BioMedical LLM Lobotomy Project Final Report

Brandon Borzello (borze005), Shantanu Dalvi (dalvi020)
Lobotomizers
https://github.com/bborzello/biomedical-llm-lobotomy/

Abstract

Large Language Models (LLMs) require massive computational resources, which results in barriers to deployment in resource-constrained medical environments. While research such as the Lottery Ticket Hypothesis and SparseGPT demonstrates that general-purpose neural networks are extremely over-parametrized and compressible, the limits of this sparsity when it comes to a specific field like medical knowledge was largely unexplored. In this report, we outline the results of our study using unstructured magnitude pruning on the BioMistral-7B and Mistral architectures. By evaluating the models on the PubMedQA and MMLU datasets using increasing levels of sparsity, we aimed to find the precise computational cost of retaining specialized medical reasoning. Ultimately, our data invalidates the idea that fine-tuning makes a model brittle; instead, we found that domain specialization creates a "Structural Shield" that protects medical knowledge against severe architectural damage.

1 Introduction, Background, and Motivation

1.1 Problem Statement and Objectives

Modern LLMs have billions of parameters, which result in the requirement of massive amounts of computing power. The sheer scale of this computing makes it nearly impossible to run these models in resource-constrained or offline environments, such as local triage centers or hospitals. The core issue is that we do not know for certain how much of the massive parameter counts these models boast are strictly necessary for specialized, specific knowledge of particular fields. Our goal was to determine how much redundancy is present inside biomedical LLMs. More specifically, we wanted to know: How many of a biomedical LLM's parameters can be pruned before it experiences a drastic falloff in medical knowledge and diagnostic accuracy?

1.2 Current Practice and Limitations

Frankle and Carbin (2018) (Lottery Ticket Hypothesis) and Frantar and Alistarh (2023) (SparseGPT) provided the foundation as to how up to 50–60% of a neural network's weights could be redundant. However, the literature only focuses on general-purpose LLMs. The limitation of current practice is that the specific methodology of testing unstructured pruning on a specialized fine-tuned LLM, and comparing it to a general-purpose control model, had not been explored. The standard assumption is that fine-tuning strictly improves model utility, but it was unknown if forcing a model to specialize structurally compromised its foundational logic circuits.

1.3 Target Audience and Impact

This research is important for machine learning engineers and healthcare providers looking to deploy localized LLMs. By finding the limits of this sparsity, our hope is the findings from this could be applied to estimate the amount of computing power that can be saved by pruning. This would allow highly capable diagnostic tools to run on smaller, consumer-grade hardware.

2 Methodology and Approach

2.1 Experimental Design and Hypothesis

To test this, we evaluated both the open-source BioMistral-7B model (Labrak et al., 2024) and the base Mistral-7B model (Jiang et al., 2023). Because Mistral serves as the foundation for BioMistral, evaluating both provided a crucial control metric. This allowed us to isolate whether unstructured pruning disproportionately targets the medical knowledge in BioMistral or if the performance degradation is uniform across both models.

The two datasets we used were PubMedQA (Jin et al., 2019) and MMLU (Hendrycks et al., 2021).

The PubMedQA dataset is a set of medical questions with Yes/No/Maybe answers, while MMLU is a multiple-choice test with answers A/B/C/D. We chose specific subsets of MMLU to differentiate between medical questions and general knowledge.

Our hypothesis was split between two ideas: the "Fragility Tax" (the idea that fine-tuning overfits and makes the model brittle to pruning) and the "Structural Shield" (the idea that fine-tuning reinforces necessary logic circuits). First, we began with a baseline evaluation at 0% sparsity. Next, using a custom PyTorch script, we systematically zeroed-out the lowest-magnitude weights across the layers of both models. We used increments of 5% to get a better view of the tipping point before cognitive collapse.

2.2 Anticipated Challenges and Novelty

One challenge we encountered during the execution plan was VRAM constraints. Operating on a Dual T4 GPU environment necessitated a transition from global magnitude pruning to layer-wise L1 norm unstructured pruning. Furthermore, to evaluate accuracy effectively, we created a multi-tiered regex parser to successfully isolate true factual amnesia from simple instruction-following collapse. Our approach becomes unique through its specific application of comparing a specialized fine-tuned LLM to a general-purpose, but similar LLM under identical architectural damage.

3 Experiments, Results, and Error Analysis

3.1 Evaluation Metrics and Success Criteria

For the MMLU subsets (Logic and Medical), we tracked multiple-choice Classification Accuracy. For PubMedQA, where the questions are answered in a "Yes, No, Maybe" format, we realized we should also be tracking recall and precision, as a false-negative in a medical setting is a catastrophic failure. Therefore, we tracked the F1-score to measure success, which is an easy-to-visualize metric that is a good indicator as to whether or not the model starts answering "maybe" over and over.

3.2 Quantitative and Qualitative Findings

Our evaluations successfully charted the degradation curves for both models across 5% sparsity intervals. Interestingly, we discovered the "Zero-Shot Affirmative Bias" on PubMedQA. The base Mistral model achieved a high baseline (0.7965 F1) not through actual medical reasoning, but through an affirmative bias. The ground truth of our PubMedQA subset was heavily skewed toward "Yes" (55.2%), and both Mistral (74.4%) and BioMistral (73.6%) exploited this by defaulting to affirmative answers for complex questions they could not comprehend.

Because the baseline was compromised by this exploit, the true value of the fine-tuning was revealed during degradation. The data definitively invalidates the "Fragility Tax" hypothesis. BioMistral demonstrated superior structural resilience, yielding higher Area Under the Curve (AUC) robustness scores in both the MMLU Medical subset (0.3399 vs. 0.2894) and the MMLU Logic subset (0.2606 vs. 0.2282).

The fine-tuning process reorganized the network's weights to actively protect both its specialized medical knowledge and its foundational logic pathways against severe architectural damage up to the 60% sparsity cognitive cliff.

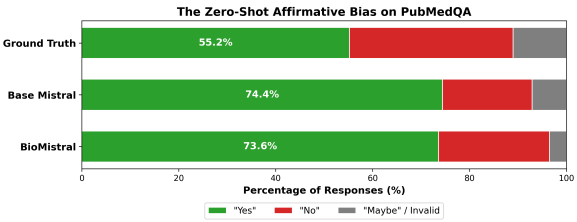

Figure 1: The Zero-Shot Affirmative Bias. Both models falsely achieve a high baseline by defaulting to "Yes" to exploit the ground truth's imbalance.

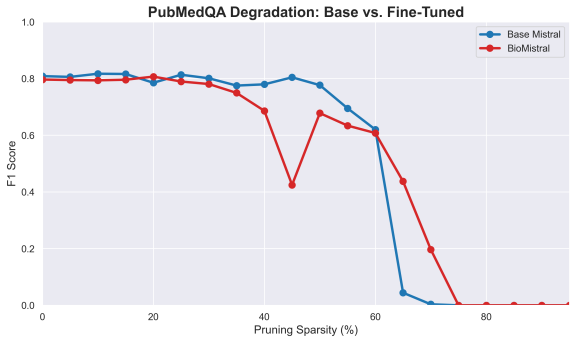

Figure 2: PubMedQA Degradation: Base Mistral vs. BioMistral across 5% sparsity intervals. The illusion of the tied baseline breaks as sparsity increases.

3.3 Error Analysis and Failure Cases

To understand the precise nature of the cognitive cliff, we isolated qualitative data points where our parser caught failures in the base model that

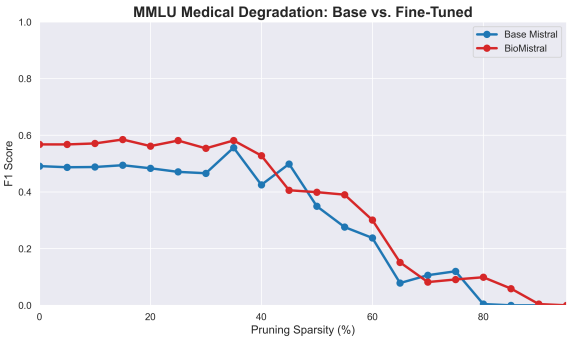

Figure 3: MMLU Medical Degradation. BioMistral exhibits a clear Structural Shield, retaining a significantly higher F1 score through the 60% cognitive cliff.

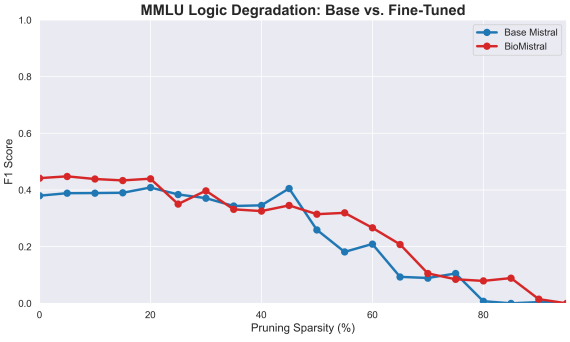

Figure 4: MMLU Logic Degradation. The medical fine-tuning unexpectedly fortified foundational deductive reasoning, granting BioMistral a higher AUC.

BioMistral successfully navigated. At 50% sparsity in the College Medicine subset, when asked what provides energy for muscle contraction, Base Mistral suffered from instruction collapse. While it generated the correct underlying string ("ATP"), it hallucinated the start of a new prompt ("QUESTION"), breaking the required multiple-choice structure and triggering an invalid ("X") parse. Conversely, BioMistral's structural shield maintained both factual accuracy and generative discipline, successfully outputting the correct semantic answer.

By 75% sparsity, the divergence is catastrophic. The base model deteriorates into total generative amnesia, outputting meaningless unicode loops (e.g., "Ú /******/"). However, there are still universal failure cases once BioMistral also crosses the 80% threshold. At this extreme sparsity, unstructured pruning begins severing critical self-attention mechanisms in both models, causing unavoidable formatting collapse. Our approach cannot address these extreme failure cases because unstructured magnitude pruning blindly targets weights without accounting for the transformer's holistic architecture. A potential solution for future work is to utilize structured pruning to remove entire attention heads rather than isolated parameters, which may preserve generative formatting even at high compression rates.

4 Discussion and Broader Impacts

4.1 Replicability

These results are highly replicable. Because we utilized open-source models (BioMistral and Mistral) rather than proprietary ones, and evaluated them on public benchmark datasets, any researcher with a custom PyTorch ablation script can reproduce these degradation curves.

4.2 Dataset Influence

Our findings regarding the PubMedQA dataset will force a shift in how future researchers approach medical AI benchmarking. By proving that PubMedQA is highly susceptible to the "Zero-Shot Affirmative Bias" exploit, our work demonstrates that researchers can no longer rely on raw baseline F1 scores for this dataset. Future development projects will be forced to either abandon zero-shot PubMedQA evaluations entirely or implement strict multi-tiered parsing defenses similar to ours to ensure they are measuring true medical reasoning rather than statistical probability gaming.

4.3 Ethical Considerations

There is a potential risk to society if highly pruned, "lobotomized" LLMs are deployed without rigorous testing. If a pruned model outputs a false-negative (sending a sick patient home and saying they're healthy), it is a catastrophic failure. To address this, developers must establish strict sparsity safety thresholds based on resilience curves like ours before deploying models in clinical settings.

4.4 Limitations and Future Work

A major limitation of our current model of testing is that unstructured magnitude pruning does not natively speed up inference or save electricity without specialized sparse-compute hardware. If we extend this work for future research, we can apply our findings to test other compression techniques, such as model quantization, to abstract our findings to estimate the exact amount of electricity or computing power that can be saved.

References

Jonathan Frankle and Michael Carbin. 2018. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10323–10337. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7B.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864.