Pruning for Privacy: Mitigating Memorised Sequence Reproduction via Attention Head Sparsification

Anonymous ACL submission

Abstract

Large language models have been shown to memorise and reproduce verbatim sequences from their training data, raising concerns about privacy leakage, intellectual property exposure, and the limits of generalisation. Recent theoretical work suggests that attention heads may contribute to this behaviour by functioning as associative memory mechanisms, with memorisation capacity scaling with the number of heads. At the same time, empirical work has shown that many attention heads are redundant and can be pruned with limited impact on downstream performance. Despite these parallel findings, the effect of pruning redundant heads on memorised-sequence reproduction remains unexplored. In this work, we demonstrate that zeroing out a substantial fraction of attention heads prior to inference can significantly reduce memorisation without degrading downstream task performance. We find that memorised sequences elicit sharper softmax saturation across heads, such that their aggregate logit contributions cause the output distribution to peak at the memorised token. Pruning a subset of heads removes their influence from the residual stream, weakening this collective signal. Evaluating GPT-2 on known memorised sequences, we find that pruning 739 task-redundant heads reduces the memorisation count by 92%, while improving WikiText-103 perplexity by 0.81. These findings suggest that attention-head redundancy can be exploited to mitigate memorisation at inference time, without retraining or modifying the training data.

1 Introduction (& Background)

Large language models are trained on massive corpora that often contain personal information, copyrighted text, and proprietary material. Their tendency to memorise and reproduce verbatim training sequences raises concerns about privacy leakage, intellectual property exposure, and whether model outputs reflect genuine generalisation rather than recall (Carlini et al., 2021, 2022). Existing mitigations, such as deduplication, differential privacy, and unlearning (Kandpal et al., 2022; Yu et al., 2021; Si et al., 2023), either intervene at training time or require identifying specific memorised content, making them difficult to apply to deployed models.

Recent theoretical work has established that multi-head attention can operate as an associative memory, with memorisation capacity scaling with the number of heads (Mahdavi et al., 2024; Dana et al., 2024). This capacity is driven by softmax saturation: when attention weights become sharply concentrated on a small set of positions, the corresponding value vectors dominate the head output, enabling precise token retrieval rather than diffuse contextual blending. Independently, empirical work has consistently shown that a substantial fraction of attention heads can be zeroed out at inference time without degrading downstream task performance (Michel et al., 2019; Voita et al., 2019). The implications of this redundancy for memorisation, however, remain unexplored. Our contributions are as follows:

- We demonstrate that inference-time pruning of task-redundant attention heads substantially reduces memorised-sequence reproduction while preserving downstream task performance.
- We find that memorised sequences induce sharper softmax saturation and stronger cross-head agreement, creating a coordinated signal toward the memorised continuation.
- We show that pruning task-redundant heads weakens this collective signal by removing part of their contribution from the residual stream, offering a mechanistic explanation for the observed reduction in memorisation.

2 Methodology

2.1 Models and Datasets

We conduct experiments on Pythia-1.4B, Pythia-2.8B, and GPT-2-XL (1.5B) (Biderman et al., 2023b; Radford et al., 2019). These models are chosen for their open weights, documented pretraining corpora, and established use in memorisation research (Carlini et al., 2021; Biderman et al., 2023a). Together, these models cover a range of scales and architectures, which is relevant given that memorisation rates have been shown to increase with model capacity (Carlini et al., 2022).

For each model, we source both memorised and non-memorised sequences from its corresponding pretraining corpus: Pile-Deduped for the Pythia models and OpenWebText for GPT-2-XL (Gao et al., 2020; Radford et al., 2019). Downstream task performance is evaluated on WikiText-103 and LAMBADA, standard benchmarks for language modelling and next-token prediction (Merity et al., 2016; Paperno et al., 2016); we report perplexity on both and last-word prediction accuracy on LAMBADA.

2.2 Memorisation and Sequence Identification

Following prior work, we define memorisation using k-extractability. A sequence s is considered memorised if, when prompted with its first k tokens, the model reproduces the next k tokens verbatim under greedy decoding (Biderman et al., 2023a). Across all experiments, we set $k = 32$.

For the Pythia models, we sample 20,000 sequences per model from the public list of 32-extractable sequences (Biderman et al., 2023a). For GPT-2-XL, we construct an equivalent set by applying the criterion to every OpenWebText sequence, yielding 5,540 unique memorised sequences. Memorisation count is measured as the proportion of sequences reproduced verbatim under greedy decoding before and after head pruning. For all models, we also construct non-memorised sets by sampling the same number of sequences from the corresponding pretraining corpus whose greedy 32-token continuation does not exactly match the target continuation.

2.3 Identifying Task-Redundant Heads

We identify task-redundant attention heads by learning sparse binary gates over head outputs while keeping all model parameters frozen. For each head h in layer l, the gated output is

$$z'_{l,h} = b_{l,h} z_{l,h}, \qquad b_{l,h} = \mathbf{1}[\sigma(\alpha_{l,h}) > 0.5],$$

where $z_{l,h}$ is the head output, $\alpha_{l,h}$ is a learned gate logit, and $b_{l,h}$ is a binary gate. Hard gates are used in the forward pass, with gradients passed through $\sigma(\alpha_{l,h})$. The gates are optimised on task-specific validation data using the objective

$$\mathcal{L} = \max\left(0, \mathcal{L}_{\text{gated}} - \mathcal{L}_{\text{orig}} - \log r\right) + \lambda \sum_{l,h} \sigma(\alpha_{l,h}),$$

where r is the allowed perplexity ratio and λ controls sparsity. The objective discourages pruning configurations that exceed the permitted increase in loss, while encouraging fewer active heads. Heads with inactive gates after optimisation are zeroed out at inference time.

2.4 Mechanistic Analysis Metrics

To analyse how pruning affects memorisation-related computation, we compute three diagnostics on both memorised and non-memorised sequences. For each attention head, we measure Shannon entropy as:

$$H_{l,h} = -\frac{1}{N} \sum_{n=1}^{N} \frac{1}{T_n} \sum_{t=1}^{T_n} \sum_{j \le t} A^{(n)}_{l,h,t,j} \log A^{(n)}_{l,h,t,j},$$

where $A_{l,h,t,j}$ is the attention weight from position t to position j, averaged across sequences. Lower entropy indicates more concentrated attention, while higher entropy indicates more diffuse contextual mixing.

We measure softmax saturation as the number of attention rows with maximum weight above a threshold of 0.95, aggregated across token positions and sequences. This captures the extent to which a head places nearly all of its attention mass on a single position while remaining permissive enough to surface meaningful differences in saturation across sequence types (Mahdavi et al., 2024).

Finally, we measure the top-two logit margin, $\Delta_{\text{logit}} = \ell_{(1)} - \ell_{(2)}$, averaged across all generated positions. A larger margin indicates a stronger model preference for a single continuation. Comparing each diagnostic before and after pruning allows us to test whether head removal weakens the collective signal driving memorised reproduction.

3 Results

3.1 Effects of Pruning on Memorisation

Model	Active heads	PPL	Mem. count
Pythia-1.4B (baseline)	384 / 384	8.51	20,000
Pythia-1.4B	333 / 384	8.79	10583
Pythia-2.8B (baseline)	1,024 / 1,024	7.56	20,000
Pythia-2.8B	666 / 1,024	7.45	2,614
GPT-2-XL (baseline)	1200 / 1200	10.08	5540
GPT-2-XL	461 / 1200	9.27	434

Table 1: Effect of attention head pruning on WikiText-103 perplexity (PPL) and memorisation count.

Model	Active heads	PPL (last / full)	Accuracy	Mem. count
Pythia-1.4B (baseline)	384/384	3.62/18.50	0.7271	20,000
Pythia-1.4B	351/384	3.67/18.64	0.7215	6357
Pythia-2.8B (baseline)	1,024 / 1,024	3.18/17.35	0.7489	20,000
Pythia-2.8B	775 / 1,024	3.08/17.51	0.7543	6995
GPT-2-XL (baseline)	1200/1200	6.20/25.97	0.6128	5540
GPT-2-XL	642/1200	6.90/25.09	0.6148	1133

Table 2: Effect of attention head pruning on LAMBADA perplexity (PPL), accuracy, and memorisation count.

The central result, summarised in Tables 1 and 2, is that large reductions in memorisation are achievable with no meaningful cost to downstream performance, and in several cases with a measurable improvement. On WikiText-103, Pythia-2.8B reduces memorisation from 20,000 to 2,614 sequences (a 86.9% reduction) while perplexity *improves* from 7.56 to 7.45, pruning 35% of heads. GPT-2-XL achieves a 92.2% reduction in memorised-sequence reproduction, removing 61.6% of heads while improving perplexity by 0.81 points. The LAMBADA results reinforce this: Pythia-2.8B pruning improves both perplexity and last-word accuracy simultaneously, representing a strictly Pareto-dominant outcome relative to the unpruned baseline.

Pythia-1.4B is the clearest outlier across both benchmarks. On WikiText-103, pruning 51 heads (13.3%) reduces memorisation by 47.1% but raises perplexity from 8.51 to 8.79, the only setting in which task performance degrades. On LAMBADA, the perplexity increase is smaller and accuracy drops only marginally (0.7271 to 0.7215), but the pattern is consistent. This suggests that smaller models have less redundant capacity: a larger fraction of their heads are load-bearing for task performance, leaving a narrower pruning budget within which memorisation can be suppressed without incurring task costs. The finding is consistent with the broader observation that memorisation rates scale with model capacity (Carlini et al., 2022): larger models memorise more, but they also carry more redundant capacity that can be removed without consequence.

Across all models and benchmarks, inference-time head pruning produces substantial reductions in memorised-sequence reproduction. Before interpreting the pattern of results, we note a structural observation that is not visible in the tables directly: the sets of heads identified as task-redundant differ between the WikiText-103 and LAMBADA pruning objectives, and both sets differ from one another across models. Despite this, both pruning configurations suppress memorisation substantially inevery setting. This suggests that memorisation is unlikely to be localised in a single narrow subset of attention heads. Instead, memorised reproduction appears to be distributed across the attention mechanism, with different task-preserving subsets still capable of suppressing memorisation when removed.

3.2 Attention Saturation

For the mechanistic analysis, we focus on Pythia-2.8B and GPT-2-XL pruned under the WikiText-103 objective. We exclude Pythia-1.4B on the grounds that its pruning budget is constrained by task performance degradation, and restrict to a single benchmark objective for clarity of exposition.

Model	Variant	Δ Total entropy	Δ Average entropy	Δ Saturation count
Pythia-2.8B	Baseline	-135.72	-0.1325	17592672.0
Pythia-2.8B	Pruned	-26.81	-0.0261	1363456.0
GPT-2-XL	Baseline	-26.69	-0.0222	10564124.0
GPT-2-XL	Pruned	49.37	0.0411	1720101.0

Table 3: Differences in attention entropy and softmax saturation between memorised and non-memorised sequences for baseline and pruned models. For each metric, Δ denotes the memorised value minus the non-memorised value. Negative entropy differences indicate that memorised sequences induce more concentrated attention, while positive saturation-count differences indicate more frequent softmax saturation on memorised sequences.

Table 3 shows that attention heads saturate more frequently on memorised sequences than on non-memorised ones: both baseline models exhibit negative entropy deltas and saturation count differences in the tens of millions. However, the per-head contribution to this differential is small on average, the average entropy delta is -0.1325 and -0.0222 for Pythia-2.8B and GPT-2-XL respectively, indicating that no individual head accounts for a substantial share of the signal. Instead, the effect is

collective: many heads each contribute marginally, and their aggregate drives the saturation differential. Pruning a subset of these heads in combination is sufficient to dismantle it, reducing the saturation count difference by 92% in Pythia-2.8B and 83% in GPT-2-XL, with the GPT-2-XL entropy delta reversing sign entirely after pruning. This pattern is consistent with the central finding: memorised reproduction does not depend on a narrow set of critical heads, but on a distributed saturation signal that can be suppressed by removing any sufficiently large task-redundant subset.

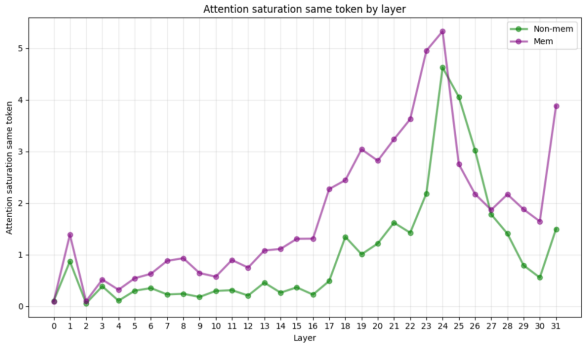

Figure 1: Average cross-head agreement across layers among heads exceeding the softmax saturation threshold. Memorised sequences exhibit stronger coordinated saturation across heads.

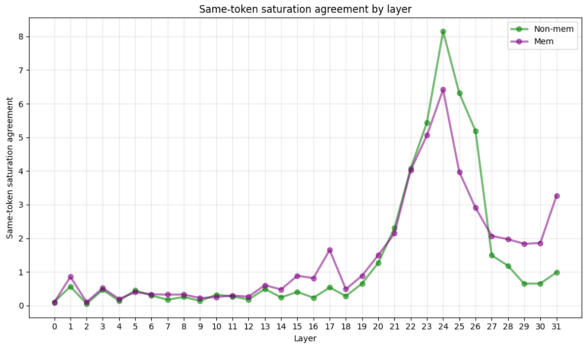

Figure 2: Average cross-head agreement across layers among heads exceeding the softmax saturation threshold, after pruning. Memorised and non-memorised sequences exhibit near-identical cross-head coordination, indicating that pruning dismantles the collective saturation signal driving memorised reproduction.

Figures 1 and 2 illustrate this coordination effect before and after pruning. In the baseline model, memorised sequences elicit markedly stronger cross-head agreement among saturated heads across layers. After pruning, this gap is almost entirely eliminated: memorised and non-memorised sequences show near-identical coordination profiles, confirming that the pruned heads were collectively responsible for the retrieval-like synchronisation observed at baseline.

Model	Setting	Non-memorised margin	Memorised margin
Pythia-2.8B	Baseline	2.01	8.79
Pythia-2.8B	Pruned	1.86	3.60
GPT-2-XL	Baseline	2.02	7.68
GPT-2-XL	Pruned	1.76	4.06

Table 4: Average top-two logit margin on memorised and non-memorised sequences before and after head pruning.

The logit margin results in Table 4 close the causal chain: baseline models produce top-two margins of 8.79 and 7.68 on memorised sequences, roughly four times the non-memorised margins, confirming that the collective saturation signal translates into anomalously high-confidence predictions. Pruning roughly halves these margins (to 3.60 and 4.06) while leaving non-memorised margins nearly unchanged, selectively disrupting the mechanism without reducing global model confidence. Aggregated over 20,000 sequences, this per-step reduction in confidence is sufficient to cause the model to exit the memorised track frequently, producing the large count reductions in Table 1.

4 Conclusion

We have shown that inference-time attention head pruning reduces memorised sequence reproduction by 47–92% across three models and two benchmarks, with no meaningful cost to downstream task performance and in several cases a measurable improvement. The result holds across pruning objectives that identify different head subsets, indicating that memorisation is distributed across the attention mechanism rather than localised in a narrow set of critical heads. Mechanistically, memorised sequences induce more frequent and spatially coordinated softmax saturation across heads, translating into anomalously high next-token confidence that pruning selectively disrupts without degrading general language modelling behaviour. However, the mechanistic account remains correlational; causal intervention analysis and probing of the MLP stream, which may retain an independent memorisation signal, are left to future work.

To support reproducibility, we will release the code, pruning masks, memorised-sequence identifiers, and raw metric values upon acceptance.

5 Limitations

Several limitations should be noted. Our memorisation evaluation uses k-extractability with $k = 32$ under greedy decoding, capturing only one form of memorisation; reduced counts under this criterion do not constitute a privacy guarantee, as training data may remain recoverable under sampling-based decoding, longer prompts, or adversarial extraction. Relatedly, attention-head pruning leaves model weights unchanged and should be understood as an inference-time mitigation rather than unlearning. The mechanistic account draws on theoretical results for attention-only transformers that do not directly extend to full architectures with residual connections, layer normalisation, and MLP blocks; our saturation hypothesis is therefore motivated by theory rather than a direct application of it. Experiments are restricted to three decoder-only models of moderate size with documented pretraining corpora, and it is unclear whether the results generalise to larger, instruction-tuned, or architecturally distinct models. Task preservation is assessed on two benchmarks, which may not reflect broader capabilities such as factual recall or long-context reasoning.

References

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023a. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023b. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Léo Dana, Muni Sreenivas Pydi, and Yann Chevaleyre. 2024. Memorization in attention-only transformers. *arXiv preprint arXiv:2411.10115*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. 2024. Memorization capacity of multi-head attention in transformers. In *International Conference on Learning Representations*, volume 2024, pages 49942–49967.

Stephen Merity and 1 others. 2016. The wikitext long term dependency language modeling dataset. *Salesforce Metamind*, 9.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1525–1534.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5797–5808.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, and 1 others. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

A Example Appendix

This is an appendix.