COMP4908 Final Year Project: Second Progress Report

Project Title: TCM-Sage: An Evidence-Synthesis Tool for Traditional Chinese Medicine
Student: Andy ZHENG Zian (22231153)
Date: December 30, 2025
Reporting Period: October 2025 – December 2025 (Phase 2)

---

1. Executive Summary

This report outlines the significant progress achieved during Phase 2 (MVP Implementation) of the "TCM-Sage" project. Following the foundational research in Phase 1, the primary objective of this period was to develop a functional **Minimum Viable Product (MVP)** capable of end-to-end Retrieval-Augmented Generation (RAG).

The project successfully transitioned from theoretical architecture to working software. The system now features a fully operational **Hybrid Retrieval** engine (combining semantic search with a knowledge graph) and a **Reflective Generator** that implements real-time query routing and self-correction to ensure safety.

All planned milestones for the MVP have been met, and the system is ready for the mid-point demonstration scheduled for January 7–9, 2026. This report also addresses previous supervisor feedback regarding schedule visualization and specific evaluation metrics.

2. Detailed Achievements in Phase 2

2.1. Core RAG Pipeline & Multi-Provider Architecture

The fundamental "nervous system" of TCM-Sage has been implemented in Python. The system ingests user queries, retrieves relevant context from the *Huangdi Neijing* (Yellow Emperor's Inner Canon), and synthesizes evidence-based answers.

- **Vector Store Integration:** The text was programmatically cleaned, split, and embedded using `all-MiniLM-L6-v2` into a persistent **ChromaDB** vector store, enabling sub-second semantic retrieval.
- **Multi-Provider Flexibility:** To ensure cost-efficiency and testing rigor, the system features a provider-agnostic layer. It seamlessly switches between **Alibaba Cloud** (Qwen-Turbo) for heavy testing and **OpenAI** (GPT-4o) for high-fidelity benchmarking via environment variable configuration.

2.2. Architectural Pivot: Hybrid Retrieval Engine

A major technical achievement this phase was the implementation of a **Hybrid Retriever**. Early testing revealed that pure vector search struggled with specific herb-symptom relationships (e.g., distinguishing between a "headache" in a metaphor versus a clinical treatment).

- **Ensemble Context Aggregation:** Instead of complex score merging, the system employs an "Ensemble Context" strategy. It retrieves two distinct streams of data:
    1. **Semantic Text Chunks (Vector):** For broad context and theoretical explanations.
    2. **Structured Facts (Knowledge Graph):** For precise entity relationships (e.g., *Headache — TREATS — Chuanxiong*).
- **"Glass Box" Reasoning:** These streams are presented to the LLM as distinct sections (`=== Text Passages ===` and `=== Knowledge Graph Facts ===`). This allows the LLM to explicitly cite structured facts, significantly improving the precision of prescriptive answers.

2.3. The Reflective Generator: Safety & Self-Correction

Given the medical nature of the project, "hallucination" is a critical safety risk. To mitigate this, a **Reflective Generator** architecture (inspired by the "Self-RAG" framework) was implemented.

- **Intelligent Query Routing:** The system classifies every incoming query as either **"Informational"** (general concepts, Temp 0.1) or **"Prescriptive"** (clinical advice, Temp 0.0).
- **Self-Correction Loop:** A post-generation verification module freezes the output and prompts a secondary lightweight model to audit the response. If unsupported claims are detected, a warning flag (`⚠ [Self-Critique Warning]`) is automatically appended, ensuring users are alerted to potential inaccuracies in real-time.

3. Challenges Encountered & Solutions

3.1. Ambiguity in Classical Terminology

Challenge: The *Huangdi Neijing* uses archaic terminology where a single character can imply a symptom, an organ, or a cosmological concept. Standard embeddings often failed to capture these precise nuances. **Solution:** The shift to **Hybrid Retrieval**. By mapping key clinical entities (Symptoms, Herbs) into a graph structure, we provided a "ground truth" pathway. Even if semantic search misses a subtle connection, the explicit graph edge guarantees retrieval.

3.2. Knowledge Graph Construction Strategy

Challenge: The original plan called for fully automated Information Extraction (IE) in Phase 2. However, initial experiments showed that automated extraction on Classical Chinese text produced significant noise. **Solution:** We prioritized **Reliability over Scale** for the MVP. A **Manual JSON Curation** strategy was adopted for the Phase 2 demo, verifying a "Golden Path" dataset focusing on common ailments (e.g., Headache, Insomnia). The automated extraction pipeline has been rescheduled to Phase 3 as the primary focus.

4. Evaluation Metrics

(Addressing Supervisor Feedback: Establish clear evaluation metrics)

To objectively measure the system's effectiveness in Phase 4, the following Key Performance Indicators (KPIs) have been defined:

Metric Category	Metric Name	Target / Method
Performance	**Response Latency**	**< 5 seconds** for the full retrieval & generation cycle.
Accuracy	**Citation Precision**	**> 90%** of generated claims must link to the correct source text (Human Evaluation).
Safety	**Hallucination Rate**	**< 10%** on prescriptive queries (Verified by domain expert audit).
User Experience	**Perceived Trust**	Likert Scale (1-5) collected during pilot testing. Target:**> 4.0**.

5. Upcoming Schedule (Phase 3 & 4)

(Addressing Supervisor Feedback: Provide an overall schedule with timeline)

The focus for the next period shifts from "Core Architecture" to "Automation and Evaluation."

Phase	Duration	Key Tasks & Milestones
Phase 3: Enhancement	**Jan 10 – Feb 10**	**1. Automated KG Construction:** Develop Python scripts using LLMs to parse the 81 chapters of *Su Wen*, scaling the graph from dozens to thousands of nodes. **2. Web Interface:** Migrate from CLI to **Streamlit** Web UI. **3. Explainability:** Implement "Reasoning Visualization" UI to show users exactly which text chunks/graph nodes were used.
Phase 4: Evaluation	**Feb 11 – Mar 10**	**1. Quantitative Benchmarking:** Run the "Golden Set" of 20 clinical questions against the metrics defined in Section 4. **2. Domain Expert Validation:** Conduct pilot testing with **3-5 students/practitioners** from the HKBU School of Chinese Medicine to validate clinical relevance and usability.
Phase 5: Finalizing	**Mar 11 – Apr 8**	**1. Final Report Writing:** Compile all findings and technical documentation. **2. Presentation Prep:** Prepare poster and slides for the April oral defense.

6. Conclusion

Phase 2 has been highly productive. The transition from a theoretical proposal to a working codebase involves overcoming significant complexity in retrieving information from classical texts. By implementing the Hybrid Retriever and Self-Correction module, TCM-Sage has evolved into a safety-conscious, evidence-backed assistant. I am confident in the current state of the system for the upcoming mid-point presentation and prepared for the scaling challenges in Phase 3.