Midterm Project Report
Agentic AI System with Evaluation-Driven Prompt Optimization

1. Introduction

This midterm project presents an Agentic AI system designed to solve complex analytical tasks through a structured multi-agent architecture. Unlike a standard chatbot, the system separates knowledge retrieval, reasoning, and synthesis, while incorporating a formal evaluation framework to ensure reliability, controllability, and measurable improvement.

2. System Architecture

The system follows a modular Agentic architecture. User queries are processed through a retrieval layer that supplies relevant context to multiple specialized agents. Each agent performs a distinct function, and their outputs are aggregated into a final response. This design enables independent evaluation and iterative improvement of each component.

3. Agent Descriptions

Market Data Agent: Retrieves and structures relevant quantitative information.
Fundamental Analysis Agent: Performs reasoning and interpretation based on retrieved data.
Portfolio Analysis Agent: Integrates insights into coherent analytical outputs.
Summarizer Agent: Produces a concise, user-facing response grounded strictly in agent outputs.

4. Methodology: ReAct and Agentic Flow

The system adopts the ReAct paradigm, combining reasoning steps with action execution. Agents explicitly reason over retrieved context before producing outputs, reducing hallucinations and enabling transparent error analysis.

5. Evaluation Framework

Evaluation is central to the system design. Three complementary evaluation methods are used: deterministic tests against known answers, LLM-based evaluation for qualitative assessment, and selective human review. Regression evaluations are applied when modifying prompts or models to ensure no degradation in performance.

6. Prompt Optimization Experiment (GEPA)

A structured prompt optimization experiment was conducted using the GEPA framework in light mode. The goal was to assess whether evaluation-driven reflection could improve generalization across increasing task difficulty.

7. Dataset and Experimental Setup

The MATH dataset was selected due to its clearly defined difficulty levels (L1–L5). Initial tuning was performed on approximately 20 Level-2 questions, followed by evaluation and refinement on Level-4 questions.

8. Results and Comparative Analysis

Agent	Training Strategy	Evaluation Level	Outcome
A	L2 Training + GEPA + Reflection	L4	Improved generalization, moderate accur
B	Direct L4 Training	L4	Higher accuracy, task-specific robustness

The comparison shows that reflection-based prompt optimization significantly improves performance on unseen difficulty levels. However, direct exposure to high-difficulty data remains necessary for optimal results.

9. Limitations and Future Work

The current system relies on prompt-based learning rather than parameter updates. Future work may include automated curriculum learning, expanded evaluation datasets, and integration with production-grade monitoring.

10. Conclusion

This project demonstrates that Agentic AI systems combined with evaluation-driven prompt optimization offer a practical path toward reliable and controllable AI behavior. The results highlight the importance of structured evaluation and iterative refinement in real-world AI systems.