Final Project Report

GitHub Repo Link: https://github.com/fpsmika/CSE476_Project/tree/main

Introduction:

- For the final project, I developed an agent that utilizes three different Inference-Time Techniques: Chain of Thought, Self Consistency, and Self Reflection
- Before using a specific technique, each question from the JSON test data file is being processed by the Question Classifier method where the LLM determines which technique would produce the best answer for that type of question.
- One technique is selected, the agent loop method runs the selected technique.
- Lastly, the "Generate Answer" pipeline loads the question from the JSON file, and outputs it into the "Answer" JSON file

Technique 1: Chain of Thought

- Prompts the model to break down the problem into smaller parts to later solve step-by-step
- Utilizes 2 api calls, one for prompting, one for answer extraction

Technique 2: Self Consistency

- Generates 3 independent solutions via temperature set to 0.7
- Normalizes answers using majority voting

Technique 3: Self Reflection

- Generates an initial solution to later critique and refine it through the maximum of 3 interactions
- At each iteration, the model calls to critique, and model calls to produce a refined solution based on the received feedback.

Question Classifier:

- Analyze the question to categorize it into either math, common_sense, or logic questions.
- Maps each of the categories to the suitable inference-time techniques
- Uses self reflection by default

Agent Loop:

- Calls the question classifier
- Returns the answer extracted from the selected technique's output

Generate Answer Pipeline:

- Loads the questions from the JSON test_data file
- Iterates through each question and calls the agent loop to generate answers. Saves checkpoints for every 20 questions for recovery.
- Validates the output results in the answer JSON file.
- Sets a limit of 4999 characters for each output.