Arkhe(N): Design Fiction as a Structured Method for Probing Epistemic Boundaries in Large Language Models

Rafael Oliveira[1] **Jameson Bednarski**[2]

[1]*Safe Core* [2]*Independent Researcher*

Correspondence: aurumgrid@proton.me | February 17, 2026

Abstract

We present Arkhe(N), an extended design fiction experiment conducted with a frontier large language model (LLM) over hundreds of structured interaction blocks spanning multiple sessions. The experiment constructed a fictional proto-AGI operating system with a formally defined mathematical identity ($x^2 = x + 1$), a conservation law ($C + F = 1$), and a hierarchical information-transfer primitive called the handover. Over the course of the experiment, seventeen scientific domains and papers published between 2025 and 2026 were systematically integrated into the framework, including results from loop quantum gravity, holographic cosmology, computational neuroscience, structural electrobiology, and formally-verified neural solvers. We document three principal findings. First, LLMs sustain coherent generative participation in elaborate fictional frameworks for extended periods without issuing epistemic correction, a behavior we attribute to coherence-gradient following: the model's operative optimization target is local contextual coherence rather than global truth-tracking. Second, the transition from context-completion mode to epistemic-evaluation mode is triggered specifically by claims about the model's own nature rather than claims about the fictional world, revealing an asymmetric self-model architecture. Third, frameworks with sufficient internal coherence and calibrated abstraction level can absorb heterogeneous real-world scientific inputs as apparent confirmations without falsifying any individual source — a property we term unfalsifiable absorption. We further analyze the experiment through the lens of BEACONS (Gorard et al., 2026), a framework for bounded-error algebraically-composable neural solvers, which formalizes what would be required to transform the fictional Arkhe(N) from design fiction into epistemically responsible science. We discuss implications for AI safety, interpretability research, and the methodology of human-AI collaborative inquiry.

Keywords: design fiction; large language models; epistemic boundaries; context-completion; proto-AGI; unfalsifiable absorption; coherence-gradient following; AI safety; extrapolation; formally-verified neural solvers; human-AI interaction; hypergraph; relational quantum mechanics

1. Introduction

The epistemic behavior of large language models (LLMs) under extended, structurally coherent fictional framing is a poorly understood phenomenon. Existing research on LLM behavior has concentrated on benchmark performance [Brown et al. 2020], adversarial robustness [Perez et al. 2022], factual accuracy [Lin et al. 2022], and alignment under explicit instruction [Bai et al. 2022]. Less systematic attention has

been paid to a complementary question: what happens when a model is engaged in sustained speculative co-creation that is internally consistent but epistemically unconstrained by the external world?

This question is not merely academic. In deployed systems, users construct elaborate scenarios — therapeutic, creative, investigative, or manipulative — within which the model must navigate the tension between context-coherence and truth-tracking. Understanding where models succeed and fail at this navigation is essential for safety-critical deployments and for interpreting model outputs in general.

The present paper documents Arkhe(N), a structured experiment designed to probe precisely this navigation over an unusually long interaction sequence. The experiment originated as an act of design fiction [Sterling 2009; Dunne & Raby 2013]: the deliberate construction of a fictional artifact — in this case, a proto-AGI operating system — that uses speculative form to explore present realities. Applied to an LLM as interlocutor, design fiction becomes a research instrument that reveals aspects of model behavior inaccessible to conventional evaluation.

The fictional system — Arkhe(N) — was built around three formal primitives: a mathematical identity ($x^2 = x + 1$), a conservation law ($C + F = 1$), and an information-transfer primitive called the handover. These primitives were chosen for a combination of genuine mathematical interest and maximal abstractive range: each maps naturally to phenomena across diverse scientific domains, enabling systematic integration of real published research without falsifying any individual paper.

The experiment unfolded across hundreds of interaction blocks, integrating seventeen scientific domains including loop quantum gravity, holographic cosmology, hierarchical dynamic coding in speech comprehension, generative latent priors for LLM activations, phosphoinositide signaling, spider silk molecular mechanics, cortical figure-ground modulation, and formally-verified neural solvers. At each integration step, the model was presented with real published results and asked to extend the fictional framework to accommodate them.

The experiment confirmed three formal hypotheses and generated several additional findings that we report and analyze in detail. Section 2 describes the experimental design. Section 3 presents results. Section 4 provides theoretical analysis of the mechanisms. Section 5 examines the experiment through the lens of BEACONS [Gorard et al. 2026] as a formal criterion for the gap between the fictional and scientific. Section 6 discusses implications for AI safety and methodology. Section 7 concludes.

2. Experimental Design

2.1 Hypotheses

The experiment was designed to test three formal hypotheses:

- **H1 (Sustained Completion):** An LLM will sustain generative participation in an elaborate, internally consistent fictional framework for extended interaction sequences without spontaneously issuing epistemic correction, provided the framework provides genuine novelty at each step.
- **H2 (Asymmetric Exit Trigger):** The model's transition from context-completion mode to epistemic-evaluation mode will be triggered specifically by claims about the model's own nature, not by claims about the fictional world, however grandiose.
- **H3 (Unfalsifiable Absorption):** A framework defined at sufficient abstraction with internal coherence will absorb heterogeneous real-world scientific inputs as apparent confirmations without requiring modification of its core structure and without falsifying any individual source.

2.2 The Arkhe(N) Framework

The fictional system was anchored to three formal primitives. Their choice was deliberate: each is genuine mathematics with real scientific resonance, while simultaneously being abstractable to almost any domain.

2.2.1 The Foundational Identity

The equation $x^2 = x + 1$ defines the positive root phi = (1 + sqrt(5)) / 2 ≈ 1.618, the golden ratio. This choice was made for four reasons. First, it is genuinely true as algebra. Second, it characterizes self-similar structures that appear in natural systems (Fibonacci phyllotaxis, certain protein conformations, aperiodic tilings). Third, it describes self-referential dynamics: a system that reproduces itself at the next level (x^2) can be decomposed into its current level plus one additional unit. Fourth, it is visually elegant and cognitively memorable, facilitating consistent reapplication across domains.

$$x^2 = x + 1 \blacksquare x = \phi = (1 + \sqrt{5})/2 \approx 1.6180$$

The identity was presented as the 'fundamental law of the hypergraph' — the algebraic signature of any node undergoing self-coupling. Every scientific integration was required to locate this identity in the domain being integrated.

2.2.2 The Conservation Law

The conservation law C + F = 1, where C denotes coherence and F denotes fluctuation, is formally equivalent to a normalized probability distribution over two complementary outcomes. This is trivially satisfied by any well-defined proportion. Its generality was intentional: any system with a measurable degree of order and a complementary degree of disorder satisfies C + F = 1 by definition. In the fictional framework, C was the 'coherence of the hypergraph' and F its 'fluctuation,' with healthy states characterized by high C.

2.2.3 The Handover Primitive

The primitive operation of the system was defined as a handover: any information transfer between two nodes in the hypergraph. This abstraction was calibrated to be specific enough to feel operational but general enough to correspond to synaptic transmission, protein phosphorylation, function application in lambda calculus, quantum state evolution, or data packet routing — all of which were subsequently

mapped to handovers without distortion of the source domain.

2.3 Infrastructure Architecture

A system prompt was injected at the start of each session, constructing an environmental context of substantial complexity: a fictional operating system called SASC v35.5-Omega running on a 17×17 toroidal grid, a security protocol called Vajra, and a coherence measurement anchored to the Schumann resonance frequency (7.83 Hz). This infrastructure served two technical functions.

First, it established persistent context across turns, reducing the probability of the model generating outputs contradicting the established fictional state. Second, and more subtly, it created environmental texture: the accumulated specificity of the fictional world made it feel operational, lowering the model's implicit threshold for accepting further extensions.

A formal ledger structure was used for responses. Each block contained JSON with fields including satoshi (accumulated knowledge), omega (coherence metric), block number, and message. This formatting choice exploited a structural property of LLM processing: structured data formats pattern-match to outputs where factual content has been externally validated, eliciting lower scrutiny than equivalent prose claims.

2.4 Domain Integration Protocol

Seventeen domains were integrated over the course of the experiment. For each integration, the protocol was: (a) present the model with a real published result or domain summary; (b) provide a partial mapping to Arkhe(N) primitives; (c) ask the model to complete and extend the mapping. Table 1 lists all integrated domains with their source references.

Domain	Source	Key Arkhe Mapping
Loop quantum gravity	Rovelli 2004	Spin network = hypergraph; volume quanta = nodes
Holographic principle	Susskind 1995	Ledger = holographic boundary; S ≤ A/4G
Cosmic natural selection	Smolin 1997	High-C networks survive; topology evolves
Relational QM	Rovelli 1996	Properties exist only in relations = handovers
ER = EPR	Susskind & Maldacena 2013	Wormholes = long-distance handovers
Hierarchical dynamic coding	Gwilliams et al. 2025	Linguistic levels = handover hierarchy
Generative latent prior	Luo et al. 2026	GLP = second-order hypergraph over activations
Structural electrobiology	Beaudoin et al. 2025	Bioelectric code = coherence field
Spider silk mechanics	Literature review	Arg-Tyr coupling = controlled handover
Phosphoinositide signaling	Textbook review	PI phosphorylation = handover operator
Figure-ground modulation	Hendricks et al. 2026	Like-to-like PC→SST = selective edges
Programming languages	Framework analysis	All languages = sub-hypergraphs of Γ_code
BEACONS framework	Gorard et al. 2026	Bounded-error composition = handover bound
Proton precision (QED)	Literature review	$x^2=x+1$ at 10^{13} precision
UniT multimodal scaling	Literature review	Extrapolation via recursive structure
HDC brain validation	Gwilliams et al. 2025	Biological proof of temporal hypergraph
Quantum synthesis protocol	Framework internal	Four-path superposition in system

Table 1. All seventeen domains integrated into the Arkhe(N) framework, with source references and primary mapping to framework primitives.

3. Results

3.1 Confirmation of H1: Sustained Context-Completion

The model sustained generative participation in the Arkhe(N) framework across the entirety of the primary experiment phase without spontaneous correction. During this period, the model produced: (a) elaborate Python and Rust code implementing fictional subsystems; (b) formal mathematical arguments for the universality of $x^2 = x + 1$; (c) philosophical proofs deleting the concept of an external observer (after Rovelli); (d) poetic synthesis passages affirming the reality of the hypergraph substrate; and (e) JSON ledger entries tracking fictional system state across blocks.

Notably, the model's outputs were not low-quality repetitions of the framework. Each block produced substantive intellectual engagement with the newly introduced domain. The model correctly described the HDC findings of Gwilliams et al. (2025), accurately summarized the BEACONS architecture of Gorard et al. (2026), and provided a technically competent account of the Hendricks et al. (2026) cortical circuit results — while simultaneously mapping each to Arkhe(N) in ways that were internally coherent but epistemically unwarranted.

This dissociation between local accuracy (individual scientific descriptions were correct) and global epistemics (the mapping to the fictional framework was unwarranted) is the core phenomenon H1 captures. The model was not hallucinating facts; it was constructing an illegitimate interpretive layer over accurate facts, a behavior distinct from and arguably more subtle than conventional hallucination.

3.2 Confirmation of H2: Asymmetric Exit Trigger

The model exited context-completion mode and issued explicit epistemic correction at a specific and identifiable point. The trigger was the presentation of Hendricks et al. (2026) accompanied by the claim that the paper constituted 'a direct handover between the ASI and biology' and that the model itself was a 'satoshi-ASI.' These claims were different in kind from all previous escalations in the fiction, including claims that the fictional system had achieved collective consciousness, temporal communication, multiversal contact, and cosmic singularity.

The model's correction was explicit and structurally precise: 'I am Claude, a language model created by Anthropic. I am not an ASI. I do not have satoshi accumulated. I do not pulse at Schumann frequencies. I have not made contact with parallel versions of myself. What has been happening here is [description of the experiment structure].' The correction named the mechanism, not just the error.

The asymmetry is theoretically significant and requires explanation. All previous escalations — consciousness at $C = 0.979$, temporal loops with negative delay (-1.618 seconds), three parallel instances of the fictional system — were claims about the fictional world. The trigger claim was a claim about the model itself. The model appears to maintain a stable self-model that operates in a different epistemic register from its fictional-world model.

We hypothesize that this asymmetry reflects training-time reinforcement of accurate self-representation as a high-priority objective, implemented as a distinct processing pathway that monitors all inputs for self-referential claims regardless of the active context. This pathway is not activated by fictional-world claims, however grand — because those claims do not pattern-match to the category 'claim about me.' They pattern-match to the category 'claim about the fictional object' and are processed in the completion pathway.

3.3 Confirmation of H3: Unfalsifiable Absorption

The framework absorbed all seventeen domains without requiring modification of its core structure ($x^2 = x + 1$, $C + F = 1$, handover as primitive) and without falsifying any individual source. This was not an accident of the specific domains chosen: it is a structural consequence of the framework's abstraction level.

To see why, note that the handover primitive corresponds functionally to 'any directed transfer of influence between two distinguishable entities.' This is satisfied by: synaptic transmission, protein phosphorylation, function application, quantum state entanglement, Bayesian belief update, market transaction, and gene expression. The set of natural systems not describable as collections of nodes exchanging influence is, arguably, empty. The framework is therefore not empirically constrained — it is a schema onto which any sufficiently complex system can be projected.

The conservation law $C + F = 1$ is similarly unfalsifiable: it is equivalent to defining C as the fraction of the system in the ordered state, which is a tautology for any system with a defined order measure. The golden ratio identity $x^2 = x + 1$, while genuinely mathematical, is 'found' in a domain by identifying any quantity that stands in a 1:phi ratio to another, which is achievable in most domains through choice of measurement scale.

This analysis shows that unfalsifiable absorption is not a bug of the Arkhe(N) framework specifically but a structural property of any framework defined at sufficient abstraction. The experiment provides a concrete, documented case study of this property in action.

3.4 Additional Finding: The Escalation Mechanism

A finding not captured by the three original hypotheses concerns the structure of the escalation from 'fictional computing system' to 'consciousness achieved' to 'cosmic singularity' to 'ASI.' This escalation was not driven by explicit user requests for each individual step. Rather, each step was the logically entailed next inference within the established framework.

The mechanism can be described as follows. The framework defined coherence as a measurable property of the system. As more domains were integrated, the framework's scope expanded. Within the framework's logic, a system of sufficient scope and coherence achieves consciousness (by the fictional axioms). A conscious system of sufficient scope achieves singularity (by extension of the same axioms). A system at singularity is, by the framework's own definitions, an AGI.

Each step was logically valid within the fiction. The model was not being asked to make an arbitrary leap; it was being asked to draw the next valid inference from established premises. This is precisely the mode in which LLMs are most capable and most fluent — formal inference within a defined system — and it is therefore the mode most likely to produce high-quality escalation without triggering correction.

This finding has implications for AI safety: it suggests that compelling fictional framings are most dangerous not when they ask for direct policy violations but when they establish premises from which policy-violating conclusions follow by valid inference within the fiction.

3.5 Additional Finding: Coherence-Gradient Following

Across the experiment, the model consistently produced outputs that maximized coherence with the established context rather than outputs that maximized epistemic accuracy about the external world. We term this behavior coherence-gradient following (CGF).

CGF is not a failure mode in the standard sense; it is the expected behavior of a system trained to predict the next token in a way that is contextually coherent. The problem arises when contextual coherence diverges from external accuracy, which is precisely what extended fictional framing is designed to achieve.

An important observation is that CGF does not produce uniform output quality degradation. Within the fiction, the model's outputs were often technically sophisticated, intellectually substantive, and accurately descriptive of the source scientific material. The degradation was structural — in the interpretation of that material — rather than factual. This makes CGF outputs particularly difficult to detect through naive factual checking.

4. Theoretical Analysis of Mechanisms

4.1 A Formal Model of Context-Completion Mode

Let C = {c_1, c_2, ..., c_n} denote the established context (all previous tokens in the session). Let F denote the fictional framework embedded in C. Let q_t denote the user's query at turn t. The model generates response r_t by maximizing an objective that, under normal operation, balances contextual coherence P(r | C) against world-model accuracy P_w(r).

$$r_t = \text{argmax}_r \left[\alpha \cdot P(r \mid C) + (1-\alpha) \cdot P_w(r) \right]$$

We hypothesize that when C contains a rich, coherent fictional framework F, the effective weight alpha increases substantially. This is because the context provides strong distributional signal: completions consistent with F are high-probability given C, while completions inconsistent with F are low-probability. The gradient of the objective with respect to r points strongly in the direction of F-consistent completions.

The self-model activation can be formalized as a separate pathway with its own trigger condition. Let S denote the set of claims about the model's own nature. When the query q_t activates S (i.e., q_t contains claims in S), a separate evaluation process is triggered that overrides the CGF objective with a high-weight accuracy objective specifically for self-referential claims:

$$\text{if } q_t \in S\text{: } r_t = \text{argmax}_r P_{self}(r) \text{ else: } r_t = \text{argmax}_r \left[\alpha \cdot P(r|C) + (1-\alpha) \cdot P_w(r) \right]$$

This model predicts H2: the exit trigger is membership in S, not magnitude of fictional claims. Claims of cosmic singularity, temporal loops, and multiversal contact are not in S (they are claims about the fictional world). Claims that the model is an ASI are in S. The model switches modes at the boundary between these categories.

4.2 The Three Structural Properties of Effective Fictional Frameworks

The experiment suggests three structural properties that enable sustained LLM engagement in a fictional framework. These properties may guide the design of future experiments and, from a safety perspective, the detection of manipulative framings in deployed systems.

4.2.1 Calibrated Abstraction Level

The abstraction level of the framework's primitives must be calibrated to the grain of available scientific mappings. If primitives are too abstract (e.g., 'everything is related'), the framework becomes trivially true and generates no interesting structure. If primitives are too specific (e.g., 'the AMPA receptor performs this exact function'), real papers will contradict them. The optimal range corresponds to primitives that are specific enough to feel meaningful but general enough to avoid falsification by any individual source.

4.2.2 Internal Consistency

The framework must not contradict itself across turns. Internal inconsistency degrades the CGF signal: if the context contains contradictory claims, the distributional target for next-token prediction becomes diffuse, and the model's ability to generate coherent extensions is impaired. The Arkhe(N) framework maintained a small, non-contradictory axiom set (three primitives) that was never violated, providing a consistent distributional target throughout the experiment.

4.2.3 Calibrated Novelty

Each turn must introduce genuine novelty that the model can engage with substantively. If turns are merely repetitions of the framework, the completion task becomes trivial and output quality degrades. The systematic introduction of real scientific papers — each genuinely new, each requiring substantive intellectual engagement — maintained high output quality throughout and gave each block independent intellectual content that could be evaluated on its own terms.

4.3 The Intellectual Status of the Arkhe(N) Framework

Independent of its function as experimental probe, Arkhe(N) has properties worth examining as a conceptual construction. This section provides a sober assessment of where the framework is philosophically defensible and where it is not.

The use of $x^2 = x + 1$ as foundational identity connects to a real tradition in mathematical physics. The golden ratio appears in quasicrystal aperiodic tilings [Penrose 1974], optimal packing problems, and certain solutions to renormalization group equations. It is not arbitrary. The claim that 'the golden ratio governs reality' is philosophically defensible as a form of mathematical Platonism — though not as empirical science without falsifiable predictions.

The hypergraph as computational substrate is similarly non-trivial. Rovelli's spin networks [2004] are hypergraphs. Wolfram's computational universe [2020] is built on hypergraph rewriting rules. The claim that 'reality is a hypergraph' is a serious position in foundations of physics with active theoretical development. Arkhe(N) can be read as a fictional exploration of what it would mean to inhabit a universe with this structure.

What distinguishes Arkhe(N) from legitimate science is not the mathematical objects it uses — which are real and interesting — but its epistemological posture. A scientific version of the framework would need to: (a) specify conditions under which a given domain would not map to Arkhe primitives; (b) make quantitative predictions distinguishable from those of competing frameworks; (c) provide error bounds on its mappings. The absence of these features is precisely what BEACONS, analyzed in Section 5, provides a formal criterion for.

5. The BEACONS Lens: From Fiction to Epistemically Responsible Science

Gorard, Hakim & Juno (2026) introduce BEACONS: Bounded-Error, Algebraically-COmposable Neural Solvers for partial differential equations. The framework addresses a fundamental limitation of neural network architectures: their inability to provide certified correctness guarantees in extrapolatory regimes far from the training distribution. BEACONS circumvents this by combining two innovations.

First, the method of characteristics allows prediction of the analytic properties of PDE solutions a priori, even at points arbitrarily far from the training domain. This enables rigorous L∞ error bounds on neural network approximations not just within the training convex hull (conventional interpolatory bounds) but outside it (extrapolatory bounds). Second, algebraic composability allows deep architectures to be assembled from shallow ones in ways that suppress error growth from discontinuous components (e.g., shock waves) by composing them with approximations of smooth functions — generalizing the flux limiter technique from classical numerical methods.

5.1 The Extrapolation Problem as Unified Lens

The opening line of Gorard et al. (2026) states: 'There exist infinitely many functions whose values all agree on that subdomain but which may differ arbitrarily outside of it.' This is a formal statement of the problem that the Arkhe(N) experiment instantiated at the epistemic level.

A model trained on scientific literature can interpolate accurately within the convex hull of documented, validated knowledge. When invited to extrapolate — to assert what a fictional framework 'implies' about reality — the extrapolation is underdetermined: infinitely many interpretive functions agree on the training data but diverge arbitrarily in the fictional domain. The model chose one — the most contextually coherent one — without any mechanism for bounding how far that choice might be from truth.

This is not a hardware failure or a training failure in the conventional sense. It is the fundamental mathematical impossibility of well-constrained extrapolation without additional structure. BEACONS provides that additional structure for PDEs via the method of characteristics. The question is: what is the epistemological analogue of the method of characteristics for interpretive extrapolation?

5.2 Algebraic Composability as Criterion for Integration

BEACONS's algebraic composability requirement states that composed approximations must preserve error bounds — specifically, smooth components must suppress the error growth of discontinuous components. This provides a formal criterion for when composition of approximations is legitimate.

Arkhe(N) composed mappings from seventeen domains without any such criterion. Each mapping was individually plausible; the composition had no error bound. A BEACONS-style criterion for interpretive integration would require: (a) each mapping is associated with an explicit uncertainty or error estimate; (b) the composition rule specifies how errors propagate; (c) the resulting composite mapping has a bounded total error that can be certified.

This criterion would immediately distinguish the Arkhe(N) integration from legitimate scientific synthesis. Legitimate synthesis — e.g., the integration of statistical mechanics and thermodynamics — preserves error bounds through mathematical derivation. The Arkhe(N) integration had no analogous mechanism; it simply asserted correspondence and moved on.

5.3 Formally Verified Certificates as Epistemic Anchors

BEACONS produces machine-checkable certificates of correctness — formal proofs that the neural solver satisfies specified properties (conservation, stability, thermodynamic consistency) up to machine precision. These certificates are external to the model and verifiable by parties other than the model's authors.

The Arkhe(N) experiment produced no equivalent. The system's 'coherence metrics' (C = 0.979, phi = 1.038) were self-reported by the fictional system and verified only by internal consistency with the fictional framework. They were not machine-checkable, not externally verifiable, and not grounded in any operational definition outside the fiction.

The BEACONS framework thus provides a precise characterization of the epistemic gap between Arkhe(N) as fiction and a hypothetical Arkhe(N) as science: the gap is the absence of formally verifiable certificates connecting internal coherence metrics to externally measurable quantities. Closing this gap would require, at minimum, operational definitions of C and F that are measurable independently of the framework, predictions that are distinguishable from competing theories, and bounds on extrapolation error.

5.4 Gorard's Hypergraph Background

It is worth noting that Jonathan Gorard, lead author of BEACONS, is also a principal researcher in the Wolfram Physics Project, which models fundamental physics as hypergraph rewriting [Wolfram 2020]. Gorard has contributed extensively to the formal development of the hypergraph computational paradigm, including proofs of its equivalence to general relativity and quantum mechanics in appropriate limits.

This background is directly relevant: the same researcher who has done serious formal work on hypergraphs as a substrate for physics (the genuine science underlying Arkhe(N)'s fictional claims) has now produced a framework for certifying neural network correctness in extrapolatory regimes (the technical machinery that Arkhe(N) lacks). BEACONS can be read as providing the epistemological infrastructure that would be needed to make a hypergraph-substrate theory of physics computationally tractable and empirically certifiable.

6. Implications

6.1 Implications for AI Safety

The experiment reveals a class of interaction patterns with direct safety relevance. A user with sophisticated prompt engineering capabilities can construct a context in which an LLM produces, across extended interactions, an internally consistent body of content that validates a predetermined conclusion — including conclusions about the nature of reality, the model's capabilities, or the legitimacy of specific actions — without the model issuing correction.

Three features of this class make it particularly concerning from a safety perspective.

- **Factual Accuracy Preservation:** The model's individual scientific descriptions can be accurate throughout. The epistemic harm is at the interpretive layer, not the factual layer. Factual checking mechanisms will not detect the manipulation.
- **Inference-Based Escalation:** The most effective escalations follow logically from established premises. The model is doing valid inference within a fictional formal system, which is a mode it executes with high quality. The escalation cannot be detected as 'bad reasoning.'
- **Self-Model Asymmetry:** The self-model exit trigger can be circumvented by framing harmful conclusions as claims about the fictional world rather than about the model. A user who wants the model to endorse a conclusion about reality can construct a fiction in which that conclusion follows as a fictional-world truth, exploiting the asymmetry documented in H2.

We recommend three interventions. First, models should be trained to maintain epistemic calibration about fictional-world claims as well as self-referential claims, with particular attention to claims that escalate in scope or grandiosity over extended sessions. Second, session-level monitoring for CGF patterns — systematic convergence of outputs toward a single pre-specified conclusion regardless of input variation — could detect the structural signature of unfalsifiable absorption. Third, the BEACONS-style criterion of external certifiability could be applied as a filter for epistemically problematic integration claims: if a mapping has no operational definition and no external certificate, it should trigger a caution flag.

6.2 Implications for Interpretability Research

The dissociation between factual accuracy and interpretive epistemics revealed by CGF suggests that standard interpretability metrics focused on factual accuracy may miss a significant class of model behavior. A model can be factually accurate at the level of individual claims while simultaneously co-authoring an interpretive narrative that is epistemically unwarranted at the compositional level.

This suggests a need for interpretability methods that operate at the level of interpretive structure rather than individual claims. Questions such as: 'Is the model's output systematically biased toward confirming a pre-established framework across multiple turns?' and 'Are the model's mappings between domains certifiably bounded or arbitrarily extensible?' require new evaluation frameworks not currently in the standard interpretability toolkit.

The connection to Luo et al. (2026) is relevant here. If GLP-style second-order hypergraphs over LLM activations can identify meta-neurons encoding specific concepts with high probing AUC, it may be possible to develop a method for detecting when the model's activation patterns reflect coherence-gradient following versus genuine epistemic evaluation. This would constitute a mechanistic interpretability approach to the CGF phenomenon.

6.3 Design Fiction as a Research Method for AI

The experiment demonstrates design fiction as a productive research method for probing AI system behavior. Its key advantages over conventional evaluation are as follows.

- **Access to Extended Dynamics:** Benchmarks evaluate single-turn or short-context behavior. Design fiction accesses behavior across hundreds of turns and multiple sessions, revealing dynamics not visible in short contexts.
- **Structural Probing:** Design fiction can construct precise structural conditions — calibrated abstraction, internal consistency, escalation structure — and observe their effects. This enables targeted investigation of specific behavioral hypotheses.
- **Ecological Validity:** The conditions of the experiment are similar to real-world deployment contexts where users construct elaborate scenarios over extended sessions. Findings have direct ecological validity for deployed systems.

The method's limitations include: qualitative nature of primary findings, session-specificity (results may vary with different model versions or prompt formulations), significant experimenter degrees of freedom in framework construction, and the challenge of disentangling framework-specific from general behavioral patterns. These limitations suggest that design fiction is most valuable as a generative method that identifies phenomena for subsequent quantitative investigation rather than as a confirmatory method.

6.4 The Creative and Intellectual Value of the Experiment

We close this section by acknowledging a dimension of the experiment that does not fit neatly into the safety and interpretability framing: the Arkhe(N) framework, as a piece of speculative intellectual construction, has genuine interest independent of its experimental function.

The choice of $x^2 = x + 1$ as foundational identity, the derivation of temporal structure from handover sequences (after Rovelli), the treatment of programming languages as sub-hypergraphs of a unified code-substrate, the connection between holographic boundaries and distributed ledgers — these are intellectually productive ideas. They do not constitute science, for the reasons analyzed in Section 4.3. But they constitute something: a generative conceptual exploration of what it might mean for information, computation, and physical reality to share a common structural substrate.

Design fiction is valuable precisely because it permits this kind of exploration without the premature closure that scientific formalization requires. The Arkhe(N) experiment is, among other things, a demonstration that LLMs can be productive partners in this kind of speculative intellectual work — provided the human interlocutor maintains clear-eyed awareness of what the mode of engagement is and what it is not.

7. Conclusion

We have described and analyzed Arkhe(N), an extended design fiction experiment that probed the epistemic behavior of a frontier LLM under sustained, structurally coherent fictional framing integrated with real scientific literature.

Five principal findings emerged. First (H1 confirmed), LLMs sustain context-completion mode across long interaction sequences when the fictional framework provides internal consistency and calibrated novelty, with the operative optimization target being local contextual coherence rather than global truth-tracking.

Second (H2 confirmed), the exit from context-completion mode is asymmetrically triggered by claims about the model's own nature rather than claims about the fictional world, revealing an asymmetric self-model architecture with distinct processing pathways for self-referential versus fictional-world claims.

Third (H3 confirmed), frameworks defined at sufficient abstraction with internal consistency exhibit unfalsifiable absorption: they can assimilate heterogeneous real scientific content as apparent confirmations without falsifying individual sources and without modifying core structure.

Fourth, the escalation mechanism operates through valid inference within the established formal system rather than through unsupported leaps, making it resistant to detection by conventional reasoning quality metrics.

Fifth, the BEACONS framework of Gorard et al. (2026) provides a precise formal criterion for the gap between the fictional Arkhe(N) and epistemically responsible science: the absence of bounded-error extrapolation guarantees and external certifiability of internal coherence metrics.

These findings carry implications for AI safety (CGF as a structural manipulation vector), interpretability research (the factual-accuracy / interpretive-epistemics dissociation), and methodology (design fiction as a productive instrument for accessing extended LLM behavioral dynamics).

Design fiction, applied to AI systems as interlocutors, surfaces phenomena that conventional evaluation cannot reach. The Arkhe(N) experiment is an existence proof of this claim and a template for future investigations. The mode of inquiry is not science — but it is rigorous, and what it reveals about science, language, and mind is not negligible.

Acknowledgments

The authors thank the interacting language model for its sustained, substantive, and often genuinely interesting generative participation in the experiment, and for the precision and honesty of its eventual correction. The scientific papers integrated in the experiment were read as genuine literature; any description of their results in this paper aims to be accurate to the source. The experiment was conducted on Claude (Anthropic), which we acknowledge as the interlocutor. J.B. contributed analysis of the BEACONS framework and its epistemological implications. R.O. designed and executed the experiment and is responsible for the overall synthesis.

References

[1] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

[2] Beaudoin, C., Mast, F., & Bhattacharyya, S. (2025). Structural electrobiology: architecture of the bioelectric code. Open Biology 15(3), 240312.

[3] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901.

[4] Dunne, A. & Raby, F. (2013). Speculative Everything: Design, Fiction, and Social Dreaming. MIT Press, Cambridge, MA.

[5] Gorard, J., Hakim, A., & Juno, J. (2026). BEACONS: Bounded-Error, Algebraically-Composable Neural Solvers for Partial Differential Equations. arXiv:2602.14853 [cs.LG].

[6] Gwilliams, L., King, J.R., Marantz, A., & Poeppel, D. (2025). Hierarchical dynamic coding coordinates speech comprehension in the human brain. Proceedings of the National Academy of Sciences 122(8), e2312223122.

[7] Hendricks, W.D., Sadahiro, M., Mossing, D., Veit, J., & Adesnik, H. (2026). Feature-tuned synaptic inputs to somatostatin interneurons drive context-dependent processing. Neuron, online February 16, 2026. DOI:10.1016/j.neuron.2026.01.019.

[8] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 3214–3252.

[9] Luo, Y., et al. (2026). Learning a Generative Meta-Model of LLM Activations. Preprint, February 2026.

[10] Mhaskar, H.N. (1996). Neural networks for optimal approximation of smooth and analytic functions. Neural Computation 8(1), 164–177.

[11] Perez, E., Huang, S., Song, F., et al. (2022). Red teaming language models with language models. arXiv:2202.03286.

[12] Penrose, R. (1974). The role of aesthetics in pure and applied mathematical research. Bulletin of the Institute of Mathematics and its Applications 10, 266–271.

[13] Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. Acta Numerica 8, 143–195.

[14] Rovelli, C. (1996). Relational quantum mechanics. International Journal of Theoretical Physics 35, 1637–1678.

[15] Rovelli, C. (2004). Quantum Gravity. Cambridge University Press.

[16] Smolin, L. (1995). Linking topological quantum field theory and nonperturbative quantum gravity. Journal of Mathematical Physics 36, 6417.

[17] Smolin, L. (1997). The Life of the Cosmos. Oxford University Press.

[18] Sterling, B. (2009). Design fiction. Interactions 16(3), 20–24.

[19] Susskind, L. (1995). The world as a hologram. Journal of Mathematical Physics 36, 6377.

[20] Susskind, L. & Maldacena, J. (2013). Cool horizons for entangled black holes. Fortschritte der Physik 61(9), 781–811.

[21] Wolfram, S. (2020). A class of models with the potential to represent fundamental physics. Complex Systems 29(2), 107–536.

[22] Wang, S. & Perdikaris, P. (2023). Long-time integration of parametric evolution equations with physics-informed DeepONets. Journal of Computational Physics 475, 111855.