My research topic: Filtering Prompt Tuning to Avoid Further Training

This project is concerned around refining the supervised learning model with prompt tuning, aiming to improve accuracy of the LLM by using a gate that predicts the quality/usability of soft prompts so that no further training is needed.

---

Useful papers:

Efficient Streaming Language Models with Attention Sinks:

Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023, September 29). *Efficient Streaming Language Models with Attention Sinks*. arXiv.org. https://arxiv.org/abs/2309.17453

**Research question:**

Can we deploy an LLM for infinite-length inputs without sacrificing efficiency and performance?
Findings:

They created StreamingLLM, an efficient framework that enables LLMs trained with a finite length attention window to generalize to infinite sequence length without any fine-tuning.

This was done by taking use of the attention sink of several initial tokens. These initial tokens are considered sink tokens because they are visible to all following tokens. A large amount of attention is being disproportionately spent on the first few tokens, and dense and window attention levels collapse after a certain training point. By reintroducing ~4 initial tokens as attention sinks, the model perplexity returns to normal and can accurately represent text over 4 million tokens.

Important findings for my research question:

The main use of this paper is their discovery of the attention sink in the first few tokens. We will be testing how adjusting the soft prompt prepended onto the frozen tokens, so knowing the importance of attention sink in initial tokens will be very useful for gauging the effect of prompt tuning with a gate.

---

P-Tuning v2:

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2021, October 14). *P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks*.

arXiv.org. https://arxiv.org/abs/2110.07602

Research findings:

Properly optimized prompt tuning can be universally effective across a wide range of model scales and NLU tasks. They can match the performance of fine-tuning, while only having to use 0.1%-3% tuned parameters.

Important findings for my research question:

The discovery that prompt tuning could possibly match the accuracy of fine-tuning is very important for my research question. Because my hope is to create an accurate LLM that only uses prompt tuning with no further refinement, this paper reinforces the possibility of this method. Since their github and datasets are public, I will be building off of their resources to experiment and hopefully find some insightful results.

---

The First Few Tokens Are All You Need for Fine-Tuning Reasoning Models:

Ji, K., Xu, J., Liang, T., Liu, Q., He, Z., Chen, X., Liu, X., Wang, Z., Chen, J., Wang, B., Tu, Z., Mi, H., & Yu, D. (2025). The first few tokens are all you need: an efficient and effective unsupervised prefix Fine-Tuning method for reasoning models. *arXiv.org*. https://doi.org/10.13140/RG.2.2.33772.07043

**Research Question:**

In this paper, they propose an unsupervised fine-tuning method that requires only a single pass of model-generated responses per question, coupled with prefix-based fine-tuning.

Findings:

Just by using prefix substrings for guidance, they were able to outperform full-token fine-tuning approaches. By taking this approach, the training time and inference time is greatly reduced.

Important findings for my research question:

Because this approach uses unsupervised fine-tuning, most methods will not be applicable to the supervised fine-tuning that we plan to conduct for our research question. However, the prefix-based fine-tuning used in this paper will provide important insights to how we can improve fine tuning in our approach.

---

ZeroTuning: Unlocking the Initial Token’s Power to Enhance Large Language Models Without Training:

Han, F., Yu, X., Tang, J., & Ungar, L. (2025, May 16). *ZeroTuning: Unlocking the initial token’s power to enhance large language models without training*. arXiv.org. https://arxiv.org/abs/2505.11739

**Research Question:**

Is it possible to improve model performance by tuning the attention to a universal and task-agnostic token without relying on task-specific token identification?

Findings:

Tuning the initial token consistently improves LLM performance more than any other token. The tuning effect propagates through all the layers, meaning that the initial and middle layers contribute more than the later ones. Tuning all the layers jointly yields the best results. Some attention heads respond positively while others respond negatively to initial token p-tuning. They found that selectively tuning them outperforms tuning them uniformly.

They created the ZeroTuning method, a training-free method that improves the LLM's performance through recalibrating the initial token's attention without requiring task-specific token identification.

Important findings for my research question:

They have found that some attention heads react positively, while other attention heads react negatively to initial token amplification. The proportion of negative versus positive heads differs based on the model, explaining the difference in uniform scaling.

This is a useful observation to consider for my research. Perhaps the gate should be able to predict whether heads have a negative or positive prediction, and for the predicted negative heads, perform a different form of initial token tuning that will result in a positive effect.

Useful models and tools:
P-tuning v2
CB and COPA datasets