Project Proposal for Team InnovAdapt

**Project Title:** Tuning Adapter for LLMs for Financial Applications

**Team Member:** Charles Kniffin, Winston Lai, Yohann Venkateswaran

**Project Summary:**

Traditionally applying a large pre-trained model for specific tasks requires a large amount of fine-tuning, which is parameter inefficient and requires large amounts of computational power. Fine tuning is typically done by taking the pre-trained model and performing updates across all parameters while training over supervised data. The motivation for adapters is to simplify and reduce the amount of parameters required to be trained/tuned by freezing the base network and instead training over a small subset of parameters that have been added to the network. Traditional adapters require architectures and as a result alternatives such as LoRA have been created. This project aims to use adapters for the Mistral 7b model designed to enhance its ability to converse about finance related topics. Using adapters for this project presents an interesting challenge as it presents the opportunity to tailor a very large model using a relatively small amount of compute for a subject area that has the potential to provide large returns.

**Approach:**

Based on our review of the literature (Ref. 1 and 2), transfer learning helps reduce the training time and the amount of parameters needed, with LoRA and QLoRA being adaptation techniques. LoRA and QLoRA help reduce the number of trainable parameters during fine tuning, and we will use these techniques to inject finance related knowledge into the model. We will be using the HuggingFace Transformers library, Oobabooga, or Predibase to conduct the training. We will then analyze the advantage of using these adapters on the network using performance metrics to judge the networks such as NER and FiQA SA. If time allows we will also bench mark the model against a LLaMa based model from AdaptLLM which has already been tuned with finance knowledge.

**Related Work:**

State of the hard for Finance tuned LLMs are likely closely held trades secret models from Hedge Funds that we do not have access to. Open source work includes the model by AdaptLLM [6] and other community models. We also intend to incorporate the recent work by Predibase [5] on how to train adapter. Data sets were also made available as part of the work by AdaptLLM. The survey of references and data below is not all encompassing and more will be incorporated over the course of the project.

**Datasets:**

<https://www.kaggle.com/datasets/Cornell-University/arxiv>

<https://huggingface.co/datasets/Open-Orca/OpenOrca>

<https://huggingface.co/datasets/GAIR/lima>

<https://huggingface.co/datasets/WizardLM/WizardLM_evol_instruct_V2_196k>

<https://allenai.org/olmo>

**Resources/Citations**

[1] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In International conference on machine learning (pp. 2790-2799). PMLR.

[2] Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3712-3722).

[3] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2020). Adapterfusion: Non-destructive task composition for transfer learning. arXiv preprint arXiv:2005.00247.

[4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[5] Wang, Zhao, Eaton, (2024), LoRA Land: Fine-Tuned Open-Source LLMs that Outperform GPT-4 (<https://predibase.com/blog/lora-land-fine-tuned-open-source-llms-that-outperform-gpt-4>)

[6] Cheng, Huang, Wei, (2023), Adapting Large Language Models via Reading Comprehension. arXiv: arXiv:2309.09530