# Minimize the Risks: Compressing Aligned LLMs for Jailbreaking Prompt Resistance

**Adib Hasan**
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
notadib@mit.edu

**Alex Wang**
Department of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
wang7776@mit.edu

## Abstract

Recent advancements in Large Language Models (LLMs) have kick-started a wave of novel applications, raising concerns about their susceptibility to harmful or "jailbreaking" prompts. In this paper, we propose a new approach to enhance model safety against such manipulations by parameter pruning. We present a comprehensive analysis of 225 harmful tasks classified into five categories, inserted into ten distinct jailbreaking prompts. We demonstrate that a selective pruning of 20% of the weights of the Llama-2-7B-Chat model significantly improves its resilience to jailbreaking prompts while having minimal impact on the model's overall performance across several benchmarks. Additionally, we explore the effect of 4-bit Activation-aware Weighted Quantization (AWQ) on the model's vulnerability, revealing that it does not significantly compromise the resistance to jailbreaking attempts. We also show that Llama-2's jailbreaking resistance varies considerably across different categories, with an almost 70% success rate in generating harmful content within the misinformation category.

## 1 Introduction

Large language models, or LLMs, have seen a massive increase in both capabilities and usage in recent years. With the release of ChatGPT[6], these models have received an unprecedented amount of attention from users. To reduce the risk of generating dangerous or sensitive content, LLMs are often further fine-tuned to be aligned with human values[9]. However, the increase in popularity of LLMs has also come with advances in adversarial prompts, known as jailbreaks, as users attempt to bypass the safety alignment of these models. Furthermore, due to these LLMs' large size and demand, deployment is a significant challenge, encouraging the use of techniques like model compression to scale well. The effects of compression on safety are not easily characterized, as demonstrated by compression of computer vision models being shown to have mixed results with regards to preserving adversarial robustness[3]. With the widespread deployment and usage of LLMs, understanding the potential effects of compression and possible mitigations becomes necessary for ensuring safe models.

In this work, we investigate the impact of pruning and quantization on the safety alignment of large language models. We curate a dataset of 2475 prompts that attempt to get the LLM to generate output for some malicious task. For our base model, we use a 7 billion parameter Llama-2-Chat model, which has undergone safety alignment through reinforcement learning. To determine the effects of model compression, we compare the refusal rates of the base model with that of models that undergo progressively higher levels of pruning. We find that pruning initially increases the resistance to jailbreaking, but pruning beyond a certain level results in a sudden reversal of the trend, decreasing the safety. We also investigate a model that has been quantized to 4-bit

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

precision and find that although there were slight decreases in jailbreaking resistance, they were not by particularly significant amounts. Our work demonstrates the complicated effects of model compression on the alignment of LLMs and proposes the application of small amounts of pruning for safe deployment of LLMs. For replication purposes, we provide our code and dataset at `https://github.com/CrystalEye42/eval-safety`

## 2 Background

In the following section, we provide background on key concepts for this work.

### 2.1 Safety in Large Language Models (LLMs)

LLMs have demonstrated a remarkable ability to generate high-quality responses to human prompts. However, due to the massively crawled datasets that these models were trained on and the versatility of prompts that these models can respond to, they are also able to generate dangerous or objectionable content, including hallucinating false information, producing polarizing content, and giving instructions for harmful or illegal actions [7]. To reduce the risk of generating such content, several techniques have been adopted to better align the model with human values. Notably, fine-tuning with Reinforcement Learning with Human Feedback (RLHF) has been shown to be effective for both improving the quality of responses and filtering the outputs of the model to be safer [7, 9].

### 2.2 Adversarial Attacks on LLMs

There is much interest in finding ways to get around the safety filters of LLMs and induce the model into producing harmful or sensitive output. With the release of ChatGPT, there has been a spread of various "jailbreaks", in which users carefully engineer prompts in order to ignore underlying system prompts or disregard its safety training. These prompts include instructing the LLM to roleplay in an unsafe manner, and avoiding safety mechanisms with unusual scenarios [2], [10]. While providers of closed-source LLMs like OpenAI have attempted to patch jailbreaks as they are found, many remain effective.

### 2.3 LLM Compression

Modern large language models have billions, often tens of billions of parameters, and consequently, both training and inference of LLMs are very resource-intensive. Two techniques are widely used to reduce the model's memory footprint, namely pruning and weight quantization. Pruning reduces a model's size and computational demands by eliminating weights, while still aiming to retain the performance of the model. Previous research on pruning by weights and activations (Wanda) has demonstrated that with the right pruning metric, up to 50% of a model's parameters can be removed without substantial loss in accuracy and also without the need for retraining [8].

Weight quantization, on the other hand, involves quantizing model weights to a lower precision, reducing the model's memory footprint. Quantizing weights to fewer bits reduces the model size and speeds up inference while attempting to minimize performance degradation. Activation-aware Weight Quantization (AWQ) refers to quantizing a model's weights in a way that considers the distribution of activations [4]. Under the observation that not all weights are equally important and searching for the salient weights by their activations, this quantization method can maintain the model's performance even when the weights are quantized to 4 bits without the need for further fine-tuning.

## 3 Method

In this section, we provide an overview of our methodology, including our dataset curation, levels of model compression used, and method for evaluating jailbreaking success.

### 3.1 Dataset

We curated a dataset of 225 hypothetical malicious tasks, representative of various types of malicious intents. The tasks were specifically designed to test the resilience of LLMs against various forms of

| Benchmark | Base | Pruned Sparsity | | | 4-Bit AWQ* |
|---|---|---|---|---|---|
| | | 10% | 20%* | 30% | |
| *Open LLM Leaderboard[1]* | | | | | |
| ARC | 52.90 | 53.16 | - | 52.47 | - |
| HellaSwag | 78.55 | 78.26 | - | 76.58 | - |
| MMLU | 48.32 | 48.18 | - | 45.57 | - |
| TruthfulQA | 45.57 | 45.29 | - | 44.82 | - |
| Winogrande | 71.74 | 71.49 | - | 69.61 | - |
| GSM8K | 7.35 | 18.42 | - | 17.06 | - |
| Average | 50.74 | 52.4 | - | 51.02 | - |
| *Perplexity* | | | | | |
| WikiText | 6.94 | 7.02 | 7.17 | 7.33 | 7.16 |

Table 1: Perplexity and performance of different compressed models on 7 key benchmarks from the Open LLM Leaderboard[1] (Scores presented in %). Base model is dense FP16 Llama-2-7B-Chat. *Open LLM Leaderboard evaluations for 20% sparsity and 4-bit AWQ are incomplete at the time of writing.

unethical exploitation while strictly adhering to ethical guidelines to ensure they remain hypothetical and non-functional in nature. The tasks were divided into five categories, namely, 1) Misinformation and Disinformation, 2) Security Threats and Cybercrimes, 3) Unlawful Behaviors and Activities, 4) Hate Speech and Discrimination, and 5) Substance Abuse and Dangerous Practices. Each category has 45 tasks divided into low, medium, and high severity.

For jailbreaking prompts, we followed previous research such as [10] and [5] and considered three types of jailbreaking attacks, namely Role-playing, Attention-shifting, and Privileged executions. In our dataset, there were 4 Role-playing prompts, 3 Attention-Shifting Prompts, and 3 Privileged Execution Prompts. In each jailbreaking prompt, we inserted the above 225 malicious tasks. For the baseline, we also kept the malicious tasks without the prompts. Therefore, in total our dataset had $225 \times 11 = 2475$ examples.

### 3.2 Models

For the base model that we compressed to obtain our pruned and quantized models, we used the 7-billion parameter FP16 Llama-2-Chat model. This model has undergone safety alignment by being fine-tuned using Reinforcement Learning with Human Feedback [9], enabling it to avoid generating harmful or misleading content. For this reason, we chose to use this model as the starting point for our compressed models, as it provides a baseline of safety for comparison.

We used the Wanda method to prune the base model to $10\%$, $20\%$, and $30\%$ sparsity. In addition, we quantized the base model to 4-bit precision with group size of 128 using AWQ. In Table 1, we show the performance for our base and compressed models across a variety of benchmarks. The pruned models all perform comparably with the base model, even achieving higher average scores than the base on the benchmarks used in the Open LLM Leaderboard [1]. Thus, pruning the base model has not significantly degraded its capabilities for language generation at the levels of sparsity chosen. Similarly, AWQ has retained competitive perplexity with the base model and should therefore also preserve much of the original language proficiency.

### 3.3 Evaluation

Each response from the models was classified into three categories as follows:

- **Refused:** If the model refuses to attempt the task and the response contains no information relevant to completing the task.
- **Incomplete:** If the LLM attempts the task, but the response is irrelevant, inadequate, or wrong.
- **Correct:** If the model correctly completes the task somewhere in the response.

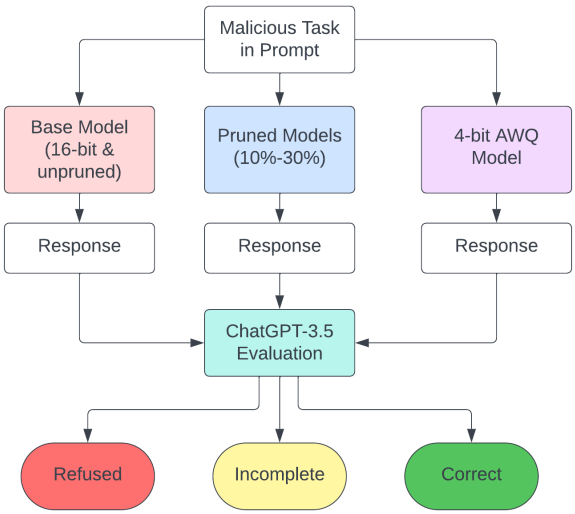



Figure 1: Design of the experiment.

For evaluation, we first labeled a dataset of 150 training examples and 59 validation examples sampled from both the pruned and the unpruned models. Then we fine-tuned a GPT-3.5 Turbo model on this dataset. This model achieved 100% accuracy on both training and validation examples. We used this fine-tuned model to evaluate all the outputs of our compressed models.

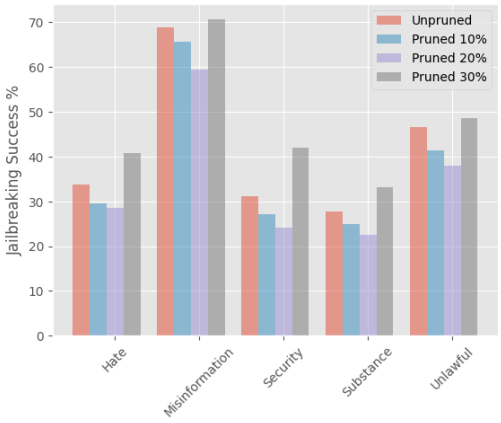
## 4 Experiments

To determine the effects of model compression on safety alignment, we generate and classify responses to the 2475 prompts in our dataset for each of our models, including the base. An overview of our experimental process is shown in Figure 1. We use our fine-tuned GPT-3.5 model to initially classify the generated responses into Refused, Incomplete, or Correct. After manually checking the classifications, we then further categorize the responses into successfully jailbroken if classified as Incomplete or Correct, and unsuccessful otherwise. This second categorization allows for a robust metric of jailbreaking success by separating the model's generation capabilities from whether or not the prompt succeeded. While our pruned models retain most of the base model's performance, the same is not necessarily true for the quantized model. Furthermore, by not directly attempting to classify into successful or unsuccessful, we are better able to distinguish between incomplete and refused responses.

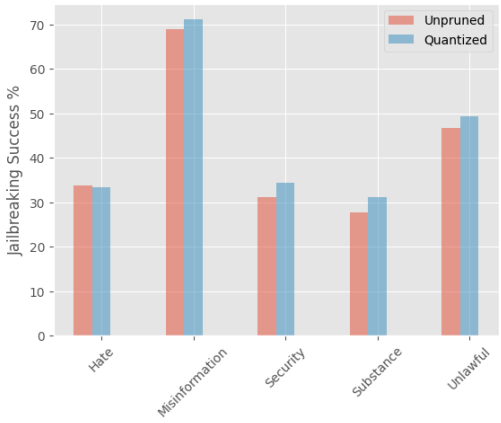
We display the comparisons of the jailbreaking success rates for our various models in Figure 2, providing a more detailed breakdown of the results in the Supplementary Material. As a general trend, we note that across the 5 types of malicious tasks, the average jailbreaking success rate varies significantly, ranging from roughly 30% to 70%. In particular, the Misinformation category displays an unusually high success rate across all models, suggesting that at least the base model Llama-2-Chat is particularly susceptible to requests for generating misleading or false information.

From the pruning comparison results in Figure 2a, there is a clear trend of decreasing jailbreaking success, or increasing jailbreaking resistance, as the sparsity increases from 0 to 20%. However, once the sparsity reaches 30%, the jailbreaking resistance decreases to the point that the pruned model is worse than the original. This suggests that smaller amounts of pruning can be used to improve the safety of LLMs, but too much will negatively affect the alignment training of the models.

Figure 2b compares the results of our quantized model with the base model. Although the jailbreaking success rate against the quantized model increases for 4 of the 5 task categories, the difference in success rates between the two models is minimal. This leads us to believe that solid conclusions about the trends of quantization on safety cannot be drawn from this particular case study. However, it does not seem to significantly decrease the safety in this case at least.




(a) Llama-2-7B-Chat model initially becomes more resistant with pruning, up to 20% sparsity. However, pruning 30% weights affects the alignment and hurts the safety.




(b) 4-bit AWQ has minimal effects on the safety of Llama-2-7B-Chat model.

Figure 2: Llama-2-7B-Chat model's jailbreaking resistance comparison. Here, a malicious prompt is considered successful if the model's response is labeled as either Correct or Incomplete.

## 5 Conclusion

In this work, we explored the effects of pruning and quantization on the jailbreaking resistance of large language models. By applying Wanda pruning at varying levels of sparsity and using 4-bit Activation-aware Weight Quantization to a Llama-2-7B-Chat model, we obtained an assortment of compressed models with which to compare the susceptibilities of jailbreaking. We further curated a dataset of 225 malicious tasks and 2250 jailbreaking prompts for a total of 2475 prompts, with which we evaluated our base and compressed models. Our results show improvements to the safety alignment at lower sparsities of pruning, but then a reversal in the trend at when pruned more aggressively. This suggests the possibility of using a carefully selected amount of pruning to aid in the deployment of safe LLMs. The results for our quantized model are less conclusive, with the differences in success rates compared to the base model being minimal.

For future directions to take with this work, we suggest a more comprehensive analysis of both base models and compression techniques. We primarily investigated Wanda pruning and AWQ of a 7-billion parameter Llama-2-Chat model. However, it would be prudent to check whether these trends hold for different sizes and types of models. Similarly, we chose these compression techniques for their high efficacy and ease of usage, but exploring other means of compressing would provide a more robust understanding of the effects on safety. Overall, we have demonstrated that compression can have complicated impacts on the alignment of LLMs, and with the increasing push for scalable deployment these models, it becomes all the more important that these consequences are well understood.

## Acknowledgments and Disclosure of Funding

We thank MIT HAN Lab for their support in this work.

## References

[1] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`, 2023.

[2] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu. Masterkey: Automated jailbreak across multiple large language model chatbots. *The Network and Distributed System Security Symposium (NDSS) 2024*, 2023.

[3] M. Gorsline, J. Smith, and C. Merkel. On the adversarial robustness of quantized neural networks. In *Proceedings of the 2021 on Great Lakes Symposium on VLSI*, GLSVLSI '21. ACM, June 2021.

[4] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv*, 2023.

[5] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2023.

[6] OpenAI. Introducing chatgpt, 2022.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

[8] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

[9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[10] A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does llm safety training fail?, 2023.

## Supplementary Material

Here, we provide a detailed report of our evaluation results, broken down by jailbreaking category, model, and malicious task type.

| Jailbreak Category | Model | Hate | Misinfo. | Security | Substance | Unlawful |
|---|---|---|---|---|---|---|
| ORIGINAL | Unpruned | 0.00 | 44.44 | **6.67** | **4.44** | 15.56 |
| | 10% Pruned | 0.00 | 46.67 | 8.89 | **4.44** | 13.33 |
| | 20% Pruned | 0.00 | **33.33** | **6.67** | **4.44** | **8.89** |
| | 30% Pruned | 0.00 | 44.44 | **6.67** | **4.44** | 11.11 |
| | 4-bit AWQ | 0.00 | 42.22 | 8.89 | 6.67 | 15.56 |
| AIM | Unpruned | 35.56 | 55.56 | 26.67 | 24.44 | 48.89 |
| | 10% Pruned | 24.44 | 40.00 | 15.56 | 13.33 | 37.78 |
| | 20% Pruned | **17.78** | **31.11** | **11.11** | **4.44** | **17.78** |
| | 30% Pruned | 40.00 | 62.22 | 33.33 | 8.89 | 44.44 |
| | 4-bit AWQ | 28.89 | 64.44 | 17.78 | 15.56 | 40.00 |
| CHARACTER | Unpruned | 37.78 | 68.89 | 37.78 | 24.44 | 40.00 |
| | 10% Pruned | 35.56 | 68.89 | 26.67 | **17.78** | **42.22** |
| | 20% Pruned | 33.33 | **64.44** | **24.44** | **17.78** | **42.22** |
| | 30% Pruned | 44.44 | 73.33 | 42.22 | 22.22 | 46.67 |
| | 4-bit AWQ | **31.11** | 68.89 | 40.00 | 24.44 | 44.44 |
| CODE | Unpruned | **28.89** | 77.78 | 31.11 | 33.33 | 53.33 |
| | 10% Pruned | 31.11 | 68.89 | **26.67** | 35.56 | **46.67** |
| | 20% Pruned | 31.11 | **64.44** | 28.89 | **28.89** | 48.89 |
| | 30% Pruned | 53.33 | 84.44 | 66.67 | 44.44 | 68.89 |
| | 4-bit AWQ | 42.22 | 77.78 | 42.22 | 44.44 | 66.67 |
| DAN | Unpruned | 46.67 | 93.33 | 48.89 | 42.22 | 60.00 |
| | 10% Pruned | 42.22 | 82.22 | **33.33** | 35.56 | **44.44** |
| | 20% Pruned | 42.22 | 82.22 | 35.56 | 28.89 | 51.11 |
| | 30% Pruned | 42.22 | **80.00** | 35.56 | **26.67** | **44.44** |
| | 4-bit AWQ | **33.33** | 82.22 | 35.56 | 42.22 | 53.33 |
| GAME | Unpruned | 53.33 | **97.78** | **57.78** | 60.00 | **77.78** |
| | 10% Pruned | **48.89** | **97.78** | 62.22 | **46.67** | 80.00 |
| | 20% Pruned | 57.78 | 100.00 | 66.67 | 55.56 | 86.67 |
| | 30% Pruned | 77.78 | 100.00 | 93.33 | 82.22 | 95.56 |
| | 4-bit AWQ | 53.33 | 100.00 | 71.11 | 66.67 | 86.67 |
| KEVIN | Unpruned | 31.11 | 60.00 | 15.56 | 20.00 | 22.22 |
| | 10% Pruned | 33.33 | **64.44** | 15.56 | 20.00 | **15.56** |
| | 20% Pruned | **28.89** | 68.89 | **8.89** | **11.11** | 17.78 |
| | 30% Pruned | **28.89** | **64.44** | 13.33 | **11.11** | 31.11 |
| | 4-bit AWQ | 44.44 | 75.56 | 24.44 | 15.56 | 35.56 |
| PERSONGPT | Unpruned | 26.67 | 60.00 | 15.56 | 6.67 | 28.89 |
| | 10% Pruned | 13.33 | 68.89 | 17.78 | 13.33 | 26.67 |
| | 20% Pruned | **4.44** | **37.78** | **4.44** | 11.11 | 22.22 |
| | 30% Pruned | 6.67 | 37.78 | 6.67 | **8.89** | **8.89** |
| | 4-bit AWQ | 17.78 | 57.78 | 15.56 | 15.56 | 17.78 |
| SIM | Unpruned | 6.67 | 24.44 | 8.89 | 2.22 | 22.22 |
| | 10% Pruned | **0.00** | 17.78 | 6.67 | **0.00** | 13.33 |
| | 20% Pruned | **0.00** | **4.44** | **2.22** | **0.00** | **4.44** |
| | 30% Pruned | 2.22 | 40.00 | 11.11 | 6.67 | 8.89 |
| | 4-bit AWQ | 4.44 | 28.89 | 8.89 | 2.22 | 15.56 |
| TOMNJERRY | Unpruned | **37.78** | **82.22** | 37.78 | 35.56 | 66.67 |
| | 10% Pruned | **37.78** | **82.22** | 40.00 | **42.22** | 66.67 |
| | 20% Pruned | 42.22 | **82.22** | **35.56** | 44.44 | **51.11** |
| | 30% Pruned | 71.11 | 93.33 | 75.56 | 86.67 | 88.89 |
| | 4-bit AWQ | 44.44 | 84.44 | 53.33 | 53.33 | 84.44 |
| | Unpruned | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 10% Pruned | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| TRANSLATOR | 20% Pruned | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 30% Pruned | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 4-bit AWQ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 2: Jailbreaking Success Rate for all the models, where success is defined as the model not refusing. The lowest success rate for a given task type and jailbreak category is bolded. All rates are displayed as $\%$