PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

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**[11].
    - **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[3333]. This involves categorizing answers as factual, hallucinated, or ambiguous to serve as training data.

2. Development of the Confidence Module

- **Objective:** Build the hybrid system that analyses both internal model states and external data quality[4].
- **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**[5].
    - **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[6].

- **Context Conflict:** Integration of Natural Language Inference (NLI) to detect contradictions in source documents[7].
- **Domain Mismatch:** Semantic similarity checks to verify relevance between query and documents[8].

3. Routing System Logic & Integration

- **Objective:** Develop the dynamic decision-making engine[99].
- **Action Items:**
  - **Score Aggregation:** Create a formula or classifier that combines the Internal and External scores into a single **Confidence Score**[10].
  - **Thresholding:** Define the thresholds that determine which path a query takes:
    - **High Confidence --> Direct Generation**
    - **Medium Confidence/Need Grounding --> RAG**
    - **High Risk/Conflict -->Human Review**.

4. System Evaluation & Benchmarking

- **Objective:** Prove the effectiveness of the system against established baselines[12].
- **Action Items:**
  - **Baseline Comparison:** Compare your "Proactive" system against "Reactive" methods (like SelfCheckGPT) and "Always-on RAG" in terms of computational cost and accuracy[1313131313131313].
  - **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.

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.