CHAPTER 7

PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

7.1. Dataset Preparation & Pre-processing

- **Objective:** Create the ground truth required to train the risk assessment model.
- **Action Items:**
    - **Acquire Benchmarks:** Download and set up existing QA benchmarks such as Natural Questions and Trivia QA.
    - **Generate Responses:** Run these questions through the selected small LLM (approx. 360M parameters) to generate initial responses[2].
    - **Labelling:** Create a custom dataset by labelling these generated responses for hallucinations. This involves categorizing answers as factual, hallucinated, or ambiguous to serve as training data.

7.2. Development of the Confidence Module

- **Objective:** Build the hybrid system that analyses both internal model states and external data quality.
- **Action Items:**
    - **Internal Signal Extraction:** Implement the logic to extract the three internal signals defined in the base paper: Semantic Alignment, Internal Convergence, and Learned Confidence.
    - **External Reality Check Implementation:** Develop the new "Context-Aware" layer to detect your three specific risk factors:
        - **Context Scarcity:** Algorithms to measure if retrieved context is too sparse.
        - **Context Conflict:** Integration of Natural Language Inference (NLI) to detect contradictions in source documents.

- **Domain Mismatch:** Semantic similarity checks to verify relevance between query and documents.

7.3. Routing System Logic & Integration

- **Objective:** Develop the dynamic decision-making engine.
- **Action Items:**
    - **Score Aggregation:** Create a formula or classifier that combines the Internal and External scores into a single Confidence Score.
    - **Thresholding:** Define the thresholds that determine which path a query takes:
        - High Confidence --> Direct Generation
        - Medium Confidence/Need Grounding --> RAG
        - High Risk/Conflict -->Large Cloud Model.

7.4. System Evaluation & Benchmarking

- **Objective:** Prove the effectiveness of the system against established baselines.
- **Action Items:**
    - **Baseline Comparison:** Compare your "Proactive" system against "Reactive" methods (like SelfCheckGPT) and "Always-on RAG" in terms of computational cost and accuracy.
    - **Performance Metrics:** Measure:
        - Accuracy: Reduction in hallucination rates.
        - Efficiency: Time taken per query and computational overhead.
        - Routing Precision: How accurately the system identifies when to use RAG vs. Direct Generation.

7.5. Documentation & Final Presentation

- **Objective:** Compile findings into a research paper and final defence.

- **Action Items:**
    - Draft the final report including the methodology, experimental setup, and results.
    - Prepare the Phase 2 Final Presentation.
    - (Optional Goal) Prepare the work for publication as mentioned in your project scope references.

Phase / Module	Objective	Key Action Items
Dataset Prep	Create ground truth for training the risk model	1. Acquire benchmarks (Natural Questions, Trivia QA) 2. Generate responses using 360M param model. 3. Label responses (Factual vs. Hallucinated)
Confidence Module	Build the hybrid internal/external analysis system	1. Extract internal signals (Convergence, Alignment) 2. Develop "Context-Aware" layer for external risks
Routing Logic	Develop the dynamic decision engine.	1. Create formula to aggregate Internal + External scores 2. Define thresholds for Direct Gen vs. RAG vs. Human Review
Evaluation	Validate system against baselines.	1. Compare against SelfCheckGPT and Always-on RAG 2. Measure Accuracy, Efficiency, and Routing Precision

Table 7.1 Plan of Work