ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

# Early Prediction of Dynamic Sparsity in Large Language Models

Reza Sedghi[1], Amit Kumar Pal[1], Anand Subramoney[2] and David Kappel[1,3] *

[1] Institut für Neuroinformatik, Ruhr Universität Bochum, Germany

[2] Department of Computer Science, Royal Holloway, University of London, UK

[3] CITEC, Bielefeld University, Germany

**Abstract**. Large language models are powerful but computationally very expensive. We investigate dynamic sparsity in attention mechanisms, using the OPT model as a case study. We explore the dynamic nature of redundancy in attention heads and analyze which components of the model provide sufficient information to predict sparsity effectively. Our findings highlight the norm of attention outputs as a reliable criterion for ranking head importance. We systematically evaluate embeddings across layers and time steps, showing that dynamic sparsity predictions can be achieved early in the model pipeline with minimal loss in accuracy. By elucidating the mechanisms underlying dynamic sparsity, this work lays a foundation for more efficient and scalable transformer models.

## 1 Introduction

Large language models (LLMs) like GPT, OPT and BERT have transformed natural language processing by leveraging the power of transformer-based architectures and attention mechanisms, enabling impressive advances in language understanding and generation [1, 2]. Their success has spurred interest in extending attention-based models to various fields, including computer vision, signal processing, and video analysis [3, 4, 5]. However, as these models grow in size [6], their computational demands also increase, making them resource-intensive and challenging to scale [7, 8, 9]. To address this problem, efficiency improvements of the models have been explored [10, 11], with sparsity emerging as a promising approach to reduce computation without sacrificing performance [12, 13, 14]. Our study specifically examines the attention mechanism within the OPT model, exploring and analyzing patterns of dynamic redundancy in attention mechanisms, aiming to identify efficient ways to reduce the amount of redundant computation while maintaining model performance. To achieve this, we trained sparsity predictors that selectively skip attention heads based on their contribution to the model's output, similar to [15]. Our research extends this approach with a detailed analysis, to better understand which information in the model can predict sparsity efficiently, both across layers and over time. Through this analysis, we highlight the potential of leveraging selective dynamic sparsity and to improve the efficiency of large-scale models.

*RS is funded by BMBF project EVENTS (16ME0733). AP is funded by BMWK project ESCADE (01MN23004D). DK is funded by project SAIL (grant no. NW21-059A). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (`www.gauss-centre.eu`) for providing computing time on GCS JUWELS at Jülich Supercomputing Centre (JSC).

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

## 2 Dynamic Sparsity in Transformer Architectures

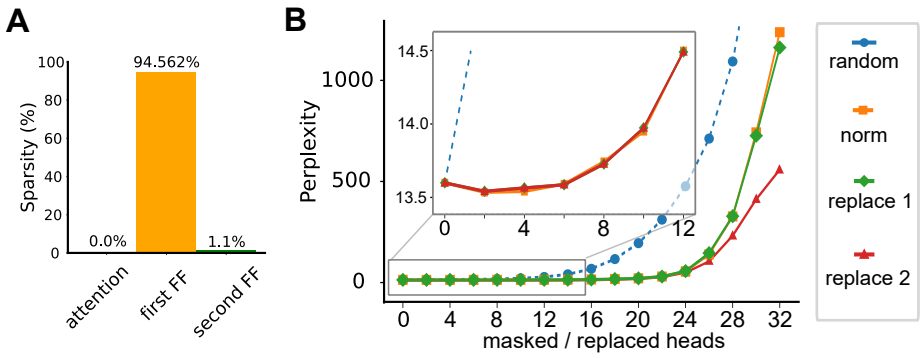


Fig. 1: **A:** Activation sparsity across three different cuts of OPT model components, averaged over layers. **B:** Impact of masking attention heads on model perplexity in a text generation task.

We evaluated dynamic sparsity on the OPT-1.3B model, where we removed or replaced unimportant attention heads using different methods. As discussed above, the attention heads do not inherently produce zeros, leading to an apparent lack of sparsity in the activations (see Fig. 1.A). Therefore, unimportant attention heads, rather than being masked, might be better replaced by static, precomputed embeddings. This matter has been studied in [16, 17], where, in most cases, the attention map or pattern of some or all heads was replaced with a static pattern. We explored whether substituting low-norm heads with precomputed static heads could offer any performance advantage over simple masking in OPT-1.3B. To do so, we ranked heads based on the Euclidean norm to determine their importance [15].

Importantly, importance scores appear to have a dynamic nature. That is, as the input context changes, the importance scores of the attention heads may vary. For replacing heads, we computed static embeddings by averaging the output of each attention head over a WikiText-103 training set, storing these averages to substitute for detected unimportant heads. Fig. 1.B shows the perplexity, $e^H$ with cross-entropy $H$, quantifies prediction uncertainty, on WikiText-103 evaluation set for different sparsity criterions, *random:* randomly masked heads (baseline), *norm:* Euclidean norm, masked heads replaced with zero vector *replace 1:* masked heads replaced with the average output of the same head, *replace 2:* refers to masked heads replaced with the average output of norm-ranked heads. The x-axis represents the number of masked or replaced heads. We observed no significant performance improvement of replacing over masking. This suggests that the model can effectively ignore non-zero but low-norm heads. In summary, our results show that Euclidean norm is a viable importance score and masking is as good as replacing, demonstrating that attention heads have a high level of dynamic sparsity.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

## 3 Predicting Sparsity

As highlighted in the previous section, the potential sparsity in attention heads exhibits a dynamic nature and can be exploited without significant loss of performance. This approach however does not provide computational savings, since calculating all head outputs remains necessary to calculate the norms. While static approaches exist for masking attention heads, they inherently disregard the dynamic changes in attention head importance [18, 19, 20]. Retaining dynamic sparsity requires predicting head importance prior to initiating the computation of attention heads. The central question addressed in this work: *Which components within the OPT model contain sufficient information to enable accurate predictions of attention head ranking, while minimizing memory and computation overhead?* Our focus is on achieving early prediction of importance scores, allowing for efficient memory management and computational savings.

### 3.1 Prediction Through Layers

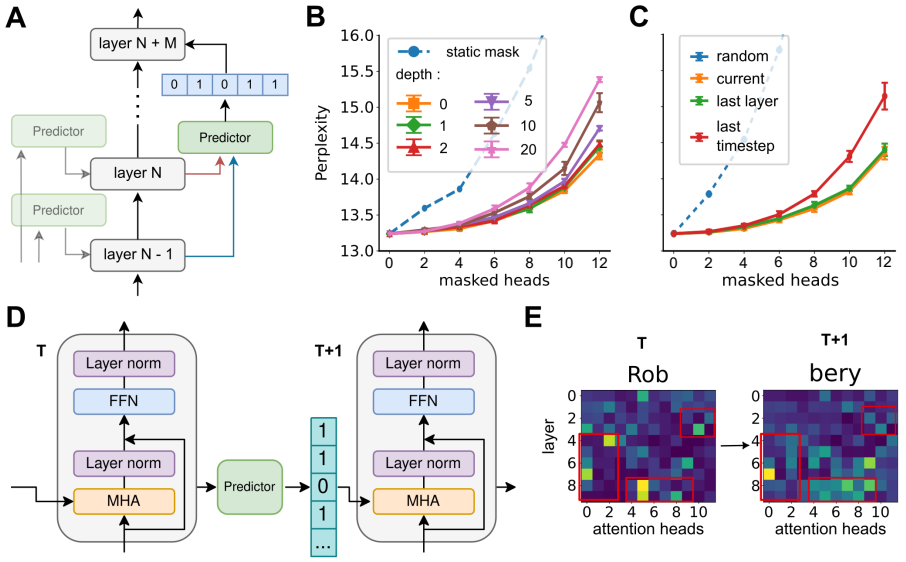


Fig. 2: **A:** Sparsity prediction: Embeddings from the $N-th$ layer guides a predictor that generates a binary mask for the $M-th$ layer. **B:** Effect of sparsity prediction on model perplexity across different depths. **C:** Effect of sparsity prediction on model perplexity. **D:** Sparsity prediction across time steps. **E:** Euclidean norms of attention heads for two consecutive tokens ("Rob-bery").

In our first approach to predict sparsity in OPT, we trained simple feedforward models. Specifically, we used a two-layer MLP with 256 hidden units per layer, ReLU activations, and a sigmoid output to take embeddings as inputs

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

and predict indices of attention heads to be masked through layers, (see figure 2).A. We experimented with different depths of the predictors (how many layers in advance are predicted), as shown in figure 2.B. For each trial, we used the embeddings from a layer $M$ levels prior to the current as input to predict the masked heads in the current layer. Figure 2.B shows the results on OPT-1.3B. The degradation in performance was minimal compared to static masking approaches, trained model with an identical matrix. Notably, for some depths $M$, such as 1, 2, or 5, the predictor's performance remains close to that achieved using embeddings from the current layer. These results suggest that all layers within the OPT model provides sufficient information for training a sparsity predictor across attention heads. Early prediction of masked heads from initial layers can have a small impact on model performance.

### 3.2 Prediction Through Time

Next, we aimed to push this approach further by exploring the feasibility of predicting attention head sparsity across time steps. Given that many LLM tasks operate in an autoregressive fashion, the model retains information from previous time steps before processing new tokens. We investigated whether this sequential information might inform sparsity predictions in advance of calculating new token embeddings. To examine this possibility, we analyzed the norm values of attention heads in the smallest OPT model (125M parameters) across consecutive tokens (Fig. 2C-E). Fig. 2D illustrates the approach. In Fig. 2.E, we illustrate the norm values for two successive tokens within the word "robbery," tokenized into "rob" and "bery." There is a noticeable, albeit subtle, similarity between the norm patterns of these consecutive tokens. This similarity motivated our approach to train a sparsity predictor over time. Figure 2.C shows the results. Although the highest accuracy is still achieved using embeddings from the current layer and, secondarily, from the last layer, the performance of timewise prediction remains promising, particularly when compared to random sparsity. This finding suggests that leveraging prior time-step embeddings to guide sparsity predictions could enable efficient head masking with minimal impact on model accuracy, further enhancing computational efficiency.

### 3.3 Results

To quantify the results, we tested the model on three datasets: WikiText, PTB, and BookCorpus, and trained sparsity predictors on OPT models with size 1.3B, 2.7B, and 6.7B. Each model was evaluated for three different sparsity ratios: 25%, 50%, and 75% (0% denotes the non-sparse baseline). Within each configuration, we compared three methods for predicting sparsity: *current:* using embeddings from the current layer; *last:* using embeddings from the immediately preceding layer; and *time:* using embeddings from the same layer but from the previous time step. The results are shown in the table 1. For all methods, the performance drop was insignificant compared to the original (0%) model for the 25% sparsity case. Larger sparsity levels resulted in performance drops. These

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

| model | method | 0% | 25% | 50% | 75% |
|---|---|---|---|---|---|
| WikiText dataset | | | | | |
| OPT 1.3B | current | 13.5 | 13.5 ±0.07 | 25.9 ±0.26 | 149.4 ±6.48 |
| | last | | 13.6 ±0.04 | 25.2 ±0.32 | 158.7 ±5.32 |
| | time | | 13.8 ±0.06 | 31.6 ±1.06 | 230.8 ±9.20 |
| OPT 2.7B | current | 11.4 | 11.5 ±0.03 | 15.8 ±0.19 | 109.2 ±5.15 |
| | last | | 11.5 ±0.02 | 16.4 ±0.33 | 128.0 ±9.47 |
| | time | | 11.6 ±0.02 | 19.6 ±0.28 | 160.5 ±11.83 |
| OPT 6.7B | current | 10.0 | 10.0 ±0.02 | 16.7 ±0.35 | 128.0 ±9.62 |
| | last | | 10.0 ±0.02 | 17.2 ±0.42 | 134.6 ±7.43 |
| | time | | 10.1 ±0.04 | 19.4 ±0.50 | 178.2 ±12.34 |
| PTB dataset | | | | | |
| OPT 1.3B | current | 13.4 | 13.5 ±0.03 | 28.1 ±1.8 | 109.48 ±10.84 |
| | last | | 13.5 ±0.02 | 26.9 ±1.07 | 109.8 ±9.10 |
| | time | | 13.7 ±0.05 | 31.1 ±2.99 | 140.3 ±4.90 |
| OPT 2.7B | current | 11.8 | 11.8 ±0.02 | 16.8 ±0.21 | 88.5 ±4.14 |
| | last | | 11.8 ±0.02 | 17.0 ±0.27 | 97.0 ±7.77 |
| | time | | 12.0 ±0.01 | 20.5 ±0.28 | 143.3 ±5.06 |
| OPT 6.7B | current | 10.4 | 10.4 ±0.03 | 17.4 ±0.26 | 108.2 ±13.23 |
| | last | | 10.4 ±0.01 | 17.4 ±0.41 | 107.41 ±7.85 |
| | time | | 10.5 ±0.02 | 19.2 ±0.14 | 147.4 ±16.00 |
| BookCorpos dataset | | | | | |
| OPT 1.3B | current | 8.1 | 8.1 ±0.01 | 19.2 ±1.38 | 78.9 ±3.81 |
| | last | | 8.1 ±0.02 | 23.1 ±0.48 | 89.3 ±2.76 |
| | time | | 8.1 ±0.01 | 20.29 ±1.05 | 77.55 ±5.01 |
| OPT 2.7B | current | 6.7 | 6.9 ±0.01 | 9.2 ±0.11 | 49.3 ±4.13 |
| | last | | 6.9 ±0.01 | 9.3 ±0.15 | 49.4 ±2.48 |
| | time | | 7.0 ±0.01 | 11.0 ±0.07 | 60.5 ±3.52 |
| OPT 6.7B | current | 6.2 | 6.2 ±0.01 | 9.7 ±0.18 | 72.9 ±4.98 |
| | last | | 6.2 ±0.00 | 9.8 ±0.15 | 79.1 ±9.04 |
| | time | | 6.2 ±0.01 | 9.9 ±0.16 | 98.8 ±8.21 |

Table 1: comparison of perplexity results across three datasets for the text generation tasks and sparsity prediction method (current, Last 1D, and last).

results confirm that Euclidean norm provides a reliable importance criterion.

## 4 Conclusion

Overall, our experiments demonstrate that both deep layer prediction and time-wise prediction have potential, performing nearly as well as predictions from the current layer, especially at lower sparsity ratios $\leq 30\%$. This indicates that OPT models, even in smaller sizes, contain sufficiently reliable information across layers and timesteps to support effective sparsity predictions, allowing selective

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

attention head masking without a substantial drop in performance. In conclusion, our results suggest that OPT models can be effectively sparsified using these prediction methods, enabling more computationally efficient models. This approach offers a viable path for implementations of resource management, that exploit dynamic sparsity while maintaining accuracy.

## References

[1] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, et al. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *NeurIPS*, 2017.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, et al. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[5] Javier Selva, Anders S. Johansen, Sergio Escalera, et al. Video transformers: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 45(11):12922–12943, 2023.

[6] Elias Frantar, Carlos Riquelme, Neil Houlsby, et al. Scaling laws for sparsely-connected foundation models. *ICLR*, (arXiv:2309.08520).

[7] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), December 2022.

[8] Markus N. Rabe and Charles Staats. Self-attention does not need $o(n^2)$ memory. *arXiv preprint arXiv:2112.05682*, 2022.

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2024.

[10] Krishna Teja Chitty-Venkata, Sparsh Mittal, Murali Emani, et al. A survey of techniques for optimizing transformer inference. *Journal of Systems Architecture*, 2023.

[11] Sehoon Kim, Coleman Hooper, Thanakul Wattanawong, et al. Full stack optimization of transformer inference: a survey. *arXiv preprint arXiv:2302.14017*, 2023.

[12] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, et al. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2022.

[13] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, et al. Sparse is enough in scaling transformers. *NeurIPS*, 34:9895–9907, 2021.

[14] Zonglin Li, Chong You, Srinadh Bhojanapalli, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. *ICLR*, 2023.

[15] Zichang Liu, Jue Wang, Tri Dao, et al. Deja vu: Contextual sparsity for efficient LLMs at inference time. *In International Conference on Machine Learning*, 2023.

[16] Shucong Zhang, Erfan Loweimi, Peter Bell, et al. Stochastic attention head removal: A simple and effective method for improving transformer based asr models. *arXiv preprint arXiv:2011.04004*, 2021.

[17] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2021.

[18] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *NeurIPS*, 32, 2019.

[19] Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. *In EMNLP*, (arXiv:2104.07012).

[20] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.