LOW-RANK IS REQUIRED FOR PRUNING LLMS

Stephen Zhang
Department of Mathematics
University of Toronto
stephenn.zhang@mail.utoronto.ca

Vardan Papyan
Department of Mathematics
University of Toronto

ABSTRACT

Post-train pruning without fine-tuning has emerged as an efficient method for compressing large language models for inference, offering a computationally cheaper alternative to other approaches. However, recent studies have revealed that, unlike quantization, pruning consistently degrades model performance as sparsity increases. We demonstrate that this degradation results from pruning's inability to preserve a low-rank structure in the model's weights, which is crucial for maintaining attention sinks. Furthermore, we show that these attention sinks play a key role in enabling the model to segment sequences—an essential mechanism for effective few-shot learning.

1 INTRODUCTION

Post-train pruning has become a widely used, computationally inexpensive method for compressing large foundation models, particularly those that do not require further fine-tuning (Benbaki et al., 2023; Frantar & Alistarh, 2023; Sun et al., 2024b; Ashkboos et al., 2024). However, these approaches assume that trained model weights can be accurately approximated by sparse matrices. Recent work by Yin et al. (2024) challenges this assumption, showing a consistent decline in performance as pruning sparsity increases. This raises doubts about whether pure sparsity is the optimal structure to impose on trained model weights. In this work, we demonstrate that:

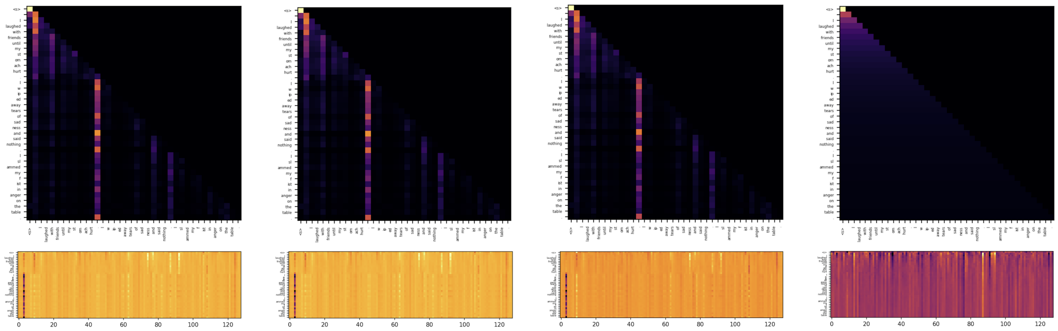
- A low-rank structure exists within trained model weights and is responsible for generating two key phenomena essential to model performance: *attention sinks* and *outlier feature dimensions*.
- Pruned models that do not preserve a low-rank structure exhibit fewer attention sinks than their dense counterparts and those that preserve a low-rank structure.
- Attention sinks coincide with sequence segmentation, which is crucial in scenarios like few-shot learning (Brown et al., 2020), where separating prompts into distinct parts is beneficial.

Based on these observations, we argue that decomposing model weights into a sparse plus low-rank structure (Candès et al., 2011; Yu et al., 2017; Thangarasa et al., 2024; Zhang & Papyan, 2024) significantly improves post-train pruning, providing a more faithful approximation of the dense model's weight matrices.

1.1 BACKGROUND: ATTENTION SINKS AND OUTLIER FEATURES

Two key phenomena observed in large language models, both relevant to model compression, are:

Attention Sinks Coined by Xiao et al. (2024), the authors found that all tokens attend strongly to the first token. Follow-up works demonstrated that sinks may also appear in later tokens (Yu et al., 2024; Sun et al., 2024a; Cancedda, 2024).



(a) Dense (b) Sparse + Low-Rank (c) $\mathbf{0}$ + Low-Rank (d) Sparse + $\mathbf{0}$

Figure 1: Comparison of attention weights and attention output for layer 2, head 5, across the dense Phi-3 Medium and Phi-3 Medium compressed by the OATS algorithm.

Outlier Feature Dimensions Feature dimensions in the activations that are significantly larger in magnitude (Kovaleva et al., 2021; Dettmers et al., 2022). Unlike attention sinks, outlier feature dimensions are consistent across tokens.

In our work, we provide evidence that these phenomena are induced by a low-rank structure in the model's weights that needs to be preserved for model performance.

2 Existence of a Low-Rank Structure

We show that the emergence of attention sinks and outlier feature dimensions in large language models can be attributed to a low-rank structure in the model's weights. We utilize a recent pruning method called OATS (Zhang & Papyan, 2024), which approximates each of the model's weight matrices as a sparse plus low-rank matrix:

$$\boldsymbol{W} = \boldsymbol{S} + \boldsymbol{U}\boldsymbol{V}^\top.$$

Figure 1 above presents the attention weights, of a Phi-3 Medium model (Abdin et al., 2024) in four configurations: a dense model, a model compressed by 50% using OATS, a model compressed by 50% using OATS where the low-rank matrices are set to $\mathbf{0}$, and a model compressed by 50% using OATS where the sparse matrices are set to $\mathbf{0}$. Attention sinks and outlier feature dimensions exist in all configurations except for the model without low-rank terms.

To explore this quantitatively, we leverage a thresholding technique by Gu et al. (2025) that counts the number of attention sinks. Denote $\boldsymbol{A}^{\ell,h}$ as the attention weights of layer ℓ, head h. Given a sequence prompt of length T, token t is deemed an attention sink if and only if:

$$\frac{1}{T-t+1}\sum_{k=1}^{T}\boldsymbol{A}^{\ell,h}_{k,t} > \epsilon \tag{1}$$

where ϵ is a designated threshold that we set to 0.1. Table 1 below depicts the number of attention sinks exhibited by compressed models in the configurations above averaged across 170 sequences.

Configuration	Llama-3 8B	Phi-3 Medium	Qwen 2.5	
			7B	14B
Low-Rank Terms Only	986	$3,334$	$1,495$	$2,317$
Sparse Terms Only	0	3	279	193

Table 1: Total number of attention sinks, computed using Equation (1), exhibited by models compressed by 50% using OATS in two configurations. Models with low-rank components exhibit attention sinks, whereas those with only sparse terms show none or few.

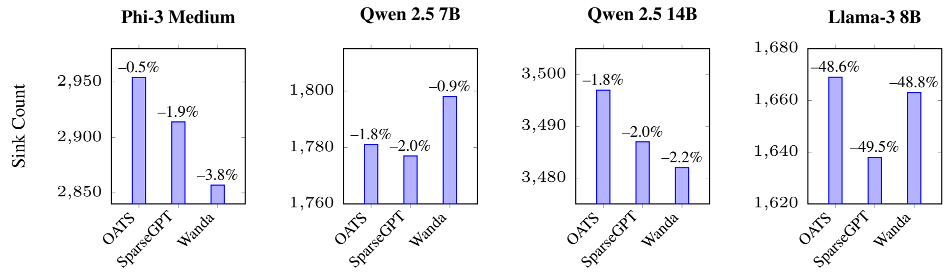


Figure 2: Total number of attention sinks exhibited by compressed models with the percent reduction from dense models shown above each bar. Compressed models exhibit fewer attention sinks than their dense counterparts with OATS preserving the most sinks most frequently.

3 Pruning Harms Low-Rank Structure

3.1 Pruning Leads to Missing Sinks

Given that a low-rank structure contributes to the emergence of these two phenomena, a natural question arises: do standard pruning methods that rely solely on sparsity inadvertently disrupt the low-rank structure responsible for generating them? To explore this, we again use Equation (1) to count the number of sinks exhibited by models compressed using OATS, as well as models pruned using SparseGPT (Frantar & Alistarh, 2023) and Wanda (Sun et al., 2024b) – both of which enforce a purely sparse structure. Figure 2 above shows the results highlighting that pruned models relying purely on sparsity exhibit fewer sinks.

3.2 Sparse Plus Low-Rank: Closer Approximation

In addition to capturing all attention sinks, representing the model's weight matrices as a sum of sparse plus low-rank matrices leads to a closer approximation of the original model's weights, as measured by the Frobenius norm:

$$\Delta_{\boldsymbol{W}} = \|\boldsymbol{W}_{\text{dense}} - \boldsymbol{W}_{\text{pruned}}\|_F^2.$$

We sum these values across all linear weight matrices in a decoder layer and normalize by the total number of parameters in that layer. Although the Frobenius norm is not a reliable indicator of pruning performance – as shown by the poor results of magnitude pruning on LLMs (Frantar & Alistarh, 2023; Sun et al., 2024b) – Figure 3 below shows that OATS often achieves a closer approximation than even magnitude pruning while outperforming sparse-only methods (Zhang & Papyan, 2024). This suggests that combining sparse plus low-rank offers a closer approximation of trained weights than sparse only.

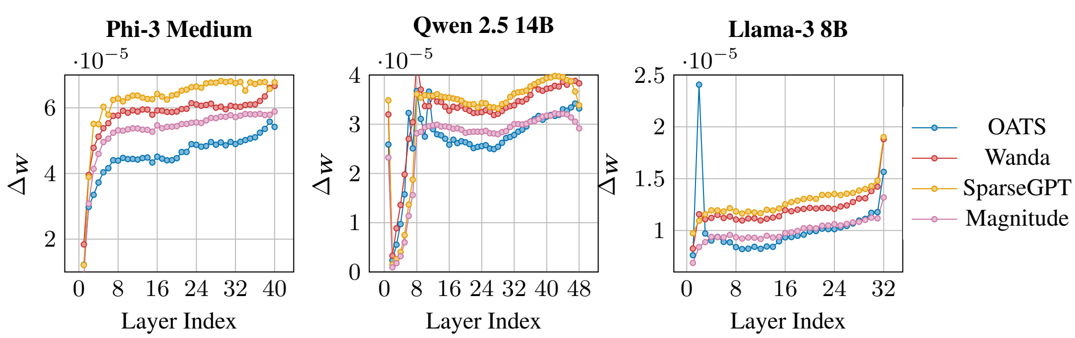
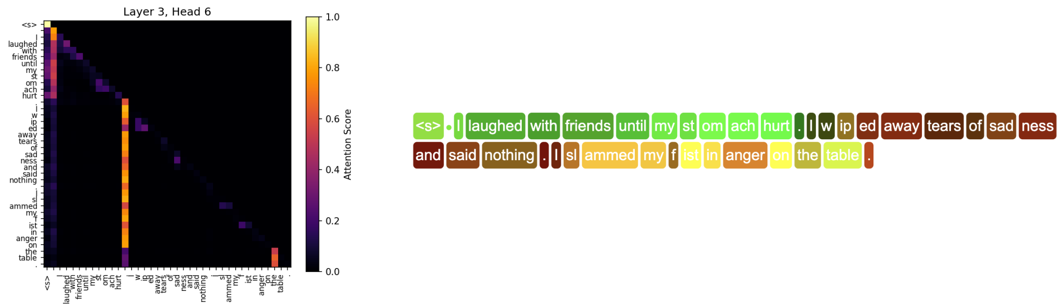


Figure 3: $\Delta_{\boldsymbol{W}}$ measured on Phi-3 Medium, Qwen 2.5 14B, and Llama-3 8b, compressed by 50% with OATS, Wanda, SparseGPT, and magnitude pruning.

4 Sinks and Outliers: Few-Shot Learning

4.1 Sequence Segmentation

We apply a similar technique to Oquab et al. (2024) to show that the model segments sequences based on the location of the attention sinks. We compute the top two principal components of the $d \times d$ covariance matrix of the attention head outputs, where d is the embedding dimension, and project the embeddings onto this basis. The projected values are then normalized to $[0, 255]$ and mapped to RGB channels. Figure 4 below depicts the results and shows that clustering aligns with the locations of the attention sinks.



(a) **Attention Weights.** Two distinct attention sinks exhibited in layer 3, head 6, of a dense model.

(b) **PCA on the residual stream in a deeper layer.** Tokens that attended to the first sink are green, while those attending to the second sink are brown and yellow.

Figure 4: **Sequence Segmentation.** Tokens in the residual stream at layer 17 in a dense Phi-3 Medium model are clustered based on the attention sink that they attend to.

4.2 Implications for Few-Shot Learning

Since attention sinks are responsible for segmenting sequences, it is likely crucial for distinguishing between examples in few-shot learning. To test this, we measure the gap in k-shot performance on the MMLU dataset (Hendrycks et al., 2021) where k ranges from 0 to 5. We compare the performance of OATS with Wanda and SparseGPT in Figure 5 on a Phi-3 Medium model that has been compressed by 50%.

While all methods perform similarly at 0-shot, where segmentation is less critical, the gap widens with k, highlighting the importance of OATS' ability to retain the low-rank structure. In contrast, the model pruned by SparseGPT shows no improvement with more examples, indicating severely compromised few-shot capabilities.

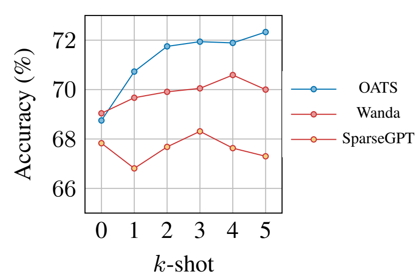


Figure 5: Impact of increasing the number of examples on model performance. Accuracy is measured on the MMLU dataset.

5 Conclusion

We present evidence that the assumption underlying pruning, which relies on a sparse representation of the model weights, is flawed and overlooks the presence of a low-rank structure inherent in the trained weights. We demonstrate that this low-rank structure is responsible for creating attention sinks, which the model leverages for sequence segmentation. Consequently, pruning methods that fail to preserve the low-rank structure not only result in a poorer approximation of the dense weights but also hinder the model's few-shot learning capabilities.

ACKNOWLEDGMENTS

We would like to thank Mustafa Khan for assisting in generating the PCA coloring figure depicted in Figure 4. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This research was supported in part by the Province of Ontario, the Government of Canada through CIFAR, and industry sponsors of the Vector Institute (`www.vectorinstitute.ai/partnerships/current-partners/`). This research was also enabled in part by support provided by Compute Ontario (`https://www.computeontario.ca`) and the Digital Research Alliance of Canada (`https://alliancecan.ca`).

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL `https://arxiv.org/abs/2404.14219`.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=vXxardq6db`.

Riade Benbaki, Wenyu Chen, Xiang Meng, Hussein Hazimeh, Natalia Ponomareva, Zhe Zhao, and Rahul Mazumder. Fast as chita: neural network pruning with combinatorial optimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks, 2024. URL `https://arxiv.org/abs/2402.09221`.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL https://doi.org/10.1145/1970392.1970395.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=dXiGWqBoxaD.

Elias Frantar and Dan Alistarh. Sparsegpt: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=78Nn4QJTEN.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.300. URL https://aclanthology.org/2021.findings-acl.300/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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. *arXiv e-prints*, 2019.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=F7aAhfitX6.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=PxoFut3dWW.

Vithursan Thangarasa, Shreyas Saxena, Abhay Gupta, and Sean Lie. Sparse-IFT: Sparse iso-FLOP transformations for maximizing training efficiency. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/thangarasa24a.html.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Lu Yin, Shiwei Liu, AJAY KUMAR JAISWAL, Souvik Kundu, and Zhangyang Wang. Junk DNA hypothesis: A task-centric angle of LLM pre-trained weights through sparsity, 2024. URL https://openreview.net/forum?id=EmUVpfrXWN.

Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 67–76, 2017. doi: 10.1109/CVPR.2017.15.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=DLTjFFiuUJ`.

Stephen Zhang and Vardan Papyan. Oats: Outlier-aware pruning through sparse and low rank decomposition, 2024. URL `https://arxiv.org/abs/2409.13652`.

A EXPERIMENT DETAILS

We use the HuggingFace Transformers library (Wolf et al., 2019) to load the Phi-3 Medium (Abdin et al., 2024), Qwen 2.5 models (Qwen et al., 2025), and Llama 3 8B model (Grattafiori et al., 2024) utilized in our experiments. To evaluate the few-shot performance of MMLU, we use LM Harness (Gao et al., 2024).

A.1 PRUNING HYPERPARAMETERS

We prune uniformly across all linear layers in the model and exclude the embedding and the language modeling head layers from pruning. This remains consistent with (Frantar & Alistarh, 2023; Sun et al., 2024b; Zhang & Papyan, 2024). As calibration data, we use 128 sequences of length 2048 from the first shard of the C4 dataset (Raffel et al., 2019). The algorithm-specific hyperparameters are:

- SparseGPT
  - Hessian Dampening: 0.1
  - Block Size: 128
- OATS
  - Iterations: 80
  - Phi-3 Medium & Qwen 2.5 7B & Qwen 2.5 14B:
    - Rank Ratio: 0.25
  - Llama 3 8B:
    - Rank Ratio: 0.3

A.2 PROMPT INFORMATION

The prompt used to generate Figure 1 and Figure 4 is:

```
I laughed with friends until my stomach hurt.  I wiped
away tears of sadness and said nothing.  I slammed my
fist in anger on the table.
```

To generate Table 1, we use 170 prompts that have been provided by (Gu et al., 2025) that can be found here: https://github.com/sail-sg/Attention-Sink/blob/main/datasets/probe_valid.jsonl. We truncate each prompt to have only 150 tokens.

To generate Figure 2, we source questions and answers from the MMLU dataset to generate 110 prompts that follow the prompt template below.

You are a helpful assistant. Answer the following multiple-choice question.

Question: <*Example question 1*>
A) <Choice A> B) <Choice B> C) <Choice C> D) <Choice D>
Answer: A

Question: <*Example question 2*>
A) <Choice A> B) <Choice B> C) <Choice C> D) <Choice D>
Answer: C

Question: <*Test question*>
A) <Choice A> B) <Choice B> C) <Choice C> D) <Choice D>
Answer: