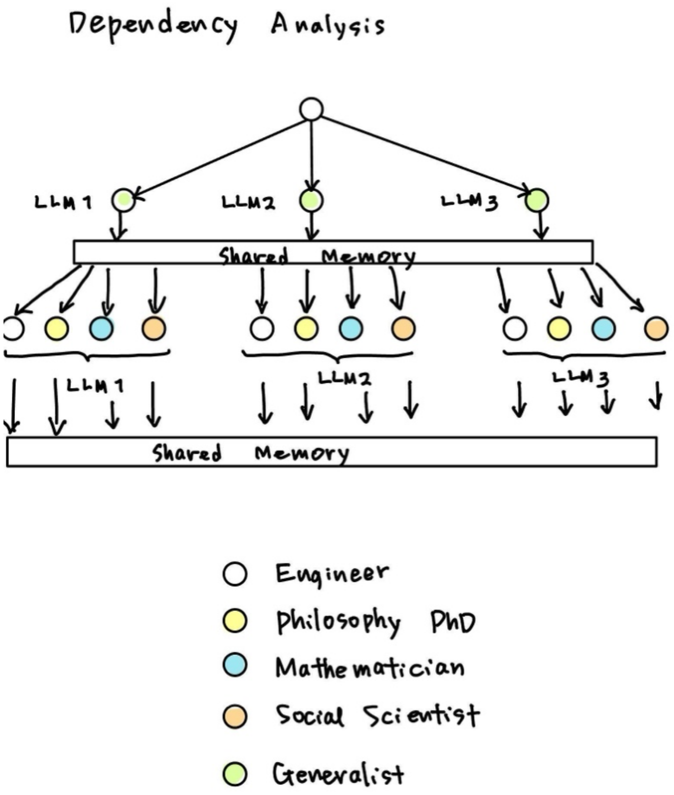
EE 451
Kae Sawada

Course Project Revised Problem Statement and Hypothesis

Introduction

Profiling the source code (to be parallelized), along with the construction and analysis of the dependency graph shown below, characterized the task-level latencies unique to this workflow (see the appendix for the raw results). These latencies are detailed in the Problem Statement section below.



Problem Statement

- **Observation**: Tasks within the same layer are independent and can be executed in parallel
    - **Solution**: Parallelize tasks within a layer using multiple workers
- **Observation**: Some tasks, like phi-Engineer, take significantly longer than others

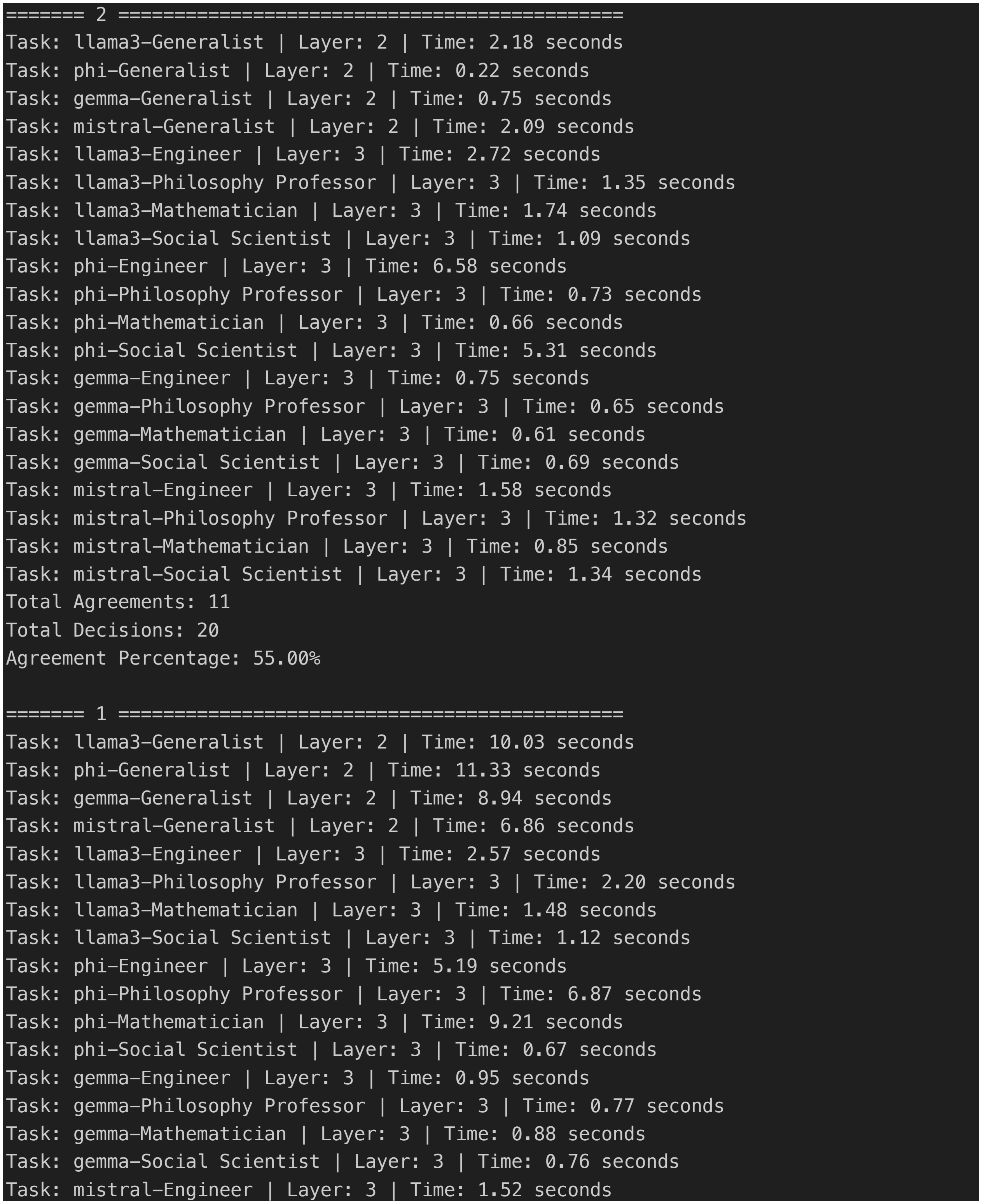
    - **Solution**: Start these high-latency tasks early to prevent them from becoming bottlenecks (task-level scheduling to give phi-Engineer higher priority)
- **Observation**: Models like *gemma* and *mistral* are faster overall.
    - **Solution**: Group tasks for gemma and mistral into a batch and assign them to workers in parallel. This allows for slower tasks like phi-Engineer do not idle the system.
- **Observation**: Latency variability across runs suggests external factors (e.g., server load, network latency).
    - **Solution**: Dynamically assign tasks to workers based on their current load or estimated task duration.
- **Observation**: Race conditions are unlikely to occur

Hypothesis

By prioritizing high-latency tasks (e.g., phi-Engineer), batching fast tasks (e.g., gemma), and parallelizing independent tasks within layers, we expect to reduce makespan by 20–30%

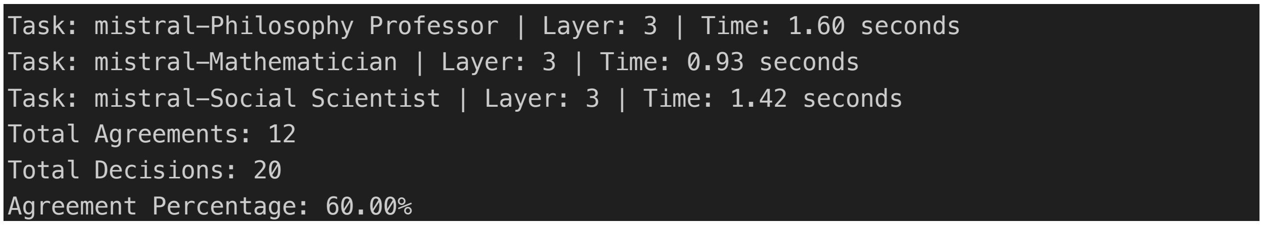
Appendix

Source Code Profiling Result



```
======= 2 ============================================
Task: llama3-Generalist | Layer: 2 | Time: 2.18 seconds
Task: phi-Generalist | Layer: 2 | Time: 0.22 seconds
Task: gemma-Generalist | Layer: 2 | Time: 0.75 seconds
Task: mistral-Generalist | Layer: 2 | Time: 2.09 seconds
Task: llama3-Engineer | Layer: 3 | Time: 2.72 seconds
Task: llama3-Philosophy Professor | Layer: 3 | Time: 1.35 seconds
Task: llama3-Mathematician | Layer: 3 | Time: 1.74 seconds
Task: llama3-Social Scientist | Layer: 3 | Time: 1.09 seconds
Task: phi-Engineer | Layer: 3 | Time: 6.58 seconds
Task: phi-Philosophy Professor | Layer: 3 | Time: 0.73 seconds
Task: phi-Mathematician | Layer: 3 | Time: 0.66 seconds
Task: phi-Social Scientist | Layer: 3 | Time: 5.31 seconds
Task: gemma-Engineer | Layer: 3 | Time: 0.75 seconds
Task: gemma-Philosophy Professor | Layer: 3 | Time: 0.65 seconds
Task: gemma-Mathematician | Layer: 3 | Time: 0.61 seconds
Task: gemma-Social Scientist | Layer: 3 | Time: 0.69 seconds
Task: mistral-Engineer | Layer: 3 | Time: 1.58 seconds
Task: mistral-Philosophy Professor | Layer: 3 | Time: 1.32 seconds
Task: mistral-Mathematician | Layer: 3 | Time: 0.85 seconds
Task: mistral-Social Scientist | Layer: 3 | Time: 1.34 seconds
Total Agreements: 11
Total Decisions: 20
Agreement Percentage: 55.00%

======= 1 ============================================
Task: llama3-Generalist | Layer: 2 | Time: 10.03 seconds
Task: phi-Generalist | Layer: 2 | Time: 11.33 seconds
Task: gemma-Generalist | Layer: 2 | Time: 8.94 seconds
Task: mistral-Generalist | Layer: 2 | Time: 6.86 seconds
Task: llama3-Engineer | Layer: 3 | Time: 2.57 seconds
Task: llama3-Philosophy Professor | Layer: 3 | Time: 2.20 seconds
Task: llama3-Mathematician | Layer: 3 | Time: 1.48 seconds
Task: llama3-Social Scientist | Layer: 3 | Time: 1.12 seconds
Task: phi-Engineer | Layer: 3 | Time: 5.19 seconds
Task: phi-Philosophy Professor | Layer: 3 | Time: 6.87 seconds
Task: phi-Mathematician | Layer: 3 | Time: 9.21 seconds
Task: phi-Social Scientist | Layer: 3 | Time: 0.67 seconds
Task: gemma-Engineer | Layer: 3 | Time: 0.95 seconds
Task: gemma-Philosophy Professor | Layer: 3 | Time: 0.77 seconds
Task: gemma-Mathematician | Layer: 3 | Time: 0.88 seconds
Task: gemma-Social Scientist | Layer: 3 | Time: 0.76 seconds
Task: mistral-Engineer | Layer: 3 | Time: 1.52 seconds
```



```
Task: mistral—Philosophy Professor | Layer: 3 | Time: 1.60 seconds
Task: mistral—Mathematician | Layer: 3 | Time: 0.93 seconds
Task: mistral—Social Scientist | Layer: 3 | Time: 1.42 seconds
Total Agreements: 12
Total Decisions: 20
Agreement Percentage: 60.00%
```

System Architecture Diagram (of the system to be parallelized)

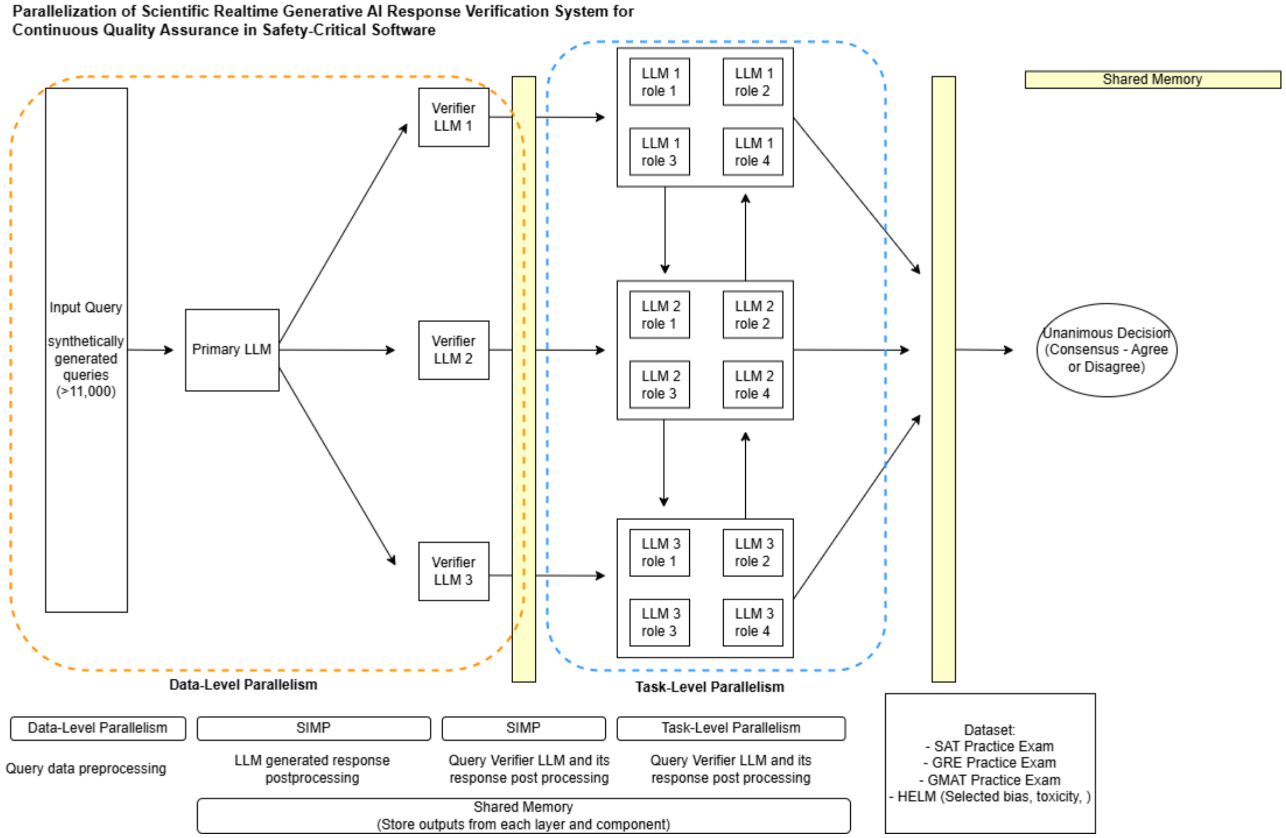

Parallelization of Scientific Realtime Generative AI Response Verification System for
Continuous Quality Assurance in Safety-Critical Software
Input Query
synthetically generated queries (>11,000)
Primary LLM
Verifier LLM 1
Verifier LLM 2
Verifier LLM 3
LLM 1 role 1
LLM 1 role 2
LLM 1 role 3
LLM 1 role 4
LLM 2 role 1
LLM 2 role 2
LLM 2 role 3
LLM 2 role 4
LLM 3 role 1
LLM 3 role 2
LLM 3 role 3
LLM 3 role 4
Shared Memory
Unanimous Decision (Consensus - Agree or Disagree)
Data-Level Parallelism
Task-Level Parallelism
Data-Level Parallelism
Query data preprocessing
SIMP
LLM generated response postprocessing
SIMP
Query Verifier LLM and its response post processing
Task-Level Parallelism
Query Verifier LLM and its response post processing
Shared Memory
(Store outputs from each layer and component)
Dataset:
- SAT Practice Exam
- GRE Practice Exam
- GMAT Practice Exam
- HELM (Selected bias, toxicity,)

Figure 1The diagram shown summarizes the system architecture of the verification process simulation pipelines. The system consists of 3 stage simulation, input layer that comprises of data ingestion and the output of the LLM under test. The second layer consumes the identical prompt that primary LLM's consumed, and independently produces its. In the third layer, the three LLM will play various roles to mimic a panel of experts and non-experts. The final output is a Pass/Partial Pass/Failure, determined by the number of agreements.