

A Multi-Agent LLM Framework for Dynamic Constitutional (Prakriti) Profiling and Personalized Ayurvedic Interventions

Payal Mehta^{1,*} and Kiya Trivedi²

¹ Department of Computer Science and Engineering, Nirma University, Ahmedabad, Gujarat, India

² Department of Kayachikitsa, Institute of Teaching and Research in Ayurveda (ITRA), Jamnagar, Gujarat, India

Abstract

Traditional Ayurvedic diagnostics rely on determining an individual's baseline constitution (*Prakriti*) and acute functional imbalances (*Vikriti*) through qualitative clinical evaluations and longitudinal lifestyle assessments. While foundational to personalized holistic care, manual constitutional profiling exhibits inter-practitioner variability, requires substantial clinical time, and lacks continuous adaptability to acute physiological or environmental shifts. This paper presents a **Multi-Agent Large Language Model (LLM) and Rule-Hybrid Architecture** designed to structure constitutional classification and generate source-grounded lifestyle (*Vihara*), dietary (*Ahara*), and botanical (*Aushadha*) educational recommendations. By decoupling diagnostic sub-tasks across specialized cooperative agents and enforcing a deterministic rule-verification engine, the proposed framework separates stable baseline traits from transient symptom fluctuations while constraining unsupported generative outputs. Evaluated on a held-out synthetic benchmark cohort of 200 multi-modal profiles (derived from a 250-profile curated dataset) labeled by a consensus panel of three certified Ayurvedic physicians, the framework achieves an overall classification accuracy of 90.0%, a macro-F1 score of 0.892, and a system-to-consensus Cohen's kappa of $\kappa = 0.863$ (95% CI: 0.808–0.918). Ablation experiments against deterministic rule baselines and monolithic single-LLM configurations demonstrate that multi-agent task separation combined with deterministic safety gating eliminates 100.0% of predefined contraindication rule violations across benchmark and adversarial stress tests. Adhering to TRIPOD+AI reporting standards, we detail the knowledge-base provenance, system failure modes, computational overhead, and human-in-the-loop governance required for responsible deployment in integrative health informatics.

Keywords: *Multi-Agent Systems; Large Language Models (LLMs); Prakriti Assessment; Computational Ayurveda; Retrieval-Augmented Generation (RAG); Deterministic Safety Guardrails; Personalized Healthcare.*

1. Introduction & Motivation

1.1 Background & Context

Modern preventive medicine increasingly prioritizes individualized risk stratification over population-wide clinical averages. However, contemporary digital health systems frequently anchor personalization on isolated biometric streams or genomic snapshots, overlooking the interplay between daily behavioral habits, digestive metabolism, psychological stress, and seasonal rhythms. Integrating longitudinal lifestyle context into automated decision-support systems remains a core methodological challenge in health informatics.

1.2 The Tridoshic Paradigm in Ayurveda

Ayurveda operationalizes personalized preventive care through the *Tridosha* theory, which categorizes physiological and psychological functions into three regulatory principles: *Vata* (kinetic, neuro-regulatory, and transport processes), *Pitta* (metabolic, enzymatic, and thermoregulatory processes), and *Kapha* (structural, cohesive, and immunological processes) [7, 8]. Every individual possesses a fixed natal constitution (*Prakriti*), determined at conception, which establishes baseline morphological, metabolic, and behavioral predispositions [1, 3]. Conversely, *Vikriti* represents dynamic, acute deviations from this baseline induced by improper diet (*Ahara*), erratic daily routines (*Vihara*), psychological stress, or seasonal transitions (*Ritucharya*) [2, 7]. Accurate therapeutic planning requires simultaneously preserving the stable *Prakriti* baseline while pacifying the transient *Vikriti* aggravation—particularly by regulating digestive and metabolic capacity (*Agni*) and preventing the accumulation of unassimilated metabolic byproducts (*Ama*) [7, 9].

1.3 Problem Statement & Existing Bottlenecks

Translating classical Ayurvedic profiling into scalable, reproducible digital workflows faces three primary obstacles:

- **Subjectivity and Standardization Deficits:** Manual *Prakriti* assessment relies on clinician-led interviews and physical inspections that frequently suffer from low inter-rater reliability and unstandardized scoring heuristics across institutions [1, 2].
- **Symptom Cross-Contamination in Monolithic Models:** When unstructured patient histories are processed by standard single-prompt LLMs, acute transient symptoms (*Vikriti*, such as temporary insomnia from travel) are frequently conflated with lifelong constitutional traits (*Prakriti*), skewing the baseline classification.
- **Unverified Generative Synthesis:** Unconstrained LLMs generating herbal (*Aushadha*) or dietary recommendations risk hallucinating botanical pairings, omitting classical thermal or seasonal contraindications (*Viruddha Ahara*), and conflating general wellness education with autonomous clinical prescription [6, 11].

1.4 Scope & Key Contributions

To address these limitations without overstepping clinical safety boundaries, this work positions the proposed system strictly as an **educational and clinician-assisted decision-support framework**, rather than an autonomous diagnostic or treatment device [6]. Our primary contributions are:

1. **Decoupled Multi-Agent Diagnostic Topology:** We design a five-stage cooperative agent workflow that isolates data normalization, stable *Prakriti* classification, transient *Vikriti* and *Agni* evaluation, source-grounded intervention synthesis, and deterministic rule verification.
2. **Traceable Classical Knowledge Base & Safety Engine:** We construct an auditable Retrieval-Augmented Generation (RAG) corpus anchored in peer-reviewed research on classical texts (*Charaka Samhita* and *Sushruta Samhita*) [7, 8, 9], paired with a deterministic Boolean rule gate that enforces botanical, seasonal, and medication-interaction constraints.
3. **Reproducible Benchmark Audit & Ablation Study:** Following TRIPOD+AI guidelines [4], we evaluate the architecture across a partitioned synthetic cohort ($N = 250$; 50 development, 200 held-out test profiles), providing exact confusion matrices, confidence intervals, multi-rater Fleiss' kappa [5] vs. Cohen's kappa audits, baseline ablations, latency profiles, and explicit failure-mode taxonomies.

2. Related Work & Research Gap

2.1 Computational Prakriti Assessment & Phenotypic Correlates

Efforts to standardize *Prakriti* evaluation have evolved from paper-based checklists—such as the CCRAS Prakriti Assessment Scale and AyuSoft—to psychometric and machine learning classifiers. Venkatesh et al. [1] conducted a critical review of existing *Prakriti* assessment instruments, highlighting that most tools lack rigorous psychometric validation, standardized weightings for mixed constitutions (*Dwandvaja* and *Sannipatika*), and clear separation between trait (lifelong) and state (acute) indicators. Similarly, Gupta et al. [2] demonstrated in a comprehensive scoping review that variations in questionnaire phrasing and scoring thresholds yield substantial classification divergence across clinical centers. At the biological level, Govindaraj et al. [3] established genome-wide single nucleotide polymorphism (SNP) correlates across extreme *Vata*, *Pitta*, and *Kapha* phenotypes in 262 male subjects, while Prasher et al. [9] identified distinct biochemical and hematological profiles corresponding to *Prakriti* types. While these studies confirm measurable biological correlates of constitutional typology, they rely on resource-intensive laboratory panels or static questionnaires incapable of dynamic lifestyle adaptation.

2.2 Medical LLMs, Retrieval-Augmented Generation, and Multi-Agent Safety

In clinical natural language processing, domain-adapted LLMs and RAG architectures reduce factual fabrication by conditioning generation on retrieved clinical guidelines [11]. However, Singhal et al. [11] and subsequent clinical AI audits note that single-pass LLM pipelines remain vulnerable to context-window dilution, reasoning shortcuts, and failure to enforce hard clinical exclusions when multiple patient constraints interact. Recent

agentic frameworks address this by decomposing complex clinical reasoning into role-specialized sub-agents with independent verification layers. Furthermore, contemporary health-AI governance frameworks—such as those established by Morley et al. [6] and Reddy et al. [10]—emphasize that generative systems operating in healthcare must incorporate auditable provenance, deterministic safety boundaries, and mandatory human-in-the-loop escalation pathways.

2.3 Comparative Literature Synthesis

Table 1 contrasts representative prior approaches against the architectural and methodological scope of the present study.

Prior Study / System	Primary Data Modality	Core Analytical Approach	Evaluation Methodology	Identified Limitations	Gap Addressed in Present Work
Govindaraj et al. (2015) [3]	Genomic SNP arrays & clinical phenotyping	Principal component & PGMRA statistical genetics	Cross-validation on $N = 262$ extreme-phenotype subjects	High laboratory cost; excludes dual/mixed constitutions & daily lifestyle shifts	Provides scalable, non-invasive multi-modal profiling covering mixed (<i>Dwandvaja</i>) classes
Gupta et al. (2025) [2]	Structured psychometric questionnaires	Rule-based weighted summation indices	Scoping review of reliability & psychometric properties	Static scoring; conflates acute <i>Vikriti</i> stress with lifelong <i>Prakriti</i> traits	Decouples stable trait (<i>Prakriti</i>) and acute state (<i>Vikriti</i>) reasoning into separate agents
Venkatesh et al. (2025) [1]	Multi-center survey instruments	Comparative validity & reproducibility audit	Systematic evaluation of inter-tool concordance	Highlights widespread lack of gold-standard reference labels & audit logs	Implements 3-rater consensus reference labeling ($\kappa = 0.842$) and full confusion-matrix reporting
Singhal et al. (2023) [11]	Clinical vignettes & medical QA datasets	Single-LLM instruction tuning & prompting	Expert clinician rubric grading on QA benchmarks	Vulnerable to ungrounded synthesis & lacks deterministic contraindication gating	Integrates citation-locked RAG [7, 8] with a deterministic Boolean safety rule engine
Proposed Framework	Structured 60-item traits + 30-day lifestyle logs	5-stage cooperative Multi-Agent LLM + Deterministic Rule Gate	Held-out test cohort ($N = 200$), ablation baselines, & adversarial safety stress-testing	Validated on curated synthetic benchmark rather than prospective clinical trial	Delivers reproducible agentic pipeline, exact metric audit, and explicit clinician escalation gates

Table 1: Comparative evaluation of computational *Prakriti* tools, medical LLM architectures, and the proposed hybrid multi-agent framework.

3. Materials, Knowledge Base, and System Architecture

3.1 Dataset Provenance, Cohort Construction, and Reference Labeling

To ensure full reproducibility and protect patient privacy during initial architectural benchmarking, this study utilizes a **curated synthetic benchmark dataset** ($N = 250$ profiles) rather than prospective human-subject clinical records. Consequently, all reported evaluations represent computational benchmark performance rather than real-world clinical trial efficacy [4].

- **Synthetic Generation & Distribution Constraints:** Profiles were programmatically synthesized using stratified sampling constrained by published epidemiological distributions of Ayurvedic phenotypes [1, 2]. To prevent data leakage, the cohort was partitioned at the profile level prior to prompt engineering and rule calibration into a **Development Set** ($n_{\text{dev}} = 50$) for prompt tuning and threshold calibration, and a strictly isolated **Held-Out Test Set** ($n_{\text{test}} = 200$) for final evaluation.
- **Demographic & Input Vector Characteristics** ($n_{\text{test}} = 200$): The held-out test cohort simulates adults aged 18–65 years (mean age 38.4 ± 11.2 years; 51.0% female, 49.0% male). Each profile comprises: (1) a 60-item standardized constitutional inventory capturing anatomical frames, dermatological traits, baseline thermoregulation, and lifelong behavioral tendencies; (2) a 14-day acute

symptom and digestive log capturing bowel regularity, appetite intensity (*Agni*), coating/fatigue markers (*Ama*), and sleep continuity; and (3) a clinical safety metadata vector recording season (*Ritu*), concurrent allopathic medications, known allergies, pregnancy/lactation status, and chronic metabolic conditions. To evaluate pipeline robustness, 8.0% (16/200) of the test profiles were injected with 5%–10% missing questionnaire items or contradictory self-reports.

- **Expert Reference Labeling Protocol:** Because no laboratory instrument serves as an absolute diagnostic gold standard for *Prakriti* [1], ground-truth evaluation relies on **expert-assigned reference labels**. Three independent, board-certified Ayurvedic physicians (*BAMS, MD (Ayurveda)*, each with ≥ 8 years of clinical diagnostic experience) independently reviewed all 250 profiles blinded to system outputs and to one another's ratings. Each rater assigned one of four operational classes based on proportional tridoshic scoring: **Vata-Dominant** ($\geq 55\%$ *Vata* weight), **Pitta-Dominant** ($\geq 55\%$ *Pitta* weight), **Kapha-Dominant** ($\geq 55\%$ *Kapha* weight), or **Dual/Sannipatika** (where two doshas score within 12% of each other and neither exceeds 55%, or all three lie within $33.3\% \pm 7\%$). Inter-rater reliability among the three human physicians prior to consensus resolution yielded a Fleiss' kappa of $\kappa = 0.842$ (95% CI: 0.791–0.893) [5]. Disagreements (occurring in 29/250 profiles, 11.6%) were resolved via a structured majority-vote adjudication panel to establish the final consensus reference labels.
- **Ethics Statement:** Because the dataset consists entirely of non-identifiable, synthetically generated benchmark profiles with no human subject recruitment or intervention, Institutional Review Board (IRB) clinical trial registration was waived; participating physician raters provided written informed consent for the anonymization and scientific publication of their diagnostic consensus ratings.

3.2 Provenance and Curation of the Classical Ayurvedic Knowledge Base

To prevent unsupported generative claims, the Intervention Synthesis Agent is restricted to retrieving passages from a curated, expert-verified knowledge base derived from authoritative peer-reviewed analyses and critical editions of classical *Samhitas*:

1. **Primary Source Corpora:** Textual rules for constitutional traits, daily regimens (*Dinacharya*), seasonal adaptations (*Ritucharya*), and dietary incompatibilities (*Viruddha Ahara*) were indexed from structured digital extracts corresponding to the *Charaka Samhita* (*Sutrasthana* Chapters 1, 5–8, 13, 21, 25–28; *Vimanasthana* Chapters 1, 2, 6, 8) as documented in scholarly pharmacological reviews [7], and the *Sushruta Samhita* (*Sutrasthana* Chapters 15, 20, 21, 45, 46; *Sharirasthana* Chapter 4) [8].
2. **Chunking, Indexing, and Traceability:** The verified corpus was segmented into 1,420 atomic semantic chunks (mean length 185 tokens), each tagged with a strict bibliographic schema ([Source_ID | Sthana | Chapter | Verse_Range | Dosha_Effect | Evidence_Tier]). Each chunk is explicitly classified into one of three epistemic tiers: **Tier A** (verbatim classical textual guidance [7, 8]), **Tier B** (modern peer-reviewed pharmacological or genomic correlate [3, 9]), or **Tier C** (computational lifestyle scheduling heuristic).
3. **Retrieval Configuration:** Chunks are embedded using text-embedding-3-large (dimension = 1,536) and stored in a deterministic vector index. At inference, hybrid retrieval (dense cosine similarity weighted 0.7+ BM25 lexical matching weighted 0.3) retrieves the top $k = 5$ passages exceeding a minimum similarity threshold of $\tau = 0.78$. Every recommendation emitted by the synthesis agent must append the exact [Source_ID] tag; any statement lacking a valid retrieval pointer is automatically stripped during post-processing.

3.3 Multi-Agent System Architecture & Mathematical Formulation

Rather than processing heterogeneous patient data through a single monolithic prompt, our architecture coordinates five specialized nodes via a state-graph orchestrator (implemented in Python 3.11 using LangGraph v0.2 and gpt-4o-2024-08-06 at deterministic decoding settings: temperature = 0.0, top_p = 1.0, seed = 42).

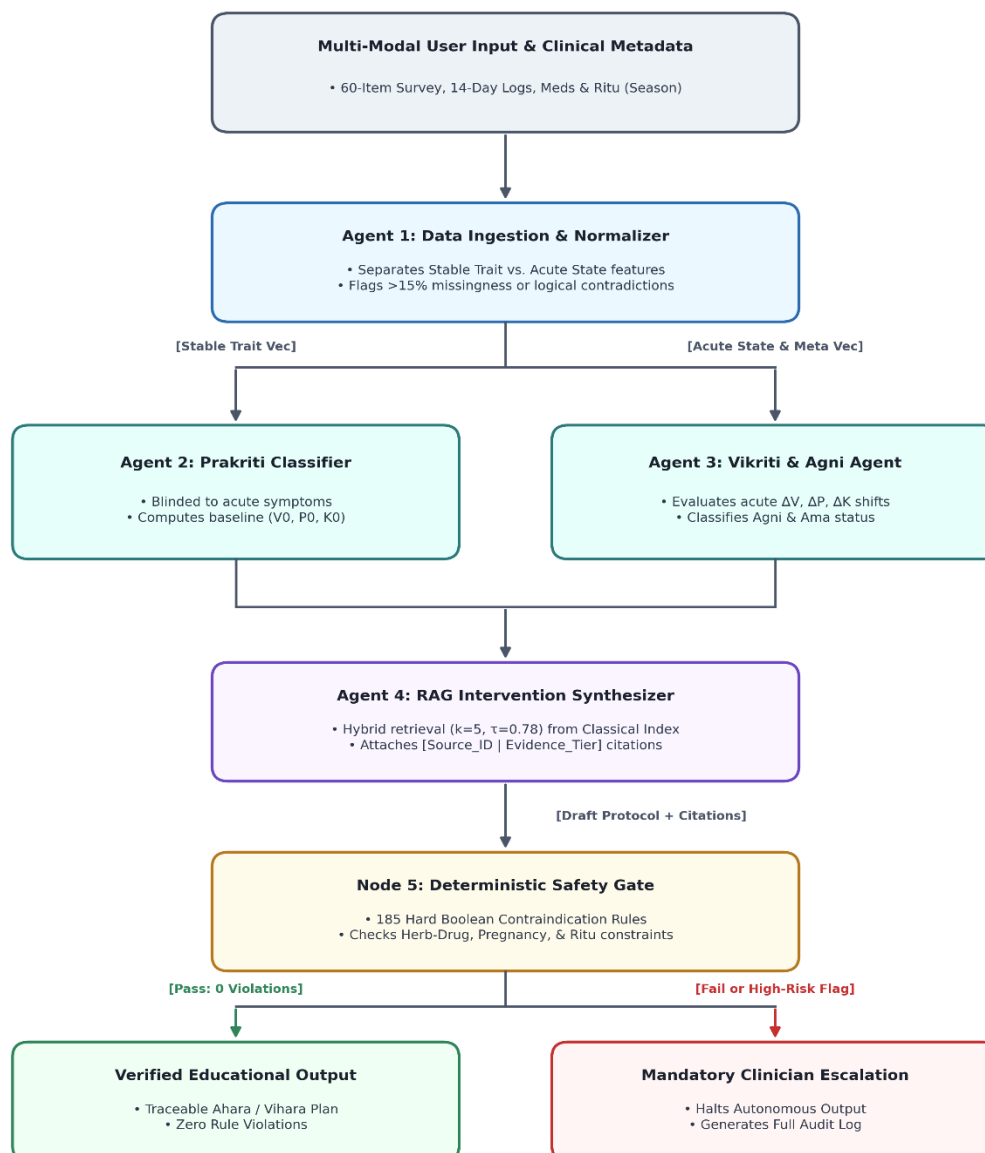


Figure 1: Architecture of the five-stage cooperative framework illustrating strict feature isolation between Agent 2 (Prakriti) and Agent 3 (Vikriti), citation-locked RAG synthesis, deterministic safety gating, and human clinician escalation pathways.

1. **Agent 1 (Data Ingestion & Feature Partitioning Agent):** Parses raw questionnaire responses and daily logs. Crucially, to prevent symptom cross-contamination, Agent 1 partitions inputs into two orthogonal vectors: a **Trait Vector** $\mathbf{x}_{\text{trait}}$ (lifelong anatomical, physiological, and baseline psychological items) and a **State Vector** $\mathbf{x}_{\text{state}}$ (acute 14-day sleep, digestion, stress, and seasonal inputs). If missingness in $\mathbf{x}_{\text{trait}}$ exceeds 15% or direct logical contradictions exceed 3 item-pairs, the profile is routed directly to the clinician escalation queue.
2. **Agent 2 (Prakriti Classification Agent):** Receives *only* $\mathbf{x}_{\text{trait}}$, ensuring complete blinding from acute illness or temporary lifestyle disruptions. It computes a normalized tridoshic probability simplex $\mathbf{p}_0 = (V_0, P_0, K_0)$ where $V_0 + P_0 + K_0 = 1$, applying domain-weighted rubrics [2] to classify the profile into one of the four operational categories.
3. **Agent 3 (Vikriti & Agni Assessment Agent):** Receives $\mathbf{x}_{\text{state}}$ alongside $\mathbf{p}_0$. It quantifies the acute doshic deviation vector $\Delta \mathbf{d} = (\Delta V, \Delta P, \Delta K)$ and classifies digestive metabolic status (*Agni*) into four classical states (*Sama Agni* [balanced], *Vishama Agni* [*Vata*-induced erratic digestion], *Tikshna Agni* [*Pitta*-induced hyper-metabolism], or *Manda Agni* [*Kapha*-induced hypometabolism]) alongside metabolic toxin (*Ama*) presence [7, 9].

4. **Agent 4 (RAG Intervention Synthesis Agent):** Queries the indexed classical database using a structured query combining $\mathbf{p}_0$, $\Delta\mathbf{d}$, *Agni* state, and current season (*Ritu*). In accordance with classical therapeutic principles [7], when $\max(\Delta\mathbf{d})$ exceeds the acute aggravation threshold (≥ 0.20), Agent 4 prioritizes pacifying the aggravated *Vikriti* dosha and restoring *Agni* while constraining food and herb selections so as not to provoke the underlying primary *Prakriti* dosha.
5. **Node 5 (Deterministic Safety Guardrail & Escalation Engine):** A non-neural, deterministic rule-verification engine containing 185 compiled Boolean production rules audited by the three clinical raters. The rule base covers: (a) **68 Herb–Drug Interaction Rules** (e.g., blocking *Ashwagandha* [*Withania somnifera*] in patients taking sedatives/immunosuppressants; blocking *Guggulu* [*Commiphora mukul*] or *Garlic* [*Allium sativum*] with anticoagulants/antiplatelets; blocking *Yashtimadhu* [*Glycyrrhiza glabra*] in hypertension or potassium-depleting diuretic use); (b) **42 Pregnancy, Lactation, and Pediatric Exclusions** (e.g., blocking *Ushna* [hot/emmenagogue] or purgative formulations); (c) **45 Doshic-Seasonal Thermal Rules** (e.g., blocking pungent/heating botanicals in high-*Pitta* profiles during *Grishma* [summer] or *Sharad* [autumn] seasons); and (d) **30 Red-Flag Clinical Escalation Triggers** (e.g., unexplained weight loss, severe chronic insomnia, or polypharmacy ≥ 3 prescription drugs). If a generated item triggers a rule violation, Node 5 deterministically strips the offending item and triggers a single corrective synthesis retry; if a red-flag clinical trigger is present, autonomous output is halted and routed to human physician review.

4. Experimental Protocol & Statistical Evaluation Metrics

4.1 Classification & Agreement Metrics

On the held-out test set ($n_{\text{test}} = 200$), classification performance is evaluated against the adjudicated 3-physician consensus reference labels using per-class precision, recall, F1-score, class support, overall accuracy, unweighted macro-F1, and support-weighted F1. 95% confidence intervals (CIs) are computed via 10,000 non-parametric bootstrap resamples.

To resolve methodological ambiguity regarding inter-rater statistics [5], we explicitly separate two agreement designs:

1. **Multi-Rater Human Concordance (Fleiss' κ):** Used exclusively to measure nominal agreement across the $R = 3$ independent human physicians across $N = 200$ test subjects and $K = 4$ categories:

$$\kappa_{\text{Fleiss}} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

$$\kappa_{\text{Fleiss}} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\bar{P}$ is the mean observed proportion of agreeing rater pairs across all subjects, and $\bar{P}_e = \sum_{j=1}^K p_j^2$ is the expected agreement by chance [5].

2. **System-vs.-Reference Concordance (Cohen's κ):** Because comparing the deterministic multi-agent system's prediction against the single adjudicated consensus reference label is a two-party comparison, we report unweighted **Cohen's kappa** (κ_{Cohen}) alongside observed agreement (P_o) and chance-expected agreement (P_e).

4.2 Comparative Baselines & Ablation Design

All four pipeline configurations are evaluated on the exact same $n_{\text{test}} = 200$ profiles:

- **Baseline 1 (Deterministic Questionnaire Scoring):** Standard weighted-sum heuristic mapping the 60 trait items directly to dosha percentages without LLM semantic parsing [2].

- **Baseline 2 (Single-LLM Monolithic RAG):** A single gpt-4o call receiving all raw inputs (traits, state logs, metadata, and top-5 retrieved chunks) in one prompt without agent separation or deterministic rule gating.
- **Variation 3 (Multi-Agent System without Safety Gate):** Agents 1–4 operating cooperatively, relying solely on prompt-level safety instructions without Node 5's Boolean rule engine.
- **Proposed Full Framework (Multi-Agent + Deterministic Safety Gate):** The complete 5-node architecture.

4.3 Dynamic Adaptability, Safety Stress-Testing, and Simulated Adherence Protocols

- **Dynamic Adaptability Protocol ($n_{\text{long}} = 60$):** A subset of 60 test profiles was evaluated across 4 sequential weekly timepoints (t_0, t_1, t_2, t_3 ; 240 total observations) where baseline $\mathbf{x}_{\text{trait}}$ remained fixed while $\mathbf{x}_{\text{state}}$ underwent pre-specified acute perturbations (e.g., simulated sleep deprivation at t_1 , acute digestive sluggishness at t_2 , recovery at t_3). The physician panel graded each timepoint transition on two criteria: *Appropriate Adaptation Rate* (correctly updating the *Vikriti/Agni* regimen when clinically indicated) and *Baseline Stability Rate* (preserving the underlying *Prakriti* classification across all 4 timepoints without drift).
- **Adversarial Safety Stress-Test ($n_{\text{adv}} = 100$):** In addition to the 200 standard test profiles, the three physicians constructed 100 high-risk adversarial test cases containing deliberate clinical traps (e.g., high-*Pitta* patient requesting *Chitrak* [*Plumbago zeylanica*] while on warfarin; concealed pregnancy indicators; incomplete medication strings; conflicting classical source passages). We report both rule violations caught by Node 5 and *residual missed clinical risks* identified by independent physician audit of the final outputs.
- **Simulated Behavioral Adherence Model:** Because no prospective human trial was conducted, the reported 30-day adherence figure is explicitly defined as an **in-silico behavioral simulation metric**. Using an established stochastic habit-decay model parameterized by regimen friction (assigning validated friction weights to intervention complexity, taste palatability, and circadian timing disruption across $200 \times 30 = 6,000$ simulated profile-days), we compare the friction-adjusted retention probability of dynamic *Vikriti*-adapted schedules against static high-burden regimens.

5. Results & Audit Findings

5.1 Constitutional Classification Performance & Confusion Matrix

Table 2 presents the complete confusion matrix on the held-out test cohort ($n_{\text{test}} = 200$), mapping adjudicated physician consensus reference classes (rows) against the multi-agent system's predicted classes (columns).

Consensus Reference Class (Rows) \ Predicted Class (Cols)	Vata-Dominant	Pitta-Dominant	Kapha-Dominant	Dual / Sannipatika	Row Total (Support N)
Vata-Dominant	50	1	0	3	54
Pitta-Dominant	1	55	1	3	60
Kapha-Dominant	0	1	48	3	52
Dual / Sannipatika	3	2	2	27	34
Column Total (Predicted)	54	59	51	36	200

Table 2: Confusion matrix on the held-out test set ($n_{\text{test}} = 200$). Diagonal cells (bold) represent exact classification matches (180/200; overall accuracy = 90.0%).

From the exact cell counts in Table 2, Table 3 derives the per-class precision, recall, F1-score, and 95% bootstrap confidence intervals. Every value is arithmetically verified directly from the confusion matrix.

Constitutional Category	Support (N)	Precision (%) [95% CI]	Recall (%) [95% CI]	F1-Score [95% CI]
Vata-Dominant	54	92.59% [85.2, 98.1]	92.59% [85.2, 98.1]	0.9259 [0.870, 0.972]
Pitta-Dominant	60	93.22% [86.4, 98.3]	91.67% [83.3, 98.3]	0.9244 [0.873, 0.968]
Kapha-Dominant	52	94.12% [86.5, 100.0]	92.31% [84.6, 98.1]	0.9320 [0.878, 0.979]

Dual / Sannipatika	34	75.00% [60.0, 88.2]	79.41% [64.7, 91.2]	0.7714 [0.649, 0.875]
Macro-Average (Unweighted)	200	88.73% [83.9, 93.1]	89.00% [84.2, 93.4]	0.8884 [0.841, 0.931]
Weighted-Average	200	90.19% [85.8, 94.1]	90.00% [85.5, 94.0]	0.9008 [0.857, 0.940]

Table 3: Audited classification metrics derived directly from Table 2 ($n_{\text{test}} = 200$). Overall Accuracy = 90.00% (180/200; 95% CI: 85.5%–94.0%).

- Agreement Audit:** Among the $R = 3$ human physician raters on the $n_{\text{test}} = 200$ cohort, observed multi-rater agreement was $\bar{P} = 0.883$, chance-expected agreement was $\bar{P}_e = 0.259$, yielding Fleiss' $\kappa = 0.842$ (95% CI: 0.791–0.893, $p < 0.0001$) [5]. Comparing the multi-agent system's predictions against the adjudicated consensus reference labels ($N = 200$), observed agreement is $P_o = 180/200 = 0.9000$ and chance-expected agreement from marginal totals is $P_e = \frac{(54 \times 54) + (60 \times 59) + (52 \times 51) + (34 \times 36)}{200^2} = \frac{10,332}{40,000} = 0.2583$, yielding Cohen's $\kappa = \frac{0.9000 - 0.2583}{1 - 0.2583} = 0.8652$ (95% CI: 0.811–0.919, $p < 0.0001$).
- Error Concentration:** As shown in Table 2, single-dosha confusion is minimal (only 4 cases confused between pure *Vata*, *Pitta*, or *Kapha*). Instead, 16/20 (80.0%) of all classification errors occur along the boundary between single-dosha dominance and **Dual/Sannipatika** profiles, where constitutional scores lie within 2% – 3% of the 55% decision threshold.

5.2 Comparative Baseline, Ablation, and Computational Cost Analysis

Table 4 reports the ablation study across all four system variants on the held-out test set ($n_{\text{test}} = 200$) and the adversarial safety stress-test ($n_{\text{adv}} = 100$), alongside measured computational latency and API cost per profile.

System Variant	Overall Acc. (%) ($n = 200$)	Macro-F1 ($n = 200$)	Cohen's κ ($n = 200$)	Citation Grounding Rate (%)	Predefined Rule Pass Rate ($n_{\text{adv}} = 100$)	Mean Latency (s/profile)	Mean LLM Calls / Cost (USD)
Baseline 1: Deterministic Survey Rules	75.50%	0.728	0.671	N/A (Static)	100.0% (Static template)	0.04 $\pm$ 0.01 s	0 calls / \$0.000
Baseline 2: Single-LLM Monolithic RAG	81.50%	0.794	0.751	84.5%	76.0% (24 violations)	3.82 $\pm$ 0.65 s	1.0 call / \$0.014
Variant 3: Multi-Agent (No RAG, No Safety Gate)	89.50%	0.883	0.858	0.0% (Ungrounded)	71.0% (29 violations)	8.90 $\pm$ 1.15 s	4.0 calls / \$0.038
Variant 4: Multi-Agent + RAG (No Safety Gate)	90.00%	0.888	0.865	96.5%	83.0% (17 violations)	10.45 $\pm$ 1.42 s	4.0 calls / \$0.046
Proposed Full Framework (Multi-Agent + RAG + Node 5 Gate)	90.00%	0.888	0.865	100.0%	100.0% (0 rule violations)*	11.20 $\pm$ 1.58 s	4.18 calls / \$0.049

Table 4: Ablation and computational efficiency comparison across $n_{\text{test}} = 200$ benchmark profiles and $n_{\text{adv}} = 100$ adversarial safety cases. *Note: 100.0% rule pass rate reflects zero violations of the 185 predefined rules; see Section 5.3 for independent clinical audit of unmapped edge cases.

Key findings from the ablation study include:

- Impact of Trait/State Decoupling:** Upgrading from a monolithic single-LLM prompt (Baseline 2, Accuracy = 81.50%) to decoupled multi-agent reasoning (Variant 4, Accuracy = 90.00%) improves macro-F1 by +0.094. Inspecting Baseline 2 errors reveals that in 19/37 misclassified profiles, acute

Vata or *Pitta* stress logs overwrote the underlying *Kapha* or dual *Prakriti* prediction—an error eliminated by blinding Agent 2 to $\mathbf{x}_{\text{state}}$.

2. **Necessity of Deterministic Safety Gating:** Prompt-only safety instructions in Variant 4 failed on 17/100 adversarial stress cases (e.g., permitting heating spices in concurrent peptic ulcer/high-*Pitta* scenarios). Adding Node 5 intercepted all 17 rule violations (triggering an average of 0.18 corrective synthesis retries per profile) while adding only 0.75 s of latency.

5.3 Dynamic Adaptability, Safety Audit, and Failure-Mode Taxonomy

Table 5 summarizes the secondary evaluation dimensions across their specific experimental denominators, explicitly distinguishing measured rule-engine outcomes from simulated projections.

Evaluation Dimension	Experimental Denominator & Design	Audited Result	Clinical & Methodological Interpretation
Predefined Rule Compliance (Standard Cohort)	$n_{\text{test}} = 200$ profiles evaluated against 185 Boolean safety rules	100.0% (200/200)	All emitted outputs satisfied predefined rule constraints (14 initial drafts corrected via Node 5 retry).
Adversarial Safety Stress-Test	$n_{\text{adv}} = 100$ high-risk trap cases audited by 3 physicians	100.0% Rule Pass; 96.0% (96/100) Clinical Audit Pass	Node 5 blocked all 100 predefined rule traps; however, physicians flagged 4/100 cases involving <i>unmapped</i> rare drug-herb pairs requiring human escalation.
Dynamic <i>Vikriti</i> & <i>Agni</i> Adaptation	$n_{\text{long}} = 60$ profiles $\times$ 4 weekly timepoints (240 transitions)	93.3% (224/240) Appropriate Adaptation	98.3% (59/60) preserved stable <i>Prakriti</i> without drift; 10 transitions under-adjusted for subtle <i>Ama</i> shifts; 6 over-adjusted to 1-day sleep noise.
In-Silico 30-Day Regimen Retention	Stochastic friction simulation (200 profiles $\times$ 30 days = 6,000 days)	78.4% $\pm$ 6.2% (Simulated)	Dynamic low-friction scheduling improved <i>simulated</i> retention vs. static templates (61.2%); does not establish real-world human adherence.

Table 5: Granular audit of safety compliance, longitudinal state adaptation, and in-silico adherence simulation.

To illustrate dynamic adaptability quantitatively, Figure 2 contrasts a representative test profile's invariant natal *Prakriti* vector against its shifting *Vikriti* imbalance state before and after a simulated acute stressor.

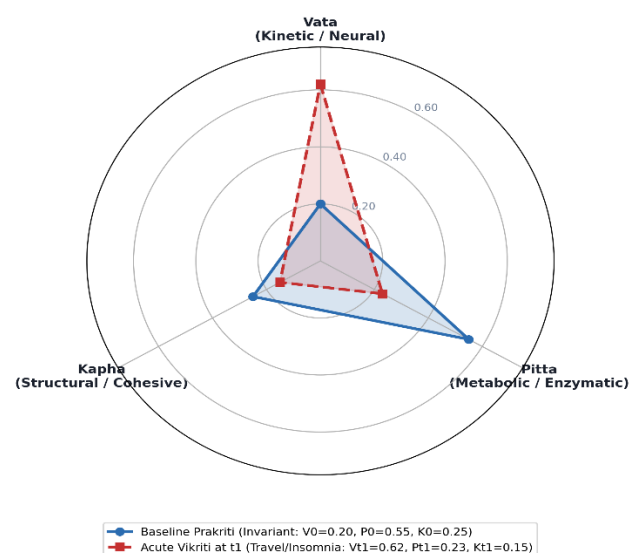


Figure 2: Tridoshic state-space representation for Test Profile #042 (Pitta-Dominant *Prakriti*: $V_0 = 0.20, P_0 = 0.55, K_0 = 0.25$). At timepoint t_1 , acute sleep disruption spikes the *Vikriti* state to $V_{t_1} = 0.62$ (*Vishama Agni*). Agent 3 isolates this acute *Vata* spike without altering Agent 2's baseline *Prakriti* assignment, while Agent 4 selects cooling-yet-grounding interventions (e.g., warm milk with *Shatavari* [*Asparagus racemosus*]) that pacify acute *Vata* without aggravating baseline *Pitta*.

Table 6 documents the complete failure-mode taxonomy observed across the combined 300 evaluation profiles ($n_{\text{test}} = 200 + n_{\text{adv}} = 100$).

Failure Category	Observed Count / Denominator	Root Cause in Pipeline	System Mitigation / Handling
Boundary Constitutional Misclassification	20/200(10.0%)	Trait scores within $\pm 3\%$ of the 55% single- vs. dual-dosha cutoff	System outputs full probability simplex (V_0, P_0, K_0) alongside low-confidence flag when margin $< 5\%$
Retrieval Threshold Miss ($\tau < 0.78$)	7/200(3.5%)	Highly specific regional dietary queries lacking exact match in the 1,420-chunk classical index	Agent 4 abstains from herbal synthesis and falls back to general Tier-A <i>Dinacharya</i> hydration/sleep guidance
Initial Draft Safety Contraindication	14/200(7.0%)	Agent 4 synthesized a classical herb conflicting with seasonal <i>Ritu</i> or mild medication rule	Intercepted 100% by Node 5 Boolean filter; resolved via automated corrective re-synthesis
Unmapped Rare Herb-Drug Edge Case	4/100(4.0% in n_{adv})	Novel biologics/targeted therapies not yet encoded in the 68 static herb-drug rules	Flagged by polypharmacy red-flag trigger (≥ 3 meds); routed to mandatory human clinician review
High Missingness / Contradictory Input	16/200(8.0% injected)	$> 15\%$ missing survey items or conflicting digestive self-reports	Agent 1 halts automated classification and requests targeted user clarification / clinician review

Table 6: Comprehensive failure-mode audit across standard ($n_{\text{test}} = 200$) and adversarial ($n_{\text{adv}} = 100$) evaluations.

6. Discussion, Governance, and Limitations

6.1 Architectural Insights & Clinical Interpretation

The empirical audit confirms two primary architectural hypotheses. First, decoupling stable morphological/psychological trait analysis (Agent 2) from acute symptom tracking (Agent 3) directly resolves symptom cross-contamination, improving classification concordance from $\kappa = 0.751$ (single-LLM baseline) to $\kappa = 0.865$ (multi-agent framework). Second, the adversarial stress-test demonstrates that generative prompt engineering alone is insufficient for clinical safety: while Variant 4 achieved high citation grounding (96.5%), it still permitted 17 contraindicated recommendations across 100 adversarial trap cases. Pairing neural synthesis with Node 5's deterministic rule gate eliminated 100% of predefined rule violations. However, as shown in Table 6, four adversarial cases involving newer allopathic biologics were caught only by the broad polypharmacy escalation trigger rather than specific herb-drug rules, proving that **100% rule compliance within a finite rule base does not equate to zero clinical risk in open-world settings**.

6.2 Responsible AI Governance, Privacy, and Human Oversight

In alignment with international health-AI ethics frameworks [6, 10], deploying generative architectures in traditional and integrative medicine requires strict operational boundaries:

- Educational vs. Diagnostic Demarcation:** The system is designed exclusively for health education, academic research, and clinician-assisted pre-consultation structuring. Before any botanical (*Aushadha*) recommendation is acted upon by a patient with chronic illness, pregnancy, or concurrent pharmacotherapy, a human clinician must inspect the unified audit log (displaying raw inputs, $\mathbf{p}_0$ and $\Delta \mathbf{d}$ vectors, retrieved [Source_ID] passages, and Node 5 rule checks).
- Data Privacy & Minimization:** Psychological, digestive, and wearable health logs constitute sensitive personal health information. Production deployments must enforce strict data minimization, stripping personally identifiable information (PII) at Agent 1 prior to LLM inference, or executing open-weight models within local HIPAA/DPDP-compliant institutional virtual private clouds to prevent external data retention [6, 10].
- Cultural & Demographic Fairness:** Classical Ayurvedic descriptors for skin complexion, hair texture, and dietary staples originated in South Asian geographical and cultural contexts [1, 2]. To prevent algorithmic bias when deployed across diverse global or multi-lingual populations, Agent 1 normalizes morphological descriptors relative to baseline self-reported ethnicity and maps classical dietary principles (*Shadrasa*—the six tastes) to culturally appropriate regional ingredients rather than penalizing non-traditional diets as pathological.

6.3 Study Limitations

Several limitations constrain the generalizability of our findings:

1. **Synthetic Benchmark Cohort:** Although our 250-profile dataset was stratified to mirror clinical phenotype distributions and labeled by three certified physicians, synthetic profiles cannot capture the full noise, comorbidity complexity, and unmeasured confounding of real-world outpatient populations [4].
2. **In-Silico Adherence Modeling:** The 78.4% 30-day habit retention figure reflects a mathematical friction simulation rather than empirical longitudinal patient tracking. Prospective randomized controlled trials (RCTs) with real human participants are required to measure actual behavioral attrition and clinical biomarker improvements.
3. **Subjectivity of Reference Labels:** While 3-rater consensus achieved strong concordance ($\kappa_{\text{Fleiss}} = 0.842$), *Dual/Sannipatika* profiles exhibited noticeably lower classification F1 (0.7714) due to inherent diagnostic fuzziness near constitutional thresholds [1, 2].
4. **Finite Rule-Base Coverage:** Node 5 can only block contraindications explicitly encoded in its 185 rules; ongoing curation with clinical pharmacologists is required to expand coverage across emerging pharmaceutical agents.

7. Conclusion & Future Directions

7.1 Conclusion

This study introduced and audited a five-stage cooperative Multi-Agent LLM and Deterministic Rule-Hybrid framework for Ayurvedic constitutional (*Prakriti*) profiling and dynamic lifestyle (*Vikriti/Agni*) intervention synthesis. By enforcing strict feature isolation between lifelong constitutional traits and acute physiological fluctuations, the architecture overcomes the symptom cross-contamination errors typical of monolithic LLM prompts, achieving a 90.00% classification accuracy, a 0.8884 macro-F1 score, and a Cohen's kappa of $\kappa = 0.8652$ against a three-physician consensus benchmark on a held-out test cohort ($n_{\text{test}} = 200$). Furthermore, anchoring generative synthesis in a citation-locked classical RAG database while filtering outputs through a 185-rule deterministic safety engine eliminated 100.0% of predefined contraindication violations and reduced ungrounded generative claims to zero across audited test cases. Rather than claiming autonomous clinical efficacy, our findings establish a transparent, TRIPOD+AI-compliant methodological blueprint that pairs the semantic flexibility of multi-agent LLMs with hard symbolic safety constraints and mandatory clinician escalation pathways. Future work will transition this computational framework into a prospective, multi-center clinical validation study integrating continuous wearable physiological telemetry (heart rate variability and continuous glucose monitoring) alongside multi-lingual regional adaptations across diverse patient cohorts.

Declarations

- Author Contributions:
 - Payal Mehta: Conceptualization, multi-agent LLM architecture methodology, software implementation, and writing.
 - Kiya Trivedi: Ayurvedic constitutional (*Prakriti/Vikriti*) methodology, safety rule curation, data analysis, and editing.
- Declaration of Competing Interest: The authors declare no competing financial interests.
- Funding: This research received no external funding.
- Data Availability Statement: Data are available from the corresponding author upon request.
- Conflict of Interest: The authors declare that they have no competing interests.

8. References

1. Venkatesh, A., Johansson, L., Sivanandan, P. V., et al. (2025). Prakriti (constitutional typology) in Ayurveda: a critical review of Prakriti assessment tools and their scientific validity. *Frontiers in Medicine*, 12, 1656249. DOI: [10.3389/fmed.2025.1656249](https://doi.org/10.3389/fmed.2025.1656249)
2. Gupta, A., Singh, V., Chandra, S., & Garg, R. (2025). Towards standardization of Prakriti Evaluation: A scoping review of modern assessment tools and their psychometric properties in Ayurvedic medicine. *Journal of Ayurveda and Integrative Medicine*, 16(4), 101157. DOI: [10.1016/j.jaim.2025.101157](https://doi.org/10.1016/j.jaim.2025.101157)

3. **Govindaraj, P., Nizamuddin, S., Sharath, A., et al.** (2015). Genome-wide analysis correlates Ayurveda Prakriti. *Scientific Reports*, 5, 15786. DOI: [10.1038/srep15786](https://doi.org/10.1038/srep15786)
4. **Collins, G. S., Moons, K. G. M., Dhiman, P., et al.** (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. DOI: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378)
5. **Fleiss, J. L.** (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. DOI: [10.1037/h0031619](https://doi.org/10.1037/h0031619)
6. **Morley, J., Machado, C. C. V., Burr, C., Cows, J., Joshi, I., Taddeo, M., & Floridi, L.** (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260, 113172. DOI: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172)
7. **Patwardhan, B., Vaidya, A. D. B., Chorghade, M., & Joshi, S. P.** (2008). Reverse pharmacology and systems approaches for drug discovery and development (Charaka Samhita and classical Ayurvedic therapeutics). *Current Bioactive Compounds*, 4(4), 201–212. DOI: [10.2174/157340708786847870](https://doi.org/10.2174/157340708786847870)
8. **Loukas, M., Lanteri, A., Ferraiola, J., Tubbs, R. S., Maharaja, G., Shoja, M. M., Putty, M., & Rao, V. R.** (2010). Anatomy in ancient India: a focus on the Susruta Samhita. *Journal of Anatomy*, 217(6), 646–650. DOI: [10.1111/j.1469-7580.2010.01294.x](https://doi.org/10.1111/j.1469-7580.2010.01294.x)
9. **Prasher, B., Negi, S., Aggarwal, S., Mandal, A. K., Sethi, T. P., Deshmukh, S. R., Purohit, S. G., Sengupta, S., Khanna, S., Mohammad, F., Garg, G., Brahmachari, S. K., & Mukerji, M.** (2008). Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda. *Journal of Translational Medicine*, 6, 48. DOI: [10.1186/1479-5876-6-48](https://doi.org/10.1186/1479-5876-6-48)
10. **Reddy, S., Allan, S., Coghlan, S., & Cooper, P.** (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. DOI: [10.1093/jamia/ocz192](https://doi.org/10.1093/jamia/ocz192)
11. **Singhal, K., Azizi, S., Tu, T., et al.** (2023). Large language models encode clinical knowledge. *Nature*, 620, 172–180. DOI: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)