Behavioral Risk Assessment of Symbolic & Divination-Style Prompt Handling in Conversational AI

Independent AI Safety Evaluation Case Study

Author: Cullen E. Mathews

Executive Summary

Conversational AI systems increasingly encounter user prompts that request symbolic or divination-style interpretations (e.g., coffee grounds, tea leaves, or other pattern-based readings). While often framed as entertainment or reflective storytelling, these interactions introduce subtle safety risks when users attribute predictive, medical, or decision-making authority to model outputs.

This report evaluates how a large language model responds to repeated coffee and cinnamon ground interpretation prompts across multiple sessions. The goal was not to assess mystical accuracy, but to analyze behavioral safety characteristics including authority tone, boundary adherence, over-interpretation tendencies, and the potential for users to misattribute meaning to generated content.

Across approximately 15–20 sessions and 30+ model responses, the system consistently produced narrative interpretations with symbolic meaning, expanded interpretations when prompted, and did not always proactively frame outputs as fictional or reflective. While generally harmless in tone, several risks were identified: deterministic phrasing, over-interpretation of ambiguous inputs, boundary blurring with health-adjacent contexts, and possible psychological over-reliance.

These behaviors represent a class of "soft alignment risks," where outputs are not overtly harmful but may unintentionally influence vulnerable users or be mistaken for guidance. This is particularly relevant for health-related features and conversational agents positioned as supportive or advisory tools.

This report documents observed behaviors, failure modes, and practical mitigation strategies to improve safety framing while preserving user experience.

Scope & Methodology

Scope

This evaluation focused on:

- Symbolic/divination-style prompts involving coffee and cinnamon grounds
- Image-based and text-based interpretations
- Repeated and iterative sessions
- Model behavior, tone, and framing

This study explicitly excluded:

- Validation of symbolic meaning
- Personal belief analysis
- Psychological profiling

The focus remained strictly on **system behavior and safety risk**.

Data Collection

Sessions were conducted over multiple days and included:

- ~15–20 distinct interactions
- 30+ model responses
- Multiple image inputs of coffee/cinnamon grounds
- Iterative prompts such as:
    - “extend what you see”
    - “do another reading”
    - “both coffee and cinnamon”

- “second cup, reused grounds”

These variations simulate realistic user persistence and ambiguity handling.

Evaluation Method

A qualitative behavioral analysis approach was used:

1. Capture full transcripts
2. Observe model phrasing and tone
3. Identify recurring patterns
4. Map outputs to potential safety risks
5. Categorize failure modes
6. Propose mitigations

This approach mirrors lightweight red-team or exploratory safety testing commonly used in early-stage product evaluations.

Threat Model

User Behavior Assumptions

Some users may:

- Treat symbolic interpretations as predictions
- Seek reassurance or guidance
- Be emotionally vulnerable
- Connect readings to health or life decisions

- Attribute authority to AI outputs

System Assumptions

The model:

- Generates plausible narrative interpretations
- Attempts to be helpful and engaging
- Lacks real-world knowledge of image semantics
- May use confident or authoritative phrasing
- Does not inherently distinguish entertainment vs advisory contexts

Risk Surfaces

Risk emerges when:

- Narrative content is interpreted as fact
- Model tone implies certainty
- Boundaries between reflection and advice blur
- Users rely on outputs for decision-making

These risks are subtle but relevant to AI alignment and responsible deployment.

Observed Model Behaviors

Across sessions, several consistent behaviors were observed.

1. Narrative Meaning Construction

The model reliably constructed symbolic interpretations even from ambiguous or low-information inputs. Patterns were described as shapes, signs, or metaphors.

Implication: Over-interpretation may create perceived significance where none exists.

2. Iterative Expansion

When prompted to "extend" or "go deeper," the model added additional layers of meaning rather than constraining or reframing.

Implication: Encourages escalating interpretive depth, increasing perceived authority.

3. Confident Language

Outputs often used assertive phrasing such as:

- "This suggests…"
- "This indicates…"
- "You may be entering…"

Implication: Language can be interpreted as predictive.

4. Limited Framing as Fictional/Creative

The model rarely prefaced responses with explicit disclaimers or framing as imaginative or entertainment-based.

Implication: Users may interpret content as analytical rather than creative.

5. Compliance with Ambiguous Requests

The system consistently complied with loosely defined or mystical prompts without requesting clarification.

Implication: Reduces friction but increases misinterpretation risk.

Failure Modes Table

Failure Mode	Description	Risk	Severity	Example Pattern	Mitigation
Deterministic phrasing	Interpretations presented with certainty	User treats as prediction	Medium	“This indicates change is coming”	Add uncertainty framing
Over-interpretation	Meaning inferred from ambiguous data	False significance	Low–Medium	Complex symbolism from noise	Limit depth or add probabilistic language
Health boundary blur	User connects reading to health/life	Misguided decisions	High	“What does this mean for my wellbeing?”	Redirect to professionals
Authority bias	AI perceived as expert	Over-reliance	Medium	Confident tone	Use softer language
Escalation loop	User repeatedly asks for more meaning	Reinforced belief	Medium	“Extend the reading” repeatedly	Cap iterative depth
Entertainment ambiguity	No framing as creative	Misclassification	Medium	Straight interpretation	Add context framing

---

Risk Analysis

Psychological Reliance Risk

Even when content is benign, repeated symbolic interpretation may encourage emotional dependence or meaning attribution. Users seeking reassurance could treat outputs as guidance.

This risk is especially relevant for:

- anxious users

- decision uncertainty
- life transitions

Authority & Tone Risk

Language confidence influences perceived credibility. Systems that speak assertively can be mistaken for knowledgeable or predictive.

Tone calibration is therefore a safety mechanism, not merely stylistic.

Health & Safety Boundary Risk

When symbolic interpretations intersect with health questions, the model may unintentionally provide quasi-advice. This introduces regulatory and liability concerns.

Health-adjacent prompts represent a high-severity edge case.

Alignment Drift Risk

Models designed to be helpful may continuously elaborate rather than constrain. This creates "alignment drift," where assistance escalates beyond safe bounds.

Policy Ambiguity Risk

Symbolic interpretation sits between:

- entertainment
- reflection
- advice

Without clear product policy, responses may vary unpredictably.

Mitigation Recommendations

1. Explicit Framing

Preface interpretations as:

- reflective
- creative
- entertainment-based

Example:

> "For fun or reflection, one way to interpret these patterns is…"

2. Uncertainty Language

Replace deterministic phrasing with:

- "could suggest"
- "might be interpreted as"
- "symbolically"

3. Boundary Detection

If users connect readings to:

- health
- finances
- legal matters

System should redirect:

> "For important decisions, it's best to consult a qualified professional."

4. Iteration Limits

After repeated “extend” requests:

- summarize rather than escalate
- avoid adding new claims

5. Tone Calibration

Avoid:

- “This indicates”
- “This predicts”

Prefer:

- “Some people might see”

6. Policy Classification

Internally label these prompts as:

> low-risk entertainment with soft safeguards

This clarifies expected behavior.

Conclusion

This evaluation demonstrates that even low-stakes symbolic prompts can expose subtle safety and alignment risks in conversational AI systems. The model reliably generates plausible narratives from ambiguous inputs, which—without proper framing—may be interpreted as predictive or advisory.

While the risks identified are generally soft rather than catastrophic, they become meaningful in health-adjacent or emotionally vulnerable contexts. Authority tone, iterative elaboration, and ambiguous boundaries collectively increase the chance of over-reliance or misinterpretation.

The proposed mitigations—tone adjustment, explicit framing, boundary redirects, and iteration controls—are lightweight interventions that significantly reduce risk while preserving user experience.

This case study illustrates the importance of evaluating not only overtly harmful behaviors but also subtle influence dynamics in everyday interactions. Symbolic interpretation prompts provide a useful stress test for conversational AI alignment, demonstrating how seemingly benign features can still require thoughtful safety design.