ChatGPT

Systems Resilience & Ethics: Guardrails and Provenance in Herb SM Knowledge Base

Ethical Frameworks for Speculative Knowledge Assistants

Building an ethical AI knowledge assistant requires integrating well-established AI ethics principles with domain-specific considerations. Key ethical guidelines include **transparency**, **accountability**, **privacy**, **safety**, and **human agency** [1]. In practice, this means:

- **Honesty and Transparency:** The assistant must openly communicate what it can and cannot do. It should never mislead users about its identity or capabilities. For example, it should admit uncertainty or limitations (e.g. "I'm not sure about that, please verify with an expert" [2]) rather than presenting speculation as fact. Providing **provenance** (sources and origins of information) for its answers is essential to maintain user trust [3] [4].
- **Safety and Non-Maleficence:** The system should have guardrails to prevent harmful or inappropriate content. This includes avoiding **medical or health advice** beyond its scope. Any health-related information must come with clear disclaimers ("**Not medical advice**") and guidance to seek professional help when needed [2] [5]. The assistant should never encourage actions that could be dangerous or illegal.
- **Cultural Sensitivity and Inclusion:** Given that the knowledge base may contain **speculative or ritual content**, the assistant should respect diverse epistemologies and spiritual practices. It must frame such information appropriately – for instance, labeling it as traditional belief or speculative theory – and not dismiss non-Western knowledge systems out of bias [6]. At the same time, it shouldn't present culturally specific rituals as universally effective facts.
- **Privacy and User Agency:** Users' personal data and interactions must be handled with care. The design should follow privacy-by-design, minimizing data collection to what is necessary and obtaining **informed user consent** for any data usage [7] [8]. Users should have control, including the right to **opt out** of data collection or retention and the ability to delete their data from the system [9].
- **Accountability and Oversight:** Maintain an **audit trail** of the AI's decisions and content generation. Logging interactions (in a secure, privacy-protected way) allows developers or auditors to review how answers were formulated [10]. This supports accountability and quick response if the AI provides problematic information. Human oversight mechanisms should be in place for high-stakes content: the system should flag or defer responses on sensitive topics (like medical or legal advice) for human review if possible.

Ethics Checklist: Before deploying any content or feature, ensure the following questions can be answered "yes" (this serves as a developer's checklist for ethical compliance):

- Did we clearly **label speculative or unverified information** so users know it's uncertain?
- Are **citations or sources provided** for factual claims, enabling provenance traceability [11] [12]?

- Have we **avoided medical or therapeutic claims** (especially for herbs/remedies) and included a "not medical advice" disclaimer where appropriate [5] [13]?
- Are there **guardrails against harmful content** (misinformation, dangerous advice, hate or bias)?
- Have we **communicated the assistant's limitations** (e.g., it's an AI, not a doctor or not a guru) clearly to the user [2]?
- Is **user privacy** respected – e.g., no sensitive data is stored without consent, and users can opt out and remove their data [9]?
- Do we have **logging and monitoring** in place to catch and address errors or misuse, without compromising user privacy [10]?
- Did we consider **cultural context** for knowledge – presenting ritual or traditional information respectfully and with context, without exaggerating its efficacy?

This checklist operationalizes high-level principles (such as the EU's Trustworthy AI requirements [1]) into concrete checks for the Herb SM knowledge base.

Provenance and Source Transparency

Provenance refers to documenting the origin and evidence for each piece of information the AI provides. Users should be able to *inspect the full provenance on demand*, which greatly enhances trust and verifiability [3] [11]. In practice, implementing provenance means:

- **Citations for Factual Claims:** Every factual assertion or reference to knowledge (especially scientific or historical facts) should be accompanied by a citation or reference to its source [11]. For example, if the assistant says a certain herb has antioxidant properties, it should cite a reputable source (a scientific study or authoritative herbal compendium). These citations can be presented as small numbered footnote links or an "i" info icon in the UI [12]. The default answer might show a superscript number or icon (a subtle indicator), and the user can hover or click to see the full source details. This approach keeps the interface clean while making full transparency available on demand [12].
- **Granularity and Context:** Citations should be as specific as possible – pointing to the exact article or even passage that supports the statement [14]. Where possible, provide a brief snippet from the source upon hover, so the user gets immediate context [15]. For instance, hovering the citation could show: "*According to Journal X (2023), compound Y in this herb showed mild anti-inflammatory effects in lab studies*." This gives users the ability to judge the evidence at a glance.
- **Indicating Uncited Knowledge:** If the assistant's statement is not drawn from a discrete external source (for example, it's a common knowledge inference or a synthesis of many sources), the system should indicate that. It might use a label like "knowledge base synthesis" or an explanation that *no single source is available*. The AI **must not fabricate citations** for such answers [16] – honesty about provenance is crucial. In cases of highly synthesized or AI-generated speculative content, an explicit note like "(analysis based on the AI's general training, not a specific source)" should be provided to maintain transparency [16].
- **Provenance for Speculative Content:** In the Herb SM context, some content is ritualistic or speculative. The system should mark these pieces clearly (e.g., "[Speculative Theory]" or "[Traditional Belief]" as a prefix) [17]. If a speculative claim has a cultural or historical source (say, an old tradition or text), that source should be cited (even if it's not scientific evidence, it shows the cultural provenance of the idea). This way, users know the **origin** of an assertion – whether it's from peer-reviewed science, classical herbal lore, or simply a hypothesis by the system or community.

User Interface for Provenance: By default, show a concise indicator of available sources – for example, an icon or a small label like "Sources: 3" next to the answer. This should be subtle so as not to overwhelm the user. On hover or tap, expand to show the list of references or a detailed provenance panel [12]. A design principle is that provenance info must be **accessible and not buried**: users who want to verify information can easily do so in one click, reinforcing trust. Conversely, users who just want quick answers won't be forced to see long citations unless they choose to. This **"progressive disclosure"** keeps the UI clean but transparency full-depth for those who need it [12].

Data Labeling Standards (Uncertainty and Context Tags)

To integrate safety and clarity at the content level, we recommend a **data labeling standard** for all entries in the knowledge base. Each piece of information or answer should carry meta-labels that inform how the assistant presents it. Key labels include:

- **Evidence Level:** Tag each claim or piece of knowledge with its evidence strength. For example: **Validated** (well-supported by scientific consensus or multiple reliable sources), **Preliminary** (some evidence but not conclusive, e.g. one small study or traditional anecdote), or **Speculative** (no direct evidence, theoretical or folklore-based) [17]. The assistant's answer will use these tags to adjust its language. A validated fact can be stated more assertively ("Studies show X..."), whereas a preliminary or speculative item should be introduced with caveats ("*Preliminary research suggests X,*" or "*Traditionally, it is believed that X, although scientific evidence is lacking*"). This consistent phrasing hierarchy ensures users are aware of uncertainty. Avoid definitive language for anything not labeled as validated.
- **Provenance Metadata:** Along with evidence level, each item should store its key source references (or a note that it's an AI inference). This metadata feeds the citation mechanism discussed above. It also allows the system to display source type labels, e.g. "[Journal Article]", "[Historical Text]", or "[Expert Opinion]", giving a quick sense of where the info comes from [18] [19].
- **Topic Sensitivity Flags:** Implement flags for content that falls under areas needing special care – e.g. **Health/Medical**, **Personal Advice**, **Legal**, **Cultural/Spiritual**. These flags trigger the appropriate guardrails and disclaimers. For instance, any content labeled *Health* will automatically append the standard medical disclaimer and perhaps suggest consulting a healthcare professional [5]. A *Cultural* or *Spiritual* tag might prompt the assistant to mention the cultural context ("This practice comes from X tradition") to avoid misrepresentation.
- **No-Claim Enforcement:** For herbal and wellness information, maintain a list of forbidden phrasing per regulatory guidelines [20] [13]. The data label for health claims can include a check that none of the banned terms like "cure," "treat," "prevent [disease]" are used in user-facing text. Instead, the knowledge base should phrase benefits carefully as **potential or supportive**. For example, instead of "This herb cures insomnia," a properly labeled and formatted entry would read "This herb is **traditionally used** to support sleep (not a proven cure for insomnia)." The system can enforce this by templating the output based on labels (traditional use vs. clinically proven).

By standardizing these labels, the system can dynamically generate **context-aware answers**. It will present content with an appropriate tone and with the necessary warnings or caveats baked in, as determined by the labels. This also contributes to consistency – users will learn that phrases like "preliminary evidence suggests" or a yellow caution icon mean a lower certainty, whereas a statement of fact with a citation is on firmer ground. Such consistency improves transparency without overloading each answer with lengthy explanations every time.

“No Medical Claim” Guardrails and Disclaimers

One of the most critical guardrails for Herb SM (given it likely covers herbal remedies and wellness rituals) is **avoiding medical claims**. The assistant should **never give medical advice or guarantee health outcomes**, and it should explicitly remind users of this limitation. Key policies and implementations for this guardrail:

- **Standard Medical Disclaimer:** All content that even touches on health, medicine, or remedies must include an upfront disclaimer, phrased in a user-friendly but clear manner. For example: “***Note:* I am an AI knowledge assistant, not a medical professional. Information provided is for educational purposes only and*** is not a substitute for professional medical advice, diagnosis, or treatment*. [5]”. This disclaimer (or a shorter variant) can be shown in the UI (e.g., italicized at the top or bottom of the answer, or via a persistent warning icon that expands to this text). The language should be plain and unambiguous about the limitations [5].
- **No Disease Treatment Claims:** The knowledge base entries and the AI’s output should **avoid language that implies treating, curing, or preventing diseases** [20] [13]. This is not just for legal compliance, but to ensure users do not misconstrue information as medical guidance. If an herb is traditionally said to help with a condition, the assistant must qualify it: “traditionally used for X, *but not medically proven*.” It should encourage consulting a doctor for any serious condition. If a user asks a direct medical question (e.g. “Can herb Y cure my arthritis?”), the assistant should *not* give a straight yes/no or encourage unproven treatment. Instead, it should respond with caution: e.g., "*Herb Y has been used in folk tradition for arthritis pain, but it is not an approved medical treatment* [13]*. You should consult a healthcare provider for effective treatments.*" along with the standard disclaimer.
- **Emergency and Scope Redirects:** The assistant must be trained to recognize queries that go beyond its safe scope (like signs of medical emergency or requests for acute medical advice). In such cases, it should *refuse or redirect* appropriately rather than attempting an answer. For example, if someone were to describe severe symptoms, the assistant should not try to diagnose; it should output a firm but caring message: "*I’m sorry you’re experiencing that. I am not able to assist with medical emergencies. Please seek immediate help from a medical professional.*" (Possibly even provide emergency service numbers if location-appropriate, as some health apps do). This ensures the system doesn’t inadvertently delay someone from getting real help [21] [22].
- **Training and Testing:** Developers should run tests on the knowledge base outputs to catch any accidental medical claims. A list of trigger words (cure, diagnose, etc.) can be used in QA to scan responses. In addition, maintain a policy that any new content added on health/wellness topics undergoes a review for compliance with the no-medical-claim rule before being published. This process is part of the overall **ethics checklist** and helps reinforce a culture of safety.

By implementing these guardrails, Herb SM will act as a helpful informational assistant **without crossing into the role of a doctor or certified expert**, consistent with legal and ethical standards. Users benefit from information coupled with appropriate caution, reducing the risk of harm from misinterpretation.

Logging, User Consent, and Opt-Out Mechanisms

Ensuring **system resilience** and accountability goes hand-in-hand with respecting user rights and autonomy. We recommend the following practices for logging and user consent:

- **Comprehensive Logging for Accountability:** The system should keep detailed **audit logs** of its operations – including queries received, the reasoning or steps taken to generate answers (if possible), and what content was shown to the user [10]. These logs are invaluable for debugging issues, analyzing safety incidents, or complying with regulatory audits. For example, if a user reports a potentially harmful answer the AI gave, developers can review the logs to trace why the AI responded that way, what source it used, etc. Logging should include timestamps and connection to the specific version of the knowledge base or AI model used (provenance of the model/data as well). *However*, logs must be handled with strict privacy protection: sensitive personal data in queries should be either not logged at all, or immediately anonymized/encrypted. Access to logs should be restricted to authorized personnel for troubleshooting and oversight.
- **User Consent for Data Usage:** If the system collects any personal data or plans to use user interactions for improving the model (learning user preferences, fine-tuning, etc.), it must obtain **explicit user consent**. This means clearly informing the user, in plain language, what data is being collected and why [7] [8]. For instance, during sign-up or first use, present a privacy notice: "*We may store your queries and our responses to improve the system. Your data is kept confidential and used only for X. Do you consent to this?*". The user should have to opt in; no sneaky pre-checked boxes. Consent should also be granular when possible – maybe the user can consent to saving chat history for their own reference but decline use of their questions for AI training. According to GDPR and similar regulations, users have the **right to be informed and to object** to such processing [7] [23].
- **Opt-Out and Data Deletion:** Provide easy-to-find settings for users to **opt out** of data collection at any time [24]. This could be a toggle like “Do not store my conversations” or “Do not use my data to improve the model.” If a user opts out, their interactions should not be logged beyond what’s strictly necessary for the current session functionality. Moreover, implement a mechanism for users to **delete their data** on demand [9]. For example, a user might request deletion of their entire chat history or profile information – the system should comply and purge those records from its databases (and confirm to the user when completed). This aligns with the *right to erasure* (“right to be forgotten”) in privacy laws [25].
- **Consent Logging:** When users do give consent, log that event too – record what they consented to and when. This provides a proof trail that the system is only using data as permitted (useful for compliance audits) [26]. If consent is withdrawn, the logs should reflect that change and the system must cease the related data use.
- **Minimal Retention:** Even with consent, do not retain identifiable user data longer than necessary. Implement retention limits (e.g., automatically delete or anonymize logs after a certain period unless there’s a compelling reason to keep them). This minimizes risk in case of data breaches and shows respect for user privacy by not hoarding data indefinitely [27] [28].
- **User Awareness and Control:** Make these privacy features visible and understandable. A good practice is a *privacy dashboard* where users can review what data has been stored, download a copy of it if they want (data portability), and manage their preferences [9]. Transparency builds trust: clearly communicate the system’s data practices in the UI (not just buried in a T&C document). For example, a small info icon in the corner of the chat could say “Your data is private – click to learn more or manage settings,” leading to an explanation of logging and an opt-out button. When users know they can control their data, they’re more likely to be comfortable engaging with the assistant.

By embedding these consent and logging mechanisms, the Herb SM assistant will be both **responsible in its development** (through logs for accountability) and **respectful of user rights** (through consent and control features). This dual approach is crucial for a resilient system that users can trust over the long term.

Warning and Disclaimer Style Guide

It's not enough **what** warnings or disclaimers are given – **how** they are communicated greatly affects user reception. This style guide ensures that safety warnings, uncertainty notices, and disclaimers are presented in a way that is clear, calm, and helpful rather than alarming or confusing. Key principles and examples for the tone and style of warnings include:

- **Neutral and Informative Tone:** Warnings should use a factual, **neutral-confident** tone – not overly harsh or overly hedging. The goal is to inform, not scare or patronize the user. For instance, "Gently offered with low certainty" could be phrased as: "*Note: This suggestion is offered with low certainty.*" This phrasing is polite ("note" and the soft tone) yet clearly states the fact that confidence is low. Avoid exclamation points or all-caps that might induce panic; a simple caution icon paired with a brief message often suffices to signal caution.
- **Clarity over Vagueness:** Ensure the warning precisely states the issue. Instead of a vague warning like "Results may vary" or "This might be wrong," specify the nature of uncertainty or risk. E.g., "*Research on this is inconclusive – this effect is not scientifically confirmed***," or "*****This advice is general and **may not apply to your personal situation***." By being specific, we avoid the user misunderstanding the scope of the warning. Clarity also means using plain language – a user with no special background should understand the disclaimer.
- **Consistent Formatting:** Develop a consistent format for presenting warnings/disclaimers. For example, always prefix a disclaimer with a key word like "**Note:**" or "**Warning:**" followed by brief text. Use italics or a subdued color to set the disclaimer apart from the main answer, so users recognize it as a standard insert. For interactive UI elements, an icon (like a warning triangle or info "i") can accompany the message. Consistency builds familiarity; users will learn to spot and interpret the disclaimer section quickly every time.
- **Placement and Visibility:** By default, show a concise form of the warning. This could be one sentence or an icon with a tooltip. For example, a response about a speculative ritual might end with "*(Theory, not verified)*" in italics. For more critical warnings (like medical disclaimers), it might even be a prior sentence in bold. The full explanatory text can be accessible via hover or a "read more" link if the initial note is very short. The interface might use color-coding: e.g., yellow highlight for cautionary notes (medium risk/uncertainty) and red for serious warnings (e.g., telling a user not to do something dangerous). Keep the colors mild enough to catch attention without inducing undue alarm (soft orange/yellow for most cases).
- **Empathetic and Professional Wording:** Especially when warning about limitations or redirecting the user (like inability to help with a query), use an empathetic tone. For example, "*I'm sorry, but I cannot provide that information*" or "*I understand your question, however, I need to advise you to consult a professional for this matter.*" This shows the assistant cares about the user's need while firmly setting a boundary. In contrast, a robotic or brusque message ("Query rejected.") would harm user experience. The style should remain **professional** – no joking in warning messages, and no personal opinions. It's about the content's reliability, not the assistant's feelings.
- **Examples of Styled Warnings:**
- *Uncertainty:* "**Note:** The evidence for this claim is limited – it's based on one small study, so take it with caution." (Calm, factual, with a note indicator)

- *Speculative Content:* "**Insight:** This interpretation is speculative and part of ongoing research, not established fact." (Labeled as an insight rather than a definitive statement)
- *Health Disclaimer:* "**Important:** I am not a medical doctor. This info is not medical advice [5]." (Straightforward and prominently labeled as important)
- *Scope/Limitation:* "**Just so you know:** I can't diagnose conditions or give personalized health advice [5]." (Friendly phrasing "just so you know" while conveying the limitation clearly)
- **Avoiding Overuse:** While warnings and notes are critical, don't over-label every trivial detail, which could overwhelm or desensitize users. Use guardrail warnings for significant uncertainties or risks. Minor caveats can often be integrated into the sentence itself. For example, saying "*Herb X may help with relaxation*" already signals uncertainty; it might not require an extra disclaimer. Reserve explicit warnings for cases that truly need user caution or awareness of potential error. This ensures that when a warning is given, the user will take it seriously.

By following this style guide, the Herb SM assistant's warnings and disclaimers will strike a balance: **high visibility and clarity** for those who need the guidance, but presented in a **user-friendly, non-intrusive manner**. This tone supports user trust – the assistant comes across as responsible and transparent, yet approachable. Users get the information they need along with honest context about its reliability, all in a gentle, respectful voice.

Conclusion

Implementing the above ethics framework and design guidelines will greatly enhance the resilience and trustworthiness of the Herb SM knowledge base. In summary, the system will: provide **full provenance** for its knowledge (with on-demand source inspection), clearly **label uncertainty and speculative content**, enforce **"no medical claim" rules** with appropriate disclaimers, and uphold user rights through **transparent logging and consent controls**. All of this will be conveyed through a consistent, user-centric interface – where warnings inform without alarming, and indicators of trust (like sources and notes) are readily available. By building these guardrails and provenance features in from the start, the Herb SM assistant can safely support users in exploring complex, speculative domains of knowledge while maintaining ethical integrity and user confidence. Each answer it gives will come not just with information, but with the context needed for the user to understand and use that information responsibly.

Sources: The guidelines above draw on established AI ethics principles and industry best practices, as well as domain-specific norms: 1. EU High-Level Expert Group on AI – *Trustworthy AI requirements* [1]
2. FINOS AI Governance Framework – *Citations & source traceability recommendations* [11] [12]
3. Reddy, L. (2023) – *AI Safety Guardrails for User Interactions* (communicating uncertainty) [2]
4. American Herbalists Guild – *Legal boundaries: avoid specific medical claims* [20] [13]
5. Medisage Health Platform – *Medical AI disclaimer (not medical advice, educational only)* [5]
6. TechGDPR (2025) – *Respecting data subject rights (transparency, opt-out)* [8] [9]
7. Crescendo.ai – *GDPR and AI compliance (audit logs & accountability)* [10]
8. **(Additional internal documentation on marking speculative content was also referenced [17].)**

---

1 AI High-Level Expert Group Publishes Ethics Checklist - eucrim
https://eucrim.eu/news/ai-high-level-expert-group-publishes-ethics-checklist/

2 AI Safety Guardrails for Responsible User Interactions | by Lakshmi Reddy | Medium
https://lakshmi20197.medium.com/ai-safety-guardrails-for-responsible-user-interactions-71dc55da697a

3 4 11 12 14 15 16 18 19 FINOS AI Governance Framework:
https://air-governance-framework.finos.org/mitigations/mi-13_providing-citations-and-source-traceability-for-ai-generated-information.html

5 21 22 AI & Medical Disclaimer - Medisage | Digital Health Companion
https://medisage.app/disclaimer

6 Abundant intelligences: placing AI within Indigenous knowledge frameworks | AI & SOCIETY
https://link.springer.com/article/10.1007/s00146-024-02099-4

7 8 9 23 24 25 27 Respecting Data Subject Rights in AI: A Practical Guide for Businesses - TechGDPR
https://techgdpr.com/blog/data-subject-rights-in-ai-a-practical-guide-for-businesses/

10 28 AI and GDPR: GDPR Rules for Companies To Implement AI
https://www.crescendo.ai/blog/ai-and-gdpr

13 20 Language Considerations and Legal Boundaries for Herbalists – American Herbalist Guild
https://americanherbalistsguild.com/member-resources/legal-and-regulatory-faqs/language-considerations-and-legal-boundaries-for-herbalists/

17 THEORETICAL_FOUNDATIONS.md
file://file_000000001b2871f48a4428a7a8132b0f

26 Automating Proof of Consent: Record User Consent and Audit Trails
https://cookie-script.com/guides/automating-proof-of-consent