Proposal 1 - Literature Review:
Aaron Wu, Lucas Li, Benjamin Mao

Our project focuses on conducting a comprehensive literature review of adaptive and agentic AI systems, centered around the four core papers in the Agentic AI chapter with additional exploration of related work. Our goal is to synthesize the key approaches, identify and refine complexities, challenges, and models, and structure a detailed analysis of the current state of agentic AI optimization techniques for presentation.

The core themes we'll explore include
1. Adaptive Prompt Tuning: Analyzing how "One Size Fits All for Semantic Shifts" approaches continual learning through adaptive prompt mechanisms and comparing it with other adaptive tuning approaches in the field.
2. Prompt Compression Techniques: Examining "Compress, Then Prompt" alongside related work in efficient LLM inference to understand the broader landscape of accuracy-efficiency trade-offs.
3. Personalization in Federated Systems: Studying "FedJETs" in the context of other federated learning approaches that tackle personalization challenges.
4. Task Adaptation Strategies: Investigating "Sweeping Heterogeneity with Smart MoPs" and its relationship to other prompt mixture approaches in the literature.

Additionally, we plan on exploring other papers that discuss similar ideas including "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts" and "Black-Box Prompt Learning for Pre-trained Language Models". AutoPrompt introduces an automated approach to prompt generation that uses gradient-based optimization to find discrete text prompts, offering insights into how we can automate the process of prompt engineering for better model performance. Black-Box Prompt Learning presents a novel framework for optimizing prompts without access to model internals, demonstrating methods for efficient prompt learning in scenarios where model parameters cannot be directly modified.

Our plan is broken down into several weeks. For weeks 1-2, we will focus on an initial deep dive into the four core papers and other relevant papers, where we will create detailed summaries and identify key technical approaches that define the foundation of agentic AI systems. For weeks 3-6, we will move onto analyzing common patterns, challenges, and potential future directions across the papers, synthesizing our findings to identify key trends and opportunities in the field. For weeks 7-8, we want to develop comparative frameworks to evaluate different approaches, creating structured methods to assess the strengths and limitations of various techniques in agentic AI. Finally, for weeks 9-10, we will focus on synthesizing our findings and preparing the final report, which involves creating comprehensive documentation of our analysis, developing clear visualizations of technical comparisons, and ensuring our review provides valuable insights for future research directions. We will document our findings in an oral presentation and LaTeX report following the ICML format, including detailed technical comparisons, theoretical foundations, and practical implications of different approaches.
https://arxiv.org/abs/2010.15980 https://arxiv.org/abs/2201.08531

Proposal 2 - Independent Research:
Aaron Wu, Lucas Li, Benjamin Mao

Our research project proposes to explore the intersection of efficient prompt compression and adaptive task specialization in large language models. Building upon the foundations laid in the Agentic AI papers, particularly "Compress, Then Prompt" and "Sweeping Heterogeneity with Smart MoPs", we aim to develop a novel approach that combines prompt compression with dynamic adaptation capabilities.

The core methodologies we'll explore are
1. Development of a hybrid approach that maintains the efficiency benefits of prompt compression while preserving the adaptability of mixture-of-prompt systems
2. Implementation of an efficient mechanism for dynamic prompt modification based on task requirements to optimize the trade-off between compression ratio and adaptation capability
3. Creation of a benchmark suite to evaluate both compression efficiency and adaptation capability to determine theoretical bounds on performance when combining the approaches

Our plan is broken down into several weeks. For weeks 1-2, we will focus on literature review and theoretical framework development, where we will thoroughly analyze existing approaches and establish the mathematical foundations for our hybrid system. For weeks 3-4, we will focus on implementing baseline systems from both papers for comparison, setting up our development environment and establishing performance benchmarks of the original approaches. For weeks 5-6, we will move onto developing our hybrid approach and conducting initial testing, where we will implement our combined compression-adaptation mechanism and perform preliminary evaluations. For weeks 7-8, we want to conduct comprehensive experimentation and performance analysis by running extensive tests across various benchmarks, comparing our approach against baselines, and analyzing the efficiency-adaptability trade-offs. Finally, for weeks 9-10, we will focus on documentation and preparation of results, which involves analyzing our experimental findings, preparing detailed technical documentation, and creating a comprehensive report of our methodology and results with reproducibility guidelines.

Our implementation will use Python with PyTorch on one GPU, focusing on creating an efficient and reproducible codebase. We will evaluate our approach using standard benchmarks from both papers and additional tasks to test generalization. Our implementation will be thoroughly documented in a GitHub repository and presented through both an oral presentation and LaTeX report following the ICML format. The documentation will include our theoretical framework, implementation details, experimental methodology, and comprehensive performance analyses. We aim to bridge the gap between prompt compression and adaptive systems in a way that advances the field while remaining accessible, providing clear explanations of our hybrid approach for those new to the topic, as well as detailed technical insights and empirical evidence for researchers looking to build upon our work. Through this project, we hope to not

only demonstrate the feasibility of combining compression efficiency with adaptive capabilities but also establish a foundation for future research in this direction.