

DeepRoot: Graph-Coordinated Multi-Agent Reasoning for Therapeutic Hypothesis Mining from Historical Medical Texts

Sijbren Kramer, Sean Wang, Zijian (Carl) Ma
Department of Bioengineering, Stanford University
{sijbren, sjw44, mazijian}@stanford.edu

Abstract

Historical medical texts encode empirical therapeutic observations, yet extracting pharmacologically actionable hypotheses from them is hindered by archaic terminology and the absence of mechanistic evidence. We present *DeepRoot*, a multi-agent LLM pipeline that addresses these challenges by constructing a typed biomedical knowledge graph in Neo4j and reasoning over it with metapath queries. Seven specialized agents collaboratively process the *Shen Nong Ben Cao Jing* (~200 AD): extracting entities and therapeutic claims, mapping historical maladies to modern disease ontologies, linking natural-product sources to chemical compounds, connecting compounds to biological targets, evaluating target–disease associations, and auditing graph quality. The resulting graph contains 7,721 active nodes and 19,217 edges. A three-pass Critic Agent evaluates therapeutic claims by combining quantitative graph scoring with LLM semantic reasoning. In blinded expert evaluation, the KG-guided Critic was preferred over text-only LLM baselines in 4 of 5 cases, producing more mechanistically grounded verdicts while using 9.76% fewer tokens.

Key Information. TA mentor: Arpandeeep Khatua. External collaborators: No. External mentor: No. Sharing project: No.

1 Introduction

Historical medical texts contain centuries of empirical observations about natural remedies and symptoms that may encode hypotheses relevant to modern pharmacology. A notable example is the discovery of the antimalarial compound artemisinin by Tu Youyou, which was inspired by traditional Chinese medical writings including the *Shen Nong Ben Cao Jing* [1]. This discovery demonstrated how insights preserved in historical sources can lead to transformative modern therapies, and Tu Youyou was awarded the 2015 Nobel Prize in Physiology or Medicine for this work [2]. However, extracting such hypotheses computationally is challenging because these texts use archaic terminology, inconsistent naming conventions for substances, and symptom descriptions that do not map cleanly to modern diseases. While recent advances in large language models (LLMs) have enabled automated information extraction and reasoning over scientific literature, single-pass LLM systems often struggle with tasks requiring multi-step reasoning, persistent state, and verifiable evidence chains, lacking mechanisms for tracing how conclusions were derived [3]. Multi-agent LLM frameworks attempt to address this by decomposing tasks across specialized roles, yet many implementations rely on agents exchanging unstructured notes rather than building explicit intermediate representations, limiting interpretability [4].

We introduce *DeepRoot*, a multi-agent LLM pipeline that addresses these limitations by using a shared biomedical knowledge graph as both the communication substrate and the reasoning artifact. Rather than passing unstructured summaries between agents, each of seven specialized agents reads from and writes to a typed Neo4j graph whose schema spans the full translational path from historical remedy to modern pharmacological mechanism. This design makes every intermediate inference—from archaic-term normalization to compound–target binding—explicit, auditable, and available to downstream agents. We demonstrate the pipeline on the *Shen Nong Ben Cao Jing* (~200 AD) and evaluate a Critic Agent that combines graph-derived evidence scoring with LLM semantic reasoning to produce mechanistically grounded verdicts on traditional therapeutic claims.

2 Related Work

DeepRoot builds on several lines of research at the intersection of multi agent systems, knowledge graphs and automated therapeutic mining of traditional medicinal texts.

Multi-Agent Systems (MAS). Classical MAS research addresses task decomposition, agent communication, and scalability [5, 6]. More recent work extends these ideas to learned cooperation: Huh and Mohapatra (2024) survey multi-agent reinforcement learning (MARL) communication protocols and decentralized learning [7], while Ruan et al. (2022) propose a graph-based coordination strategy

that factorizes joint policies into a graph generator and coordinated policy [8]. These foundations underpin our multi-agent text-mining approach.

Knowledge Graphs and MAS Knowledge graphs have emerged as semantic backbones to coordinate and share information among agents. In SciAgents (Ghafarirollahi and Buehler 2024), a KG of scientific concepts is coupled with multiple LLM-based agents to autonomously generate hypotheses, demonstrating that a shared ontological graph can organize distributed reasoning [9]. In industrial settings, Qin and Lu (2025) develop KG-enhanced MARL (KGMARL) for smart manufacturing scheduling: agents learn cooperative policies guided by domain knowledge encoded in a shared production KG, which accelerates training and improves efficiency [10]. These works illustrate that embedding a KG in a MAS facilitates richer communication and planning: agents effectively read from and write to the graph, enabling collective inference and task coordination.

Minining Therapeutics from Historical Texts Several AI systems have targeted mining medicinal knowledge from historical or traditional texts. Li et al. (2024) introduce Evi-BERT, a BERT-based information extraction model tailored to Traditional Chinese Medicine (TCM) trial literature: combining deep learning with rule-based postprocessing, it automatically extracts structured evidence from thousands of TCM clinical trials with high precision [11]. Zhang et al. (2025) survey the use of large language models in TCM, noting that LLMs can analyze ancient medical texts to identify symptom–syndrome–treatment patterns and build unified digital TCM knowledge platforms [12]. In a related direction, Zhao et al. (2025) propose an agent-based question-answering system for TCM compounds: their hybrid Retrieval-Augmented Generation (RAG) framework integrates LLMs with structured database queries, allowing natural-language queries about herbal compounds to retrieve accurate information [13]. On the knowledge graph side, Joshi et al. (2026) develop GRAYU, a comprehensive graph linking over 1,000 Ayurvedic formulations with tens of thousands of plants, phytochemicals, and diseases: this graph unifies traditional recipe data with modern disease ontologies, enabling complex queries that trace ancestral remedies to molecular targets [14]. Likewise, Zhang and Hao (2024) construct a TCM knowledge graph via LLM-driven extraction to organize entities and relations for downstream inference [15]. These systems demonstrate how NLP and KG methods can unlock therapeutic insights from classical sources, guiding our approach to extract drug leads from ancient medical corpora.

Novelty. Unlike prior work that either constructs curated biomedical KGs or applies LLMs to modern corpora, DeepRoot performs end-to-end extraction and evaluation from unstructured historical sources. The knowledge graph is explicitly structured for hypothesis generation—not just information storage—with typed edges enabling metapath-based validation, while coordinated specialized agents resolve the semantic ambiguity inherent in archaic medical terminology.

3 Methods

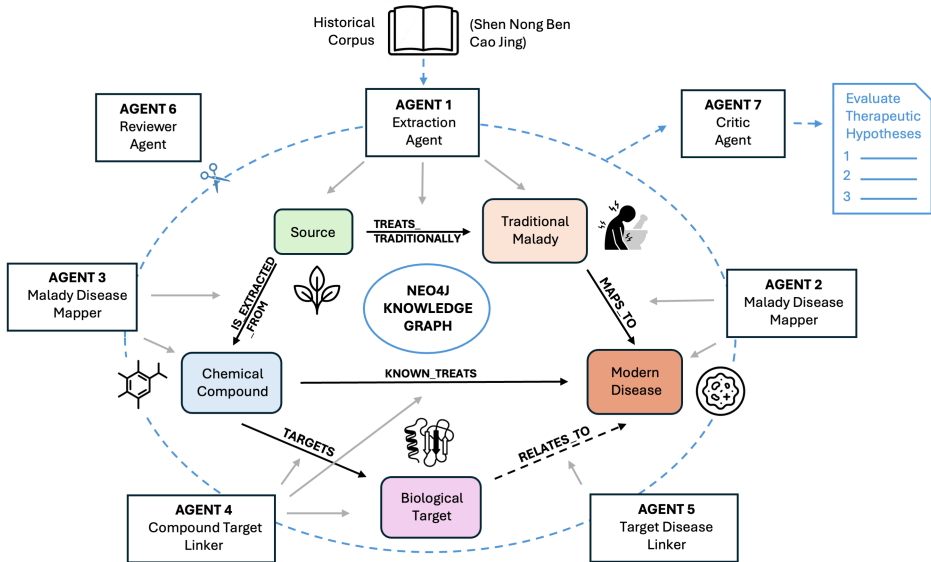


Figure 1: Schematic of the DeepRoot multi-agent system. Graph nodes and edges are represented by rounded rectangles and black arrows, respectively. Gray arrows indicate the creation of specific nodes and edges in the

graph by particular agents. Blue dotted arrows and schematic elements denote logical operations performed over the graph as a whole.

3.1 Graph Ontology

Our pipeline is organized into a knowledge construction and parser/critique phase backed by a Neo4j AuraDB cloud graph database as persistent shared state. We define six node types (Source, Traditional_Malady, Modern_Disease, Chemical_Compound, Biological_Target and six directed edge types, each carrying a `confidence_score` from its responsible agent. Edges span the full translational path from historical text to modern pharmacology: LLM-extracted therapeutic claims and preparations, ontology-matched disease mappings, database-sourced compound-source and compound-target links (COCONUT, ChEMBL), and evidence-scored target-disease associations (Open Targets). Full schema in Appendix A. All code is written by our team: no existing code base was used.

3.2 Knowledge Graph Construction

The graph is populated by five specialized agents with defined dependencies. All graph writes use MERGE (not CREATE) for idempotency, and Pydantic models enforce a shared type contract.

Agent 1: Extraction Agent. Processes the *Shen Nong Ben Cao Jing* [1] corpus (72 chunks, ~2,000 tokens, 200-char overlap) using **Gemini 3 Flash** with structured JSON output constrained to a three-entity schema (sources, maladies, relationships), enforcing extraction of only explicitly stated information with canonical botanical names, original-language aliases, and tiered confidence scores (0.9–1.0 explicit, 0.5–0.7 implied).

Agent 2: Malady→Disease Mapper. Maps 2,863 Traditional_Malady nodes to standardized Modern_Disease nodes (ICD-10, MeSH, SNOMED CT) via a four-step pipeline: (1) normalize through 136 curated archaic-to-modern synonyms (e.g., “ague” → “malaria”); (2) query three ontology APIs (NLM Clinical Tables, MeSH RDF, EBI OLS4) for *exact matches only*—fuzzy matching was abandoned after producing semantic errors (e.g., “warm malaria” → “Warm autoimmune hemolytic anemia”); (3) batched LLM mapping via Gemini 3.1 Flash Lite (30 items/batch, structured JSON) for unmatched maladies; (4) exact-only API validation of LLM-suggested disease names using concurrent lookups (20 threads) against all three ontology APIs. Confidence tiers: exact API = 0.95, LLM-batch-verified (API-confirmed codes) = 0.80, LLM-batch-remap (unverified) = 0.60, LLM-disambiguated = 0.50–0.85, LLM-generated = 0.30–0.60.

Agent 3: Source→Compound Mapper. A direct lookup first maps Source to a known plant organism in the COCONUT2.0 [16] database. If unsuccessful, Gemini clarifies Source to canonical nomenclature. Sources not mapping to a known plant (animals and minerals) are assigned to an “unknown” class and excluded downstream. Edge scores are computed as a normalized average of COCONUT2.0’s annotation score and natural product likeness score. The top 30% of compounds and capped by 30 total compounds by score are added to the graph.

Agent 4: Compound→Target Linker. Queries ChEMBL [17] for each compound: resolves by SMILES/name, expands salt-form hierarchies to capture all forms, fetches mechanisms and activities (pChEMBL ≥ 5), enriches targets with UniProt IDs and gene symbols from ChEMBL’s target endpoint, and filters to single-protein targets only. It uses a high-performance batch client with ThreadPoolExecutor for parallel lookups.

Agent 5: Target→Disease Linker. Performs the key logical step that completes the graph’s closed-loop architecture by evaluating whether a scientifically justified relationship exists between the biological target and the modern disease within a given evidence chain. It constructs RELATES_TO edges between Biological_Target and Modern_Disease nodes, by: (1) finding target-disease pairs connected by the evidence chain (Compound → Source → Malady → Disease); (2) querying the Open Targets Platform [18] GraphQL API for real-world evidence scores (genetic association, known drug, etc.); and (3) using Gemini to assess the biological plausibility of this link and generating a short rationale and confidence score. If modern disease names are not found in Open Targets, they are mapped to broader disease categories (hematologic disease, liver disease) via a Gemini fallback.

Agent 6: Reviewer Agent. Two-pass graph quality audit ensuring downstream agents operate on clean data. **Pass 1 (Rules-based, zero API cost):** deterministic pattern matching across five categories—(a) OCR corruption detection via regex for invalid characters; (b) generic source filtering against pharmaceutical category names (“herbs,” “medicine,” “powder”) and pattern matches like “Cold medicinals”; (c) orphan node detection (Source nodes with degree = 0); (d) metaphysical

malady detection matching supernatural patterns (demon, ghost, spirit); (e) invalid disease detection via a hardcoded set of non-diseases and over-mapped disease detection (nodes with >50 incoming MAPS_TO edges). **Pass 2 (LLM, batched Gemini):** ambiguous survivors—short-name sources with degree ≤ 2 , diseases with 20–50 incoming mappings—are sent to Gemini with structured JSON output requiring a keep/archive verdict and reasoning string. Nodes are *archived* (not deleted) with metadata (archived=true, archive_reason) for provenance; edges to archived nodes are cascade-archived. All Phase 2 queries include archive filters to exclude noise.

3.3 Graph-Guided Claim Evaluation

Agent 7: Critic Agent. The Critic Agent evaluates individual (Source, Traditional_Malady) pairs via a three-pass architecture separating graph scoring, LLM semantic reasoning, and compound nomination.

Pass 1: Subgraph extraction and graph scoring. A SubgraphExtractor lazy-loads two global caches—*target genericity* (distinct diseases per Biological_Target) and *compound ubiquity* (active sources per Chemical_Compound)—then executes a Cypher query pulling all Compound $\rightarrow$ Target $\rightarrow$ Disease paths from the query source with full edge properties (pChEMBL, Open Targets, COCONUT scores, LLM rationales). Paths are tiered as Gold (specific disease, OT score > 0), Silver (broad category), or Junk (zero OT score), and flagged as *closed loops* when the terminal disease matches the malady’s MAPS_TO target. Two cross-source signals are also collected: shared compounds across sources treating the same malady, and target convergence across different sources.

From these paths, 14 discriminative statistics are aggregated into a composite graph score $s_g \in [0, 1]$:

$$s_g = (0.15 \tilde{L} + 0.20 \tilde{D} + 0.15 \tilde{P} + 0.15 \tilde{T} + 0.10 \tilde{M} + 0.10 \tilde{C} - 0.10 \tilde{G} - 0.05 \tilde{U}) \times (0.5 + 0.5 f_{\text{gold}})$$

where $\tilde{L}$, $\tilde{D}$, $\tilde{P}$, $\tilde{T}$, $\tilde{M}$, $\tilde{C}$ are normalized closed-loop count, target diversity, best pChEMBL, best Open Targets score, mapping quality, and cross-source overlap, respectively; $\tilde{G}$ and $\tilde{U}$ are genericity and ubiquity penalties; and f_{gold} is the Gold-tier closed-loop fraction. $s_g = 0$ if no closed loops exist.

Pass 2: LLM semantic evaluation. A structured evidence package—comprising the claim header, malady $\rightarrow$ disease mapping, graph summary, top 5 ranked closed-loop paths, and weak-signal caveats—is sent to Gemini (temperature = 0, structured JSON). The model returns four scores (mechanistic_coherence, specificity, mapping_quality, evidence_strength $\in [0, 1]$), a weakest_link string, and a verdict $\in \{\text{Validated, Plausible, Weak, Unsupported}\}$.

Pass 3: Compound nomination. Compounds from closed-loop paths are ranked by:

$$s_c = \frac{p}{10} \cdot r_{\text{tier}} \cdot (1 + 0.5 \mathbf{1}_{\text{novel}}) \cdot \max\left(0.1, 1 - \frac{u}{100}\right)$$

where p is best pChEMBL, r_{tier} weights RELATES_TO confidence by OT tier ($1 \times$ Gold, $0.3 \times$ Silver, 0 for Junk), $\mathbf{1}_{\text{novel}}$ flags absence of a KNOWN_TREATS edge, and u is compound ubiquity. Results are returned as ranked CompoundCard objects with SMILES, targets, diseases reached, and loop closure counts.

4 Graph Construction Experiments, Evaluation, and Results

4.1 Source Data

The *Shen Nong Ben Cao Jing* (Divine Farmer’s Materia Medica, ~ 200 AD) [1], was used as our test historical corpus, containing 365 herb entries with nature, flavor, toxicity, and indications. The English translation was obtained from Archive.org (public domain) and chunked into 72 passages of $\sim 2,000$ tokens each.

4.2 Experimental Details

All LLM calls for graph construction use Gemini 2.5 Flash (temperature 0.0, structured JSON output). Agents process sequentially: extraction (72 chunks), malady mapping (13s intervals across 3 ontology APIs), compound–target linking (5 parallel ChEMBL workers), and review (rules in ~ 30 s, 3 LLM batches in ~ 3 min); total wall time ~ 50 min on shared Neo4j AuraDB.

4.3 Knowledge Graph Construction Results

After all Phase 1 agents, the active graph contains **7,721 nodes** (741 Sources, 2,807 Maladies, 1,716 Diseases, 2,063 Compounds, 233 Targets, 161 Preparations) and **19,217 edges** (4,329

TREATS_TRADITIONALLY, 2,942 MAPS_TO, 5,626 IS_EXTRACTED_FROM, 456 TARGETS, 5,549 RELATES_TO, 182 PREPARED_AS, 133 KNOWN_TREATS). Edge confidence distributions are in Appendix E. An additional 1,448 noise nodes and 421 edges were archived by the Reviewer Agent (see Appendix C for full breakdown).

Malady→Disease mapping. The mapper achieved **98% coverage** (2,832/2,863 mapped; 165 skipped as too vague for any mapping). The initial database-first pipeline produced exact ontology matches for 19.1% of maladies (9.4% MeSH exact, 7.8% SNOMED exact, 1.9% ICD-10 exact), but 76% of edges used fuzzy SNOMED matching with low confidence (≤ 0.52) and frequent semantic errors. The `remap_fuzzy()` refactoring replaced all 2,157 bad fuzzy edges with batched LLM mapping validated against ontology APIs, yielding the final distribution: 54% `llm_batch_remap`, 17.5% `llm_batch_remap_verified` (API-confirmed codes), 9.4% `mesh_exact`, 7.8% `snomed_exact`, 4.6% `llm_disambiguated`, 4.0% `llm_generated`, 1.9% `icd10_exact`. Of 1,620 Modern_Disease nodes, 575 (35.5%) now carry API-verified ontology codes. The 136-entry archaic synonym dictionary remains critical for bridging historical TCM terminology to searchable modern terms. Full method breakdown in Appendix B. Representative corrected mappings are shown in Appendix H.

Reviewer Agent. The two-pass review archived 1,448 nodes: 1,282 orphan Sources (degree = 0), 67 OCR-corrupted, 46 metaphysical maladies, 36 generic categories, 9 too-generic terms, 4 LLM-flagged borderline cases, 3 invalid diseases, and 1 empty name. The rules pass handled 99.7% of archival decisions in ~ 30 seconds; the LLM pass reviewed 51 ambiguous nodes, archiving only 4 (manually verified). The most impactful cleanup was archiving the Qigong disease node, which had incorrectly absorbed 70+ maladies via fuzzy mapping, cascade-archiving 150 MAPS_TO edges.

Compound→Target and Target→Disease. ChEMBL lookup linked 2,063 compounds to 233 biological targets (gene symbols: PTGS1, HSD17B10, NFKB1, etc.) via 456 TARGETS edges. The Target→Disease linker created 5,549 RELATES_TO edges, and a standalone KNOWN_TREATS pass matched 133 compound-disease pairs via ChEMBL indications (MeSH ID and name matching).

5 Claim Evaluation Experiments, Evaluation, and Results

5.1 Evaluation Method & Experimental Details

We evaluate the Critic Agent along three axes: automated metrics across 50 source-malady pairs, automated pairwise comparison via LLM-as-Judge, and blinded human expert review of 5 pairs. The evaluation corpus contains parsed text regions from the *Shen Nong Ben Cao Jing*, each corresponding to one source-malady pair. We compare three conditions: (1) **KG + Gemini 3.1 Flash** (the full Critic pipeline); (2) **Text-only + Gemini 3.1 Flash** (the relevant corpus section provided directly, no graph); and (3) **Text-only + Gemini 3.1 Flash + Tool Calling** (Gemini with autonomous API access). To ensure comparability, we provide the relevant corpus section as input to the text-only baseline, as the number of nodes in the KG subgraph is proportional to the length of the corresponding text passage. A fairer baseline would use autonomous API tool calling, but this was infeasible ($> 70,000$ tokens per source-malady pair).

Automated metrics. We measure token efficiency (input + output tokens), verdict distribution, and robustness to graph corruption (progressive edge shuffling at 0–100%). We also compare verdict consistency across three LLMs (Gemini 2 Flash, Gemini 3 Flash, Llama-3.1-8B) on the same KG inputs. It should be noted that a fair comparison would be to compare the Critic with an LLM that can use API tool calling without formulating a KG. However, allowing Gemini 3 Flash to make API tool calls autonomously was unfeasible due to the high token count ($> 70,000$ tokens) per source-malady pair and time out due to exponentially growing number of tools calls as it reasons through target, source, and chemicals (more than 50 tool calls per run).

Human expert evaluation. For 5 source-malady pairs, we generated Critic outputs under KG-guided and text-only conditions and showed them (generation type hidden) to 2 Stanford Bioengineering PhD students familiar with the project.

LLM-as-Judge. To scale evaluation beyond the 5 manually reviewed pairs, we conducted an automated pairwise comparison across all 50 source-malady pairs. For each pair, the KG-guided Critic output and the text-only baseline output were presented to a judge LLM (Gemini 2.5 Flash, temperature = 0) in blinded, randomized order (labeled “Response A” and “Response B”). The assignment was randomized independently per pair to mitigate position bias. The judge evaluated each pair on five criteria: Evidence Traceability (specific, verifiable evidence vs. broad chemical classes),

Falsifiability & Rigor (testable claims vs. unverifiable assertions), Calibrated Confidence (verdict severity matching evidence strength), Actionability (whether a medicinal chemist could design follow-up experiments), and Intellectual Honesty (acknowledging gaps vs. masking uncertainty). For each criterion and overall, the judge selected a winner (A, B, or TIE). All outputs were collected as structured JSON with per-criterion explanations.

5.2 Quantitative Results

Token efficiency and verdict distribution. We compared token usage between the KG-guided and text-only conditions across all 50 source–malady pairs. The KG condition used 9.76% fewer total tokens (input plus output) than the text-only baseline while producing more conservative verdict distributions (**Figures 2A, 2B**).

Robustness to graph corruption. To verify that the Critic genuinely relies on graph structure rather than hallucinating plausible-sounding evidence, we progressively shuffled 0–100% of graph edges and re-ran the Critic on all 50 pairs. The graph score decreased monotonically with increasing perturbation, and the verdict distribution shifted steadily toward UNSUPPORTED (**Figure 2C**).

Cross-LLM consistency. We replaced the Critic’s LLM backbone with three different models—Gemini 2 Flash, Gemini 3 Flash, and Llama-3.1-8B—while holding the KG input constant. Gemini 2 and Gemini 3 Flash produced similarly conservative verdict distributions, while Llama-3.1-8B was notably more lenient. This discrepancy is likely attributable to the 8B model’s weaker reasoning capabilities compared to the Gemini models (**Figure 2D**).

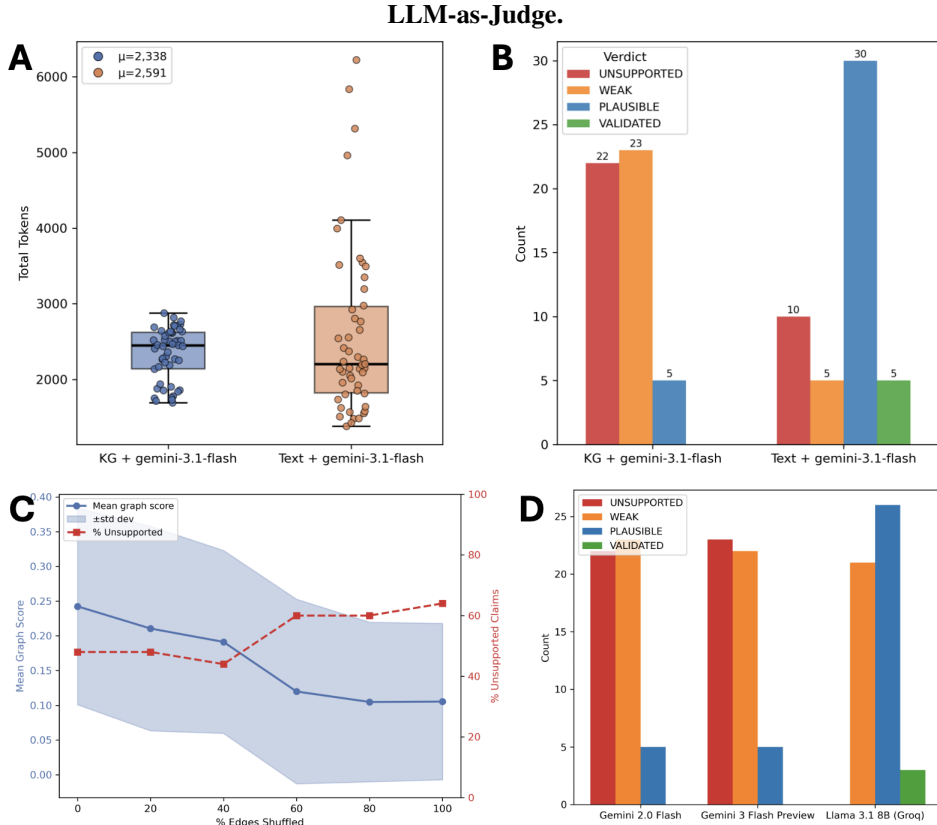


Figure 2: Critic outputs. (A) Total amount of input and output tokens used by the critic agent (KG + gemini) vs gemini on text only. (B) Verdicts by both models. (C) Critic agent output in response to KG edge perturbation. (D) Effect of different LLMs on verdicts.

LLM-as-Judge. The KG-guided Critic won decisively overall (37–13), with the largest margins on falsifiability & rigor (37–8) and actionability (36–14) (Table 1). The text-only baseline’s 13 wins concentrated in pairs where the KG had zero closed loops and thus no evidence to present.

Criterion	KG wins	Text wins	Tie
Evidence Traceability	30	20	0
Falsifiability & Rigor	37	8	5
Calibrated Confidence	25	3	22
Actionability	36	14	0
Intellectual Honesty	24	5	21
Overall	37	13	0

Table 1: LLM-as-Judge pairwise comparison (drug-discovery rubric, $n = 50$). The KG-guided Critic wins on all five criteria, with the largest margins on falsifiability and actionability.

Evaluation 1: Fritillaria bulb (Source) and retained placenta (Malady)

KG + G3.1F Verdict: **UNSUPPORTED**

Weakest Link: The target-to-disease association (**Open Targets score 0.000**)¹ for all identified compounds.

Reasoning: While some compounds like Hederagenin show high binding affinity to targets like F3, there is **no established biological link between these targets and the physiological process of placental detachment**². The evidence chains fail to provide a plausible mechanism, as the targets are either involved in unrelated processes like iron storage or general coagulation rather than uterine contractility or placental separation.

1. Clear leveraging of knowledge graph edge information.
2. Understands that placenta retainment can not be solved with a drug since it is a physically-induced condition.

TO + G3.1F Verdict: **PLAUSIBLE**

Weakest Link: The assumption that isosteroidal alkaloids, primarily known for respiratory smooth muscle relaxation, exert a sufficiently strong and specific uteronic effect to expel placental tissue.

Reasoning: Fritillaria contains various isosteroidal alkaloids, such as peimine and verticine, which are known to modulate smooth muscle activity through calcium channel and muscarinic receptor interaction. While the text emphasizes its role in lung disorders, the pharmacological potential for these compounds to influence uterine contractions provides a plausible mechanism for treating retained placenta, particularly when used in the described 'emergency' combinations intended to break physical blocks.

1. Able to obtain chemicals, and have reasonable biochemical connections. However, does not understand that you can't use a drug to treat a physically-induced condition.

Evaluation 2: Cynanchum (Source) and wind stroke (Malady)

KG + G3.1F Verdict: **PLAUSIBLE**

Weakest Link: The high pleiotropy of identified targets, particularly PTPN1 and GPRBAR1, which connect to over 200 diseases each, reducing the specificity of the stroke-related claim.

Reasoning: The claim is supported by a **high-potency interaction between Caryophyllene oxide and FTH1 (pChEMBL 9.0)**¹, which is mechanistically relevant to **oxidative stress in stroke**². While the majority of compounds in the source remain uncharacterized and the identified targets are generic, the **alignment of multiple gold-tier evidence chains**² regarding vascular health and neuroinflammation provides a plausible pharmacological basis.

1. Clear leveraging of knowledge graph edge information for pChEMBL score.
2. Clear biological reasoning.
3. Synthesis of multiple claims for a better answer.

TO + G3.1F Verdict: **PLAUSIBLE**

Weakest Link: The assumption that systemic anti-inflammatory and diuretic-like effects (precipitating water qi) are sufficient to reverse acute neurological deficits and loss of consciousness in a stroke event.

Reasoning: Cynanchum contains steroidal saponins and phenanthroindolizidine alkaloids which exhibit anti-inflammatory and potential neuroprotective properties. These bioactivities align with the historical text's description of 'draining heat' and 'precipitating water qi,' which may help reduce secondary neuroinflammation and cerebral edema associated with modern stroke pathology.

1. Able to obtain chemicals, but they are broad classes instead of specific natural products. Tries to do a mechanistic link but is not specific.

TO + G3.1F + Tool Call Verdict: **PLAUSIBLE**

Weakest Link: The direct clinical evidence linking Cynanchum-specific steroidal saponins to stroke outcomes is less robust than the evidence for its secondary metabolites like paeonol.

Reasoning: Cynanchum species, particularly *C. paniculatum*, contain bioactive compounds like **paeonol and apocynin which target MAPK1 and NADPH oxidase (NOX2/NOX4)**¹, respectively. These targets are **highly associated with the pathophysiology of stroke (ischemia and neuroinflammation) in modern databases**², providing a credible biochemical basis for the traditional claim.

1. Clearly identified natural products and targets.
2. Tries to do a mechanistic link but is not specific as those symptoms are common with many other brain diseases.

Figure 3: Representative Critic Agent outputs for Evaluations 1 and 2. Green notes indicate strengths; red notes indicate weaknesses; blue notes indicate neutral observations. G3.1F= Gemini 3.1 Flash; KG= knowledge graph-guided; TO= text-only.

5.3 Human Expert Evaluation of Critic Agent

To evaluate response quality, we generated Critic outputs for 5 source-malady pairs under two conditions — KG-guided and text-only — and showed them (generation type hidden) to 2 Stanford Bioengineering PhD students familiar with the project (**Figures 3, S1**). Reviewers preferred the Critic in 4 of 5 cases, consistently citing more mechanistically specific reasoning.

Evaluation 1 (Figure 3) illustrates the Critic’s ability to detect false positives. "Retained placenta" is a physical complication requiring surgical intervention, not pharmacological treatment. Because no biological target could be confidently linked to placental detachment, the Critic correctly returned UNSUPPORTED. The text-only LLM instead inferred plausibility from the general smooth-muscle activity of *Fritillaria* alkaloids — directionally coherent but clinically invalid.

Evaluation 2 (Figure 3) shows the Critic’s advantage of mechanistically grounding claims. For *Cynanchum* and wind stroke, all three conditions returned PLAUSIBLE, but the reasoning quality diverged sharply. The Critic based its verdict to a specific high-potency interaction (caryophyllene oxide–FTH1, pChEMBL 9.0) with direct relevance to oxidative stress in ischemic stroke, synthesizing multiple evidence chains across neuroinflammation and vascular health. The text-only LLM cited only broad chemical classes without binding evidence, while the tool-augmented condition named specific compounds but cited biological targets MAPK1, NOX2/NOX4 that are associated with neuroinflammation generally rather than stroke specifically.

The selected evaluation examples demonstrate the robustness of the KG in steering the LLM toward more mechanistic and evidence-grounded reasoning. However, **Evaluation 5 (Figure S1)** reveals a current limitation: the Critic failed to support the *Epimedium*–impotence pair, which the text-only LLM correctly identified as plausible. This failure stems from two compounding constraints—*Epimedium* has over 787 associated compounds in COCONUT2.0, yet the pipeline caps ingestion at 30 compounds per source, and the Critic evaluates only the top 5 evidence paths per query. Together, these filters likely excluded the pharmacologically relevant icariin–PDE5 axis entirely, leaving the Critic with insufficient graph signal to reach a supportable verdict.

5.4 Qualitative Case Studies

Fritillaria subgraph validation. To validate the generated evidence paths, we examined the subgraph rooted at the *Fritillaria* Source node, an East Asian flowering plant genus whose bulbs have been used in traditional Chinese medicine for centuries (**Appendix L, Figure S2**). Its extracted traditional malady, "accumulations and binding," was mapped by Agent 2 to the modern disease "Abdominal mass." Traversing the full evidence graph from this source, the most strongly connected modern diseases were "Diabetes Mellitus," "Goiter," and "Abdominal Pain"—a distribution that reflects both the metabolic and structural pathologies plausibly encoded in the original TCM indication. The Critic Agent nominated the top five plausible compounds by composite score and provided reasoning for the overall decision (**Figure S2**). The top nomination of "cortosolic acid" is supported by independent literature where it improves insulin sensitivity [19]. Betulin and Ursolic acid were jointly nominated for their shared activity on PTPN1 (protein tyrosine phosphatase 1B), a negative regulator of insulin signaling. While emerging studies suggest betulin may modulate insulin receptor signaling [20], the clinical evidence remains sparse compared to cortosolic acid, consistent with the agent’s lower composite scores for these compounds. Taken together, this case study demonstrates that DeepRoot can potentially surface mechanistically grounded relevant compound–target associations.

6 Conclusion

DeepRoot shows that a knowledge graph can discipline LLM reasoning about historical therapeutic claims—replacing hallucinated plausibility with database-grounded evidence chains that experts and automated judges consistently prefer. But this discipline comes at a cost: the system can only evaluate what the graph contains, and sparse target coverage means most claims receive conservative verdicts. The text-only baseline, unconstrained by evidence requirements, occasionally outperforms the Critic precisely because it is free to hallucinate. Whether this trade-off favors precision or recall depends on the downstream application: in drug discovery, where false positives waste years of effort, we argue that verifiable skepticism is more valuable than confident speculation.

More broadly, this work suggests that multi-agent LLM pipelines benefit from shared structured state—not just for coordination, but as an epistemic constraint that forces each agent’s outputs to be explicit, auditable, and available for downstream verification. The graph is not merely a data store; it is the mechanism by which the system distinguishes what it knows from what it can plausibly guess.

Future work should expand the translational coverage of the graph and test whether this architecture generalizes to other historical medical traditions and modern biomedical corpora.

Team Contributions. All three members jointly brainstormed the project concept, iterated on the system architecture, and conducted regular check-ins throughout the quarter. Individual contributions: Sijbren — background research, related work, figure illustrations, and target-related agents (compound–target linker, target–disease linker). Sean — human expert evaluation, automated evaluations (evals 1–4), case studies, and the chemical compound agent. Carl — extraction agent, malady–disease mapper, updated compound–target linker, target – disease linker, known treat agents, reviewer agent, subgraph pooling, critic agent, LLM-as-Judge evaluation, and Neo4j infrastructure.

Late Days We are pooling 6 late days (2 from Sean and 4 from Sijbren)

References

- [1] Shou-Zhong Yang. *The Divine Farmer’s Materia Medica: A Translation of the Shen Nong Ben Cao Jing*. Blue Poppy Press, 1998.
- [2] Youyou Tu. The discovery of artemisinin (qinghaosu) and gifts from chinese medicine. *Nature Medicine*, 17(10):1217–1220, 2011.
- [3] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore, December 2023. Association for Computational Linguistics.
- [4] Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. Talk Structurally, Act Hierarchically: A Collaborative Framework for LLM Multi-Agent Systems, February 2025. arXiv:2502.11098 [cs].
- [5] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. Multi-Agent Systems: A Survey. *IEEE Access*, 6:28573–28593, 2018.
- [6] Peter Stone and Manuela Veloso. Multiagent Systems: A Survey from a Machine Learning Perspective. *Auton. Robots*, 8(3):345–383, June 2000.
- [7] Dom Huh and Prasant Mohapatra. Multi-agent Reinforcement Learning: A Comprehensive Survey. 2023. Publisher: arXiv Version Number: 2.
- [8] Jingqing Ruan, Yali Du, Xuantang Xiong, Dengpeng Xing, Xiyun Li, Linghui Meng, Haifeng Zhang, Jun Wang, and Bo Xu. GCS: 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022. In *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, pages 1128–1136. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), May 2022.
- [9] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):e2413523, 2024.
- [10] Zhaojun Qin and Yuqian Lu. Knowledge graph-enhanced multi-agent reinforcement learning for adaptive scheduling in smart manufacturing. *Journal of Intelligent Manufacturing*, 36(8):5943–5966, December 2025.
- [11] Yizhen Li, Zhongzhi Luan, Yixing Liu, Heyuan Liu, Jiaying Qi, and Dongran Han. Automated information extraction model enhancing traditional Chinese medicine RCT evidence extraction (Evi-BERT): algorithm development and validation. *Frontiers in Artificial Intelligence*, 7:1454945, August 2024.
- [12] Ruyi Zhang, Dingcheng Tian, Yu Wang, Ruyi Zhang, Dingcheng Tian, and Yu Wang. Large Language Models in Traditional Chinese Medicine: A Short Survey and Outlook. *AI Medicine*, 2(1):3, June 2025. Publisher: Scilight Press.
- [13] Feifan Zhao, Qianjin Li, Meng Wang, and Xingchuang Xiong. An AI Agent-Based System for Retrieving Compound Information in Traditional Chinese Medicine. *Information*, 16(7):543, July 2025. Publisher: Multidisciplinary Digital Publishing Institute.

- [14] Sarthak Joshi, Aditi Pathak, Dheemanth Reddy Regati, Revathy Menon, Deepthi S. Ajith, Aditya Sheshadri, Neerja Viswanathan, Poulomi Ray, Vimarishi Koul, Pratishruti Panda, Shriya Anand Bhambore, Shailya Verma, Ananya Sinha, K. Mohamed Shafi, Murugavel Pavalam, and Ramanathan Sowdhamini. GRAYU: graph-based database integrating Ayurvedic formulations, medicinal plants, phytochemicals and diseases. *Frontiers in Pharmacology*, 16, January 2026. Publisher: Frontiers.
- [15] Yichong Zhang and Yongtao Hao. Traditional Chinese Medicine Knowledge Graph Construction Based on Large Language Models. *Electronics*, 13(7):1395, January 2024. Publisher: Multidisciplinary Digital Publishing Institute.
- [16] Venkata Chandrasekhar, Kohulan Rajan, Sri Ram Sagar Kanakam, Nisha Sharma, Viktor Weißenborn, Jonas Schaub, and Christoph Steinbeck. COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Research*, 53(D1):D634–D643, January 2025.
- [17] David Mendez, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María P. Magariños, Juan F. Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019.
- [18] David Ochoa, Andrew Hercules, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*, 49(D1):D1302–D1310, 2021.
- [19] Toshihiro Miura, Satoshi Takagi, and Torao Ishida. [retracted] management of diabetes and its complications with banaba (*lagerstroemia speciosa* L.) and corosolic acid. *Evidence-Based Complementary and Alternative Medicine*, 2012(1):871495, 2012.
- [20] Feyisayo O. Adepoju, Ksenia V. Sokolova, Irina F. Gette, Irina G. Danilova, Mikhail V. Tsurkan, Alicia C. Mondragon, Elena G. Kovaleva, and Jose Manuel Miranda. Protective effect of betulin on streptozotocin–nicotinamide-induced diabetes in female rats. *International Journal of Molecular Sciences*, 25(4), 2024.

A Graph Ontology Details

Node Type	Description	Example
Source	Remedy organism/substance	<i>Artemisia annua</i>
Traditional_Malady	Historical symptom/disease	“Intermittent fever”
Modern_Disease	Standardized disease	Malaria (ICD-10: B54)
Chemical_Compound	Active ingredient	Artemisinin
Biological_Target	Molecular target	PTGS1 (P23219)
Preparation_Method	Preparation/administration	Decoction

Table 2: Node types in the knowledge graph ontology.

Relation	Direction	Confidence Source
TREATS_TRADITIONALLY	Source → Malady	LLM extraction
MAPS_TO	Malady → Disease	Ontology matching
IS_EXTRACTED_FROM	Compound → Source	COCONUT2.0
TARGETS	Compound → Target	ChEMBL bioassay
RELATES_TO	Target → Disease	Open Targets + Gemini
PREPARED_AS	Source → Preparation	LLM extraction
KNOWN_TREATS	Compound → Disease	ChEMBL indications

Table 3: Relation types with confidence provenance.

B Malady Mapping Method Breakdown

Mapping Method	Count	%
llm_batch_remap	1,537	54.0%
llm_batch_remap_verified	501	17.5%
mesh_exact	269	9.4%
snomed_exact	222	7.8%
llm_disambiguated	131	4.6%
llm_generated	114	4.0%
icd10_exact	55	1.9%

Table 4: Malady→Disease mapping method distribution. 19.1% via exact ontology APIs; 71.5% via batched LLM mapping (17.5% API-verified); 8.6% LLM fallback.

C Reviewer Agent Archival Breakdown

Node Type	Active	Archived	% Trimmed
Source	741	1,388	65.2%
Traditional_Malady	2,807	56	2.0%
Modern_Disease	1,716	4	0.2%
Chemical_Compound	2,063	0	0.0%
Biological_Target	233	0	0.0%
Preparation_Method	161	0	0.0%
Total	7,721	1,448	15.8%

Table 5: Graph node statistics after Phase 1 (active vs. archived).

Archive Reason	Count
Orphan (degree=0)	1,282
OCR corrupted	67
Metaphysical	46
Generic category	36
Too generic	9
LLM-reviewed	4
Not-a-disease	3
Empty name	1
Total	1,448

Table 6: Reviewer Agent archival breakdown by reason.

D Metapath Schemas

Task	Metapath
A (Validate)	Source $\leftarrow$ IS_EXTRACTED_FROM $\leftarrow$ Compound $\rightarrow$ TARGETS $\rightarrow$ Target $\rightarrow$ RELATES_TO $\rightarrow$ Disease $\leftarrow$ MAPS_TO $\leftarrow$ Malady
B (Novel)	Disease $\leftarrow$ MAPS_TO $\leftarrow$ Malady $\leftarrow$ TREATS_TRAD $\leftarrow$ Source $\leftarrow$ IS_EXTRACTED_FROM $\leftarrow$ Compound $\rightarrow$ TARGETS $\rightarrow$ Target $\rightarrow$ RELATES_TO $\rightarrow$ Disease

E Edge Confidence Distributions

Edge Type	Count	Mean	Median	Min	Max
TREATS_TRADITIONALLY	4,720	0.943	0.950	0.50	1.00
IS_EXTRACTED_FROM	5,626	0.595	0.626	0.13	0.81
MAPS_TO	2,955	0.803	0.850	0.30	1.00
TARGETS	456	0.608	0.572	0.50	1.00
RELATES_TO	5,549	—	—	—	—
KNOWN_TREATS	133	1.000	1.000	1.00	1.00

Table 7: Confidence score distributions across edge types. TREATS_TRADITIONALLY and MAPS_TO scores are assigned by the extraction and mapping agents respectively. IS_EXTRACTED_FROM scores are derived from COCONUT annotation and natural-product-likeness scores. TARGETS scores reflect pChEMBL-based confidence. RELATES_TO edges carry Open Targets association scores (stored separately as `ot_overall_score`) rather than a unified confidence field. KNOWN_TREATS edges are binary (ChEMBL indication match = 1.0).

F Node Degree Distribution

Node Type	Count	Mean Degree	Max Degree
Source	741	13.8	68
Traditional_Malady	2,807	2.7	35
Modern_Disease	1,716	5.0	110
Chemical_Compound	2,063	3.0	61
Biological_Target	233	25.8	335

Table 8: Degree distribution of active (non-archived) nodes. Degree counts all incident edges regardless of direction. Biological_Target nodes have the highest mean degree due to dense RELATES_TO connections (each target links to many diseases via Open Targets). The maximum degree of 335 for a Biological_Target is PTPN1 (protein tyrosine phosphatase non-receptor type 1), a highly pleiotropic drug target. The Critic Agent’s genericity penalty (Section 3.3) down-weights evidence paths through such hub targets.

G Target Genericity Analysis

Gene Symbol	Full Name	Diseases	TARGETS Edges
PTPN1	Protein tyrosine phosphatase 1B	313	—
POLB	DNA polymerase beta	219	—
GPBAR1	G-protein coupled bile acid receptor	213	—
PTPN2	Protein tyrosine phosphatase 2	212	—
FTH1	Ferritin heavy chain 1	186	—
F2	Prothrombin (Factor II)	168	—
GAA	Acid alpha-glucosidase	158	—
AKR1B10	Aldo-keto reductase 1B10	152	—
PTGS1	Prostaglandin H synthase 1 (COX-1)	147	—
PYGM	Glycogen phosphorylase, muscle	134	—
NFKB1	NF-kappa-B p105 subunit	105	—
NOS3	Endothelial nitric oxide synthase	—	—

Table 9: Top biological targets by number of connected Modern_Disease nodes via RELATES_TO edges. High-genericity targets like PTPN1 (313 diseases) provide weak evidence for any specific therapeutic claim. The Critic Agent penalizes paths through such targets via the $\tilde{G}$ genericity term in the graph score.

H Malady Mapping Examples

Traditional Malady	Before (Fuzzy)	After (LLM + API)
warm malaria	Warm autoimmune hemolytic anemia	Malaria (ICD-10: B54)
women’s diseases	Women’s hairdresser	Gynecological Diseases (MeSH: D005831)
dream seminal emission	—	Nocturnal emission (SNOMED: 2216005)
tongue sores	—	Glossitis (ICD-10: K14.0, MeSH: D005928)
vaginal discharge disease	—	Leukorrhea (MeSH: D007973)
mounting pain	—	Abdominal pain (ICD-10: R10.9)
urinary stones	—	Urolithiasis (MeSH: D052878)
postpartum blood disorders	—	Postpartum hemorrhage (MeSH: D006473)
original yang vacuity and cold	—	Hypothyroidism (ICD-10: E03.9)
miasmic evils	—	Malaria (SNOMED: 61462000)

Table 10: Representative malady→disease mappings before and after the LLM+API validation pipeline. The “Before” column shows incorrect fuzzy SNOMED matches (confidence 0.52) that were replaced. Dashes indicate maladies that had no prior mapping. The “After” column shows LLM-suggested names with API-verified ontology codes. Mappings marked `llm_batch_remap_verified` have codes confirmed by exact match against ICD-10 (NLM), MeSH (NLM RDF), or SNOMED CT (EBI OLS4) APIs.

I Cross-Source Agreement

Sources Treating Same Malady	Number of Maladies
1	majority
2–5	common
6–10	moderate
11–15	5
16–20	6
21–33	3
33 (max)	1

Table 11: Distribution of maladies by number of distinct sources that claim to treat them. Most maladies are mentioned by a single source, but some (e.g., “pain”, “swelling”) are claimed by 20+ sources—a cross-source agreement signal that the Critic Agent uses to boost confidence ($\tilde{C}$ term in the graph score).

J Compound–Target Coverage Analysis

Category	Compounds	%
Has SINGLE PROTEIN targets (in graph)	181	8.8%
In ChEMBL, no qualifying bioassay data	~540	~26%
In ChEMBL, ORGANISM target only	~275	~13%
Not found in ChEMBL	~1,067	~52%
Total	2,063	100%

Table 12: Breakdown of compound–target linking coverage. Only 8.8% of compounds have TARGETS edges because the pipeline filters to SINGLE PROTEIN target types in ChEMBL. Over half of COCONUT-sourced natural product compounds are not indexed in ChEMBL at all. The ORGANISM-only category includes anti-parasitic compounds like artemisinin, whose ChEMBL target (CHEMBL364 = *P. falciparum*) is classified as a whole organism rather than a protein.

K Additional Critic Agent Evaluations

Evaluation 3: Lithospermum Root (Source) and Qi Deficiency (Malady)

KG + G3.1F Verdict: **WEAK**

Weakest Link: The lack of a specific mechanistic link between the identified targets (FTH1, NOS3, NFKB1) and the pathophysiology of **Postviral fatigue syndrome**¹.

Reasoning: While several compounds show **high binding affinity**² to biological targets, these **targets are highly pleiotropic and lack specific relevance**³ to postviral fatigue syndrome. The mapping of 'Qi deficiency' to this modern diagnosis is **speculative**³, and the evidence chains rely on general inflammatory and metabolic pathways rather than disease-specific mechanisms.

1. Clear identification of a reasonable modern disease.
2. Reference to biologically meaningful binding affinity.
3. Mechanistic reasoning that the binding affinity is weakly convincing but speculative in context with the disease.

TO + G3.1F Verdict: **UNSUPPORTED**

Weakest Link: The conceptual contradiction between clearing 'evil qi' (pathogenic excess) and treating 'Qi deficiency' (vitality depletion).

Reasoning: The historical text describes Lithospermum as a cooling, draining, and purging agent used to eliminate 'evil qi' and treat jaundice or edema. In traditional pharmacology, these actions are used to treat excess conditions and are generally contraindicated for 'Qi deficiency,' which requires tonifying rather than clearing.

1. Nonsensical evaluation because the contradiction that is cited doesn't even exist.

Evaluation 4: Ophiopogon (Source) and Vomiting (Malady)

KG + G3.1F Verdict: **PLAUSIBLE**

Weakest Link: The functional relevance of generic targets like NFKB1 and F2 to the specific emetic reflex is weak and indirect.

Reasoning: The high-affinity interaction between (-)-**Epicatechin** and **NOS3**¹ provides a plausible pathway for modulating the vomiting center, **supported by secondary links to dopamine regulation via SNCA**². However, the overall evidence is tempered by the **high pleiotropy of the targets**³ and the lack of data for the majority of the herb's chemical constituents.

1. Identification of a drug and target.
2. Added discussion of physical symptom alleviation.
3. Downsides of the analysis mentioned.

TO + G3.1F Verdict: **PLAUSIBLE**

Weakest Link: The direct anti-emetic mechanism is less established than its general mucosal-soothing and anti-inflammatory properties.

Reasoning: The text attributes vomiting to 'stomach fire' and a 'damaged center,' which maps well to **modern gastritis or acid-related irritation**¹. Ophiopogon contains **steroidal saponins (ophiopogonins) and polysaccharides**² that exhibit **anti-inflammatory and cytoprotective effects on mucosal linings**³, providing a plausible biochemical basis for soothing the gastric environment and reducing the urge to vomit.

1. Mapping vomiting to potential other causes like gastritis.
2. Broad chemical classes without specific natural products named.
3. Connected steroidal compounds to a biochemical basis for alleviation.

Evaluation 5: Epimedium (Source) and impotence (Malady)

KG + G3.1F Verdict: **UNSUPPORTED**

Weakest Link: Zero Open Targets scores for all identified targets.

Reasoning: The evidence fails at every level: the disease mapping is a category error (education vs. pathology), the targets identified (PTPN1, Polb, PYGM) have no known relevance to erectile function, and the binding affinities are weak. Furthermore, 29 of the 30 compounds in the source have no target data, and the single compound analyzed (Ursolic acid) is a ubiquitous phytochemical with non-specific activity.

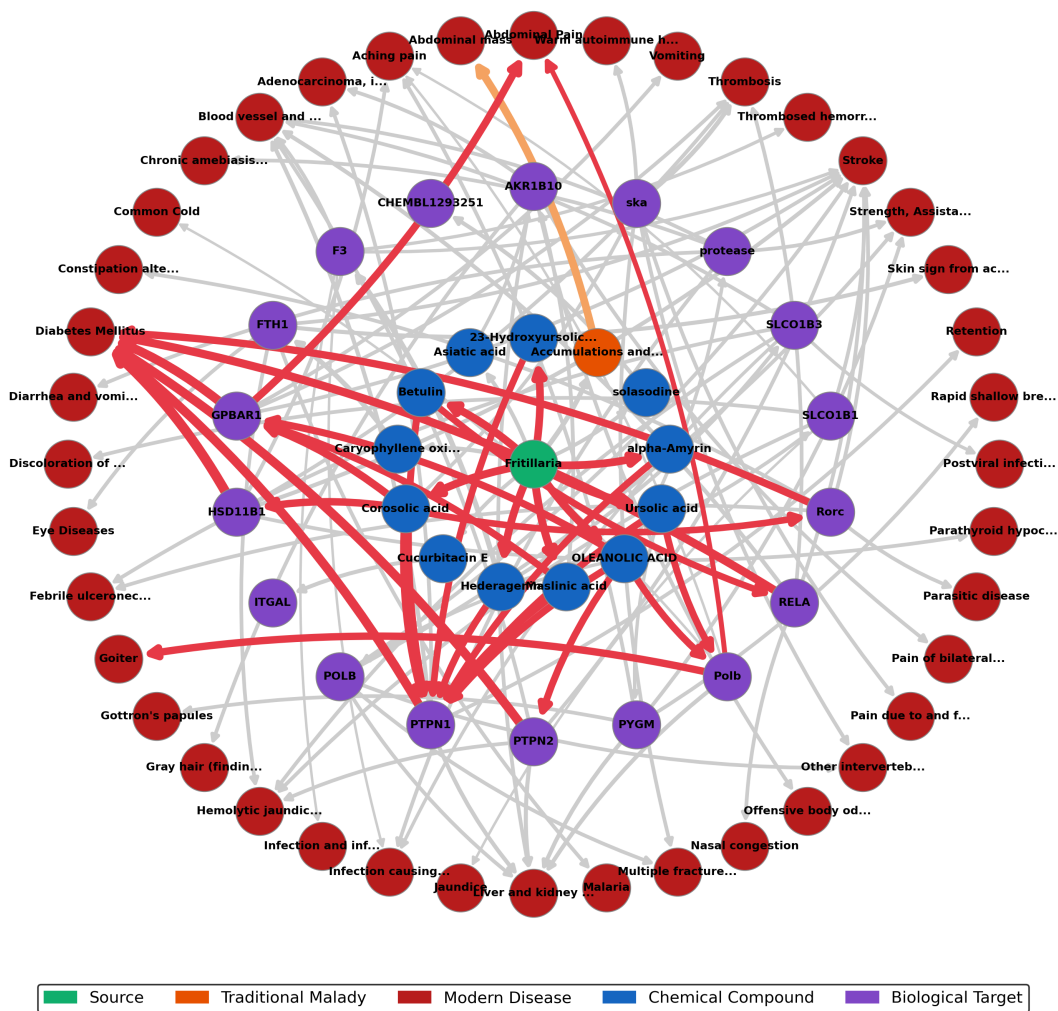
TO + G3.1F Verdict: **VALIDATED**

Weakest Link: The oral bioavailability and relative potency of icariin compared to synthetic PDE5 inhibitors like sildenafil.

Reasoning: Epimedium contains icariin, a bioactive flavonol that functions as a phosphodiesterase type 5 (PDE5) inhibitor, which is the same mechanism used by modern erectile dysfunction medications. The historical text's focus on the 'network vessels' as 'passageways' for essence aligns remarkably well with the modern understanding of vascular-mediated tumescence and nitric oxide-driven vasodilation.

Figure S1: Critic Agent outputs for Evaluations 3–5. Green notes indicate strengths; red notes indicate weaknesses; blue notes indicate neutral observations. G3.1F=Gemini 3.1 Flash; KG=knowledge graph-guided; TO=text-only.

L Fritillaria Network



Rank	Compound	Score	Rel. Conf	pChEMBL	OT Tier	Top Targets
1	Corosolic acid	0.561	0.900	6.3	gold	HSD11B1
2	Maslinic acid	0.491	0.800	6.2	gold	GPBAR1
3	Oleanolic acid	0.447	0.800	5.7	gold	GPBAR1
4	Betulin	0.437	0.800	5.5	gold	PTPN1
5	Ursolic acid	0.436	0.800	5.5	gold	PTPN1

Gemini 3 Flash Verdict: PLAUSIBLE. The evidence identifies multiple high-confidence targets such as HSD11B1 and PTPN1 that are central to glucose metabolism and insulin signaling. While the mechanistic routes are highly coherent and supported by strong Open Targets scores, the relatively low binding affinity of the compounds suggests they may act as weak modulators rather than potent therapeutic agents.

Figure S2: Visual example of graph traversal for the Fritillaria source. The network was pruned to be sparse for visualization purposes as the Fritillaria source branches to 424 nodes and 5329 edges. Orange arrows indicate the modern disease mapping to the traditional malady. The red arrows pointing to Diabetes Mellitus, Goiter, and Abdominal Pain are other modern diseases similar to the mapped "Abdominal Mass" modern disease. The table and semantic reasoning outputs are the ranked compounds and answers returned by the Critic Agent.