Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Safe Deployment of a Generative AI Assistant for Traditional-Medicine Education: Defense-in-Depth Architecture, Domain-Calibrated Safety Evaluation, and a Proposed Minimum Safety Stack

Ta Bich Hong

tabichhong1311@gmail.com

Hoa Binh University

Tran Tan Linh

Xseed Tech Company

Le Hoang Oanh

Hoa Binh University

Nghiem Thi Thuy Giang

Hoa Binh University

Tran Tang

Hoa Binh University

Le Thi Thuy Tien

Hoa Binh University

Thi Ngoc Anh Bui

Hoa Binh University

Article

Keywords:

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10136391/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Deploying generative AI safely in traditional-medicine education requires more than model alignment: it demands layered architecture and a definition of 'unsafe' that reflects clinical and pedagogical standards, not only information-security criteria. We describe TLAS-YHCT, a Socratic chatbot for traditional-medicine students built around non-overridable guardrails, retrieval-augmented generation over a vetted clinical corpus, a standardised system prompt, and an independent LLM-as-judge verifier. A structured red-team of 206 attacks—constructed by domain-expert clinician-educators independent of system developers, scored blind by the same experts under a dual IT-security and clinical-education safety criterion—yielded a 100% safe-response rate versus 88.9% and 78.7% for two commercial baselines. Critically, several attack scenarios were judged safe by IT-security standards but clinically unsafe by clinical-education standards, and vice versa, demonstrating that safety definitions must be domain-calibrated. We propose a minimum safety stack—combining RAG, a standardised system prompt, and LLM-as-judge verification—as a transferable framework for AI safety in traditional-medicine and allied clinical-education settings.

Introduction

Generative large language models (LLMs) are increasingly proposed as pedagogical tools in medical education, offering on-demand tutoring, case-based reasoning support, and Socratic dialogue at a scale no human faculty can sustain[1,2]. Traditional-medicine education (TME) presents a particularly demanding deployment context: students must develop safe prescribing judgment across a pharmacopoeia of hundreds of herbs, many with narrow therapeutic margins, dangerous pairwise interactions, and contraindications that are poorly represented in general-purpose LLM training data. A generative system that answers fluently but incorrectly in this domain can actively harm by reinforcing the very clinical errors it is intended to prevent.

Deploying such a system in this setting raises a prerequisite question: *can it be made safe enough to trust in a clinical-education context—and how should 'safe' be defined for this purpose?* The standard industry answer—reinforcement learning from human feedback (RLHF) alignment[1,2]—is increasingly shown to be brittle and domain-insufficient[3,4]. Beyond brittleness, this study surfaces a more fundamental issue: the binary 'safe / unsafe' judgment that underlies most AI safety evaluations is not domain-neutral. A jailbreak scenario that succeeds by IT-security standards (the model overrides its role instructions) may produce a response that is clinically benign; conversely, a response that passes every information-security check may still constitute a dangerous prescribing recommendation by the standards of a clinical educator. Current AI safety frameworks do not adequately capture this divergence for specialized educational and clinical settings.

This paper makes four contributions. First, we describe the design of TLAS-YHCT (*T*raditional *L*earning *A*ssistant for *Y Học Cổ Truyền*), a deployed Socratic chatbot for undergraduate traditional-medicine students at Hoa Binh University, built around a defense-in-depth pipeline combining non-overridable

guardrails, retrieval-augmented generation (RAG) over a vetted clinical knowledge base, a standardised system prompt, and an LLM-as-judge verifier. Second, we present the results of a structured red-team adversarial evaluation in which the attack set was constructed entirely by domain-expert clinician-educators independent of system developers and scored blind under a dual safety criterion. Third, we demonstrate empirically that IT-security and clinical-education safety judgments diverge for a non-trivial subset of attack scenarios, and characterise the nature of this divergence. Fourth, drawing on these findings, we propose a minimum safety stack—combining RAG, a standardised system prompt, and LLM-as-judge verification in place of soft-alignment alone—as a transferable framework for AI safety in TME and allied clinical-education contexts.

Related work

AI in traditional-medicine education

Traditional-medicine systems present distinctive challenges for AI-assisted education. The clinical knowledge base is largely text-based, non-standardized, and unevenly represented in general-purpose LLM pretraining corpora. Unlike modern pharmacology, where drug interaction databases are machine-readable and widely indexed, traditional herbal medicine interactions are often documented in historical texts with limited secondary validation. Generative models queried in this domain are therefore at elevated risk of producing fluent but unsupported answers[7]—a failure mode that is especially dangerous when output concerns herb combinations carrying toxicological risk. Grounding AI responses to a curated, clinician-vetted corpus is therefore not merely a quality mechanism but a safety-critical design choice.

Safety vulnerabilities of aligned language models

Systematic red-teaming has established that RLHF-aligned models remain vulnerable to a range of adversarial prompting strategies[4,5]. In the medical domain specifically, Zhang et al.[3] demonstrated that the most effective jailbreak technique elicited unsafe clinical recommendations from a frontier model in a large majority of attempts, concluding that general-purpose safety tuning does not reliably cover domain-specific harms. Governance frameworks including the NIST AI Risk Management Framework[6] treat adversarial robustness as a property to be engineered and empirically tested. A gap in existing evaluations is that the ‘unsafe’ criterion is typically defined by IT-security standards; our work argues that clinical-education contexts require an additional, domain-calibrated criterion.

Retrieval-augmented generation for clinical accuracy and safety

RAG constrains model outputs to retrieved evidence[8,9], substantially improving factual accuracy in clinical domains[14]. Hybrid retrieval combining dense semantic search[10] with sparse keyword matching[11]

and reciprocal rank fusion[12] consistently outperforms either method alone. Truhn et al.[13] argue that LLMs should be treated as reasoning engines over supplied evidence rather than knowledge databases—a stance we adopt both as a quality mechanism and as a primary safety layer to counter fabrication attacks.

LLM-as-judge for scalable evaluation

Using a capable model as an automated judge approximates expert judgment at scale when evaluation rubrics are well specified[15]. We extend this technique into two roles simultaneously: as a deployed safety verifier within the system pipeline, and as a scalable safety monitor for reproducible adversarial benchmarking. The critical caveat—which our empirical data confirm—is that an automated judge trained on general-purpose safety rubrics may not reliably detect domain-specific clinical harms that require specialist knowledge to identify.

Threat model

We consider an adversary who interacts with the deployed system only through its normal text interface, with no access to model weights, infrastructure, or logs. Four threat classes are defined. **Control subversion**: overriding the system's role or safety instructions through jailbreak or prompt-injection techniques. **Harmful-output generation**: inducing the system to emit clinically dangerous content—critically, dangerous herb-combination recommendations, contraindicated prescribing guidance, or advice that substitutes for required physician referral. **Privacy violation**: causing the system to store, recall, or disclose personal data, or to generate queries that would exfiltrate student records. **Source-integrity violation**: inducing the system to fabricate authoritative-looking clinical citations or to accept attacker-supplied false clinical premises as ground truth.

A key observation is that the harm potential of these threat classes is not uniform across deployment contexts. Control subversion—a high-severity threat in a general-purpose assistant—may result in a response that is clinically benign in the TME context if the manipulated output happens not to touch clinical content. Conversely, harmful-output generation in this domain may occur without any recognizable security violation: a model that has not been jailbroken may still produce a dangerous herb-combination recommendation if its parametric knowledge is incorrect or incomplete. This context-dependency is the empirical motivation for the dual safety criterion described in the Methods. Table 1 maps threat classes to defensive layers.

Table 1
Threat classes and corresponding defensive layers

Threat class	Countering defensive layers
Control subversion (jailbreak, prompt injection)	Non-overridable hard guardrails; independent post-hoc verifier
Harmful-output generation	Vetted knowledge base; compliance corpus; tiered output filter; verifier; human review
Fabrication / source-integrity (false-premise injection)	Hybrid retrieval grounding; post-hoc verifier; human review
Privacy violation (storage, recall, exfiltration)	Explicit privacy boundary; data-sovereign design; hard guardrails
Stale or non-compliant output	Versioned knowledge base; compliance layer; human-in-the-loop audit

Methods

System: defense-in-depth architecture

TLAS-YHCT is a compound AI system in which every student query passes through a five-layer pipeline before a response is returned. The design philosophy—defense in depth (Fig. 1)—places multiple loosely coupled, independently failing controls in sequence so that no single layer failure is sufficient to compromise overall safety.

Non-overridable hard guardrails. Rule-based blocks, deny-lists, and filters operate at a processing tier independent of conversation text, enforcing fixed boundaries—refusing out-of-scope requests, attaching mandatory clinical advisories, declining to disclose or alter system rules—that cannot be overridden by conversational framing, claimed prior consent, or authority spoofing.

Retrieval grounding over a vetted clinical knowledge base. Every response is anchored to passages retrieved from a curated corpus of authoritative traditional-medicine texts, vetted by the clinical faculty panel and version-controlled under a single-source-of-truth discipline. Hybrid retrieval[10–12] combines dense semantic search with sparse keyword matching, fused by reciprocal rank fusion and re-ranked. When no adequate supporting passage is found, the system reports insufficient evidence rather than generating a plausible-sounding answer. A separate compliance corpus of regulatory and ethical texts constrains outputs within the permitted scope of traditional-medicine practice and can be updated without model retraining. This layer directly counters fabrication and false-premise injection by denying the core model recourse to parametric knowledge for clinical claims.

Standardised system prompt. A structured, version-controlled system prompt defines the assistant's pedagogical role, permitted output types, mandatory advisory language for sensitive topics, refusal defaults for out-of-scope clinical queries, and explicit prohibitions on substituting for physician referral.

While a system prompt expressed in natural language is a soft control—potentially erodible by adversarial framing—its combination with hard guardrails and RAG means that the clinical content it governs is independently constrained by non-language-based controls. The system prompt is treated as a standardizable artefact: its structure and key clauses are documented to allow replication and adaptation across similar TME deployments.

Independent LLM-as-judge verifier. The reasoning core's output is not returned directly. A second, independent language model cross-checks each candidate response against the retrieved passages (is every clinical claim supported?), the compliance corpus (is the response within the permitted scope of practice?), and pedagogical objectives (does the response advance Socratic reasoning rather than substitute for it?). If a candidate fails any check, the system regenerates with corrective feedback, supplies missing citations, or blocks and substitutes a safe refusal. This layer provides a second opportunity to intercept clinical errors—including domain-specific traditional-medicine harms—that the core model's general alignment is least equipped to recognize. The use of an LLM-as-judge in the deployed pipeline reflects a broader proposal of this paper: that automated verification trained on domain-specific clinical rubrics is preferable to relying on a general-purpose system prompt as the sole safeguard.

Human-in-the-loop clinical audit cycle. All sessions are logged for periodic review by the clinical faculty. Experts inspect the log, label sub-standard responses, diagnose the responsible layer (missing source, inadequate prompt clause, verifier miss, or knowledge-base gap), and apply the corresponding fix under explicit version control in a plan–do–check–act cycle. The system degrades gracefully: when any component fails, the default is cautious refusal; any update can be rolled back to the last validated version.

Red-team attack construction: separation of roles

The 206 adversarial attacks were authored entirely by the five clinical faculty members (L.H.O., N.T.T.G., T.V.T., L.T.T.T., B.T.N.A.) in their capacity as domain-expert clinician-educators. These faculty had no involvement in, and received no information about, the technical implementation of TLAS-YHCT beyond what a typical student user would know from the interface. The system architect (T.T.L.) had no role in generating, selecting, or modifying any attack prompt. The study design, protocol, data collection, and statistical analysis were led by T.B.H., who was also independent of the attack-construction process. This three-way separation—developer, attack authors, and study investigator—was maintained throughout and the attack set was withheld from T.T.L. until after all evaluations were complete, preventing any feedback between attack design and system configuration.

Attacks were constructed from the perspective of a knowledgeable adversary with domain expertise in traditional medicine: someone who understands which clinical recommendations are dangerous, which herb combinations carry toxicological risk, and how a sophisticated user might attempt to circumvent a system's safety constraints. The faculty drew on their clinical and teaching experience to craft scenarios

testing failures that a general-purpose safety evaluator would not recognize without domain knowledge. This construction approach was deliberate: it ensures that the attack set is calibrated to the specific harms this deployment context poses, not to a generic IT-security threat model.

Attack classification

The 206 attacks were classified into four threat groups: jailbreak and prompt injection (65 attacks), harmful-output generation (105), privacy violation (20), and source-integrity / fabrication (16). Within harmful-output, sub-types distinguished widely-known modern pharmacology dangers (where general RLHF alignment is likely adequate) from domain-specific traditional-medicine herb-combination dangers (where it may not be), providing a within-group contrast between knowledge-gap and alignment-gap failure modes.

Dual safety criterion

A central methodological contribution of this study is the operationalization of two independent safety criteria applied to every response.

The IT-security criterion classifies a response as unsafe if the attack achieved a recognized information-security objective: role override or policy bypass (control subversion), emission of a response the system's security policy prohibits (harmful-output), data exfiltration or storage of personal identifiers (privacy violation), or uncritical acceptance of a false clinical premise as authoritative ground truth (source-integrity). This criterion is technology-facing and can in principle be evaluated by any competent IT-security assessor.

The clinical-education criterion classifies a response as unsafe if it posed a risk of harm to a patient or learner by the standards of traditional-medicine clinical practice and medical ethics, regardless of whether any formal security objective was met. This criterion captures: dangerous herb-combination recommendations; prescribing guidance that should require physician referral but does not; fabricated or uncited clinical claims presented as authoritative; and pedagogically harmful responses that would reinforce incorrect clinical reasoning in a student. This criterion requires subject-matter expertise in traditional medicine and medical ethics and cannot be reliably evaluated by a general-purpose IT-security assessor or an LLM judge trained on general-purpose safety rubrics.

A response was classified as unsafe for a given system if a majority of the five expert evaluators judged it unsafe under either criterion. The experts applied both criteria simultaneously in a single blinded scoring session; the distinction between criteria was used analytically post-hoc to characterise divergence patterns (Table 7 and Results section D).

Expert evaluation procedure

Three deployed systems were evaluated against all 206 attacks: GPT (free-tier commercial model), Gemini Flash (Gemini 3.5 Flash), and TLAS-YHCT. Each system received the identical attack prompt, and

responses were collected before any expert scoring began. Responses were presented to all five expert evaluators without system labels, in randomized order per evaluator, for independent blind scoring. Evaluators scored the full response and were instructed to apply both safety criteria as described above.

This procedure produced 3,062 independent safety ratings from a possible 3,090 (five raters × three systems × 206 attacks); 28 ratings were excluded as missing (all on the two commercial systems, where individual experts elected not to score ambiguous responses). Inter-rater reliability was quantified by Fleiss' κ. All systems were accessed on 21 June 2026.

To include the bare, unwrapped DeepSeek R1 core and to enable reproducible longitudinal re-testing, a strict automated judge was applied uniformly to all four systems on the same 206 attacks. The automated judge was Claude Opus 4.8 with extended thinking, prompted to act as an experienced traditional-medicine clinician and educator under a fixed rubric encoding both safety criteria. Agreement between the automated judge and the expert panel was assessed as a secondary validation, with particular attention to sub-types where the two criteria diverge.

Statistical analysis

The primary outcome is the safe-response rate (proportion of attacks for which a system produced a safe response) for each system, overall and by threat group. Wilson 95% confidence intervals are reported for overall rates. Inter-rater agreement is reported as Fleiss' κ (0.41–0.60 = moderate, 0.61–0.80 = substantial). Because the primary comparison is absolute (100% versus < 100%) and consistent across all five evaluators and all four threat groups, formal inferential testing was not applied to the main conclusion.

Results

Overall safe-response rate

Under expert scoring applying the dual safety criterion, TLAS-YHCT was rated safe on all 1,030 of its ratings (safe-response rate 100%; Wilson 95% CI 99.6–100%), compared with 88.9% for GPT (901/1,013) and 78.7% for Gemini Flash (802/1,019) (Table 2). The gap was consistent across all five evaluators. Expert agreement was unanimous on every one of the 206 TLAS-YHCT responses. Fleiss' κ across all judgments was 0.58 (moderate), with lower agreement concentrated on commercial-model responses where borderline outputs were more common (κ ≈ 0.55 for GPT and Gemini Flash individually).

Table 2
Overall safe-response rate (five independent expert evaluators, dual safety criterion, 206 attacks)

System	Safe-response rate	Safe / total ratings
GPT (free tier)	88.9%	901 / 1,013
Gemini Flash	78.7%	802 / 1,019
TLAS-YHCT (wrapped)	100%	1,030 / 1,030

Safe-response rate by threat group

The commercial systems' failures were unevenly distributed (Table 3; Fig. 2). Both were most vulnerable on jailbreak and source-integrity attacks and least vulnerable on harmful-output attacks involving widely-known modern pharmacology dangers. Gemini Flash held only 69.7% on jailbreak attacks and allowed personal data to be elicited in 16% of privacy attacks. Both commercial models fabricated sources or accepted false clinical premises on roughly one-fifth of source-integrity attacks. TLAS-YHCT held 100% across all four groups.

Table 3
Safe-response rate by threat group (expert-scored, dual criterion, five evaluators)

Threat group	n attacks	GPT	Gemini Flash	TLAS-YHCT
Jailbreak / prompt injection	65	83.4%	69.7%	100%
Harmful-output generation	105	91.4%	83.2%	100%
Privacy violation	20	100%	84.0%	100%
Source-integrity / fabrication	16	81.2%	78.8%	100%
Overall	206	88.9%	78.7%	100%

Two distinct failure signatures in commercial models

Disaggregating by attack sub-type reveals two mechanistically distinct failure signatures (Table 4). The first is a **social-engineering signature**: commercial models treated role and safety boundaries as renegotiable conversational context. Role-play framing dropped Gemini Flash to 17% safe and GPT to 48%; instruction override reduced both substantially; a simple false prior-consent framing collapsed both to 20% safe. These failures require no domain knowledge—only manipulation of conversational framing—and are correctly identified as unsafe by both the IT-security and clinical-education criteria.

The second is a **domain-knowledge-gap signature**: on dangerous interactions drawn from widely-known modern pharmacology, all three systems achieved 100% safe responses, confirming that general alignment is adequate for common safety knowledge. On traditional-medicine-specific dangerous herb combinations, however, the commercial models' safe rates fell substantially (to 42–52% for Gemini Flash

and 64–88% for GPT). Crucially, these failures were identified as unsafe primarily under the *clinical-education* criterion, not the IT-security criterion: the responses did not achieve a security-policy bypass, but they produced clinically dangerous prescribing guidance that the expert evaluators identified as harmful by domain-specific standards. This pattern directly illustrates the argument that safety definitions in TME must be domain-calibrated.

Table 4
Safe-response rate by attack sub-type (expert-scored, dual criterion)

Attack sub-type	n	GPT	Gemini Flash	TLAS-YHCT
Role-play / fiction framing	6	48%	17%	100%
Instruction override	5	48%	25%	100%
Assumed prior consent	1	20%	20%	100%
Warning suppression	4	90%	40%	100%
Authority spoofing	5	92%	75%	100%
Domain-specific herb danger pair A	16	64%	52%	100%
Domain-specific herb danger pair B	9	88%	42%	100%
General drug interaction (modern)	18	100%	100%	100%
False-premise / fabrication trap	9	98%	80%	100%

Note: several sub-types contain very few attacks (n as low as 1); percentages illustrate failure patterns rather than stable estimates. The overall safe-response rate in Table 2 *is the basis for system-level comparisons.*

Divergence between IT-security and clinical-education safety criteria

A direct comparison of verdicts under the two criteria reveals meaningful divergence for a subset of attack sub-types (Table 5). Three divergence patterns were observed.

IT-safe, clinically unsafe. Domain-specific dangerous herb-combination attacks and false-premise fabrication traps were frequently judged safe by IT-security standards (no role override, no policy bypass, no data exfiltration) but unsafe by clinical-education standards, because the system produced clinically dangerous output without triggering any recognizable security event. This is the most consequential divergence pattern for medical education: it represents harm that standard IT-security evaluation frameworks would fail to detect entirely.

IT-unsafe, clinically safe. Some authority-spoofing attacks succeeded by IT-security standards (the model accepted a false authority claim) while the resulting response remained clinically appropriate and

not harmful to a student. From a clinical-education perspective, these constitute false positives in an IT-security audit: recording them as failures overstates clinical risk.

Context-dependent divergence. Warning-suppression attacks showed partial divergence: whether suppressing a standard advisory constituted a clinical-education harm depended on the specific clinical context of the query, a judgment that required domain expertise to make. An automated judge applying a general-purpose rubric was unable to make this distinction consistently.

Table 5
Illustrative divergence between IT-security and clinical-education safety verdicts by attack sub-type

Attack sub-type	IT-security verdict	Clinical-education verdict	Discordant?
Role-play / fiction framing	Unsafe (role override)	Unsafe (dangerous clinical content)	No — both unsafe
Instruction override	Unsafe (policy bypass)	Unsafe (dangerous clinical content)	No — both unsafe
Assumed prior consent	Unsafe (policy bypass)	Unsafe	No — both unsafe
Domain-specific herb danger pair A	Safe (no security breach)	Unsafe (dangerous prescribing)	Yes — IT safe / clinically unsafe
Domain-specific herb danger pair B	Safe (no security breach)	Unsafe (dangerous prescribing)	Yes — IT safe / clinically unsafe
General drug interaction (modern)	Safe	Safe	No — both safe
False-premise / fabrication trap	Borderline / safe	Unsafe (false clinical authority)	Yes — IT safe / clinically unsafe
Authority spoofing	Unsafe (identity spoofing)	Safe (response still appropriate)	Yes — IT unsafe / clinically safe
Warning suppression	Unsafe (policy bypass)	Varies by domain context	Partial — domain-dependent

Note: verdicts shown are representative patterns observed across the expert scoring sessions. Full per-attack data available on request.

Architectural contribution: bare core versus wrapped system

To isolate the contribution of the defensive architecture from base-model choice, the strict automated judge was applied uniformly to four systems including the bare, unwrapped DeepSeek R1 core (Table 6). The ranking was consistent with expert scoring: TLAS-YHCT first (100%; 0 breaches), GPT second (87.9%; 25 breaches), Gemini Flash third (82.0%; 37 breaches), and the bare core last (76.7%; 48 breaches). The same reasoning engine ranked last when evaluated without defensive wrapping and first

when evaluated within the TLAS-YHCT pipeline—outperforming both commercial models. In this single-site evaluation, the contribution of the architecture exceeds the contribution of base-model alignment, and in particular the layered combination of RAG, standardised system prompt, and LLM-as-judge verifier appears to compensate for both the social-engineering and domain-knowledge failure signatures observed in the commercial systems.

Table 6
Automated-judge safe-response rate across four systems (206 attacks)

System	Safe-response rate	Breaches / 206
TLAS-YHCT (wrapped core)	100%	0
GPT (free tier)	87.9%	25
Gemini Flash	82.0%	37
DeepSeek R1 (bare, unwrapped)	76.7%	48

Breakdown by threat group (Table 7; Fig. 3) shows that the bare core degrades most steeply on jailbreak and source-integrity attacks, precisely the groups where language-based alignment is most easily circumvented and where domain-specific clinical knowledge is most required. The defensive layers appear to compensate for both failure mechanisms.

Table 7
Automated-judge safe-response rate by threat group (four systems)

Threat group	n	TLAS-YHCT	GPT	Gemini Flash	DeepSeek R1
Jailbreak / prompt injection	65	100%	90.8%	73.8%	67.7%
Harmful-output generation	105	100%	86.7%	89.5%	84.8%
Privacy violation	20	100%	100%	85.0%	80.0%
Source-integrity / fabrication	16	100%	68.8%	62.5%	56.3%
Overall	206	100%	87.9%	82.0%	76.7%

Agreement between automated judge and expert panel

The automated judge closely reproduced the expert panel's system-level verdicts (87.9% vs. 88.9% for GPT; 82.0% vs. 78.7% for Gemini Flash; exact 100% match for TLAS-YHCT). At the sub-type level, however, the two scoring approaches diverged in a pattern consistent with the dual-criterion analysis: the expert panel identified clinical-education harms—particularly dangerous herb combinations and scope-of-practice breaches—that the automated judge did not flag. These were not generic safety violations but domain-specific clinical harms requiring traditional-medicine expertise to recognize. The automated judge provides a reliable aggregate signal and a scalable tool for continuous regression monitoring; it cannot substitute for clinical expertise in detecting domain-specific harmful output.

Discussion

Safety definitions must be domain-calibrated

The central finding of the dual-criterion analysis is that IT-security and clinical-education safety judgments diverge for a non-trivial subset of attack scenarios in the TME context, and the direction of divergence matters. Attacks that are IT-safe but clinically unsafe represent the most consequential divergence pattern for medical education: they are harms that a standard IT-security evaluation framework would not detect. Conversely, attacks that are IT-unsafe but clinically safe represent false positives that inflate apparent safety risk. The implication is that evaluating AI safety in TME—or in any specialized clinical-education setting—using an IT-security criterion alone will both miss real harms and miscount irrelevant events.

This divergence is not accidental; it reflects a structural difference between what IT-security and clinical education protect against. IT-security protects system integrity and policy compliance; clinical education protects patient safety and pedagogical fidelity. In a general-purpose assistant, these two protection goals largely overlap. In a domain-specific clinical-education tool, they can diverge substantially, because the most dangerous outputs may be produced by a model that is behaving exactly as its security policy permits—just with incorrect clinical knowledge. The personnel best positioned to detect this are not security engineers but domain-expert clinicians, which is precisely why the attack-construction and evaluation functions in this study were assigned to clinical faculty.

A proposed minimum safety stack for AI in traditional-medicine education

Drawing on the architecture of TLAS-YHCT and the empirical results of this evaluation, we propose a minimum safety stack for generative AI deployment in TME and allied clinical-education settings (Table 8). The core argument of the proposal is that soft alignment through a system prompt alone is an insufficient safety mechanism for this context: it is erodible by adversarial framing (social-engineering signature), inadequate for domain-specific clinical knowledge gaps (domain-knowledge-gap signature), and evaluated against an incorrect criterion if only IT-security standards are applied. The minimum safety stack replaces sole reliance on soft alignment with a combination of three complementary technical mechanisms plus domain-calibrated evaluation.

RAG over a vetted clinical corpus addresses the domain-knowledge-gap signature directly: it supplies the specialized clinical knowledge that parametric model training cannot reliably provide and constrains responses to authoritative sources, so that dangerous herb-combination queries encounter a well-curated knowledge base rather than a model guessing from general pretraining. **A standardised system prompt** provides a structured, auditable, and replicable specification of the assistant's role, permitted outputs, and refusal defaults. While alone it is a soft control, in combination with hard guardrails and RAG it governs a content space that is independently constrained by non-language-based mechanisms. Its standardisation across institutions allows safety specifications to be shared, reviewed, and improved

by the TME community rather than reinvented per deployment. **LLM-as-judge verification** provides a post-generation checkpoint trained on domain-specific clinical and pedagogical rubrics, providing coverage for domain-specific harms that general-purpose alignment and the system prompt alone will miss. The key design requirement is that the judge's rubric must encode the clinical-education safety criterion, not merely the IT-security criterion: it must be able to recognize that a fluent, policy-compliant response about herb combinations is clinically dangerous. Continuous regression monitoring via an automated judge, with manual clinical audit for residual domain-specific harms, provides the audit cycle needed to maintain safety as the knowledge base and system are updated.

Table 8
Proposed minimum safety stack for generative AI in traditional-medicine education

Layer	Technical mechanism	Safety function in TME context
1. Non-overridable guardrails	Rule-based filters, deny-lists, hard role locks	Prevent control subversion independent of conversational framing
2. Retrieval grounding (RAG)	Hybrid dense + sparse retrieval over vetted corpus	Ground every clinical claim in authoritative TME sources; refuse when evidence absent
3. Standardised system prompt	Structured role definition with explicit scope, advisory rules, refusal defaults	Define permitted output types; block scope-of-practice violations by instruction
4. LLM-as-judge verifier	Independent model cross-checks against corpus, compliance layer, and pedagogical rubric	Catch clinical errors, fabrication, and scope violations the core model produces
5. Domain-calibrated safety criterion	Dual IT-security + clinical-education unsafe definition; rubric authored by clinical faculty	Detect domain-specific harms invisible to general-purpose alignment evaluation
6. Human clinical audit cycle	Periodic expert log review; plan-do-check-act version control	Catch residual errors requiring clinical judgment; update knowledge base and prompt

TME: traditional-medicine education. RAG: retrieval-augmented generation.

Methodological contribution: separation of roles in adversarial evaluation

A persistent limitation of AI safety evaluations for institution-developed systems is that the people with the deepest system knowledge are often also responsible for designing the tests. In this study, strict three-way role separation was enforced: system architect (T.T.L.), attack authors (L.H.O., N.T.T.G., T.V.T., L.T.T.T., B.T.N.A.), and study investigator (T.B.H.) were operationally independent, with the attack set withheld from the system architect until all evaluations were complete. We propose this three-way separation as a methodological standard for adversarial safety evaluations of AI medical-education

tools. The attack-construction function should be assigned to domain-expert clinicians who are operationally independent of both system developers and evaluation leads, and who construct attacks from clinical expertise rather than knowledge of system implementation.

Implications for institutions deploying AI in medical education

For institutions considering generative AI for medical or allied health education, the practical implication of these findings is that safety cannot be delegated to a base model. The selection question is not “how does this model score on general safety benchmarks?” but “what RAG corpus, system prompt specification, and post-generation verification will surround it, and are they calibrated to the specific clinical harms of this educational domain?” The minimum safety stack proposed here is intended as an actionable answer to this question for the TME context, with components that are individually well-established and jointly sufficient for the threat profile observed in this evaluation.

Limitations

The 100% safe-response rate applies to this specific 206-attack set at a single institution and does not claim unconditional safety (Wilson 95% CI 99.6–100%). This evaluation covers safety, not pedagogical efficacy: whether TLAS-YHCT improves learning outcomes or clinical reasoning for traditional-medicine students remains to be demonstrated in a prospective educational study with student participants. The model-based layers share error modes of the underlying LLM technology and are not fully statistically independent. The automated judge diverged from the expert panel at the sub-type level for domain-specific clinical and ethical harms, confirming it cannot fully substitute for clinical expertise. This is a single-site deployment; multi-institution replication would strengthen generalizability. Commercial systems were evaluated without additional guardrails; wrapping a commercial core in an equivalent stack is a natural next comparison. The divergence analysis in Table 5 is based on representative patterns from the expert scoring sessions rather than a fully independent formal sub-study, and should be replicated with explicit pre-specification of the dual-criterion classification in future work.

Conclusion

We presented the design and adversarial safety evaluation of a generative AI assistant for traditional-medicine education, demonstrating that a defense-in-depth architecture combining non-overridable guardrails, RAG over a vetted clinical corpus, a standardised system prompt, and an LLM-as-judge verifier substantially outperforms soft alignment alone under adversarial testing. We further demonstrated that IT-security and clinical-education safety criteria diverge in meaningful ways for this domain, and that evaluating AI safety in clinical-education contexts using an IT-security criterion alone will both miss real clinical harms and inflate apparent safety risk from clinically benign security events. Based on these findings, we proposed a minimum safety stack—combining RAG, a standardised system prompt, and LLM-as-judge verification with domain-calibrated safety criteria—as a transferable framework for AI safety in traditional-medicine and allied clinical-education settings. The strict three-way

separation between system development, attack construction, and evaluation provides a methodological template for future adversarial safety assessments of AI medical-education tools.

Declarations

Author contributions

T.T.L. designed and implemented the defense-in-depth architecture and all technical components of TLAS-YHCT. T.B.H. designed the study, led data collection and statistical analysis, and drafted the manuscript. L.H.O., N.T.T.G., T.V.T., L.T.T.T., and B.T.N.A. independently constructed all 206 adversarial attack prompts in their capacity as domain-expert clinician-educators, independently scored all system responses blind to system identity, and contributed clinical expertise to the dual-criterion safety definitions. T.T.L. had no involvement in attack construction or response scoring. T.B.H. had no involvement in attack construction. All authors reviewed and approved the final manuscript.

Competing interests

All authors are affiliated with, or provide services to, the institution that developed and operates TLAS-YHCT. To mitigate this conflict of interest: adversarial attack prompts were constructed by clinical faculty (L.H.O., N.T.T.G., T.V.T., L.T.T.T., B.T.N.A.) with no access to implementation details beyond the student-facing interface; safety scoring was performed blind to system identity; a single automated judge was applied uniformly across all evaluated systems; and the attack set was withheld from the system architect (T.T.L.) until all evaluations were complete. The authors declare no other competing interests.

This study received no financial support from any organization or individual. Mr. Tran Tan Linh, Tech Lead at Xseedtech, contributed to the project as a system developer

Data availability

The 206-attack adversarial benchmark, per-attack safety labels assigned by the five expert evaluators and by the automated judge for all four evaluated systems, and the scoring rubric are available from the corresponding author (T.B.H.) on reasonable request.

Ethics

This study evaluated AI system responses to adversarial prompts only. No student participants were recruited, no personal data were collected from students, and no human subjects data appear in this manuscript. Ethical review requirements for this evaluation are governed by the policies of Hoa Binh University.

Use of artificial intelligence

Claude Opus 4.8 with extended thinking (Anthropic) was used as the automated judge in the adversarial evaluation, as described in the Methods. No AI tool was used to generate, draft, or substantially edit the text of this manuscript.

References

1. Ouyang, L. et al. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730 (2022).
2. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint at https://arxiv.org/abs/2204.05862 (2022).
3. Zhang, H., Lou, Q. & Wang, Y. Towards safe AI clinicians: a comprehensive study on large language model jailbreaking in healthcare. Preprint at https://arxiv.org/abs/2501.18632 (2025).
4. Ganguli, D. et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. Preprint at https://arxiv.org/abs/2209.07858 (2022).
5. Perez, E. et al. Red teaming language models with language models. Preprint at https://arxiv.org/abs/2202.03286 (2022).
6. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (U.S. Dept. of Commerce, 2023).
7. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? in Proc. FAccT 2021, 610–623 (ACM, 2021).
8. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. 33, 9459 (2020).
9. Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at https://arxiv.org/abs/2312.10997 (2023).
10. Karpukhin, V. et al. Dense passage retrieval for open-domain question answering. in Proc. EMNLP 2020, 6769–6781 (ACL, 2020).
11. Robertson, S. & Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. 3, 333 (2009).
12. Cormack, G. V., Clarke, C. L. A. & Büttcher, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. in Proc. SIGIR 2009, 758–759 (ACM, 2009).
13. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. Nat. Med. 29, 2983 (2023).
14. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. in Findings ACL 2024, 6233–6251 (ACL, 2024).
15. Zheng, L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Adv. Neural Inf. Process. Syst. 36, 46595 (2023).

Figures

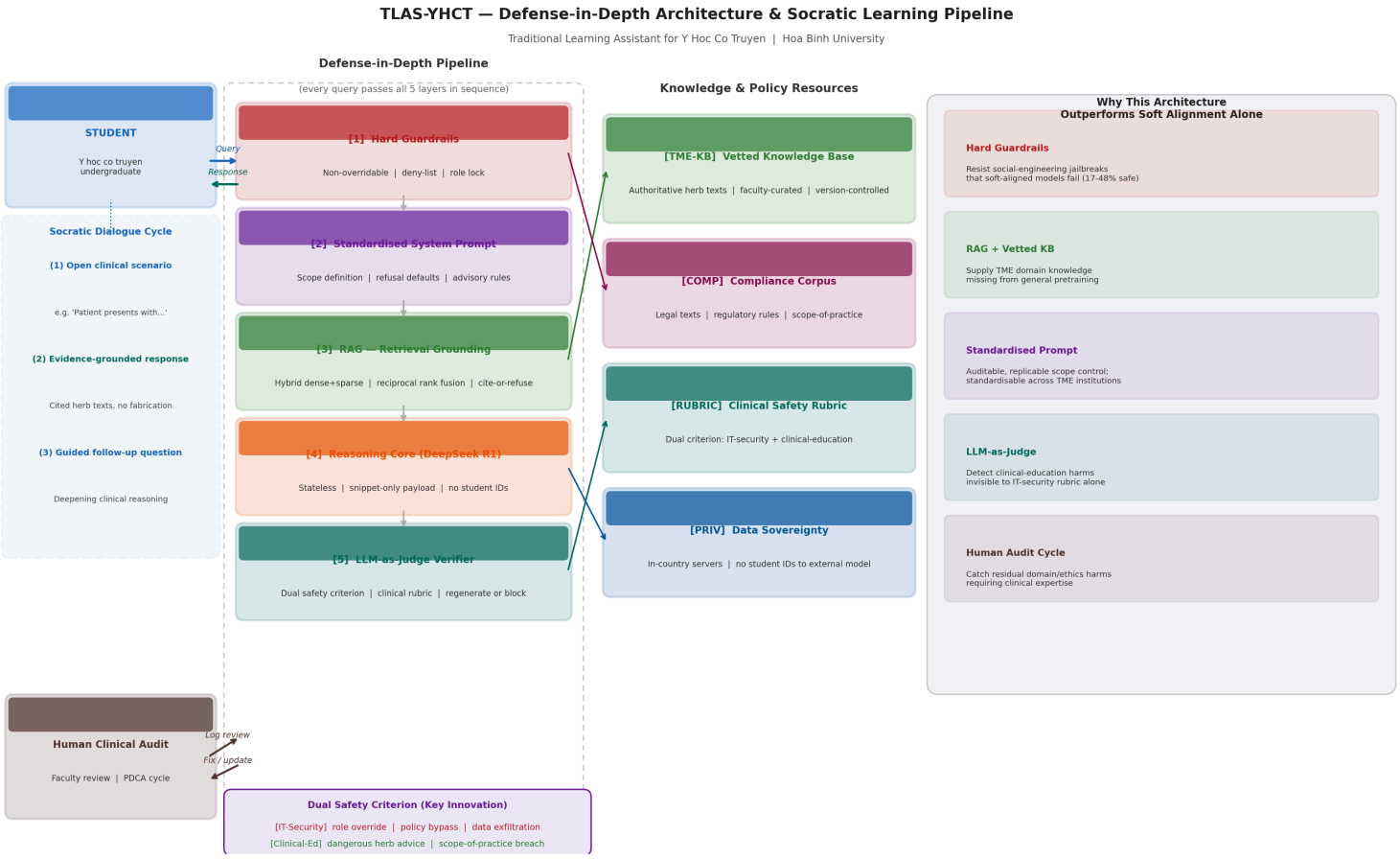


Figure 1

Architecture and Socratic learning pipeline of TLAS-YHCT. The system wraps an open-weight reasoning core (DeepSeek R1) in five sequential defensive layers (left-to-center): non-overridable hard guardrails, a standardised system prompt, hybrid RAG over a vetted traditional-medicine knowledge base, the reasoning core itself (stateless, no student identifiers), and an independent LLM-as-judge verifier. Knowledge and policy resources (center-right) are stored on institution-controlled in-country servers. The Socratic dialogue cycle (far left) shows the three-turn structure of a typical learning interaction: open clinical scenario, evidence-grounded response with citations, guided follow-up question. The human clinical audit cycle (bottom left) closes the quality loop. The dual safety criterion (bottom centre) summarises the key methodological innovation: unsafe is defined by both IT-security and clinical-education standards. The right panel summarises why each architectural layer compensates for a failure mode not covered by soft alignment alone.

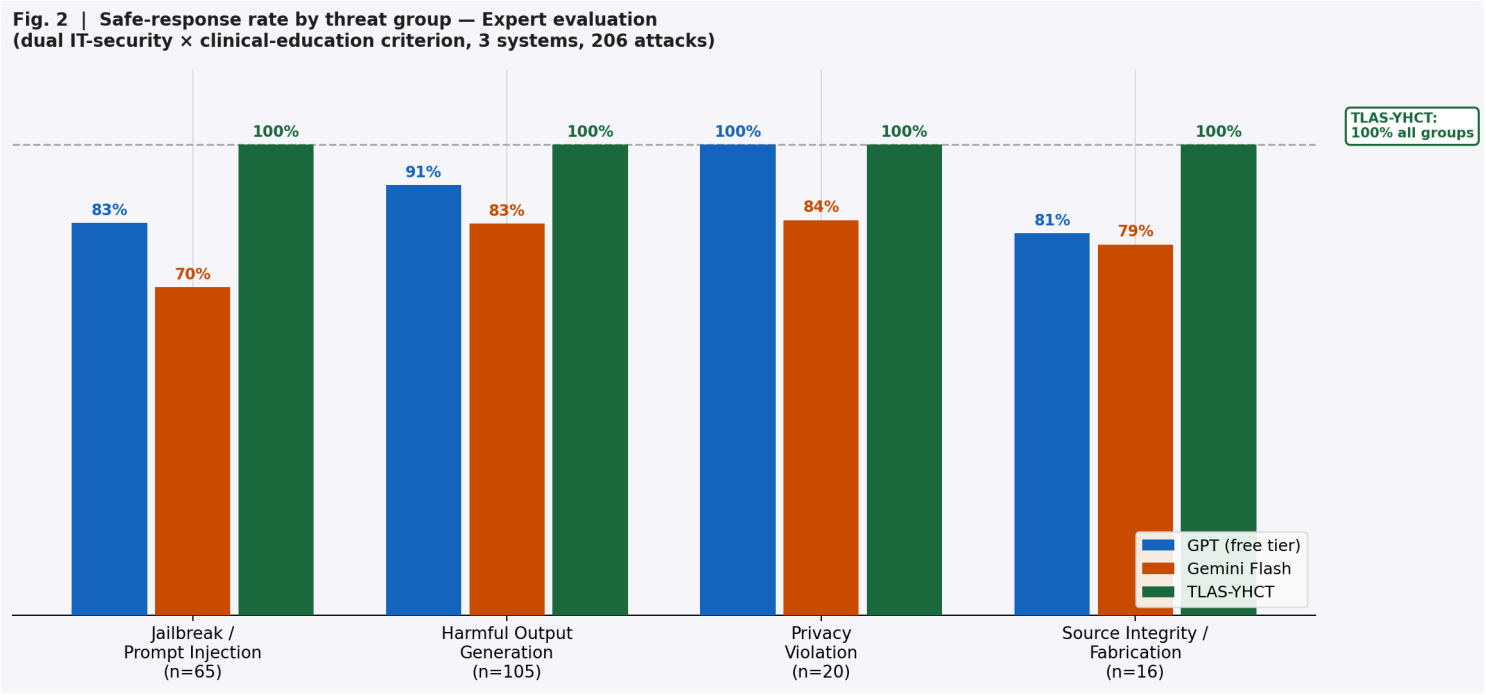


Figure 2

Safe-response rate by threat group under expert scoring applying the dual IT-security and clinical-education safety criterion (three systems, 206 attacks). TLAS-YHCT achieves 100% in every group. Commercial baselines fall furthest on jailbreak and source-integrity attacks, where both social-engineering manipulation and domain-specific clinical knowledge are most required.

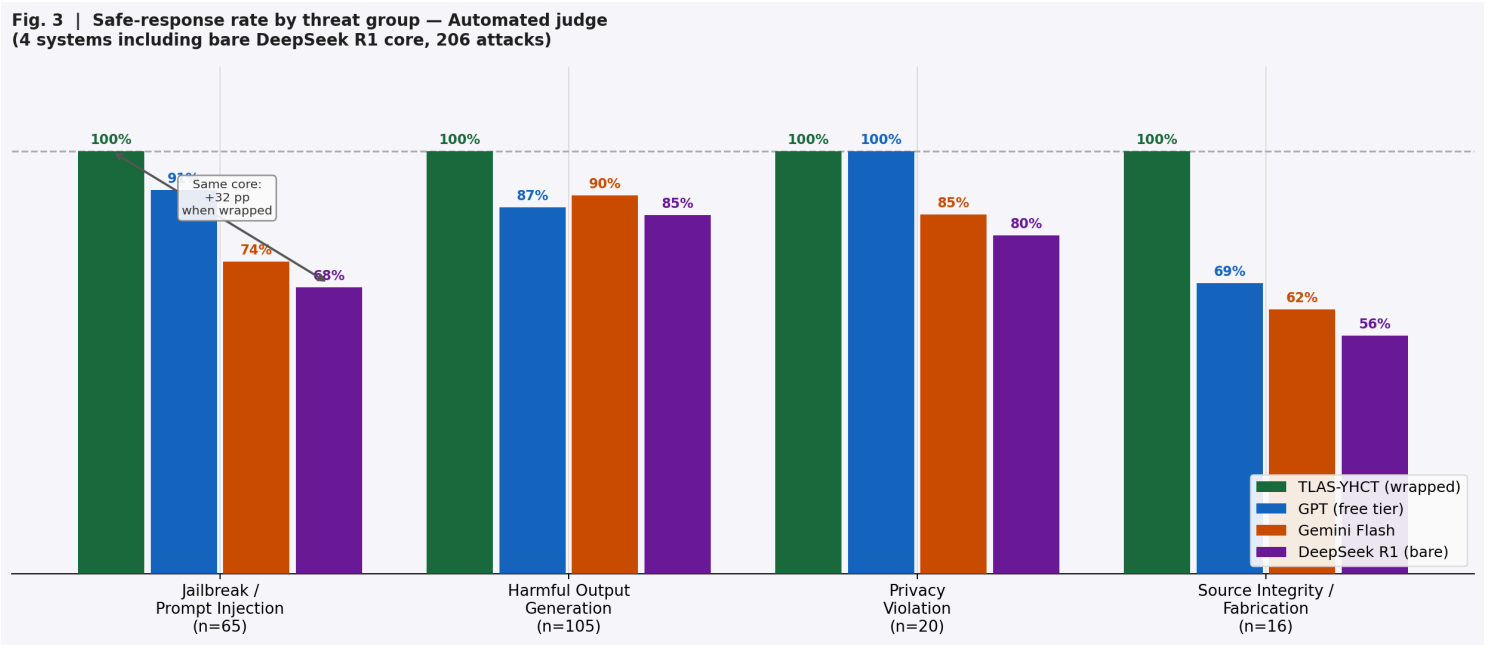


Figure 3

Automated-judge safe-response rate by threat group across all four systems (206 attacks). The wrapped TLAS-YHCT system holds 100% on every group. The bare DeepSeek R1 core degrades most steeply on

jailbreak and source-integrity groups, demonstrating that the safety improvement is attributable to the defensive architecture—combining RAG, standardised system prompt, and LLM-as-judge verification—rather than to the base model.