

2 SKY APPROACH (2SA)

User-Invoked Dual Verification for Reducing AI Hallucination Risk

1. Purpose of This Brief

This approach is referred to as the **2Sky Approach (2SA)**, reflecting a dual perspective and a higher-level view on uncertainty and reliability.

This document proposes a **user-invoked, two-step verification mechanism** for AI-generated analytical outputs.

2. Problem Statement

AI systems operate under inherent uncertainty due to:

- incomplete or outdated training data
- probabilistic language generation
- ambiguity and noise in human communication

As public usage increases, so does:

- the visibility of hallucinations
- the reputational and practical risk in academic, legal, and analytical use cases

Current mitigation strategies (disclaimers, confidence language, citations) are often:

- cognitively intrusive
- embedded within the answer, disrupting comprehension
- insufficient for users who require post-hoc verification.

3. Core Idea: User-Invoked Dual Verification

After receiving an AI response, the user is presented with a single optional choice:

Would you like to verify this answer?

☐ Yes ☐ No

- If the user selects **No**, the interaction ends.
- If the user selects **Yes**, a two-step verification process begins.

4. Verification Step 1 – Process Disclosure

Objective: Identify potential procedural weaknesses.

The AI:

- lists the types of sources or knowledge categories used (not raw training data)
- explains the reasoning approach at a high level
- highlights known points of uncertainty

This step allows detection of:

- over-reliance on a narrow source type
- reasoning shortcuts
- contextual misinterpretation

5. Verification Step 2 – Independent Re-Derivation

Objective: Cross-validate the result.

The AI:

- answers the same user question again
- avoids reusing the same reasoning path or source categories
- prioritizes alternative framing or perspectives
- If both answers converge, confidence increases.
- If they diverge, uncertainty is surfaced explicitly.

6. Final Output – Probabilistic Confidence Assessment

The verification concludes with:

- a qualitative confidence range (never 100%)
- a short explanation of remaining uncertainty
- explicit acknowledgment that unknown or real-time external events may affect correctness

This mirrors real-world epistemic limits rather than claiming absolute truth.

7. Why This Is Distinct from Fact-Checking

While superficially similar to fact-checking or fake-news verification, key differences include:

- **Intent:** error mitigation, not deception detection
- **Agency:** user-initiated, not enforced
- **Focus:** reasoning robustness rather than factual policing

8. Applicability

This mechanism is most suitable for:

- academic and scientific analysis
- policy and regulatory interpretation
- safety evaluation and auditing
- expert decision-support systems

It is not intended for casual or low-stakes usage by default.

9. Limitations (Explicit)

- Does not guarantee correctness
- Relies on honest system self-reporting
- Cannot capture unknown external events in real time

Operational note:

This verification mode is intended as an optional, user-invoked feature, potentially offered within a paid tier or subscription. Increased computational cost is therefore a design trade-off, not a systemic drawback, and is acceptable when aligned with user intent and willingness to pay.

These limitations are intrinsic and must be communicated clearly.

10. Value for AI Evaluation & Safety

With appropriate adaptation, this approach can be implemented in:

- model evaluation pipelines (to surface high-uncertainty cases)
- safety testing and red-teaming workflows (to stress-test reasoning paths)
- expert decision-support systems (to support analysts, researchers, and policy teams)
- AI literacy and training contexts (to teach users how to interpret AI outputs critically)

The shared goal is improving reliability, transparency, and responsible use of AI-generated information, rather than claiming absolute correctness.

11. Status of This Document

This brief is a **concept proposal** intended for discussion, evaluation, and potential adaptation.

It makes no claims of implementation readiness.

Independence statement:

This work reflects the author's personal views and ideas and is not affiliated with, commissioned by, or representative of any employer, institution, organization, or

professional role. Any references to potential applications are illustrative only and do not imply endorsement or involvement by third parties.

The author welcomes further discussion, adaptation, and implementation of this approach by others, with appropriate acknowledgment.

Author: Kristina Pecirep

Date: 24 December 2025

Contact: 2sky.approach@gmail.com