CHAPTER V
CONCLUSION

5.1 Summary

Based on the experimental results conducted in this paper, we present a summary of our findings.

1. Pruning larger models is more advantageous than pruning smaller ones, as the latency improvements and efficiency gains are more pronounced in higher-complexity models. This is likely due to the already small model size and reduced computational demands of the T5-small model.
2. While pruning effectively reduces overall LLM inference latency, it introduces a trade-off in terms of Time to First Token (TTFT). The increase of TTFT, ranging from 25% to 27% can be seen on both models, which may impact usability in latency-sensitive applications such as those deployed on edge devices.
3. Pruning can introduce instability in generation performance, leading to increased variability in latency depending on the input, which may impact real-time responsiveness. This effect is evident in the higher standard deviation observed in the pruned T5-base and T5-small variants compared to their fine-tuned counterparts.
4. Although pruning causes a slight degradation in model performance, the accuracy loss—typically under 2%—is generally acceptable when balanced against the significant latency improvements, particularly in larger models.
5. Pruning contributes to a modest reduction in computational resource usage, with observed decreases ranging from 2% to 10% in RAM, CPU, and occasionally GPU utilization. This reduction facilitates more efficient execution of large LLMs, especially in resource-constrained environments.
6. For larger models, pruning leads to a notable decrease in power consumption. reducing the power consumption by 14%, despite the average operating temperature remaining unchanged.

5.2 Suggestion

This research serves as a foundational step toward completing the researcher's thesis, and therefore is not without its limitations. Several areas offer opportunities for future exploration and improvement. The following suggestions are proposed as potential directions for the next research:

1. Explore More Complex Architectures: Future work could investigate the use of more complex encoder-decoder architectures combined with state-of-the-art pruning techniques to better understand Time to First Token (TTFT) behavior and its underlying factors.
2. Focus on Decoder-Only Models: Since much of the recent research has centered on decoder-only architectures, such as LLaMA and GPT, examining the pruning characteristics in these models could provide valuable insights that differ from encoder-decoder counterparts.
3. Incorporate Network Latency for Real-World Scenarios: To better simulate real-world deployment conditions, future studies should incorporate end-to-end latency measurements, including network latency, to assess model responsiveness in practical use cases.
4. Bigger Datasets: Expanding the evaluation to include larger and more diverse datasets can help validate the generalizability of the pruning approach. This includes testing on various input lengths, languages, and domains to understand how model compression scales and performs under different workloads.