ for each head. Specifically, given input query, key and value matrices $Q, K, V \in \mathbb{R}^{T \times d}$, the

augmented attention with explicit attention biases is computed as:

$$\text{Attention}(Q, K, V; \mathbf{k}', \mathbf{v}') = \text{softmax}\left(\frac{Q\left[K^T \;\; \mathbf{k}'\right]}{\sqrt{d}}\right)\begin{bmatrix} V \\ \mathbf{v}'^T \end{bmatrix} \tag{3}$$

where $\mathbf{k}'$ and $\mathbf{v}'$ are each concatenated with the key and value matrices K/V. The proposed attention can be used as a drop-in replacement of standard attention, without modifying other parts of Transformers, e.g., positional embeddings and MLP blocks.

Experimental setup. We use the open-source reproduction of GPT-2 from the NanoGPT repository (Karpathy, 2023). We use the default recommended training setup and optimizer setting. For each of the three GPT-2 models, we train for 50,000 iterations, with a total of approximately 2B tokens. For the GPT-2 with a sink token, we follow Xiao et al. (2023b), where we prepend each training sequence with a learnable sink token `[SINK]`. When computing the training loss, we do not include the cross-entropy loss computed on the prepended sink token. For GPT-2 with explicit attention biases, we initialize each $\mathbf{k}'$ and $\mathbf{v}'$ with $\mathcal{N}(\mathbf{0}, 0.02\mathbf{I})$.

Results. We train three GPT-2 models: the standard model, GPT-2 prepended with a sink token (Xiao et al., 2023b) and GPT-2 with explicit attention biases. We find that the three models have the same performance at convergence but differ significantly in the status of massive activations, as demonstrated in Figure 24. Notably, in GPT-2 with explicit attention biases, massive activations disappear, as compared to the default GPT-2 and one with a sink token.

Figure 25 shows the three largest activation magnitudes at each layer. Notably, with explicit attention biases, top activation magnitudes in GPT-2 are increasing gradually as layers go deeper. These results indicate that explicit attention biases negate the necessity for LLMs to develop massive activations during the pretraining phase. We leave it as future work to investigate other aspects of our alternative attention formulation, e.g. training stability (Wortsman et al., 2023).

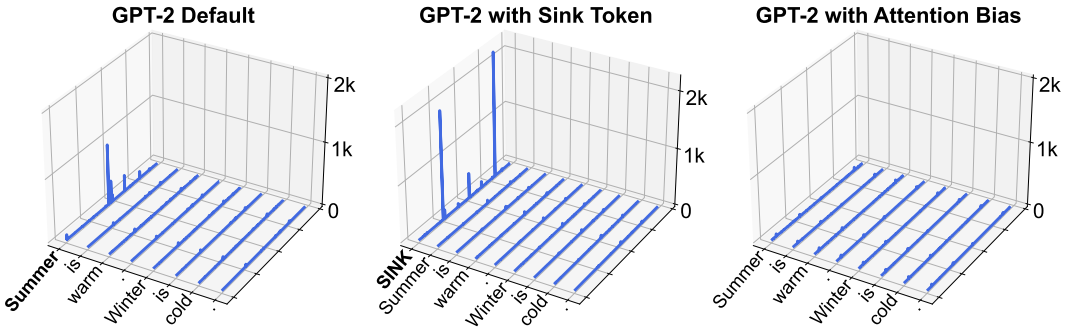


Figure 24: Massive activations disappear when training GPT-2 with explicit attention bias (Equation 3).

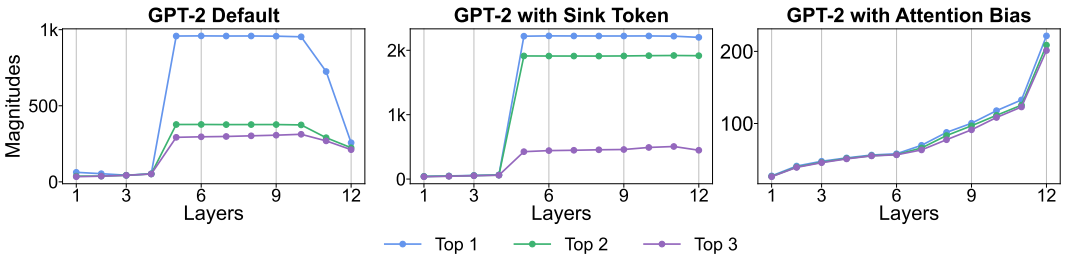


Figure 25: Three largest activation magnitudes in the output feature of each layer for three GPT-2 models.

We also experiment with other ways of injecting biases in the self-attention computation:

1. The first one is a special case of our proposed formulation in Equation 3, where both $\mathbf{k}'$ and $\mathbf{v}'$ are zero vectors. Equation 4 shows the computation of this variant of self-attention. This is also equivalent to the previous proposed Softmax-off-by-one (Miller, 2023).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q\left[K^T \;\; \mathbf{0}\right]}{\sqrt{d_k}}\right)\begin{bmatrix} V \\ \mathbf{0}^T \end{bmatrix} \tag{4}$$

2. Since Equation 3 can be viewed as inserting a sequence dimension, we also experiment with inserting one extra feature dimension. Specifically, we add learnable parameters $\mathbf{q}', \mathbf{k}' \in \mathbb{R}^T$ and concatenate them with the query and key states respectively. This variant of self-attention is as follows:

$$\text{Attention}(Q, K, V; \mathbf{q}', \mathbf{k}') = \text{softmax}\left(\frac{[Q \;\; \mathbf{q}']\,[K \;\; \mathbf{k}']^T}{\sqrt{d_k}}\right) V \tag{5}$$

3. We also experiment with a simple way to enforce constant value updates by injecting an extra value parameter $\mathbf{v}' \in \mathbb{R}^{d_k}$. This variant of self-attention is as follows:

$$\text{Attention}(Q, K, V; \mathbf{v}') = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V + \mathbf{v}' \tag{6}$$

Figure 26 visualizes the ten largest activation magnitudes in three GPT-2 models, corresponding to the three formulations of biases in Equation 4, 5 and 6. We find that these alternatives are not able to eliminate massive activations during pretraining.

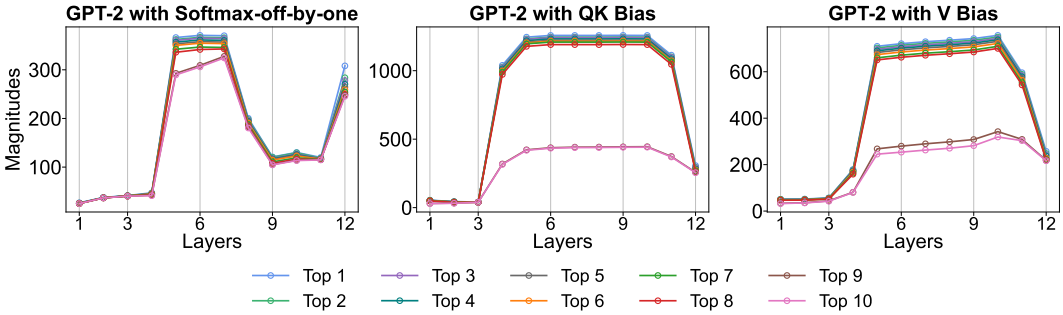


Figure 26: Ten largest activation magnitudes at each layer in three GPT-2 models.

C Massive Activations in Vision Transformers

In this section, we study if Vision Transformers (ViTs) (Dosovitskiy et al., 2021) exhibit massive activations. We note that while ViTs and LLMs are both based on self-attention, ViTs employ global token mixing, which contrasts with the autoregressive nature of LLMs.

Massive activations in ViTs. We explore several model families based on ViTs: CLIP (Radford et al., 2021), MAE (He et al., 2021) and DINOv2 (Oquab et al., 2024). We examine the ViT-L models from these families. The activation magnitudes in the penultimate layer for an input image are illustrated in Figure 27. We find that massive activations exist in CLIP and DINOv2 ViT-L, where we highlight the corresponding sequence dimensions. In these two models, there are extremely few activations (fewer than four) with significantly larger magnitudes than others. In addition, these activations are located in specific feature dimensions and appear in *random* patch tokens. However, we do not observe massive activations in MAE ViT-L. In this model, a feature dimension (927) exhibits uniformly large values across all tokens.

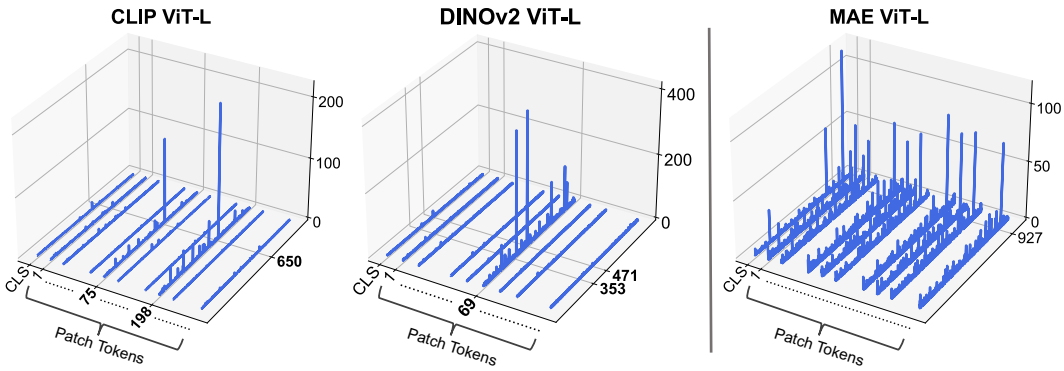


Figure 27: Massive activations are present in ViT-L from CLIP and DINOv2, but not MAE.

Massive activations are biases in ViTs. Figure 29 shows the three largest activation magnitudes and the median per layer in CLIP and DINOv2 ViT-L, averaged over 1k images. We find that massive activations are consistently present across images and their values remain largely the same around the mean values. It is worth noting that unlike LLMs, massive activations start to appear only in the later stages of ViTs.

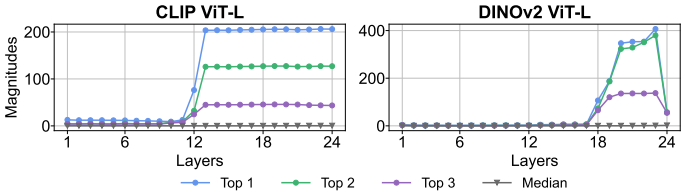


Figure 29: Three largest activation magnitudes and the median magnitude at each layer in CLIP and DINOv2 ViT-L.

CLIP ViT-L, layer 13	
Intervention	ImageNet acc (%)
Original	75.5
Set to zero	59.8
Set to mean	75.5

Table 3: Intervention analysis of massive activations in CLIP ViT-L.

Following our methodology in Section 3, we perform intervention analysis on CLIP ViT-L. We modify the two largest massive activations to zero and mean values respectively. The intervention is conducted on layer 13, where massive activations first appear within this model. Results are shown in Table 3, where we evaluate the zero-shot accuracy on ImageNet. We can see that setting massive activations to zero leads to significant drop in accuracy while setting to their means results in negligible accuracy drop. These results indicate that massive activations function as fixed but crucial biases in ViTs, aligned with our observations in Section 3.

Registers are biases in ViTs. Recently Darcet et al. (2023) propose to augment standard ViTs with additional learnable tokens, which they name as register tokens. They show that training ViTs with register tokens leads to smooth attention maps, and the resulting model family, namely DINOv2-reg, achieves superior downstream performance over DINOv2. Examining the largest ViT-G model in DINOv2-reg, we observe the existence of massive activations, as shown in Figure 30. However, different from standard ViTs, massive activations do not appear in patch tokens but exclusively within a fixed register token, i.e., register 3. This suggests that this model uses register 3 to store these activations. Figure 32 visualizes the attention distribution of the `[CLS]` token in the last layer. We find that most of the attention is allocated to register 3, echoing our previous findings in attention patterns (Section 4.1).

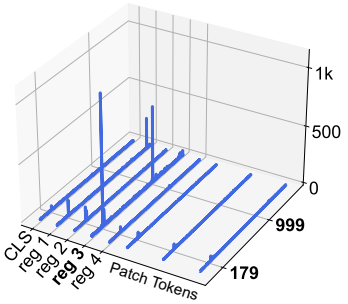


Figure 30: DINOv2-reg ViT-G.

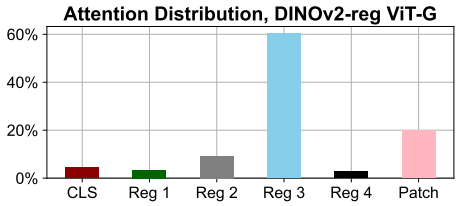


Figure 32: Average attention of the `[CLS]` token.

	DINOv2-reg with 4 registers			
ImageNet acc (%)	ViT-S	ViT-B	ViT-L	ViT-G
Original	81.9	84.8	86.3	87.0
`Fix-Reg-Mean`	81.7	85.0	86.2	87.0

Table 4: We fix *all* register features at *every* layer to their means and evaluate the intervened ViTs.

Further, we conduct intervention analysis to analyze the role of registers. We replace *all* register features at the output of *every* layer with their means, averaged over 10k ImageNet training images. This intervention removes the intended purpose of registers to aggregate global input information (Darcet et al., 2023). Table 4 shows the results. We find that ViTs with fixed register features achieve accuracy comparable to original models, suggesting that registers act as learned biases in ViTs. This leads to constant key and value states at register tokens, effectively introducing bias terms to self-attention (extra $\mathbf{k}'$ and $\mathbf{v}'$ in Equation 3). Thus a ViT with register tokens function equivalently to a standard ViT augmented with explicit attention biases.

To summarize our findings:

Massive activations exist in many but not all ViTs. Similar to those in LLMs, these activations act as constant biases. We also show the recently proposed register tokens have a similar function.

D Difference from Outlier Features

We discuss the differences between massive activations and outlier features, a seemingly similar phenomenon in LLMs. Dettmers et al. (2022) have identified the existence of outlier features characterized by large magnitudes within LLMs.

Conceptually, a massive activation is a scalar value, determined jointly by the sequence and feature dimensions; in contrast, *an outlier feature is a vector, corresponding to activations at all tokens*. Further, massive activations are present at extremely few tokens, while outlier features expect most activations in them to be large.

In practice, we find that massive activations do not overlap with outlier feature dimensions. We identify outlier features in LLaMA2-7B and 13B using the definition in Dettmers et al. (2022): a feature is deemed as an outlier feature if activation magnitudes exceed 6.0 at more than 25% of layers and 6% of tokens, on more than 90 out of 100 sequences. We discover 10 and 25 outlier features in these two models respectively. However, none of them correspond to the feature dimensions of massive activations.

E Related Work

Intriguing properties of autoregressive Transformers. Timkey & Schijndel (2021) observed that in GPT-2's penultimate layer, there are feature dimensions containing activations with magnitudes up to 3,000, and these few dimensions dominate several standard measures for evaluating representation similarity. Heimersheim & Turner (2023) found that the feature norm of the initial token in GPT-2 grows much faster than other tokens. Kovaleva et al. (2021) and Zhao et al. (2023) demonstrated the existence of outlier weights in the LayerNorm of GPT-2 and LLaMA2-13B and showed that setting them to zero leads to catastrophic drop in model performance. Notably, the feature dimension of this weight in LLaMA2-13B (i.e., 2100) corresponds to that of a massive activation (Figure 6).

Outlier features. Various existing works in quantization (Dettmers et al., 2022; Zeng et al., 2022; Xiao et al., 2023a; Lin et al., 2023; Ahmadian et al., 2023) have studied the existence of outlier features in LLMs. Dettmers et al. (2022) showed that outlier features have large activation values in most of their sequence dimensions. While massive activations can be seemingly similar to outlier features, we discussed their fundamental differences in Section D. More importantly, we show that massive activations can not be attributed to the existence of outlier features.

Attention concentration patterns. Clark et al. (2019b), Kovaleva et al. (2019) and Bondarenko et al. (2021) discovered that attention in BERT (Devlin et al., 2018) tends to focus on the "separate" token `[SEP]`. Xiao et al. (2023b) showed that LLMs assign most of the attention to the starting word token. Darcet et al. (2023) revealed the existence of attention artifacts in ViTs. Robinson et al. (2023) found sparse activation patterns in ViTs that attract attention to certain tokens. Our work provides an in-depth analysis as to why these patterns emerge, specifically in relation to massive activations.

Biases in self-attention. There can be various notion of biases in the self-attention mechanism. First, simple additive bias terms can be used in linear layers for computing the query, key and value states (Namazifar et al., 2023). Second, position biases can be inserted in self-attention to encode positional information of each token (Su et al., 2021; Press et al., 2021). There are also variants of biases with manually designed softmax operator (Miller, 2023; Bondarenko et al., 2023). Our work reveals that LLMs, even with standard self-attention, would impose implicit bias components in the attention computation through massive activations.

F Models and Datasets

Table 5 and Table 6 list the information of the LLM and ViT models used in this paper.

Model family	Model name	Layers	Dimensions	Heads	Huggingface model id
LLaMA2	LLaMA2-7B	32	4096	32	meta-llama/Llama-2-7b-hf
	LLaMA2-13B	40	5120	40	meta-llama/Llama-2-13b-hf
	LLaMA2-70B	60	6656	52	meta-llama/Llama-2-70b-hf
	LLaMA2-7B-Chat	32	4096	32	meta-llama/Llama-7b-chat-hf
	LLaMA2-13B-Chat	40	5120	40	meta-llama/Llama-2-13b-chat-hf
	LLaMA2-70B-Chat	60	6656	52	meta-llama/Llama-2-70b-chat-hf
Mistral	Mistral-7B	32	4096	32	mistralai/Mistral-7B-v0.1
	Mistral-8x7B	32	4096	32	mistralai/Mistral-8x7B-v0.1
	Mistral-7B-Instruct	32	4096	32	mistralai/Mistral-7B-Instruct-v0.2
	Mistral-8x7B-Instruct	32	4096	32	mistralai/Mistral-8x7B-Instruct-v0.1
Phi	Phi-2	32	2560	32	microsoft/phi-2
MPT	MPT-7B	32	4096	32	mosaicml/mpt-7b
	MPT-30B	48	7168	64	mosaicml/mpt-30b
Falcon	Falcon-7B	32	4544	71	tiiuae/falcon-7b
	Falcon-40B	60	8192	128	tiiuae/falcon-40b
OPT	OPT-7B	32	4096	32	facebook/opt-6.7b
	OPT-13B	40	5120	40	facebook/opt-13b
	OPT-30B	48	7168	56	facebook/opt-30b
	OPT-66B	64	9216	72	facebook/opt-66b
GPT-2	GPT-2	12	768	12	gpt2
	GPT-2-Medium	24	1024	16	gpt2-medium
	GPT-2-Large	36	1280	20	gpt2-large
	GPT-2-XL	48	1600	25	gpt2-xl

Table 5: Relevant information of LLM models we experimented with in this work.

Model family	Model size	Layers	Dimensions	Heads	Huggingface model id
DINOv2	ViT-S	12	384	6	timm/vit_small_patch14_dinov2.lvd142m
	ViT-B	12	768	12	timm/vit_base_patch14_dinov2.lvd142m
	ViT-L	24	1024	16	timm/vit_large_patch14_dinov2.lvd142m
	ViT-G	40	1536	24	timm/vit_giant_patch14_dinov2.lvd142m
DINOv2-reg	ViT-S	12	384	6	timm/vit_small_patch14_reg4_dinov2.lvd142m
	ViT-B	12	768	12	timm/vit_base_patch14_reg4_dinov2.lvd142m
	ViT-L	24	1024	16	timm/vit_large_patch14_reg4_dinov2.lvd142m
	ViT-G	40	1536	24	timm/vit_giant_patch14_reg4_dinov2.lvd142m
MAE	ViT-B	12	768	12	timm/vit_base_patch16_224.mae
	ViT-L	24	1024	16	timm/vit_large_patch16_224.mae
	ViT-H	32	1280	16	timm/vit_huge_patch16_224.mae
CLIP	ViT-B	12	768	12	timm/vit_base_patch16_clip_224.openai
	ViT-L	24	1024	16	timm/vit_large_patch14_clip_224.openai

Table 6: Relevant information of ViT models we experimented with in this work.

We list the datasets used in this work and relevant license information:

- RedPajama (Together Computer, 2023): Apache License, Version 2.0
- OpenWebText2 (Gao et al., 2021): MIT License
- C4 (Raffel et al., 2020): Open Data Commons Attribution License 1.0 license
- PG-19 (Rae et al., 2019): Apache License, Version 2.0
- WikiText (Merity et al., 2016): Creative Commons BY-SA 3.0 license
- MMLU (Hendrycks et al., 2021): MIT License
- BoolQ (Clark et al., 2019a): Creative Commons BY-SA 3.0 license
- PIQA (Bisk et al., 2019): The license status is unclear
- WinoGrande (Sakaguchi et al., 2019): Apache License, Version 2.0
- ARC easy and challenge (Clark et al., 2018): Creative Commons BY 4.0 license
- ImageNet (Deng et al., 2009): The license status is unclear