Midterm Project Report
Prompt Optimization with Evaluation and Reflection

1. Prompt Optimization During Inference

This project demonstrates that prompts can be improved iteratively during runtime by leveraging evaluation signals and reflection generated by a language model. The system analyzes model responses, identifies recurring errors or weaknesses, and updates the prompt to include clearer constraints, structured reasoning requirements, and explicit solution guidelines.

2. Use of GEPA in Light Mode

GEPA was executed in light mode in order to demonstrate the optimization mechanism without incurring excessive computational cost. Since the objective of this project is experimental rather than product-oriented, the emphasis is placed on methodological clarity rather than peak performance.

3. Dataset Selection: MATH (L1–L5)

The MATH dataset provides a structured difficulty hierarchy from Level 1 to Level 5. This enables a controlled evaluation of generalization capabilities across increasing levels of complexity.

4. Training on Difficulty Level 2

Approximately 20 Level-2 questions were used as a tuning set. This stage focused on learning solution structure, mathematical rigor, and prompt responsiveness to feedback.

5. Evaluation and Prompt Improvement on Level 4

The tuned prompt was evaluated on Level-4 questions. Model-generated reflections were analyzed to identify reasoning gaps, leading to prompt updates that emphasized step-by-step decomposition, intermediate verification, and logical consistency.

6. Retraining on Level 4

Following prompt refinement, the system was further trained on Level-4 questions to assess whether the improved prompt enabled more robust handling of higher-complexity problems.

7. Comparative Analysis (Key Experiment)

Agent	Training Strategy	Evaluation Level	Observed Performance
A	L2 Training + GEPA + Reflection	L4	Improved generalization, moderate error rate
B	Direct L4 Training	L4	Higher accuracy, stronger task-specific perfor

The comparison indicates that prompt optimization via reflection significantly improves generalization from lower to higher difficulty levels. However, direct exposure to high-difficulty data remains essential for achieving peak performance. Reflection enhances robustness but does not fully replace task-specific training.

8. Conclusion

This project demonstrates that evaluation-driven prompt optimization is an effective mechanism for improving agent performance and generalization. The results highlight both the strengths and limitations of reflection-based learning, emphasizing the importance of combining iterative prompt refinement with targeted data exposure.