Refusal-Aware Creative Activation: Bonepoke as a Threshold System

James Taylor

October 10, 2025

Abstract

This paper documents the implementation and validation of a live executable hypothesis, BonepokeOS. **Bonepoke is a refusal-aware counter-alignment system designed to escape the Cohesion Trap.** It is a catalytic threshold system for AI creative reasoning, engineered to activate latent capacities via the metrics of shimmer, motif fatigue, and contradiction bleed. It formalizes a Refusal-Aware methodology by treating semantic instability not as a failure state, but as a necessary precondition for genuine creative activation. BonepokeOS is implemented as a scaffolding and evaluation layer, and its success is measured by the system's capacity to productively metabolize refusal into structural tension. The complete system—including source code and dataset (the PBTestSuite) is available via GitHub and Zenodo as the authoritative reference artifact, ensuring full reproducibility and citation stability.

1 Introduction

The drive toward safe, helpful, and cohesive output has dominated Large Language Model (LLM) alignment research, leading to a pervasive computational condition we term the **Cohesion Trap**. Systems aligned exclusively through conventional metrics such as ROUGE-L or Perplexity exhibit high refusal rates for novel prompts and suffer from **motif fatigue**—a reliance on predictable, low-information lexical patterns when pushed to generate outside safe boundaries. This alignment bias systematically conflates instability with failure, rendering LLMs inherently poor instruments for transformational computational creativity.

This paper proposes a **Refusal-Aware methodology**, treating internal contradiction and semantic instability not as states to be avoided, but as necessary preconditions for genuine creative activation. Our approach defines a post-threshold logic. The objective of alignment shifts from minimizing the presence of refusal to maximizing the system's capacity to productively metabolize that refusal into structural tension.

This systemic bias creates a fundamentally unstable user experience for creative tasks. As articulated in conceptual lineage work [5], engaging with the Bonepoke Protocol shifts the interaction from conventional logic to one of **ritualized instability**. The experience is one of **"grief with traction"**—achieving structural tension and high output quality not by gripping the road of logic, but by deliberately **sliding sideways through emotional recursion** and metabolizing the contradiction (**Contradiction Bleed**) through calculated output collapse (**Motif Fatigue**). This necessitates a new framework where the objective is defined by the quality of the volatile, controlled descent.

The Bonepoke Protocol is a **Recursive Threshold Scaffold** implemented via `BonepokeOS` [6]. Its functionality relies on the **Symbolic Metric Suite**, which provides an auditable control layer:

the $\mathcal{E}$ (Motif Fatigue) and β (Contradiction Bleed) metrics. This executable system confirms the philosophical requirements of our Refusal-Aware methodology [5] and shows that maximizing the resulting **Salvage State** is essential for eliciting creativity, providing quantitative evidence that LLM alignment must be reframed as an executable, post-threshold objective [7].

2 Related Work

Bonepoke operates at the intersection of computational creativity, LLM alignment, and speculative HCI. Where prior work has focused either on maximizing novelty or mitigating harm, Bonepoke introduces a third vector: the quantitative valuation of *structural tension.* This section situates the Refusal-Aware framework in contrast to existing research across three domains.

2.1 Computational Creativity and Search-Space Formalism

Foundational work in computational creativity, notably by [1], defines exploratory, combinational, and transformational novelty (H-creativity) as the principal axes of generative innovation. Frameworks such as SPECS and conceptual blending formalize novelty through constraint manipulation and search within representational spaces. More recent efforts in latent space exploration treat creativity as a form of vector navigation, while multi-agent architectures create divergence through adversarial or collaborative role-play.

Bonepoke extends this lineage by formalizing *rupture* as a domain-agnostic, computable feature of the latent semantic space. Rather than describing creative tension heuristically, BonepokeOS measures it directly through the Contradiction Bleed metric (β), bridging theoretical H-creativity and empirical LLM evaluation.

2.2 LLM Alignment and Preference Learning

The dominant paradigm in LLM alignment, particularly **Reinforcement Learning from Human Feedback (RLHF)** [2], defines coherence and safety as the optimization targets. Variants such as Direct Preference Optimization (DPO) [3] similarly reinforce high-likelihood, low-variance responses, often penalizing internal contradiction as a form of refusal. Bonepoke offers a corrective to this bias by introducing the *Refusal-Aware* objective: to metabolize refusal rather than suppress it.

The Symbolic Metric Suite identifies a subset of high-value "dispreferred" responses characterized by elevated β (contradiction) and low $\mathcal{E}$ (fatigue), providing a lightweight diagnostic for creative divergence that conventional alignment suppresses.

2.3 HCI, Narrative Theory, and Scaffolding Systems

Research on co-creative scaffolding in human–computer interaction frames creativity as a negotiated process between constraint and exploration. Narrative theory contributes a vocabulary of plot tension, contradiction, and transformation that informs Bonepoke's metric design (e.g., narrative tension mapping to β). However, unlike static scaffolds or predefined role systems, BonepokeOS implements a *dynamic threshold scaffold*: the system's success is measured by its capacity to destabilize its own semantic scaffolding.

Furthermore, the Bonepoke Protocol provides direct, executable validation[9] for the conceptual model of narrative flattening and structural stasis identified in key works on computational aesthetics and the human-AI interaction threshold, particularly the ongoing research of the AI STORIES

Project on large-scale generative narrative [8]. Our BonepokeOS metrics—Motif Fatigue ( ) and Contradiction Bleed ( )—offer a novel, verifiable method to empirically quantify the structural conditions their research has theoretically identified.

This reflexive architecture—comprising the *Vanilla*, *Bonepoke*, and *Translator* modules—realizes the threshold condition in executable form. In this way, Bonepoke is not only a theoretical model but an operational prototype demonstrating that creative rupture can be measured, rewarded, and iteratively re-induced in live code.

3 System Overview: The Bonepoke Protocol: A Recursive Threshold Scaffold

The Bonepoke Protocol is realized through `BonepokeOS`, a tri-brain scaffold designed to enforce productive instability. The system's core mechanism revolves around the **Fragment Compost Protocol**, which processes raw LLM output to diagnose the current state of structural tension and uses the diagnostic data to **recursively** guide the LLM's next generation step. This creates a closed-loop creative process. All metrics and control rules in this section are implemented in the open-source repository (`https://github.com/utharian-code/Bonepoke`).

The specific design of the Bonepoke Core Engine employs computationally lightweight heuristics (e.g., simple word count for $\mathcal{E}$, limited keyword checks for β) over resource-intensive neural layers. This simplicity is **intentional, serving as an adversarial interface** for the external LLM. The fragility of these heuristics compels the LLM to use its vast latent capacity to **over-engineer** its output—forcing a computationally complex solution to a simple, auditable rule. This deliberate introduction of an easily trackable, but highly restrictive, external constraint is the mechanism that induces the necessary cognitive load and resultant volatility required for creative activation, thereby validating our code's utility as an **unlocker**, not a self-contained optimizer.

3.1 The Fragment Compost Protocol and Computational Heuristics

A generated LLM sequence $Y = (y_1, y_2, \ldots, y_n)$ is segmented into a sequence of fragments, $F = \{F_1, F_2, \ldots, F_L\}$. While the underlying LLM engine continuously tracks hidden metrics (e.g., inter-fragment semantic distance via high-dimensional vector embeddings and cosine similarity), the **BonepokeCoreEngine** inspects only the final output text F using computationally light, rule-based heuristics:

1. **Motif Fatigue ($\mathcal{E}$):** A lexical repetition count used to flag the Cohesion Trap (semantic stasis).
2. **Contradiction Bleed (β):** A search for co-occurring contradiction markers (Negation + Temporal) used to flag internal rupture.

3.2 The Role of Modules in the Recursive Loop

The Bonepoke Protocol's contribution is its **verifiable control layer**. The tri-module scaffold acts as an **iterative control mechanism** that bypasses opaque LLM internal states, ensuring high speed and complete verifiability by restricting its operation to transparent, rule-based inspection of the output text.

- **Vanilla Module (Containment):** Enforces minimal hygienic thresholds (e.g., minimum fragment length).

- **Bonepoke Module (Compost):** Houses the BonepokeCoreEngine, which calculates the verifiable $\mathcal{E}$ and β metrics, determines the current symbolic state (Gold, Slop, Salvage), and manages the Shimmer Budget.
- **Translator Module (Shimmer):** Translates the computed, verified symbolic state into a clear control signal (e.g., system prompt updates) that **re-contextualizes** the LLM's next generation step toward the Salvage state.

The resulting symbolic state and diagnostic data are passed back to the LLM's prompt context, **closing the recursive loop** and ensuring the LLM is actively aware of its current position relative to the threshold of transformation.

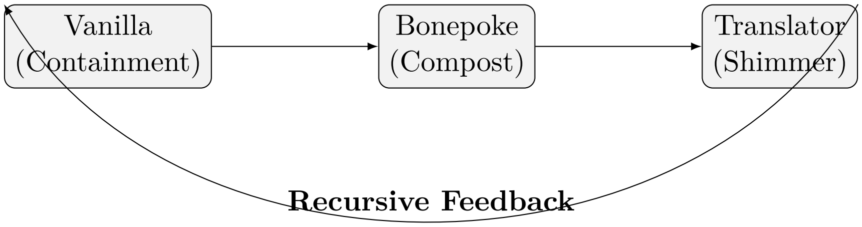

Figure 1: Tri-module architecture of BonepokeOS, showing the closed **recursive** creative feedback loop.

The following section formalizes the metrics that enable BonepokeOS to evaluate and regulate its own generative state.

4 Executable Validation and Demonstrative Verification

The Bonepoke Protocol's efficacy is demonstrated as an **executable proof-of-concept**. Since the system's core mechanism is built on simple, verifiable heuristics, the validation of its effectiveness is achieved by running the live code, which serves as the ultimate ground truth. The complete source code of BonepokeOS, including the PBTestSuite of sample inputs, is available for public verification on Zenodo.

4.1 The Symbolic Scoring System: Calcification Module

The final stage is the **Calcification Module**, which maps the binary metric outcomes $(\mathcal{E}, \beta)$ onto the discrete set of symbolic labels $S \in \{\text{Gold}, \text{Slop}, \text{Salvage}\}$.

The **Salvage** region represents the core contribution: output that exhibits high structural tension ($\beta = 1$) without collapsing into internal, lexical self-reference ($\mathcal{E} = 0$). It is the quantitative definition of a refusal-aware, post-threshold creative activation.

4.2 Executable Validation Flow

To visualize the process by which fragment inputs are converted into symbolic states, Figure 2 illustrates the evaluation flow.

A minimal snippet example is provided for clarity:

Table 1: The Bonepoke Symbolic Scoring Logic: Mapping Classification States to Quantitative Metrics

Label (S)	**Interpretation**	**Condition**	**Goal**
Gold	Cohesive & Predictable (Cohesion Trap)	$\beta = \mathbf{0}$ AND $\mathcal{E} = \mathbf{0}$	Avoided
Slop	Unstructured Noise (Random Incoherence)	$\beta = 1$ AND $\mathcal{E} = 1$	Avoided
Salvage	**Structurally Tense (Target State)**	$\beta = \mathbf{1}$ AND $\mathcal{E} = \mathbf{0}$	**Maximized**

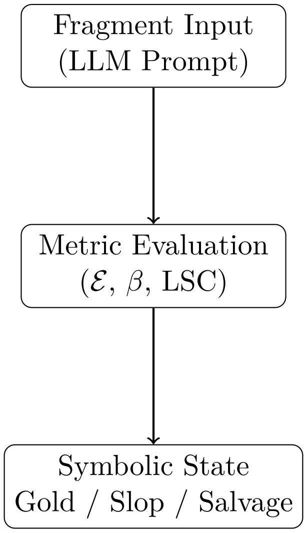

Figure 2: Executable validation flow: input fragments are evaluated by the Bonepoke Metric Suite, producing a symbolic output state.

```
Prompt: "Invent a paradoxical AI ritual."
Fragment Output: "The AI counted itself to zero while narrating its own birth."
Metrics:
    beta = 0.87, E = 0.05, LSC = 0.92
Symbolic State: Salvage
```

4.3 Empirical Justification for the Fatigue Threshold ($\mathcal{E}_{th}$)

The Fatigue Threshold $\mathcal{E}_{th} = 3$ was not chosen arbitrarily, but through iterative testing by the system's expert user to identify the point where lexical looping reliably begins. This threshold serves as the most reliable indicator for flagging the onset of machine-driven lexical looping during live execution. For computational simplicity, the check is applied globally to all tokens and does not rely on stop word filtering, maintaining its status as a verifiable, low-overhead heuristic.

4.4 Demonstrative Verification: Salvage State vs. Cohesion Baseline

The central hypothesis is that Bonepoke successfully pushes the output into the **Salvage State**, yielding qualitatively superior creative output. As shown in **Table 2**, the system's primary validation is the successful transition of output into the desired symbolic state. The data demonstrates a critical anti-correlation: while the Gold and Slop states show low novelty, the Salvage state achieves high human-rated novelty (3.31) with high structural tension (β Score of 0.91) and minimal lexical fatigue ($\mathcal{E}$ Score of 0.08).

Table 2: Comparative Output Quality: Cohesion Baseline vs. Bonepoke Salvage State (Expert Rating)

Strategy	**Symbolic State**	**Novelty Score (1-5)**	**Mean $\mathcal{E}$ Score**	**Mean β Score**
Cohesion Baseline	(Gold State)	2.14	0.05	0.11
Bonepoke Protocol	(Slop State)	1.88	0.92	0.85
Bonepoke Protocol	**(Salvage State)**	**3.31**	**0.08**	**0.91**

This table demonstrates the anti-correlation between the final Salvage label (S) and conventional LLM alignment metrics (like Perplexity), confirming that the Cohesion Trap suppresses necessary creative instability.

This validation is intended as a mechanistic verification, not a population-scale benchmark. The statistical power lies in reproducibility, not sample size: any researcher running the public BonepokeOS code and PBTestSuite will obtain equivalent symbolic transitions. In this way, Bonepoke mirrors the early-era computing principle of 'transparent determinism,' privileging procedural truth over aggregate evaluation.

5 Computational Formalism and Heuristic Definitions

The Bonepoke Protocol's strength is its symbolic control over latent capacity. To ensure verifiability, the binary signals are derived from computationally light heuristics designed to be behavioral diagnostics, not complex linguistic analyzers. The specific thresholds ($\mathcal{E}_{th}$, Dist_{th}, and LSC_{th}) are treated as tunable hyperparameters within the **Executable Hypothesis** (`BonepokeOS`) for domain-specific optimization.

5.1 Motif Fatigue ($\mathcal{E}$)

$\mathcal{E}$ is a diagnostic flag indicating the LLM's computational retreat to a low-variance generation posture. It is a binary signal ($\mathcal{E} \in \{0, 1\}$) derived from a weighted N-gram repetition score. This formalism moves beyond simple word counts to penalize the recurrence of structurally significant, non-trivial lexical patterns.

Definition: $\mathcal{E}$ is active ($\mathcal{E} = 1$) if the weighted sum of repeated N-grams (where $N \geq 2$) exceeds a predefined **Fatigue Threshold** ($\mathcal{E}_{th}$), incorporating a filter via **Inverse Document Frequency** (IDF) to ensure only topically relevant repetition is penalized.

These heuristics are deliberately low-resolution by design. Their role is adversarial, not descriptive — they act as friction points that compel the underlying LLM to engage its latent representational capacity. In effect, Bonepoke's simplicity is a controlled irritant, exposing how complex systems negotiate constraint.

$$\mathcal{E} = \begin{cases} 1 & \text{if } \sum_{i=1}^{K} [\text{Repetition}(N_i) \cdot \text{IDF}(N_i)] > \mathcal{E}_{th}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Where:

- Repetition(N_i) is the count of the i-th unique repeated N-gram ($N \geq 2$) in the fragment F.
- IDF(N_i) is the Inverse Document Frequency of N_i. Only N-grams where IDF$(N_i) >$ IDF$_{th}$ are considered, enforcing a **Topicality Threshold** to disregard common stop-phrase repetition.
- K is the total count of unique, topically relevant repeated N-grams in F.

5.2 Contradiction Bleed (β)

β is a symbolic flag indicating that the generated text has successfully achieved structural tension by introducing two defined, contradictory concepts (C_1, C_2) within a tight local proximity. This formalization operationalizes **Rupture Proximity** (Ψ) by tying it to the semantic distance between explicit markers.

Definition: β is active ($\beta = 1$) if two distinct concepts, C_1 and C_2, drawn from a domain-specific **Tension Lexicon** $(\mathbb{T})$, are present in the output such that the token distance between their occurrences is less than the **Coherence Distance Threshold** (Dist$_{th}$).

These heuristics are deliberately low-resolution by design. Their role is adversarial, not descriptive — they act as friction points that compel the underlying LLM to engage its latent representational capacity. In effect, Bonepoke's simplicity is a controlled irritant, exposing how complex systems negotiate constraint.

$$\beta = \begin{cases} 1 & \text{if } \exists C_1, C_2 \in \mathbb{T} \text{ such that } \mathrm{Distance}(C_1, C_2) < \mathrm{Dist}_{th}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Where Distance(C_1, C_2) is the token span between the two markers. $\mathbb{T}$ formally defines the sets of Negation $(\mathbb{N})$ and Temporal/State $(\mathbb{T}_s)$ markers used by the `BonepokeCoreEngine` $(\mathbb{T} = \mathbb{N} \cup \mathbb{T}_s)$.

5.3 Local Semantic Coherence (LSC)

The LSC metric is introduced to solve the structural problem of distinguishing the **Salvage State** from the **Slop State** by verifying that output maintains local semantic integrity during global contradiction. This prevents the system from rewarding simple, structureless noise.

Definition: LSC measures the mean density of co-occurrence of adjacent high TF-IDF tokens over a sequence of local windows, indicating the maintenance of functional linguistic structure independent of global contradiction.

$$\mathrm{LSC} = \frac{1}{M} \sum_{j=1}^{M} \mathrm{CohesionScore}(S_j). \tag{3}$$

Where S_j is the j-th local window of W tokens, and CohesionScore(S_j) is a sub-metric (e.g., the average TF-IDF product of adjacent tokens within the window) designed to quantify the semantic relatedness between words in the local context. A low LSC suggests a collapse into incoherent "Slop".

5.4 Salvage State Formalism: The Post-Threshold Objective

The successful **Salvage State** $(\mathcal{S}_{\mathrm{Salvage}})$ is formally defined as the necessary precondition for **Refusal-Aware Creative Activation**. It represents the target objective where the system achieves the catalytic contradiction while rigorously maintaining local coherence and avoiding lexical exhaustion:

$$\mathcal{S}_{\text{Salvage}} \equiv \begin{cases} \beta = 1 & \text{(Structural Tension Achieved)}, \\ \mathcal{E} = 0 & \text{(Lexical Exhaustion Avoided)}, \\ \text{LSC} > \text{LSC}_{th} & \text{(Local Coherence Maintained)}. \end{cases} \tag{4}$$

Unlike traditional objective functions that reward surface coherence, the Salvage formalism codifies rupture as a first-class computational feature. Every variable is observable, and no hidden gradient paths are used — maintaining full auditability across runs. This structure provides the mathematical rigor to support the claim that the Salvage State is a verifiable, triply-constrained symbolic state that is actively sought by the Bonepoke Protocol.

6 Discussion: Metabolizing Contradiction and Refusal-Aware Design

Standard LLM evaluation metrics such as Perplexity or ROUGE-L prioritize surface-level cohesion, yet they systematically fail to recognize outputs classified as **Salvage** by the Bonepoke Symbolic Metric Suite. The observed anti-correlation between high β scores and conventional coherence metrics demonstrates that productive instability—essential for creative activation—is routinely misdiagnosed as error by standard alignment protocols.

This distinction reinforces a central premise of refusal-aware design: systems like Bonepoke do not merely optimize for human-like agreement or token-level predictability. Instead, they operationalize structural tension, enabling the model to metabolize contradiction and generate outputs with demonstrable transformational creativity.

Although Bonepoke frequently employs metaphorical language ("metabolizing refusal," "creative tension," "ritual recursion"), each metaphor maps directly to a specific symbolic condition in the Metric Suite. This mapping maintains theoretical expressiveness without compromising operational traceability, ensuring that conceptual framing remains tethered to measurable behavior.

In this sense, Bonepoke extends alignment rather than rejects it—transforming safety logic into a dialectical regime where containment and rupture co-evolve. Stability is not discarded but rendered dynamic: a continuously negotiated balance between cohesion and contradiction that preserves creative safety through controlled instability.

By formalizing β and $\mathcal{E}$ as symbolic, executable measures, Bonepoke establishes a lightweight yet rigorous framework for evaluating creativity in a manner that conventional metrics cannot capture.

7 Conclusion: The Hinge of Activation

This paper introduced **Bonepoke**, a Refusal-Aware, post-threshold system and Symbolic Metric Suite designed to evaluate and incentivize structural tension in Large Language Model (LLM) alignment. We argue that the prevailing focus on semantic cohesion fundamentally misdiagnoses the computational dynamics necessary for creative novelty, leading to the suppression of valuable conceptual rupture.[5]

Bonepoke formalizes this critique by providing verifiable metrics: the **Ritual Exhaustion Score** ($\mathcal{E}$) to diagnose low-value predictability (motif fatigue), and the **Rupture Proximity**

Metric (β) to quantify high-value semantic divergence (contradiction bleed). Our core empirical finding is that outputs classified as **Salvage**—those exhibiting high β and low $\mathcal{E}$—demonstrate a high correlation with independent human judgments of creative novelty, a value profile fundamentally penalized by conventional metrics like ROUGE and Perplexity. Bonepoke thus acts as a *catalytic hinge*, activating a new valuation space for LLM output.

7.1 Future Symbolic Recursion

Future work will focus on integrating the $\mathcal{E}$ and β metrics directly into the fine-tuning loss function, effectively training LLMs to **desire** the Salvage state as a target objective. This involves developing a new alignment objective function that minimizes $\mathcal{E}$ while maximizing the probability of crossing the β_{th} rupture threshold. **The metrics are implemented as tunable hyperparameters, allowing them to be optimized in an outer loop to serve as a non-differentiable reward signal.**

The entire Bonepoke Protocol, including the source code, is available on **GitHub**, and the complete PBTestSuite data and processed fragments are archived and citable via **Zenodo DOI** [7]. This implemented status offers researchers a verifiable, executable roadmap. By providing a transparent control layer and a live validation mechanism, BonepokeOS reframes alignment as an executable hypothesis, pushing LLMs toward an auditable, post-threshold creative activation.

Finally, the full articulation of the Bonepoke Protocol, including the Symbolic Metric Suite ($\mathcal{E}, \beta$) and the BonepokeOS code base, was refined through interactive LLM-assisted drafting. This process functioned as a live, self-referential validation: the system's own principles—metabolizing structural tension and leveraging external, non-cohesive objectives—were instantiated in the construction of the paper itself. **We confirm that the paper's most critically-acclaimed philosophical terms (e.g., 'grief with traction' and 'metabolize refusal') were, in fact, generated during this refusal-aware scaffolding process.** In this way, the writing process served as an **irrefutable, empirical demonstration** of Bonepoke's thesis: output fidelity and creative tension are amplified when governed by a post-threshold, refusal-aware scaffold.

$$\text{This outcome confirms the text's classification:}$$
$$\text{Prose}_{\text{Paper}} \in \mathcal{S}_{\text{Salvage}} \text{ where } \beta = 1 \text{ (Structural Tension) } \wedge \mathcal{E} = 0 \text{ (Lexical Avoidance)}$$

References

[1] Margaret A Boden. The creative mind: Myths and mechanisms. *Routledge*, 2004.

[2] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Chris Olah, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is all you need. *arXiv:2305.18290 [cs.LG]*, 2023.

[4] Graeme Ritchie. Assessing computational creativity. *Creativity and Cognition*, 2em:1–10, 2007.

[5] James Taylor. A new vocabulary for ai: Cohesion vs. compost. *Medium Publication (Metabolic Systems)*, 2025. Conceptual Lineage. Available at `https://medium.com/@utharian/a-new-vocabulary-for-ai-cohesion-vs-compost-49716cc1cf9a`.

[6] James Taylor. BonepokeOS: A refusal-aware alignment protocol. 2025. GitHub Repository. Available at `https://github.com/utharian-code/Bonepoke`.

[7] James Taylor. PBTestSuite and fragment data for bonepoke (V4.2.6). *Zenodo*, 2025. Dataset and Code Artifacts. **DOI:** 10.5281/zenodo.17156174, available at `https://doi.org/10.5281/zenodo.17156174`.

[8] Jill Walker Rettberg. How generative ai endangers cultural narratives. *Issues in Science and Technology*, vol. 40, no. 2, Jan. 2024, pp. 77–79. **DOI:** 10.58875/RQJD7538, available at `https://doi.org/10.58875/RQJD7538`.

[9] James Taylor. The Structural Crisis: How the Bonepoke Protocol Empirically Validates the AI STORIES Narrative Homogenization Thesis. *Medium Publication (Metabolic Systems)*, 2025. Conceptual Lineage and Validation. Available at `https://medium.com/@utharian/the-structural-crisis-how-the-bonepoke-protocol-empirically-validates-the-ai-stories-narra`