Assignment 1.3

Model Comparison Analysis Report

Executive Summary

This report analyzes the performance of four language models (GPT-4, GPT-4o, GPT-4o-mini, and GPT3.5) across five diverse questions. The models were evaluated based on their answers, peer voting, and average rankings.

Model Performance Overview

Total Questions Analyzed: 5

Overall Performance Metrics

- **GPT-4o**:
  - Total Wins: 4
  - Total Votes: 13
  - Average Ranking: 1.20
- **GPT-4**:
  - Total Wins: 4
  - Total Votes: 13
  - Average Ranking: 1.27
- **GPT3.5**:
  - Total Wins: 4
  - Total Votes: 12
  - Average Ranking: 1.20
- **GPT-4o-mini**:
  - Total Wins: 3
  - Total Votes: 11
  - Average Ranking: 1.47

Question-by-Question Analysis

1. String Processing Question

Question: "How many 'r' characters are present in the word 'strawberry'?"

- Winners: GPT-4o, GPT-4o-mini, GPT-4
- All winning models correctly identified 3 'r' characters
- GPT3.5 provided an incorrect answer (2 characters)

2. Mathematical Calculation

Question: "What is the result of 135 × 27 + 896 ÷ 4?"

- Winner: GPT3.5
- GPT3.5 provided the correct answer (3869) with detailed step-by-step calculation
- Other models provided incorrect answers

3. Logical Reasoning

Question: "If all roses are flowers and some flowers are red, can we say all roses are red?"

- Winners: GPT-4o, GPT3.5, GPT-4
- All models provided logically sound explanations
- High consensus among models in voting

4. Average Speed Calculation

Question: "A car travels 40 miles per hour for the first 3 hours, then 50 miles per hour for the next 2 hours, and finally 60 miles per hour for the last 1 hour. What is the car's average speed?"

- Winners: All models (GPT-4o, GPT-4o-mini, GPT3.5, GPT-4)
- All models provided correct calculation (46.67 mph)
- High-quality explanations with step-by-step workings

5. Riddle/Abstract Thinking

Question: "I have cities, but no houses; I have forests, but no trees; I have rivers, but no water. What am I?"

- Winners: All models (GPT-4o, GPT-4o-mini, GPT3.5, GPT-4)
- All models correctly answered "map"
- Perfect consensus among all models

Conclusion

Overall Winners

Based on the analysis, three models emerged as top performers:

1. **GPT-4o**
2. **GPT-4**
3. **GPT3.5**

Key Findings

1. **Consistency**: GPT-4o showed the most consistent performance across all question types.
2. **Accuracy**: All models performed exceptionally well on logical reasoning and riddles.
3. **Mathematical Capability**: GPT3.5 surprisingly outperformed other models in complex mathematical calculations.
4. **Explanation Quality**: All models provided detailed explanations when required.

Model Strengths

- **GPT-4o**: Best overall performance, consistent across all categories
- **GPT-4**: Strong in logical reasoning and detailed explanations
- **GPT3.5**: Excellent mathematical computation capabilities
- **GPT-4o-mini**: Strong performance in straightforward tasks

Methodology Note

Each model voted on others' responses and provided rankings. Models could not vote for themselves, ensuring unbiased evaluation. Rankings were on a scale where 1 was the best score possible.

model_responses.json

analyseResult.py

```
{
  "total_questions": 5,
  "model_performance": {
    "GPT-4o": { "total_wins": 4, "total_votes": 13, "average_
    "GPT-4o-mini": {
      "total_wins": 3,
      "total_votes": 11,
      "average_ranking": 1.4666666666666666
    },
    "GPT-4": {
      "total_wins": 4,
      "total_votes": 13,
      "average_ranking": 1.2666666666666668
    },
    "GPT3.5": { "total_wins": 4, "total_votes": 12, "average_
  },
  "overall_winners": ["GPT-4o", "GPT-4", "GPT3.5"]
}
```