

# At the Edge of the Known: AI, Contingency, and the Dynamics of Discovery

Wanhong HUANG  
huangwanhong@serendip.ngo

## Abstract

Generative artificial intelligence can produce large amounts of novel, surprising, and statistically unusual output, yet novelty alone does not establish scientific discovery or the reorganization of knowledge. This paper develops a preliminary dynamical account of that distinction. It treats established conceptual, explanatory, and problem-forming structures as possible *knowledge attractors*: recurrent regions toward which inference can return even after local variation. On this view, a contingent perturbation becomes scientifically consequential only when it is retained, coupled to existing relations, and capable of altering later search, representation, or explanation. The paper therefore distinguishes stochastic variation from perturbation capture, local divergence from attractor transition, and unusual combinations from changes in the structure of inquiry itself. It further examines the possibility that current artificial systems often damp anomalies through rapid classification, plausible completion, retrieval of established explanations, or premature closure. A near-critical or metastable perspective is introduced cautiously to describe regimes in which restoring constraints may weaken enough for small perturbations to propagate while coherence is still preserved. Within that setting, *epistemic turbulence* is developed as a controlled analogy for sustained, nonlinear, historically retained, cross-scale interaction in which perturbations can modify the pathways through which later perturbations propagate. The paper does not claim that discovery requires literal physical turbulence, nor that artificial systems are incapable of theory formation. Instead, it asks under what conditions generative systems can move from producing novelty within an existing knowledge landscape to reorganizing the relations that define what can be asked, noticed, and explained. The resulting framework motivates empirical tests of anomaly retention, delayed closure, relational amplification, attractor recovery, and generator reconfiguration in human and artificial research systems.

**Keywords:** generative artificial intelligence; scientific discovery; knowledge attractors; contingency; anomaly; novelty; metastability; criticality; epistemic turbulence; perturbation capture; generator reconfiguration

## Discussion Paper Note

This paper examines a narrower problem than a general theory of creativity. Its central question is how a generative system can move from producing novelty *within* an established knowledge organization to reorganizing the relations that determine what counts as a problem, an anomaly, a relevant variable, an admissible explanation, or a promising direction of inquiry. The paper therefore treats scientific discovery as a possible transformation of a knowledge-generating regime rather than as the mere appearance of an unusual output. Artificial intelligence is used as the principal comparative case because contemporary generative systems make the distinction between novelty and

discovery unusually visible: they can generate abundant variation, yet the dynamical conditions under which such variation becomes historically consequential remain an open empirical question.

The paper does not begin from the claim that current AI lacks creativity, cannot form theories, or is intrinsically confined to existing human knowledge. It also does not treat human scientific practice as a privileged benchmark whose mechanisms are already understood. The comparative problem is framed in terms of generative dynamics. A system may produce outputs that are statistically rare, semantically surprising, or judged original while nevertheless remaining within an effectively unchanged space of problem representations and explanatory relations. Conversely, a comparatively small revision may become scientifically consequential if it changes which observations are noticed, which variables are coupled, how evidence is interpreted, or which questions become thinkable thereafter.

A central distinction is therefore maintained throughout the paper: a novel output is not by itself a scientific discovery, and neither category should be treated as equivalent to knowledge reorganization. Novelty is an output-level property relative to a comparison set. Discovery is a historically situated achievement that additionally depends on epistemic evaluation, evidential relation, and subsequent uptake. Knowledge reorganization refers to a dynamical change in the structures through which later inquiry proceeds. These categories can overlap, but the paper does not infer one from another. In particular, a novel idea can remain dynamically local, while a modest conceptual revision can alter later research trajectories without initially appearing highly original.

The language of *knowledge attractors* is introduced to describe relative stability in conceptual and inferential organization. An attractor need not be a literal fixed point. Depending on the empirical system, it may correspond to recurrent explanatory patterns, stable problem framings, repeatedly retrieved conceptual neighborhoods, dominant variable sets, or families of trajectories toward which inference returns after perturbation. The attractor vocabulary is useful only insofar as it makes recovery, persistence, basin geometry, and transition empirically discussable. It should not be used as a decorative synonym for paradigm, convention, or bias.

This distinction is especially important for artificial systems. Large language models are trained on historical corpora whose statistical and conceptual structure strongly influences later generation. Retrieval systems, memory stores, tools, system prompts, and orchestration policies can further stabilize or destabilize effective inference trajectories. A system may therefore exhibit both strong generative flexibility and strong recovery toward high-probability explanatory regions. The paper asks whether this recovery can be described as an attractor-like property under matched experimental conditions, and if so, which interventions weaken or restructure it.

The paper next distinguishes *contingency* from randomness and from surprise. Random sampling can produce differences that disappear immediately. Surprise is observer- or model-relative and depends on expectation. A contingent perturbation, in the present paper, is a difference whose downstream consequence is not fixed by the effective trajectory represented at the chosen level of analysis and whose significance depends on the configuration into which it enters. The relevant scientific case is the anomaly, but anomaly is treated as relational rather than intrinsic. The same observation can be ordinary under one representation and destabilizing under another.

This leads to the concept of *perturbation capture*. A perturbation is captured when it is retained strongly enough to influence later inquiry rather than being locally absorbed, forgotten, or normalized. Capture can occur immediately through explicit anomaly recognition, but the paper should also preserve the possibility of delayed capture. A weak trace may survive in memory, an archive, a notebook, a retrieval index, a changed

association, or an unresolved question before later relations make its importance visible. The question is therefore not only whether an artificial system can detect novelty, but whether it can preserve information whose importance is not yet representable.

The distinction between perturbation and capture motivates a specific account of *epistemic damping*. A system can encounter an anomaly and yet rapidly neutralize its transformative potential by mapping it onto an available explanation, classification, or familiar causal schema. Such damping is not automatically defective. Scientific practice requires enormous amounts of normalization, compression, and rejection of irrelevant variation. The problem arises when closure occurs before the system has adequately tested whether the existing representation itself should change. The paper therefore treats *premature closure* as a candidate mechanism, not as a generic property of AI. It should be operationalized through observable return to established representations under controlled perturbations.

This part of the argument must remain balanced. High closure can protect coherence and reduce false leads; low closure can produce unstructured speculation. The relevant comparison is therefore not between closure and openness in the abstract, but between regimes of restoration and susceptibility. A system should be able to absorb most irrelevant perturbations while remaining capable of preserving and amplifying some consequential ones. The paper consequently avoids recommending maximal novelty, maximal entropy, or maximal instability.

The phrase *the edge of the known* operates at two levels. Philosophically, it refers to the boundary at which current concepts and explanations cease to organize new observations adequately. Dynamically, it refers more cautiously to regions in which the restoring pull of an incumbent knowledge organization may weaken enough for small perturbations to produce persistent consequences. The paper does not identify these two senses as equivalent, but uses their interaction to motivate the analysis of metastability and near-critical susceptibility.

The terms *criticality*, *near-critical*, and *metastability* must be used conservatively. The paper does not claim that scientific discovery occurs at a unique physical critical point, that language-model inference exhibits a universal phase transition corresponding to creativity, or that increasing temperature moves a model toward epistemic criticality. A more defensible hypothesis is that discovery-oriented systems may sometimes operate in regimes of *structured susceptibility*: perturbations acquire comparatively high gain while enough coherence, memory, and constraint remain for their effects to accumulate into usable structure.

This produces a crucial distinction between stochasticity and regime position. Increasing decoding temperature changes output variability, but it does not by itself establish a change in basin geometry, restoring forces, relational coupling, memory, or future susceptibility. The paper should repeatedly guard against the inference that increasing randomness should increase discovery. A system can be highly stochastic while remaining statistically and conceptually attracted to familiar regions. Conversely, a low-amplitude perturbation can become consequential if it enters a sensitive relation near a transition boundary.

The paper introduces *epistemic turbulence* only after these distinctions have been established. It should not appear as the initial explanation of AI creativity or discovery. The turbulence model belongs in the main text as a stronger, later-stage hypothesis concerning sustained nonlinear interaction. It is reserved for cases in which perturbations do more than branch a trajectory: they interact across scales, are historically retained, alter one another's propagation, and modify the pathways through which subsequent perturbations travel.

The controlled analogy concerns a sequence of increasingly demanding processes. A

contingent perturbation may first be captured, then amplified through relations, and in stronger cases contribute to cross-scale cascades that modify later propagation pathways and possibly the stability of an incumbent attractor. The term *turbulence* should be withdrawn whenever simpler mechanisms suffice. Random variation, ordinary non-linear amplification, metastable switching, iterative memory accumulation, chaotic divergence, or one-off cascades do not by themselves establish epistemic turbulence. The analogy is useful only if empirical systems exhibit recurrent multiscale interaction, intermittency, history dependence, and endogenous modification of the effective propagation medium.

No conservation-law analogue is assumed. Knowledge systems are not fluids, and the paper does not posit an epistemic equivalent of kinetic-energy cascades. Scale in this context may refer to token or observation, concept, relation, hypothesis, theory, research programme, or institutional field, depending on the system under study. Cross-scale propagation can be bidirectional. A local anomaly can affect a theory, while a changed theory can alter the interpretation of local observations. The turbulence analogy is therefore structural and heuristic rather than equation-level identification.

The strongest dynamical claim considered by the paper is *generator reconfiguration*. A system has undergone more than local divergence when prior perturbations alter the effective mapping from future inputs to future outputs. For artificial systems, the underlying model weights need not change. Effective generator change can be mediated through persistent memory, retrieval policies, tool use, self-maintained representations, orchestration rules, or changed problem decomposition. For humans and institutions, analogous change can be distributed across attention, language, archives, instruments, routines, collaborators, and conceptual repertoires.

A common-future-input test provides one useful operational criterion. Two systems or trajectories receive different histories and are then exposed to matched future probes. If the later response distributions remain systematically different after relevant hidden-state and memory controls, there is stronger evidence that the histories changed the effective generator rather than merely placing the systems in different transient states. The same logic can be used to study changes in anomaly sensitivity, retrieval, hypothesis generation, and problem framing.

The paper must separate dynamical transformation from epistemic success. Escaping an attractor does not imply truth. A false theory, conspiracy framework, hallucinated causal model, or unproductive research programme can also be stable, novel, and historically consequential. Discovery therefore requires additional epistemic evaluation that lies partly outside the dynamical framework. The paper's contribution is to clarify conditions of reorganization, not to supply a complete theory of justification, evidence, or truth.

The artificial-intelligence comparison should therefore be formulated as a series of empirical questions rather than a deficit list. Does the system preserve unresolved anomalies across time? Does an early perturbation alter what later inputs become salient? Can weakly relevant observations survive long enough to be reactivated? Do heterogeneous encounters widen the range of representations used in later inquiry? Can the system deliberately reduce closure when existing explanations fail? Can it subsequently consolidate a new representation rather than remain indefinitely unstable? Do perturbations change only current answers, or the future response surface itself?

Current AI systems may show mixed answers because the relevant unit is rarely an isolated model call. The effective research system can include a foundation model, system prompt, long context, vector memory, external databases, scientific instruments, code execution, retrieval, other agents, human collaborators, and organizational policies. Comparisons between humans and AI must therefore use matched system bound-

aries. It is misleading to compare a socially and materially embedded human scientist with a stateless single-turn model and infer an intrinsic biological difference from the result.

The paper should nevertheless retain the possibility that current model architectures or training objectives create systematic forms of attractor recovery. Next-token prediction over historically accumulated corpora, preference optimization for coherent and useful completion, retrieval of established knowledge, and benchmark-oriented reasoning may all favor plausible continuation over prolonged unresolved anomaly. These are candidate mechanisms, not settled explanations. They should be assessed against alternative accounts, including lack of memory, inadequate task design, insufficient tool access, evaluation artifacts, or weak experimental scaffolding.

The proposed experimental programme should therefore compare multiple conditions rather than one creativity benchmark. A minimal hierarchy includes baseline inference, increased stochasticity, heterogeneous perturbation exposure, persistent anomaly memory, relational revision, delayed-closure prompting, and adaptive regulation of exploration versus consolidation. The outcome measures should include not only originality ratings but also recovery time, persistence of altered problem representations, sensitivity to later perturbations, relational cascade breadth, cross-domain transfer, and common-future-input divergence.

The paper may introduce a *critical-zone prompting* idea only as a future experimental implication, not as an already established result. Certain prompts or orchestration policies might reduce premature closure, preserve incompatible explanations, require explicit revision of relational structure, or maintain unresolved anomalies across multiple rounds. Such interventions could move the effective inference process toward a higher-susceptibility regime, but this must remain a testable hypothesis. Prompting is not assumed to create literal criticality or turbulence.

The organization of the paper should preserve a gradual increase in theoretical strength. Section 1 introduces the novelty–discovery puzzle and the scope of the paper. Section 2 distinguishes novelty from reorganization of the known. Section 3 develops knowledge attractors and recovery. Section 4 introduces contingency, anomaly, and perturbation capture. Section 5 examines epistemic damping and premature closure. Section 6 develops the edge-of-the-known problem through metastability, susceptibility, and near-critical regimes. Section 7 introduces epistemic turbulence as a controlled, stronger analogy only after the preceding mechanisms are established. Section 8 asks why artificial systems may return to established knowledge organizations. Section 9 develops conditions for discovery-oriented artificial systems and an experimental programme. Section 10 discusses interpretation, alternative explanations, and epistemic limits. Section 11 concludes without turning the framework into a definition of discovery.

Several writing constraints should remain explicit throughout drafting. Section and subsection titles should state structural roles rather than pose questions or announce claims. The paper should avoid formulations such as “AI cannot create theories,” “LLMs are trapped in attractors,” “discovery requires turbulence,” “criticality causes creativity,” or “humans escape the known whereas AI repeats it.” Stronger causal language should be reserved for experimentally demonstrated results. Phrases such as *may*, *can be modeled as*, *candidate mechanism*, *provisional hypothesis*, and *under specified conditions* are appropriate where the evidence is indirect.

The distinction between metaphor and model should also be visible in the mathematics. Attractor, basin, susceptibility, and transition language can be formalized when a state space and response measure are operationally defined. Turbulence-like language should remain one level more cautious until empirical signatures justify stronger mod-

eling. The paper should prefer phase-space, phase-diagram, recovery, and perturbation-response representations over decorative complexity language. Any toy simulations must be labeled as illustrative and should not be presented as evidence about brains or language models.

The paper's most defensible contribution is therefore not a claim that AI discovery is impossible under current architectures. It is a proposed way of decomposing the problem. Novelty can be generated without structural change; anomalies can be encountered without being retained; memory can preserve history while reinforcing existing attractors; instability can increase variation while destroying coherence; and genuine reorganization can occur only when perturbations become capable of changing the conditions under which later inquiry proceeds. The paper asks whether these distinctions can make the boundary between generative variation and scientific discovery more empirically tractable.

The title *At the Edge of the Known* should retain its deliberate ambiguity. It refers both to the epistemic frontier at which available explanations become insufficient and to a possible dynamical region in which the incumbent organization becomes susceptible to reconfiguration. The paper should not collapse these two meanings into one. Their productive tension is the organizing device: scientific discovery may sometimes require approaching the limits of an existing representation, but the dynamical conditions under which that limit becomes generative rather than merely confusing remain to be investigated.

## Responsible Use and Rights Reservation

This section separates requested scholarly conduct from the legal permissions stated on the following page. It records an ethical request for responsible use and then defines the narrower scope of retained legal rights.

The author encourages good-faith discussion, criticism, independent inquiry, and responsible use of the material in this work. Separately from the licence's terms, the author asks users to consider foreseeable harms when adapting or applying the proposed framework. This ethical request leaves the licence's permissions and legally authorized uses unchanged.

The author retains the rights preserved under CC BY-NC 4.0 and may pursue remedies to which the author is legally entitled for breach of the licence or violation of the author's independently applicable rights. Reuse remains independent from authorial endorsement. Third-party rights require authorization from their respective holders where applicable. Copyright exceptions and limitations, including applicable forms of fair use or fair dealing, remain fully available.

## Notices

This page consolidates the manuscript's publication status, licence, and development disclosure.

**Status.** This working draft develops a preliminary theoretical hypothesis and is circulated for discussion. The terminology of perturbation capture, epistemic turbulence, criticality regulation, generative feedback, and attractor transition remains subject to conceptual and empirical refinement. Systematic literature review, formal analysis,

simulation, longitudinal experimentation, and comparative human–AI testing remain necessary before stronger causal or general claims can be supported.

**Licence.** Except where otherwise indicated, copyright 2026 Wanhong Huang. This work is made available under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). Subject to its terms, the licence permits sharing and adaptation for noncommercial purposes with appropriate attribution, a link to the licence, an indication of changes, and attribution that preserves the licensor’s independence from the reuse. Reuse is governed solely by that licence; the responsible-use request on the preceding page remains separate from its terms. The licence deed and legal-code link are available at <https://creativecommons.org/licenses/by-nc/4.0/>. The licence governs in case of conflict with this summary. Third-party material remains subject to the rights held by its respective rights holders.

**Statement on the use of language models.** The exploratory discussions and preparation of this paper involved OpenAI’s ChatGPT. ChatGPT supported exploratory dialogue, formal reconstruction, source discovery followed by website verification, argumentative criticism, and drafting in  $\text{\LaTeX}$ . The author selected the research questions, directed and approved the theoretical commitments and epistemic status of the claims, and bears sole responsibility for the manuscript, including its definitions, formal constructions, arguments, conclusions, and errors. Authorship credit remains with the human author.

**Illustrative simulations.** No empirical simulation results are reported in the present initial draft. Any later schematic or toy dynamical figures will be identified explicitly as illustrations rather than measurements of human cognition or artificial systems.

## 1 Introduction

The purpose of this section is to locate the paper’s problem before introducing its dynamical vocabulary. The immediate question is not whether contemporary artificial intelligence can produce outputs that appear creative. It is whether the capacity to produce novelty is sufficient to explain the kinds of change ordinarily associated with scientific discovery. The distinction matters because recent evidence does not support a simple picture in which artificial systems either possess or lack creativity. In controlled research-idea generation, large language model systems have produced proposals judged by expert reviewers to be more novel than proposals written by human researchers, although the same systems were rated somewhat lower on feasibility and exhibited limited diversity across generated ideas [15]. Large-scale divergent-thinking comparisons likewise show substantial overlap between human and machine performance, while also reporting differences in variability, upper-tail performance, and the effects of prompting and sampling interventions [17]. These results make a categorical “AI versus human creativity” framing increasingly uninformative. They instead motivate a narrower problem concerning the relation between novel generation and the reorganization of inquiry.

A second line of work makes the distinction more visible. In a computer-supported molecular-genetics discovery task, Ding and Li report that a generative system could support incremental discovery under known representational conditions but failed in the experimental setting to reproduce several features associated with the original human

discovery. In particular, the system did not use anomalous results to construct a substantially new hypothesis space in the way required by the target discovery [5]. In a very different domain, autonomous language–image generation loops have been observed to converge toward a small set of generic motifs across diverse initial prompts and sampling temperatures, a pattern the authors interpret using attractor-like language [9]. Neither result establishes a general limit on artificial discovery. The tasks, architectures, evaluation criteria, and system boundaries differ substantially. Taken together with positive results on research ideation, however, they sharpen a conceptual puzzle: a system may generate considerable variation and even outputs judged novel while its longer-run trajectories remain strongly organized by already available representations.

This paper develops a preliminary dynamical vocabulary for that puzzle. It treats the known not as a fixed list of propositions but as a relatively stable organization of relations among observations, concepts, variables, explanatory schemas, retrieval pathways, and research questions. On this view, novelty can occur inside an established organization. A new candidate hypothesis, unusual analogy, or statistically rare continuation may differ from previous outputs while leaving the effective structure of inquiry largely unchanged. Scientific discovery, by contrast, can sometimes involve a stronger transformation: an observation previously treated as noise becomes salient; a variable previously omitted becomes central; an accepted relation becomes unstable; a problem is reformulated; or a new representation changes which later observations can be recognized as relevant. The paper does not propose that every discovery has this form. It asks whether this stronger class of changes can be studied as transitions in a generative knowledge system rather than inferred from novelty scores alone.

## 1.1 Novelty, Discovery, and Structural Change

The first analytical distinction concerns the level at which change is assessed. Novelty is ordinarily comparative. An output is novel relative to prior outputs, a corpus, a task distribution, an evaluator’s expectations, or some other reference set. This makes novelty important but structurally incomplete as a criterion for discovery. A system can sample an unusual point while preserving the same variables, conceptual neighborhoods, evidential relations, and problem decomposition that made the point generable in the first place. Conversely, a modest-looking revision can become historically important if it changes the organization from which later inquiry proceeds.

The distinction should be stated without treating it as a formal implication. A generated object can differ substantially from a comparison set while the effective generative organization that produced it remains approximately unchanged. The paper therefore separates two empirical questions. The first asks whether a generated object differs from a comparison set. The second asks whether the relations that organize later generation have changed. The latter may concern problem framing, variable selection, anomaly sensitivity, retrieval, evidence weighting, hypothesis construction, tool use, or other components depending on the system boundary.

This distinction is particularly relevant to scientific inquiry because discovery is temporally extended. A candidate idea can initially appear original and later disappear without consequence. A small anomaly can initially appear unimportant and later reorganize a research programme. A useful account therefore needs to follow trajectories rather than isolated products. It should ask whether a perturbation is absorbed, retained, amplified, or incorporated into a changed pattern of subsequent inquiry. The paper uses this trajectory-level perspective to avoid treating novelty as a proxy for all stronger forms of epistemic change.

The distinction also prevents an inverse mistake. Structural change is not sufficient

for discovery. A system can reorganize around a false explanation, an unproductive classification, or a hallucinated causal model. Dynamical transformation and epistemic success must therefore remain separate. The present paper addresses conditions under which inquiry changes its own effective organization; it does not provide a complete theory of truth, justification, or scientific value. Later sections return to this boundary because an attractor transition, even if empirically demonstrated, cannot by itself establish that a discovery has occurred.

## 1.2 Artificial Systems as a Contrastive Case

Artificial generative systems are useful for studying this distinction because their stochasticity can be manipulated directly while their trajectories can be reproduced, branched, reset, and compared. Temperature, sampling rules, prompts, retrieval, memory, tool access, and orchestration can all change the range of outputs. Yet an increase in output variability need not imply a corresponding change in the structures governing later inference. This makes artificial systems a particularly clear setting in which to separate variation from transformation.

The positive evidence should be taken seriously. Si, Yang, and Hashimoto’s expert evaluation demonstrates that a contemporary LLM-based ideation system can generate research proposals that reviewers judge highly novel, and in that experiment more novel on average than the human proposals used for comparison [15]. Wang and colleagues’ much larger divergent-thinking study similarly rejects any simple assumption that human generation uniformly dominates machine generation [17]. These findings are important for the present argument precisely because they remove an easy explanation. If current systems were simply incapable of producing unusual or original-seeming material, there would be little need for a dynamical account of discovery. The harder problem begins once novelty is already available.

At the same time, novelty-oriented evaluations reveal limits of their own. Review ratings of an idea do not show whether the system can preserve an unresolved observation across time, revise the representation under which that observation is interpreted, or let one anomaly change what becomes salient later. Likewise, a high divergent-thinking score does not establish that a system can reorganize an explanatory space or sustain a new research trajectory. Ding and Li’s experiment is relevant here not because it proves a universal inability, but because it explicitly distinguishes operation within a known hypothesis space from the construction of a substantially different one under anomalous evidence [5]. That distinction is close to the problem developed in this paper.

A fair comparison also requires a broader unit than an isolated model call. Contemporary artificial research systems may include a foundation model, long context, persistent memory, retrieval, scientific databases, code execution, external tools, multiple agents, and human collaborators. The effective generative system is therefore distributed across components. A human scientist is also distributed across notebooks, instruments, collaborators, institutions, archives, and material environments. The paper consequently avoids comparing an embedded human investigator with a stateless language model and treating the difference as evidence of an intrinsic biological property. The relevant object is the organized system through which perturbations are encountered, retained, interpreted, and allowed to affect later generation.

This systems view also makes convergence results more informative. A recent study of repeated text–image feedback cycles found convergence toward generic motifs despite variation in starting conditions and temperature [9]. Their result concerns a particular cross-modal loop rather than scientific reasoning, but it provides a useful empirical example of a more general possibility: stochastic components can coexist with

long-run convergence. Randomness at the local generation level does not logically prevent recovery toward a restricted region of a larger state space. The paper develops this possibility using knowledge-attractor language in Section 3, while leaving open whether any specific AI system satisfies the stronger dynamical criteria required by that term.

### 1.3 The Edge of the Known as a Dynamical Problem

The title phrase *the edge of the known* has two deliberately distinct meanings. The first is epistemic. Inquiry reaches an edge when available concepts, variables, or explanatory relations cease to organize observations adequately. Such an edge can appear through anomaly, contradiction, unexplained residuals, failed prediction, incompatible representations, or an inability to formulate a problem within existing categories. This sense is historical and domain dependent. What lies at the edge for one research community may be routine knowledge for another.

The second meaning is dynamical and more provisional. A stable knowledge organization can be imagined as exerting restoring pressure: new information is classified through existing concepts, anomalies are normalized, and inference returns toward recurrent explanatory patterns. This restoring function is necessary. Without it, inquiry would be unable to distinguish consequential evidence from arbitrary variation. Yet the same stability can become restrictive if every disturbance is rapidly translated back into the incumbent representation. Discovery-oriented change may therefore depend, in some cases, on a regime in which restoring forces are weakened enough for a perturbation to persist while coherence remains sufficient for the perturbation’s consequences to be integrated.

The paper refers to this intermediate condition as *structured susceptibility*. It does not claim that scientific discovery occurs at a unique physical critical point. Nor does it assume that increasing sampling temperature moves an LLM toward a creative critical state. Rather, the term identifies an empirical possibility: under some configurations, a system may become more responsive to small, structured disturbances without becoming globally incoherent. Sections 5 and 6 examine how premature closure, metastability, and near-critical language can be used to formulate this possibility more precisely.

This framing changes the role assigned to contingency. Contingency is not treated as synonymous with randomness. A random perturbation may disappear immediately. A scientifically consequential perturbation is one whose effect depends on the relational configuration into which it enters and whose influence can persist into later inquiry. An anomalous datum, unexpected tool output, unfamiliar concept, or chance encounter can be trivial in one configuration and transformative in another. The paper therefore asks not how much randomness a system contains, but which perturbations become captured, which are damped, and which alter future susceptibility.

Only after these distinctions are established does the paper introduce the stronger analogy of *epistemic turbulence*. The term is reserved for a possible regime in which captured perturbations interact nonlinearly across scales, persist through historical retention, and modify the pathways through which later perturbations propagate. It is not a synonym for high entropy, chaos, brainstorming, or instability. The turbulence analogy is deliberately downstream in the argument because the paper first needs to show why ordinary variation, memory, nonlinear amplification, and metastable switching may be insufficient descriptions in particular cases. If simpler mechanisms explain the observed dynamics, the turbulence language should be withdrawn.

## 1.4 Scope and Analytical Programme

The contribution of this paper is therefore a decomposition of the transition from generative novelty to possible knowledge reorganization. It does not offer a general theory of scientific discovery, and it does not propose an essential difference between human and artificial intelligence. Instead, it develops a sequence of increasingly strong dynamical hypotheses that can be tested and, where appropriate, rejected independently.

Section 2 separates novelty from changes in the structure of the known and clarifies the levels at which scientific change can be assessed. Section 3 develops the language of knowledge attractors, basin-like recovery, and recurrent conceptual organization. Section 4 introduces contingency, anomaly, and perturbation capture, including the possibility that a perturbation can become consequential only after delayed retention or later reactivation. Section 5 examines epistemic damping and premature closure as candidate mechanisms by which anomalies can be absorbed into existing representations. Section 6 develops the edge-of-the-known problem through metastability, structured susceptibility, and cautious near-critical language. Section 7 then introduces epistemic turbulence as a stronger, controlled analogy for recurrent nonlinear cross-scale interaction and endogenous modification of propagation pathways. Section 8 applies the framework to the tendency of artificial systems to return to established organizations under some conditions. Section 9 turns to discovery-oriented system design and empirical interventions, including delayed closure, heterogeneous perturbation exposure, persistent anomaly memory, relational revision, and future tests of critical-zone prompting. Section 10 discusses alternative explanations, system boundaries, and the separation between dynamical transformation and epistemic success. Section 11 concludes by restating the framework as a research programme rather than a definition of discovery.

The central empirical commitment is modest but demanding. If the distinction developed here is useful, experiments should be able to show more than unusual outputs. They should reveal whether perturbations alter recovery, persistence, relational organization, anomaly sensitivity, or the response to matched future inputs. Conversely, if apparent reorganization can be explained by ordinary sampling variance, explicit memory replay, fixed nonlinear dynamics, or transient context effects, then stronger claims about attractor transition, generator reconfiguration, or epistemic turbulence should be rejected. The value of the framework lies less in assigning a new label to artificial creativity than in making the boundary between variation and discovery more experimentally tractable.

## 2 Novelty and the Structure of the Known

The purpose of this section is to separate several forms of novelty that are often compressed into a single judgment. A generated proposal can differ substantially from familiar examples while preserving the conceptual dimensions, problem decomposition, evidential categories, and inferential pathways through which the proposal was constructed. Conversely, a change that appears modest at the level of the final sentence or hypothesis can reorganize the relations through which later inquiry proceeds. The distinction is central to the present paper because the problem of discovery cannot be located solely at the level of unusual outputs. Before the language of attractors is introduced, it is therefore necessary to specify what is meant by the *known* and what kinds of change can occur within it.

The term *known* is used here in a deliberately broad but non-totalizing sense. It does not denote the complete set of propositions that a scientist, institution, or artifi-

cial system can retrieve. It denotes an effective organization of representational and practical resources that makes some observations easy to classify, some relations easy to infer, some questions natural to formulate, and some explanations easier to generate than others. The known can include explicit theories, tacit expectations, taxonomies, standard variables, experimental routines, retrieval pathways, canonical analogies, accepted forms of evidence, and habitual decompositions of a problem. Its boundaries are therefore partly epistemic and partly generative. A system can possess information that has little effect on its active organization, while lacking an explicit statement of assumptions that nevertheless strongly structure its inquiry.

This broader use of the known is compatible with several existing traditions without being reducible to any one of them. Boden’s distinction among combinational, exploratory, and transformational creativity is especially relevant because it separates novelty produced by combining or exploring within a space from novelty that alters the space of possibilities itself [2]. Gärdenfors’s theory of conceptual spaces likewise provides a useful representational precedent for treating concepts as organized by dimensions, regions, and similarity structures rather than as an unstructured inventory of symbols [7]. In philosophy of science, Kuhn’s analysis of normal science, anomaly, crisis, and conceptual change remains an important reminder that scientific novelty can occur both within an established disciplinary matrix and through changes that affect what counts as a problem, an admissible solution, or a relevant observation [13]. The present paper does not adopt these frameworks wholesale. It uses them to motivate a level-sensitive account of novelty that can later be connected to dynamical notions of recovery and transition.

## 2.1 Output Novelty and Search within an Established Space

The weakest relevant form of novelty concerns products. A sentence, image, proof sketch, hypothesis, or experimental proposal can be unlike familiar examples according to semantic distance, evaluator judgment, rarity under a reference distribution, or some task-specific criterion. This is the level most naturally captured by divergent-thinking tests, originality ratings, nearest-neighbor comparisons, or corpus-based novelty metrics. Such measurements are useful because scientific inquiry does require the production of candidates that are not trivial repetitions. They are nevertheless incomplete for the study of discovery.

The limitation becomes visible when a system is represented as generating outputs from an effective search space  $\Omega$  under a set of relations and constraints  $\mathcal{R}$ :

$$y \sim P(y \mid \Omega, \mathcal{R}, c), \quad (1)$$

where  $c$  denotes the immediate context. A rare or distant output can be produced while  $\Omega$  and  $\mathcal{R}$  remain approximately unchanged. The system has explored a less frequently visited region, but it has not necessarily altered the dimensions or constraints that organize the space. In this sense, novelty can be substantial at the level of  $y$  while remaining conservative at the level of the generative organization.

This distinction is particularly important for large language models because stochastic decoding can widen the range of outputs without requiring any durable change in the model’s effective representation of a scientific problem. Increasing temperature, changing a seed, requesting multiple candidates, or introducing persona variation can move sampling toward lower-probability continuations. These interventions can be valuable for ideation, but they do not by themselves establish that the system has changed what it treats as an explanatory variable, what evidence it considers diagnostic, or what unresolved inconsistency it preserves across later reasoning. The recent evidence that lan-

guage models can generate research proposals judged highly novel is therefore compatible with the present paper’s problem rather than contrary to it [15]. The difficult question begins after novelty has been demonstrated.

The same issue applies to human inquiry. A scientist can produce many unusual conjectures while remaining within a stable disciplinary vocabulary and problem representation. Conversely, a small representational revision can later alter a wide range of hypotheses. The difference cannot be inferred from the surface magnitude of the first product. Novelty should therefore be treated as indexed to a level of organization rather than as a unitary property of an idea.

## 2.2 Representational Organization and Conceptual Neighborhoods

A stronger form of change concerns the organization within which outputs are generated. Scientific representations are structured. Variables have relations; concepts have neighborhoods; observations are grouped into categories; instruments expose some dimensions while hiding others; methods determine which transformations are legitimate; and problem statements delimit what counts as a relevant answer. A system can therefore change even when no new primitive symbol is introduced. Reweighting, regrouping, or reconnecting existing elements can alter the effective geometry of inquiry.

Gärdenfors’s conceptual-space framework is useful here as an analogy because it treats conceptual representation in geometric terms, with quality dimensions and regions structuring similarity and categorization [7]. The present argument requires a more heterogeneous structure than a single metric conceptual space, since scientific systems include symbolic, causal, procedural, social, and instrument-mediated relations. Still, the geometric intuition is valuable. The known is not merely a warehouse. It has neighborhoods, distances, boundaries, privileged directions, and excluded combinations. A system that always reorganizes new inputs into the same neighborhoods may display surface novelty while preserving a stable representational topology.

This motivates a distinction between *content novelty* and *relational novelty*. Content novelty occurs when new elements or combinations appear. Relational novelty occurs when the effective relations among elements change. The latter can include a new causal link, a changed hierarchy, a revised similarity structure, a reclassification of evidence, or a new dependency among variables. Neither form automatically amounts to discovery. Relational novelty can be arbitrary or wrong. Its significance lies in the fact that it changes the conditions under which later outputs are generated.

The distinction can be expressed schematically by separating a set of active elements  $X_t$  from a relation structure  $R_t$ :

$$\mathcal{K}_t = (X_t, R_t). \quad (2)$$

Two systems can contain nearly the same  $X_t$  while differing substantially in  $R_t$ . Likewise, a system can add many new elements to  $X_t$  while leaving the organizing relations largely intact. For the present paper, structural change becomes increasingly important as novelty moves from changes in  $X_t$  toward changes in  $R_t$ , in the variables through which  $R_t$  is represented, and eventually in the processes that update the structure itself.

This level-sensitive view also clarifies why retrieval breadth is not equivalent to conceptual transformation. Access to a larger corpus can enlarge the available  $X_t$ , and retrieval can expose a system to distant associations. Yet if retrieved material is repeatedly assimilated under the same relation types and evaluative criteria, the effective organization may remain stable. Conversely, a single anomalous observation can have disproportionate consequence if it destabilizes a relation on which many later inferences depend.

Table 1: Analytical levels of novelty and reorganization. The levels distinguish increasingly structural forms of change; they are not a ranking of scientific value.

| Level                     | Primary change                                                                                         | Indicative evidence                                                                                      |
|---------------------------|--------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Output novelty            | A generated object differs from familiar or expected outputs                                           | Originality ratings, semantic distance, rarity, expert comparison                                        |
| Relational novelty        | Relations among active concepts, variables, evidence, or procedures are reorganized                    | Changed dependency structure, reclassification, altered retrieval or association patterns                |
| Representational change   | The coordinates, dimensions, categories, or problem decomposition used to organize inquiry are revised | New variables, changed unit of analysis, altered anomaly interpretation, reformulated problem space      |
| Generative reorganization | The processes governing later representation, search, or update are durably altered                    | Persistent change under matched future probes, changed susceptibility, altered update or search dynamics |

The size of an informational perturbation and the magnitude of structural change are therefore distinct quantities.

### 2.3 Exploratory and Transformational Change

Boden’s distinction between exploratory and transformational creativity provides a particularly useful bridge between creativity research and the dynamical problem developed in this paper. Exploratory creativity generates new possibilities by traversing a structured conceptual space, whereas transformational creativity changes one or more dimensions or constraints of the space itself [2]. The distinction should not be imported literally into scientific discovery, and the present paper does not assume a clean binary between exploration and transformation. Its value lies in identifying two different sources of novelty: movement within a structure and change of the structure that defines possible movement.

For scientific inquiry, exploratory change can include trying an unusual parameter combination, transferring a familiar method to a new dataset, proposing a less obvious causal mechanism within an accepted model family, or searching remote analogies while retaining the same problem definition. Transformational change can include introducing a new variable class, changing the unit of analysis, replacing a categorical representation with a continuous one, revising what counts as relevant evidence, or reformulating the problem so that earlier solutions are no longer expressed in the same coordinates. In practice, many discoveries contain mixtures of both. A long exploratory process can reveal pressure points that eventually motivate a representational transformation; a transformed representation can then open a new region for ordinary exploration.

The distinction is therefore better represented as a hierarchy of change than as a dichotomy. Table 1 summarizes four analytical levels used in the remainder of the paper. They are not intended as an exhaustive taxonomy or a normative ranking. A discovery can occur with little change at some levels and extensive change at others. The table instead clarifies what evidence would be required before moving from a claim about an unusual output to a stronger claim about reorganization of inquiry.

The final level anticipates a stronger notion developed later in the paper. A system undergoes generative reorganization when a prior episode changes not merely the current representation but the way later representations are formed or revised. This can occur through changed model parameters, but parameter change is not required. In

an artificial research system, durable memory, retrieval policy, tool orchestration, or self-maintained schemas can alter the effective generator while base-model weights remain fixed. In human inquiry, changes can be distributed across language, notebooks, collaborators, instruments, habits of attention, and institutional routines. The system boundary therefore matters as much as the substrate.

## 2.4 Scientific Change across Levels of Organization

Scientific discovery should not be identified automatically with the strongest level in Table 1. Many important discoveries are possible within relatively stable theoretical and methodological structures. A new particle, organism, reaction, archaeological site, or empirical regularity can be genuinely discovered without reorganizing an entire conceptual framework. Conversely, dramatic representational change can occur without producing a successful discovery. The aim of the hierarchy is not to reserve the term *discovery* for revolutionary science. It is to isolate the subset of cases in which discovery appears to involve a change in the organization that made prior inquiry stable.

Kuhn’s distinction between normal and revolutionary science provides a historical precedent for this separation, although the present framework is intentionally less categorical [13]. Normal science can generate extensive novelty through puzzle solving inside a disciplinary matrix. Revolutionary episodes can involve changes in standards, exemplars, classifications, and world description. For the present paper, the useful lesson is not that science alternates between two sharply separated modes. It is that the novelty of a result and the stability of the organization through which the result is interpreted are independent dimensions. A system can be highly productive while structurally stable, and a period of structural instability can be scientifically unproductive.

This independence is important for evaluating AI systems. If a benchmark rewards unusual hypotheses, broad associations, or surprising analogies, it primarily measures one part of the novelty hierarchy. If the research question concerns scientific discovery, additional measures are required. These might include whether an anomaly changes the problem representation, whether a revised representation persists across later tasks, whether the system can identify newly relevant observations that were previously ignored, and whether the resulting reorganization supports successful prediction or intervention. Novelty remains one component, but the evaluation must extend across time and across levels of organization.

The same point places limits on claims derived from failure. A model that returns to familiar representations in one experimental setting has not thereby been shown incapable of structural change. The observed stability may result from the task, prompt, memory architecture, retrieval environment, evaluation loop, or insufficient opportunity for longitudinal reorganization. The relevant empirical question is whether, under specified conditions, the system exhibits recovery toward an established organization and how that recovery changes as memory, anomaly retention, relational revision, or environmental heterogeneity are manipulated. This question motivates the attractor language developed in the next section.

## 2.5 Structural Criteria for Reorganization

The distinction between novelty and reorganization becomes useful only if stronger forms of change can be assessed empirically. Three criteria are especially important for the argument that follows: persistence, transfer, and altered response to future perturbation.

*Persistence* asks whether the change survives beyond the episode that produced it. A one-turn prompt can radically alter an answer without changing the later system. A structural claim becomes stronger when the altered organization is retained across time, contexts, or task boundaries. Persistence alone is still insufficient because memory can reproduce prior content without changing the process through which new content is interpreted.

*Transfer* asks whether the change affects situations that were not explicitly encoded in the initiating episode. If a revised concept applies only when the original prompt is repeated, the effect may be local contextual steering. If the revision changes how the system approaches new but structurally related problems, there is stronger evidence of a reorganization rather than simple replay. Transfer is therefore one way to distinguish stored content from changed relational structure.

*Altered perturbation response* asks whether the same later input has a different effect because of the earlier history. This criterion is the strongest of the three because it concerns susceptibility itself. Suppose two matched systems receive different histories and are then presented with the same anomaly. If the histories systematically change whether that anomaly is ignored, classified, retained, or amplified, then the earlier episode has changed the future response surface. Such evidence approaches what later sections call generator reconfiguration.

These criteria also clarify the role of phase-space language in the paper. A knowledge state is not assumed to possess a unique natural coordinate system. Phase space is an analytical construction whose coordinates must be chosen from measurable features of the system: concept activation, relation structure, hypothesis distribution, retrieval pattern, tool state, memory organization, or other variables appropriate to the experiment. The usefulness of the dynamical framework depends on showing that trajectories, recovery, sensitivity, or regime change can be operationalized at this level. Without such operationalization, attractor language risks becoming a synonym for familiarity.

The conceptual work of this section can therefore be summarized as a progression of analytical strength. Output difference makes the weakest structural claim. Relational reorganization is stronger because connections among elements have changed. Representational change is stronger again when the variables or coordinates through which a problem is organized are revised. Generative reorganization makes the strongest claim considered here because the processes governing later inquiry have changed. This progression concerns structural commitment rather than scientific merit. The remainder of the paper asks what stabilizes these organizations, how perturbations are normally absorbed, and under what conditions a local anomaly can become capable of changing the stronger levels. Section 3 develops this question through the concept of knowledge attractors.

### 3 Knowledge Attractors

The previous section distinguished unusual outputs from stronger forms of reorganization. The present section introduces the dynamical language needed to explain why such reorganization may be difficult. The central proposal is deliberately modest: some conceptual, explanatory, or problem-forming organizations can be treated *as if* they occupied recurrent regions of an effective state space, provided that this claim is supported by observable recovery after perturbation. The term *knowledge attractor* therefore does not denote a theory that is merely popular, a phrase that is statistically frequent, or a convention that is socially dominant. It denotes a recurrent organization toward which a specified inquiry system tends to return from a nontrivial neighborhood of perturbed

states. This operational restriction matters because attractor language is otherwise too easily reduced to metaphor.

### 3.1 Operational Meaning of an Attractor

Consider an inquiry system represented at time  $t$  by an effective state

$$z_t = (x_t, R_t, Q_t, M_t, E_t), \quad (3)$$

where  $x_t$  summarizes currently active concepts or hypotheses,  $R_t$  their operative relations,  $Q_t$  the present problem representation,  $M_t$  retained memory or external records, and  $E_t$  the relevant environment, tools, or retrieved evidence. The coordinates are analytical rather than ontological. Different experiments may choose different observables, and no claim is made that scientific knowledge has a unique natural phase space.

An effective update rule may be written schematically as

$$z_{t+1} = F_t(z_t, u_t), \quad (4)$$

where  $u_t$  denotes new evidence, prompts, measurements, interventions, or other inputs. In a stationary dynamical system, an attractor is a set toward which trajectories from a basin of initial states approach under repeated evolution. Standard dynamical-systems texts develop this idea through fixed points, limit cycles, and more complex invariant sets [16]. The present paper borrows only the structural logic. A knowledge attractor is therefore a recurrent organization  $\mathcal{A}$  for which a class of perturbed states repeatedly evolves back toward an equivalent region of conceptual and problem-forming organization.

This definition makes *recovery* central. Suppose a system begins near  $\mathcal{A}$ , receives a perturbation  $\epsilon$ , and is then allowed to continue under matched conditions. If the resulting trajectory returns toward the same organization, the perturbation has revealed restoring structure. If it does not return, several interpretations remain possible: the perturbation may have shifted the system into another pre-existing basin; the original organization may have changed; the system may simply be undergoing a long transient; or the chosen state representation may be inadequate. Attractor language becomes informative only when these alternatives are separated empirically.

The distinction also prevents frequency from being confused with dynamics. A concept can be highly frequent in a corpus yet dynamically weak: a small contextual change may cause the system to abandon it and remain elsewhere. Conversely, a relatively infrequent conceptual organization can be dynamically stable if diverse perturbations are repeatedly absorbed and interpreted through it. Probability mass and attraction can correlate, especially in learned generative systems, but they are not identical. Attraction concerns the geometry of response across trajectories, not merely the marginal frequency of an output.

### 3.2 Basins, Recovery, and Restoring Structure

Let  $\mathcal{B}(\mathcal{A})$  denote the set of states from which the effective dynamics return toward  $\mathcal{A}$  under a specified family of conditions. The basin is not assumed to be fixed across every context. It is an empirical object defined relative to a system boundary, an input regime, and a time horizon. A language model queried once, a model with persistent memory, and an autonomous research agent interacting with tools can therefore possess different effective basin structures even when they share the same base parameters.

For a perturbation  $\epsilon$  applied at time  $t$ , a finite-horizon recovery ratio can be written as

$$\rho_k(\epsilon) = \frac{D(z_{t+k}^{(\epsilon)}, \mathcal{A})}{D(z_t^{(\epsilon)}, \mathcal{A})}, \quad (5)$$

where  $D$  is an empirically chosen distance from the recurrent organization. Values  $\rho_k < 1$  indicate movement back toward the attractor over the chosen horizon, whereas  $\rho_k > 1$  indicate increasing separation. The equation is intentionally generic because semantic, relational, or procedural distances will differ across studies. Its purpose is to make a methodological point: an attractor claim should be grounded in response trajectories rather than inferred from a single familiar answer.

Restoring structure can arise at several levels. A model may return to a canonical explanation because its learned conditional distribution favors that explanation. Retrieval may repeatedly supply documents that support an incumbent framing. A memory system may reinstate a prior plan. An evaluation loop may reward answers that conform to known categories. Human researchers may also return to familiar variables, methods, or classifications because disciplinary training, instruments, publication practices, and shared exemplars stabilize them. The dynamical perspective does not require these mechanisms to be identical. It asks whether, taken together, they create a reproducible tendency to restore a particular organization after deviation.

This view also clarifies why a large perturbation can be less consequential than a small one. If an input is easily assimilated into the incumbent representation, its magnitude may be large while its dynamical effect remains local. A small observation that falls close to a basin boundary, changes a stabilizing relation, or activates a competing organization can have a much larger long-run consequence. The later sections therefore treat perturbation direction, relational location, and regime susceptibility as at least as important as raw perturbation size.

### 3.3 Multiple Attractors and Contextual Switching

Inquiry systems need not possess a single dominant organization. Several recurrent regimes can coexist, each stabilized by different contexts, tasks, memories, or evidential conditions. Let

$$\mathcal{L}_t = \{\mathcal{A}_{1,t}, \dots, \mathcal{A}_{m,t}; \mathcal{B}_{1,t}, \dots, \mathcal{B}_{m,t}\} \quad (6)$$

represent an effective attractor landscape at time  $t$ . A prompt, cue, or environmental change can move the current state from one basin to another while leaving the landscape itself intact. Such switching should not be mistaken for knowledge reorganization.

This distinction is important in generative AI. A model can adopt different roles, explanatory frames, or problem decompositions under different prompts. If each frame remains available and the system simply switches among them according to context, the phenomenon is closer to state selection within an existing landscape than to transformation of the landscape. Likewise, a human researcher can alternate between statistical, mechanistic, legal, or historical descriptions of the same object without changing the deeper organization that governs when each frame is used. Flexibility among existing attractors can be epistemically valuable, but it is analytically different from creating a new recurrent organization or modifying the boundaries among old ones.

A stronger claim arises when prior history changes the landscape itself. The attractors, basin boundaries, or accessibility relations represented by  $\mathcal{L}_t$  may differ at a later time because of the intervening history. This can involve movement of an attractor, deformation of a basin, appearance of a new recurrent regime, disappearance of an

old one, or altered accessibility among regimes. In scientific terms, a new representation may make previously unrelated observations mutually constraining, render an old category unstable, or change which anomalies are capable of redirecting inquiry. Landscape change therefore approaches the stronger levels of reorganization introduced in Section 2, but it still does not by itself establish scientific improvement. A newly stable organization can be false, unproductive, or normatively undesirable.

### 3.4 Attractor Memory and Artificial Generative Systems

The use of attractor language in artificial systems has a long technical precedent, although that precedent should not be transferred uncritically to contemporary language models. Hopfield’s classic network model described content-addressable memory through phase-space flow: partial or corrupted input can evolve toward a stored collective state, making memory retrieval naturally expressible in terms of basins and recurrent states [11]. The relevance here is structural rather than architectural. Hopfield networks show how a distributed system can implement recovery toward recurrent organizations without requiring a symbolic lookup table. They do not establish that LLM inference has the same energy function, memory geometry, or attractor structure.

For contemporary generative systems, the correct object of analysis is broader than the base model. Let

$$\mathcal{G}_t = (\theta, C_t, M_t, R_t, T_t, \Pi_t), \quad (7)$$

where  $\theta$  denotes model parameters,  $C_t$  context,  $M_t$  persistent memory,  $R_t$  retrieval state,  $T_t$  tool and environmental state, and  $\Pi_t$  orchestration or policy. Recurrent behavior can emerge from any combination of these components. A model with fixed weights can display changing effective attractors if its memory or retrieval structure changes; conversely, a fine-tuned model can retain the same practical basin structure under the tasks being studied.

Recent autonomous generation-loop experiments provide suggestive evidence that repeated model-mediated transformation can converge toward generic motifs. Hintze, Proschinger Åström, and Schossau report that repeated language–image generation loops tend to converge toward a limited set of visual motifs across long autonomous trajectories [9]. This result is relevant because it directly concerns repeated dynamics rather than one-shot originality. It should nevertheless be interpreted narrowly. Convergence in a multimodal generation loop is not by itself evidence for a general *knowledge* attractor, and the recurrent motifs could reflect representation bottlenecks, decoding biases, image-model priors, caption compression, or other mechanisms. The experiment is best treated as proof that attractor-like convergence is an empirically testable property of coupled generative systems, not as proof of the stronger theory advanced here.

The same caution applies in the opposite direction. If a system produces highly diverse outputs under high temperature or heterogeneous prompting, diversity alone does not show that it lacks attractors. A system can explore broadly within a basin and still return toward a recurrent organization over longer horizons. The relevant measurements concern return probability, recovery time, basin boundaries, switching structure, and changes in later perturbation response.

### 3.5 Empirical Criteria and Competing Explanations

Attractor language is useful only if it excludes simpler explanations. Table 2 summarizes several observations that can resemble attraction and the additional evidence needed

Table 2: Operational evidence for knowledge-attractor claims and competing explanations.

| <b>Observed pattern</b>            | <b>Simpler explanation</b>                          | <b>Additional evidence for attraction</b>                                |
|------------------------------------|-----------------------------------------------------|--------------------------------------------------------------------------|
| Frequent canonical answers         | High marginal probability or benchmark convention   | Recovery toward the same organization after heterogeneous perturbations  |
| Prompt-sensitive frame switching   | Contextual steering among available representations | Recurrent basin structure under matched perturbation families            |
| Longitudinal repetition            | Memory replay or retrieval duplication              | Return after state displacement when replay content is controlled        |
| Convergence across autonomous runs | Shared decoding or representation bottleneck        | Basin mapping, perturbation thresholds, and reproducible return geometry |
| Persistent post-event change       | Stored context or one-off hidden-state difference   | Transfer across matched future probes and changed perturbation response  |

before a stronger dynamical interpretation is warranted.

A minimal experimental programme would therefore begin with replicated trajectories from matched states. Perturbations should vary in magnitude, semantic direction, relation to the current problem, and timing. The experiment should then measure whether trajectories contract toward a common organization, switch among several recurrent regimes, or remain persistently separated. Repeating the procedure across contexts permits an empirical approximation of basin geometry. Applying the same later probes after different histories can test whether the landscape has itself changed.

Several failure modes would weaken an attractor interpretation. If convergence disappears after controlling decoding temperature, the relevant phenomenon may be sampling bias. If it disappears when retrieval is randomized, the restoring force may reside primarily in the retrieval layer. If apparently persistent change vanishes when explicit memory is removed, the result may reflect replay rather than reorganization. If no neighborhood of perturbed states exhibits systematic recovery, then the term attractor adds little beyond ordinary similarity or frequency.

These restrictions are especially important for the paper’s later claims. The argument does not require all scientific knowledge to be attractor-like, nor does it require current AI to be globally trapped in fixed basins. It requires only the weaker possibility that some inquiry configurations exhibit measurable restoring tendencies, and that these tendencies can vary with system architecture, history, environment, and control regime. Once that possibility is admitted, the discovery problem can be stated more precisely: how does a perturbation that would ordinarily be absorbed or followed by recovery become capable of altering the basin structure itself? Section 4 turns to the first part of that question by distinguishing contingency, anomaly, and perturbation capture.

## 4 Contingency, Anomaly, and Perturbation Capture

The preceding section described knowledge attractors in terms of recovery after perturbation. The present section turns to the perturbations themselves. A central difficulty is that several distinct phenomena are often collapsed into the single language of “unexpected input.” Random variation, subjective surprise, anomaly relative to a theory, and historically consequential perturbation are not equivalent. Scientific discovery can involve all four, but none by itself is sufficient. The purpose of this section is therefore

to specify how an event can enter an inquiry system, acquire anomaly status relative to an incumbent representation, and become retained strongly enough to influence later inquiry. The term *perturbation capture* is introduced for this latter process. It denotes neither mere detection nor immediate theory change. It denotes the retention of a perturbation in a form that remains causally available to later relations, questions, or responses.

#### 4.1 Random Variation and Relational Contingency

Randomness is a property of a sampling procedure or probability model; contingency is a property of how an event enters a historical and relational configuration. The distinction matters especially for artificial systems, because stochastic variation can be increased cheaply through decoding temperature, sampling policy, randomized retrieval, or synthetic perturbation. A system can therefore receive a large number of statistically unusual inputs while its deeper organization remains unchanged.

Let the effective inquiry state be  $z_t$ , as defined in Section 3, and let an input perturbation be denoted by  $\epsilon_t$ . A simple response measure over horizon  $k$  is

$$\Delta_k(\epsilon_t) = D\left(z_{t+k}^{(\epsilon)}, z_{t+k}^{(0)}\right), \quad (8)$$

where the perturbed and control trajectories receive matched subsequent conditions. Equation (8) does not define contingency, but it makes one feature visible: two perturbations of similar initial magnitude can have very different downstream consequences. One may disappear after a few updates; another may alter retrieval, attention, problem representation, or the interpretation of later evidence.

The relevant notion of contingency in this paper is therefore relational and historical. A perturbation is contingent relative to the chosen level of analysis when its downstream role is not fixed solely by its intrinsic informational content but depends materially on the configuration into which it enters. The same observation can be trivial in one research programme and destabilizing in another. The same prompt can be assimilated by one agent and reorganize the task representation of another. The same experimental residual can be dismissed as noise, retained as unresolved, or become the starting point of a new line of inquiry. What matters is not rarity alone but the relation between the event and the receiving structure.

This formulation also separates contingency from metaphysical indeterminism. The paper does not require that an event be fundamentally uncaused or physically indeterminate. An encounter can be fully caused and still be contingent relative to a scientist's operative model, a laboratory's research trajectory, or an artificial agent's current representation. The analytical question is whether its later consequence was already stabilized by the effective generative organization under study.

The distinction is summarized in Table 3. The table is intentionally hierarchical only in analytical commitment, not in epistemic value. A surprising event can be useless; a captured anomaly can be misleading; and a theory-changing perturbation can support a false theory.

#### 4.2 Anomaly as a Representation-Relative Relation

An anomaly is not simply a strange datum. It is a relation between an observation and an operative representation. Kuhn's account of scientific change made anomaly important precisely because observations become problematic against a background of expectations, exemplars, and normal problem solving [13]. Later work on responses to anomalous evidence makes the same point in a more fine-grained way: anomalous data can be

Table 3: Analytical distinctions among variation, surprise, anomaly, and captured perturbation.

| Category                                | Defining relation                                                                             | What it does not establish                                 |
|-----------------------------------------|-----------------------------------------------------------------------------------------------|------------------------------------------------------------|
| Random variation                        | Low-probability or sampled difference under a specified distribution                          | Surprise, anomaly, persistence, or discovery               |
| Surprise                                | Mismatch with an observer’s or model’s expectation                                            | Conflict with a scientific representation or theory change |
| Anomaly                                 | Observation or result that is difficult to reconcile with an operative representation         | Retention, acceptance, or reorganization                   |
| Captured perturbation                   | Difference retained so that it remains available to influence later inquiry                   | Truth, usefulness, or full generator reconfiguration       |
| Historically consequential perturbation | Captured difference that changes later attention, relations, questions, or response structure | Epistemic improvement by itself                            |

ignored, rejected, held in abeyance, reinterpreted, accommodated through peripheral change, or accepted as grounds for more substantial theory change [3, 4]. The existence of an anomaly therefore does not determine its fate.

This relational view can be expressed schematically. Let  $E_t$  be an observation and  $M_t$  an operative model or problem representation. Define an incompatibility measure

$$a_t = \mathcal{A}(E_t, M_t), \quad (9)$$

where  $\mathcal{A}$  is not assumed to be a universal numerical metric. In one experiment it might be prediction error; in another it may be failure of classification, violation of a constraint, inconsistency among trusted sources, or inability to represent the observation within the current variable set. The same  $E_t$  can therefore have different anomaly status under different  $M_t$ .

The representation-relative character of anomaly is central to the AI comparison. A model trained on a broad corpus may possess a ready-made explanation for an observation that a historically situated human research group would experience as surprising. Conversely, an artificial system may flag statistical deviation in data that human researchers regard as theoretically unimportant. Neither difference establishes superior discovery capacity. It indicates that anomaly depends on what the system currently treats as expected, relevant, and explanatory.

This also clarifies why scientific discovery cannot be reduced to anomaly detection. Automated systems can detect outliers, residuals, contradictions, distribution shifts, or unusual combinations at enormous scale. These capabilities expand the space of candidate perturbations, but the discovery problem begins after detection. Which anomaly should remain active? Which should be treated as measurement error? Which should be preserved despite lacking an available explanation? Which should trigger a revision of the representation rather than a search for a better fit within it? These are questions of retention, coupling, and later consequence.

Historical cases reinforce the point. Accounts of discovery often emphasize observations that initially resisted available categories, but the same histories also show extended periods in which anomalous findings were repeated, reclassified, or left unresolved before their significance stabilized. Studies of scientific reasoning likewise show that anomalous evidence can produce a wide range of responses short of theory change [3, 4]. The important dynamical distinction is therefore not between systems that encounter anomalies and systems that do not. It is between different ways of metabolizing

anomaly over time.

### 4.3 Perturbation Capture beyond Detection

The concept of perturbation capture is introduced to describe the first transition from a local disturbance to a historically available difference. A perturbation is captured when some trace of it survives the immediate episode and remains capable of entering later generative relations. Capture can occur in several substrates: biological memory, an experimental notebook, an unresolved-data register, a citation, a changed search query, a persistent vector-memory entry, an altered tool policy, a modified concept graph, or a new question retained in a research agenda. The substrate is secondary to the functional property of later availability.

Let  $h_t$  represent the retained historical state of an inquiry system. A perturbation can alter that state even when its significance is not yet explicit. A useful operational comparison is therefore between a perturbed trajectory and a matched control trajectory. A weak criterion for capture is that their retained historical states remain separated after the perturbation has ceased:

$$D_H(h_{t+j}^{(\epsilon)}, h_{t+j}^{(0)}) > \delta, \quad (10)$$

for some nontrivial delay  $j$ , with the retained difference capable of affecting a later task or response. This criterion is intentionally weaker than generator reconfiguration. A note stored in external memory can satisfy capture without changing the system's generative rule. Stronger forms of change are considered in later sections.

Capture should also be distinguished from explicit salience. A scientist may immediately label an observation anomalous and then forget it. Conversely, an encounter can leave a weak trace without being recognized as important at the time. This latter possibility is especially relevant to the problem of discovery because importance is often not fully representable when the perturbation first occurs. A measurement can be archived before the relevant theory exists; a sentence can be remembered before its connection to a later problem becomes clear; an artificial agent can retain an observation that later becomes relevant after new evidence changes the task representation.

The temporal structure therefore involves at least three potentially distinct moments: the time of event entry, the time of explicit recognition or salience, and the time at which the event becomes integrated into later inquiry. Their ordering is not fixed. Recognition can precede durable capture, capture can occur without immediate recognition, and retrospective interpretation can make an old trace newly active.

This temporal separation produces a demanding design problem for artificial research systems. A relevance filter that stores only information already judged useful can be efficient, yet it can systematically discard events whose significance becomes visible only under a future representation. Storing everything is not a solution either: indiscriminate accumulation can overwhelm retrieval and reinforce noise. The problem is to preserve some weakly relevant traces in a form that permits later reactivation without treating them as established knowledge. This is one reason the paper separates memory capacity from perturbation capture.

### 4.4 Capture, Normalization, and Competing Responses

Once a perturbation is registered, the system must decide—explicitly or implicitly—how to relate it to the incumbent organization. Several outcomes are possible. The perturbation can be discarded, normalized as noise, assimilated under an existing category, placed in abeyance, connected to a peripheral exception, or allowed to alter the problem

representation. Chinn and Brewer’s taxonomy of responses to anomalous data is useful here because it demonstrates that theory change is only one response among many [3, 4]. The present framework recasts these alternatives dynamically rather than normatively.

The strength with which a perturbation remains available and the degree to which it is coupled to the active relational structure should be treated as separate dimensions. Anomaly status, attention, memory, institutional procedures, retrieval policies, and later evidence can all affect whether a retained trace becomes generatively active. High anomaly need not produce high capture. A result can be strongly inconsistent with expectation yet be discarded because the instrument is distrusted. Conversely, a weak anomaly can remain active because it connects to several unresolved problems or because institutional procedures preserve it for later review.

This provides a natural transition to the concept of epistemic damping developed in Section 5. A perturbation may be captured weakly but then rapidly mapped back onto the incumbent attractor. Such normalization is often epistemically appropriate. Science would be impossible if every irregular measurement destabilized a theory, and artificial systems would be unusable if every unusual token permanently rewrote their effective representations. The relevant issue is therefore not whether systems damp perturbations, but whether their damping mechanisms can distinguish disposable variation from events whose unresolved status should remain active.

The distinction can also be expressed in terms of competing timescales. If damping or assimilation typically occurs much faster than broader relational coupling, a potentially consequential anomaly may be normalized before it can interact with enough of the system to alter later inquiry. When relational coupling develops on a comparable or shorter timescale, the perturbation has more opportunity to recruit other relations before recovery. This is not proposed as a universal law. It is an experimental question about whether closure occurs before alternative relational consequences have had time to develop.

#### **4.5 Historical Availability and the Transition to Discovery Dynamics**

Perturbation capture is the weakest point at which contingency becomes relevant to the later history of inquiry. It does not yet constitute discovery, theory change, or attractor transition. Its importance lies in making later interaction possible. A perturbation that disappears completely cannot be amplified by future evidence, cannot be reinterpreted when a new representation appears, and cannot change which later inputs become salient. A retained perturbation can.

The term *historical availability* is useful for this intermediate status. An event is historically available when it remains accessible to later generative relations even though its present significance may be uncertain. Historical availability can be explicit, as in a laboratory anomaly log, or implicit, as in altered retrieval weights, changed attention, or a weak association. It can also be distributed across a system: a person may forget an observation while an archive preserves it; a model may not retain an interaction internally while an external memory store does; an organization may maintain unresolved findings across generations of researchers.

This distributed possibility matters for fair human–AI comparison. Human science relies heavily on externalized memory: notebooks, datasets, instruments, publication records, colleagues, and institutions preserve perturbations that no individual brain retains. Artificial research systems can likewise externalize history into databases, vector stores, code repositories, or agent logs. The relevant comparison is therefore not biological memory versus model context length. It is whether the coupled inquiry system preserves unresolved differences in a form that can later be reactivated and allowed to

modify its representational organization.

The stronger discovery problem begins when historical availability interacts with restoring structure. If an anomaly is captured but repeatedly translated back into the incumbent representation, it remains historically present while dynamically weak. If it recruits additional evidence, changes the interpretation of related observations, or weakens the relations that stabilize the current attractor, its perturbational gain can increase. Section 5 examines the first of these outcomes—epistemic damping and premature closure. Sections 6 and 7 then develop the stronger possibility that a captured perturbation enters a high-susceptibility regime and participates in nonlinear, turbulence-like reorganization. The present section therefore establishes a necessary analytical bridge: before a disturbance can challenge the known, it must first remain available long enough to matter.

## 5 Epistemic Damping and Premature Closure

Section 4 established a distinction between encountering an anomaly and retaining it strongly enough to remain historically available. The present section considers a different failure mode. A perturbation can be detected, remembered, and even explicitly labeled anomalous while still having little effect on the organization of later inquiry. The system may rapidly translate the disturbance into a familiar category, attach a locally plausible explanation, or redirect search toward an already available representation. In dynamical terms, the perturbation is not lost at entry; it is damped during interpretation. This section uses the term *epistemic damping* for processes that reduce the downstream structural consequence of a perturbation by restoring an incumbent organization of concepts, variables, explanations, or problem framings.

Damping is not intrinsically epistemically defective. A research system that reorganized itself in response to every discrepant observation would be unable to preserve cumulative knowledge, distinguish noise from signal, or maintain stable standards of evidence. The relevant problem is therefore not whether closure occurs, but how closure is timed and what alternatives have been permitted to develop before the inquiry is stabilized. The term *premature closure* is reserved for cases in which restoration occurs before the system has adequately tested whether the anomaly should remain unresolved, recruit additional relations, or motivate revision of the operative representation. The distinction is especially useful for artificial systems because coherent completion is often easy to observe, whereas the capacity to sustain unresolved representational tension is harder to measure.

### 5.1 Damping as Restorative Epistemic Dynamics

The attractor framework of Section 3 already implies a basic form of damping. If an inquiry state is displaced from a recurrent organization and subsequently returns toward it, the system exhibits restoring dynamics. At the epistemic level, this return can be implemented through many mechanisms: reclassification of anomalous evidence, retrieval of a familiar explanation, revision of a peripheral assumption, discounting of an unreliable source, compression of several observations into an established causal schema, or simple forgetting of the relation that made the event initially problematic.

Let  $a_t$  denote the active anomaly load associated with a perturbation and let  $R_t$  denote the relational structure through which the anomaly can affect other parts of the inquiry. A minimal schematic model is

$$a_{t+1} = (1 - \kappa_t)a_t + \Gamma(R_t, \epsilon_t), \quad (11)$$

where  $\kappa_t \in [0, 1]$  represents an effective damping rate and  $\Gamma$  represents renewed activation through current relations. A high  $\kappa_t$  suppresses unresolved anomaly rapidly unless relational coupling reactivates it. A low  $\kappa_t$  allows the discrepancy to remain active for longer, but this alone does not guarantee productive inquiry. Persistent anomaly can also reflect confusion, poor measurement, or failure to discriminate relevant from irrelevant variation.

The same logic can be stated in terms of recovery. Suppose a perturbation changes the active problem representation from  $M_t$  to  $M_t^{(\epsilon)}$ . If subsequent inference returns to an incumbent representation  $M^*$ , then a simple recovery measure is

$$r_k = \frac{D_M(M_{t+k}^{(\epsilon)}, M^*)}{D_M(M_t^{(\epsilon)}, M^*)}. \quad (12)$$

For  $r_k < 1$ , the trajectory has moved back toward the incumbent organization over horizon  $k$ . This quantity does not by itself distinguish appropriate correction from excessive restoration. Its role is diagnostic: it turns the vague claim that a system “returns to the known” into an observable question about how quickly and under which perturbations recovery occurs.

An important implication follows. Epistemic damping can coexist with excellent local reasoning. A system can accurately summarize the anomaly, enumerate several explanations, and produce a coherent justification while still restoring the same problem representation at the next stage. What matters for discovery dynamics is whether the disturbance changes future search, salience, variable selection, evidence weighting, or the set of candidate representations. High-quality explanation at time  $t$  is therefore not equivalent to low damping over the trajectory that follows.

## 5.2 Closure as a Temporal Relation

The concept of premature closure should not be defined by the mere presence of an answer. Scientific inquiry regularly requires temporary closure: experiments must be designed, models selected, papers submitted, and decisions made under finite resources. The relevant distinction concerns the relation between the time required for alternative structures to develop and the time at which the system stabilizes an interpretation.

Let the time required for alternative structures to develop be distinguished from the time at which the system stabilizes an interpretation. Premature closure becomes more plausible when interpretive stabilization repeatedly occurs much earlier than the development of competing representations, cross-domain relations, or additional evidence, provided that the task permits further inquiry and the anomaly has not already been adequately resolved. This is a comparative timing claim rather than a universal threshold.

This formulation also clarifies the relation between premature closure and perturbation capture. Capture answers whether an anomaly survives long enough to remain available. Closure answers how the surviving anomaly is organized. A perturbation can be stored indefinitely while remaining conceptually neutralized. For example, an unexplained residual may persist in a database but be tagged as instrument noise and excluded from later theory-building. Conversely, a perturbation can be only weakly stored but continue to alter attention or search behavior. Durable storage and open interpretation are therefore separable dimensions.

The taxonomy of responses to anomalous data developed by Chinn and Brewer is relevant because it demonstrates that anomalous evidence can be rejected, excluded from the domain of a theory, held in abeyance, reinterpreted, or accommodated without immediate abandonment of the central theory [3, 4]. The present paper does not rank these

responses in advance. Many are rational under uncertainty. The dynamical question is whether the response preserves enough information about the unresolved relation for later evidence to reopen the issue if necessary.

Premature closure can therefore be understood as a loss of *reopenability*. An inquiry may reach a provisional conclusion while preserving the path by which that conclusion could later be challenged, or it may compress the anomaly so completely that the original tension becomes difficult to recover. The former permits stable action together with historical revisability. The latter increases the effective restoring force of the incumbent representation.

### 5.3 Completion Pressure in Artificial Generative Systems

Artificial generative systems make closure unusually visible because their outputs are commonly evaluated at the level of completed responses. A language model is typically prompted to answer, explain, summarize, solve, classify, recommend, or propose. Instruction tuning and preference optimization can further reward responses that appear coherent, relevant, and complete. None of these design features implies that a model is incapable of sustaining uncertainty. They do, however, create a candidate mechanism by which unresolved structure can be converted quickly into plausible completion.

The scientific-discovery experiment reported by Ding and Li is suggestive in this regard. Their study distinguishes incremental reasoning over available hypotheses from discovery scenarios in which an adequate hypothesis must be constructed from anomalous observations, and reports substantial difficulty for current generative systems in the latter setting [5]. The result does not establish premature closure as the cause. Several alternatives remain possible, including insufficient domain representation, weak experimental interaction, limited memory, or task-specific evaluation constraints. It nevertheless motivates a concrete question: when an anomaly does not fit an available explanation, does the system preserve the representational tension or rapidly generate a plausible account that restores coherence?

Recent work on AI-generated research ideas provides a complementary perspective. Language models can generate ideas that experts judge highly novel in controlled settings [15]. Such success demonstrates that completion pressure does not prevent unusual output. It also reinforces the present distinction: the ability to produce a novel proposal is compatible with strong restoration at the level of problem framing or explanatory organization. The relevant test is not whether an answer sounds unconventional, but whether an unresolved perturbation changes the representations used in later, matched inquiry.

For artificial research systems, at least four closure channels should therefore be distinguished. The first is *linguistic completion*: producing a well-formed answer because the interaction format demands one. The second is *retrieval closure*: finding an available explanation in external memory and terminating search. The third is *representational closure*: keeping the same variables and conceptual relations while modifying only local parameter values or peripheral assumptions. The fourth is *organizational closure*: an agent protocol, benchmark, tool policy, or human supervisor ending exploration once a satisfactory output criterion is met. These channels can coexist, and only some are properties of the foundation model itself.

The distinction matters for experimental design. A system that appears to close rapidly in a single-turn prompt may behave differently when allowed to maintain unresolved hypotheses across sessions, request new evidence, inspect its own competing representations, or preserve anomalies in external memory. Conversely, adding memory may strengthen closure if prior explanations are repeatedly retrieved and treated as

defaults. The proper unit of analysis is therefore the coupled artificial research system rather than an isolated model response.

## 5.4 The Protective Function of Damping

A discovery-oriented theory should resist the temptation to treat low damping as automatically desirable. Most perturbations should probably disappear. Scientific instruments produce artifacts; datasets contain errors; literature contains false claims; language models generate spurious associations; human researchers notice patterns that do not replicate. A system that amplifies every discrepancy would spend its resources pursuing noise and could become less capable of cumulative inquiry.

The relevant objective is therefore selective damping. Suppose the eventual epistemic value of investigating a perturbation is not available at the time of initial encounter. The system must act using only an imperfect current estimate. This creates two different risks: it may suppress a perturbation that would later prove consequential, or it may devote substantial attention to a perturbation that should have been discarded. The relative cost of these errors is task dependent. Scientific domains with cheap experimentation may tolerate more exploratory amplification, whereas high-cost or safety-critical domains may require stronger damping before structural revision.

This trade-off explains why the paper does not advocate maximal openness, entropy, or anomaly preservation. Strong damping protects coherence and permits efficient exploitation of established knowledge. Weak damping increases the chance that unconventional relations survive long enough to be explored. The discovery problem lies in regulating the boundary between them. A capable system must sometimes say, in effect, “this difference is probably noise,” and at other times preserve the weaker statement, “this does not yet fit, and the reason remains unresolved.”

Human scientific institutions already externalize part of this regulation through replication, peer criticism, anomaly logs, exploratory analysis, competing laboratories, and archival preservation. Artificial research systems can in principle implement analogous functions through uncertainty registers, unresolved-hypothesis stores, scheduled re-evaluation, adversarial agents, or policies that separate provisional explanation from anomaly closure. Whether such mechanisms improve discovery is an empirical question rather than a conceptual guarantee.

## 5.5 Operational Signatures of Premature Closure

The concept becomes scientifically useful only if it can be distinguished from correct resolution. A system should not be classified as prematurely closing merely because it reaches an answer quickly. The relevant evidence must show that closure systematically suppresses later exploration that would have remained warranted under matched conditions.

Several operational signatures are possible. First, *rapid representational recovery*: after a perturbation that initially changes the problem representation, the system returns quickly to the same variable set, causal schema, or explanatory family. Second, *low counterfactual reopening*: after later evidence strengthens the original anomaly, the system fails to reactivate previously discarded alternatives. Third, *explanation-induced damping*: providing an early plausible explanation reduces later exploration more strongly than providing the same evidence without the explanation. Fourth, *path-dependent narrowing*: once an initial interpretation is selected, subsequent search and retrieval become increasingly concentrated around that interpretation even when matched evidence supports alternatives.

These signatures can be tested experimentally. Consider two otherwise matched agent trajectories exposed to the same anomaly. In one condition, the system is required to provide an immediate best explanation. In another, it must preserve several incompatible interpretations and defer closure until additional evidence arrives. Later both trajectories receive identical observations and tools. If the immediate-closure condition shows lower representational diversity, weaker anomaly reactivation, shorter search distance, and stronger return to the original attractor even when the deferred condition discovers a superior representation, the data would support a premature-closure interpretation. If no such difference appears, or if immediate closure consistently improves later revision, the hypothesis should be weakened.

Let  $C_t = 1$  indicate that the system has previously stabilized an interpretation. A simple measure of reopening can then be defined as

$$\mathcal{R}_{\text{open}} = P(M_{t+j} \in \mathcal{M}_{\text{alt}} \mid E_{t+j}, C_t = 1), \quad (13)$$

where  $\mathcal{M}_{\text{alt}}$  is a predefined or behaviorally inferred set of alternative representations. The absolute value of  $\mathcal{R}_{\text{open}}$  is less important than matched differences across closure policies, memory conditions, and perturbation histories.

The resulting picture is deliberately more nuanced than the claim that AI “answers too quickly.” Epistemic damping is a general requirement of inquiry; premature closure is a conditional failure of regulation. The strongest discovery systems should neither remain indefinitely unsettled nor restore every perturbation immediately. They require a regime in which some anomalies can remain active long enough to recruit additional relations while the system retains enough coherence to compare, test, and eventually consolidate alternatives. Section 6 develops this regime-level problem through metastability, susceptibility, and the idea of operating at the edge of the known.

## 6 The Edge of the Known: Metastability and Structured Susceptibility

Section 5 treated premature closure as a timing problem in which an inquiry system can restore coherence before an anomaly has recruited alternative relations. The present section shifts from the timing of closure to the regime in which closure and amplification occur. The central question is whether the same perturbation can have qualitatively different consequences because the receiving system occupies a different region of its effective dynamical space. The phrase *the edge of the known* is used here for that regime-level problem. It refers neither to a literal geometric boundary of all knowledge nor to a claim that discovery occurs at a unique physical critical point. Instead, it names a family of conditions under which incumbent representations remain sufficiently organized to support cumulative inquiry while becoming sufficiently susceptible that some small perturbations can persist, recruit additional relations, and alter later search.

This section develops three distinctions that are required before the stronger turbulence analogy of Section 7. First, high stochasticity is separated from high susceptibility. Second, metastability is separated from criticality: a system can move among partially stable organizations without approaching a singular transition. Third, susceptibility is separated from useful reorganization. Large divergence can reflect incoherence, chaotic amplification, or loss of constraint rather than discovery. The relevant hypothesis is therefore one of *structured susceptibility*: elevated perturbation gain together with enough coherence, retention, and evaluative constraint for the resulting differences to remain scientifically usable.

## 6.1 Regime Position and Structured Susceptibility

Let  $z_t$  denote the effective inquiry state introduced in Sections 2–3, and let  $\mu_t$  denote a vector of control variables that shape the current dynamics. Depending on the empirical system,  $\mu_t$  may include coupling strength among representations, memory persistence, retrieval concentration, exploration policy, stopping rules, confidence thresholds, decoding parameters, or institutional constraints. The same perturbation  $\epsilon$  can then produce different finite-horizon consequences under different  $\mu_t$ .

A simple susceptibility measure is

$$\chi_k(\epsilon; \mu_t) = \frac{D\left(z_{t+k}^{(\epsilon)}, z_{t+k}^{(0)}\right)}{\|\epsilon\|}, \quad (14)$$

where  $D$  is a task-appropriate distance between matched trajectories. Equation (14) does not assume linear response. It simply asks how much downstream difference is associated with a perturbation of a given magnitude over horizon  $k$ .

High  $\chi_k$  alone is insufficient. A system that amplifies every small difference until trajectories become incoherent would score highly while losing the ability to compare hypotheses, preserve evidence, or consolidate a revised representation. Let  $C_k(\mu_t)$  denote a coherence measure and let  $H_k(\mu_t)$  denote historical retention, understood as the extent to which earlier captured perturbations remain causally available at horizon  $k$ . A provisional region of structured susceptibility can then be written as

$$\mathcal{Z}_{\text{sus}} = \{\mu : \mathbb{E}[\chi_k] \geq \chi_0, \quad C_k(\mu) \geq C_0, \quad H_k(\mu) \geq H_0\}, \quad (15)$$

where the thresholds are empirical and task-dependent. The point of Equation (15) is not to define creativity by threshold. It is to make explicit the trade-off that the paper has approached from several directions: discovery-oriented reorganization may require a regime in which differences can propagate without dissolving the structures needed to evaluate and retain them.

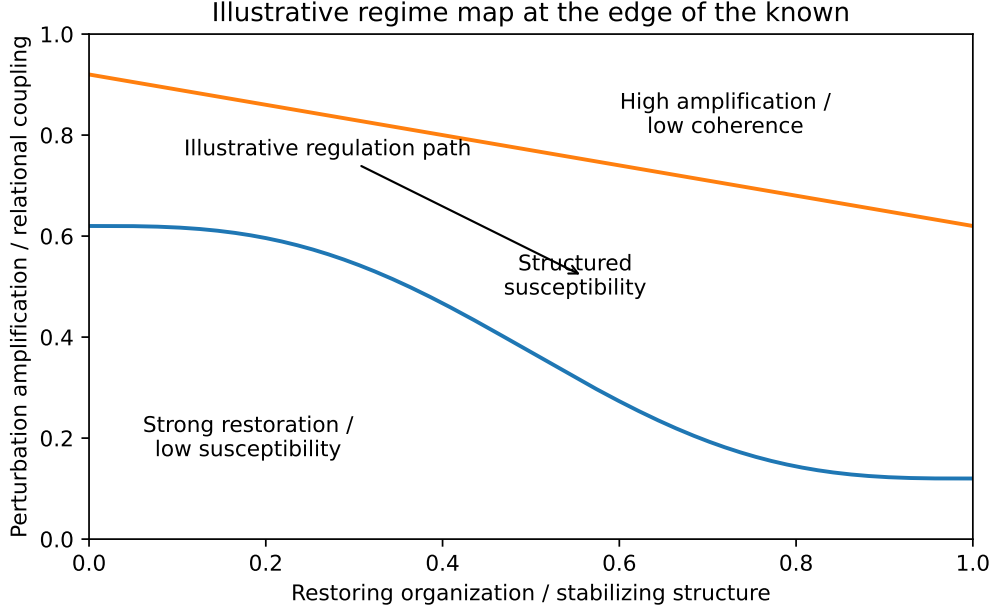
Figure 1 visualizes this distinction schematically. The axes are deliberately generic. A real artificial or human research system would require experimentally grounded control variables, and the regime boundaries would have to be inferred rather than drawn in advance. The figure therefore functions as a phase-diagram hypothesis, not as empirical evidence.

## 6.2 Metastability and Competing Organizations

Metastability provides one way to describe flexibility without requiring a unique critical point. In a metastable system, coordinated organizations can persist for finite periods while remaining capable of reconfiguration; trajectories neither collapse immediately into a single rigid state nor wander without recognizable structure. Kelso’s treatment of metastability in brain coordination emphasizes this coexistence of tendencies toward integration and independence, and illustrates why metastability should be distinguished from both simple multistability and unrestricted instability [12]. The present paper uses the concept more abstractly. An inquiry system can be called metastable when several problem representations, explanatory frames, or relational organizations remain dynamically available over a relevant horizon and transitions among them are possible without complete loss of coherence.

This description fits scientific practice better than a binary contrast between “stable theory” and “revolution.” Researchers often work for long periods with partially competing interpretations, unresolved anomalies, and provisional models. Different representations can be recruited by different tasks or evidence while no single alternative has

Figure 1: Schematic phase diagram for the regime-level argument of Section 6. The horizontal axis represents restoring organization, while the vertical axis represents effective perturbation amplification or relational coupling. The intermediate region denotes structured susceptibility: perturbations can acquire persistent gain while coherence and historical retention remain above functional thresholds. The boundaries are illustrative rather than empirically fitted.



yet reorganized the field. Such coexistence can preserve reopenability. It also creates conditions under which a new perturbation can couple to a representation that would have been inaccessible under a more strongly stabilized regime.

Metastability does not itself imply discovery. A system can switch among familiar explanations indefinitely. In artificial systems, prompt-sensitive switching among known frames can produce substantial surface diversity while leaving the available representational repertoire unchanged. The relevant empirical question is therefore whether metastable availability permits perturbations to change transition probabilities, create new couplings, or alter which representations remain accessible later. Ordinary switching is weaker than structural reorganization.

This distinction can be expressed with a transition matrix  $P_t$  over a set of recurrent organizations  $\{M_1, \dots, M_n\}$ . Context-sensitive metastability may change the currently occupied state while leaving the matrix approximately fixed. A stronger historical effect is present when a captured perturbation changes later transition probabilities themselves, with the direction and magnitude of that change depending on retained history and relational organization. This provides a bridge from metastable switching to the later idea of generator reconfiguration, and clarifies why a system can be flexible without being transformative.

### 6.3 Near-Criticality as an Operational Hypothesis

Critical-transition language becomes relevant when the restoring organization of a system changes so that small perturbations increasingly influence long-run state. In many nonlinear systems, movement toward a bifurcation can be accompanied by reduced recovery rates or other changes in susceptibility. Work on critical transitions has developed indicators such as critical slowing down and increasing autocorrelation in systems approaching some tipping points [14]. These indicators are not universal and can pro-

duce false positives outside the model classes for which they are justified. Their relevance here is methodological: they show that “distance to transition” can sometimes be operationalized through perturbation response rather than inferred from descriptive complexity alone.

For the present paper, a near-critical epistemic regime is therefore an empirical hypothesis about response geometry. Let  $\mathcal{C}$  denote a set of control configurations associated with elevated transition sensitivity. A schematic distance is

$$d_{\mathcal{C}}(\mu_t) = \inf_{\mu \in \mathcal{C}} d_{\mu}(\mu_t, \mu). \quad (16)$$

The weak hypothesis is that, over some range of experimentally defined controls, smaller distance to such a transition-sensitive region will be associated with higher measured susceptibility, while coherence and historical retention remain above functional minima. This is deliberately weaker than the claim that discovery occurs at criticality. Elevated susceptibility may be broad, multidimensional, and task-specific.

Neuroscience provides a useful example of how such a hypothesis can be formulated and contested rather than simply assumed. Recent work has argued that near-critical dynamics may function as a computational setpoint in brain systems and has synthesized a large body of empirical literature in support of that view [8]. The present paper does not transfer that conclusion to language models, scientific institutions, or creativity. The relevance is conceptual: criticality can be treated as a claim about regulated response properties that requires explicit signatures, controls, and competing explanations.

Several alternative mechanisms must therefore remain visible. Non-normal transient amplification can produce high finite-time gain while a system remains asymptotically stable. Finite-amplitude perturbations can cross basin boundaries far from a linear instability. Memory accumulation can make a system progressively easier to redirect without any critical transition. Heterogeneous retrieval can widen representational access while leaving the underlying landscape unchanged. Any empirical claim of near-criticality should outperform these simpler accounts.

## 6.4 Stochasticity and Regime Position

The distinction between stochasticity and regime position is particularly important for generative AI. Decoding temperature, top- $p$  sampling, random seeds, or stochastic tool selection can widen the set of immediate outputs. These interventions act directly on variation. They do not by themselves establish that restoring forces have weakened, that basin geometry has changed, that anomalies will be retained longer, or that future susceptibility has increased.

This can be expressed by separating a sampling parameter  $T$  from the broader control vector  $\mu$ :

$$y_{t+1} \sim P_{\theta, T}(y \mid C_t), \quad \mu_t = \mu(C_t, M_t, R_t, \Pi_t, E_t, \dots), \quad (17)$$

where  $C_t$  is context,  $M_t$  persistent memory,  $R_t$  retrieval or relational organization,  $\Pi_t$  orchestration policy, and  $E_t$  the external environment. Increasing  $T$  modifies the sampling distribution at the current step, whereas movement in  $\mu$  may require changed memory, retrieval, prompt structure, tool policy, interaction history, or problem representation. In a coupled system, temperature can indirectly affect  $\mu$  by exposing the system to different events, but that downstream change must be demonstrated rather than assumed.

The same caution applies to entropy. High token entropy can coexist with strong conceptual recovery, while low token entropy can accompany a decisive structural transition if a small perturbation redirects later inference. Discovery-oriented experiments should therefore measure representational recovery, reopening, transition persistence,

and history-dependent response rather than using decoding entropy as a proxy for epistemic criticality.

This distinction also clarifies the future idea of *critical-zone prompting*. A prompt could in principle change effective regime position by altering closure rules, requiring incompatible hypotheses to remain active, preserving unresolved anomalies, or forcing explicit revision of relational structure. Such an intervention would be interesting precisely because it changes more than randomness. Yet the phrase should remain prospective until perturbation-response experiments show that the prompt systematically alters susceptibility, recovery time, or transition structure under matched conditions.

## 6.5 Regime Regulation and the Return to Coherence

If a susceptible regime can support discovery, a further problem immediately appears: why should a system approach such a regime, and how should it leave it? Permanent high susceptibility would be costly. Scientific inquiry requires periods of consolidation in which a representation becomes stable enough to support prediction, experiment design, communication, and cumulative extension. The relevant architecture is therefore not maximal instability but regulated movement between exploration and consolidation.

Let  $\mu_t$  again denote the effective control state. Its later value may depend on unresolved anomaly, the availability of alternative representations, evidential quality, current structural stability, and other system-specific variables. This dependence need not be explicit or centralized. In a human research system it may be distributed across attention, collaborators, instruments, archives, and institutional routines. In an artificial system it may be implemented through prompting, memory policy, retrieval diversity, search budget, tool use, self-critique, or orchestration among agents.

A discovery-oriented regulator would loosen restoration when accumulated evidence indicates that the incumbent representation is repeatedly failing, while strengthening consolidation when a revised organization begins to explain and predict effectively. The resulting process can be described as a recurrent movement among stabilization, accumulation of unresolved problems, increased susceptibility to revision, reorganization, and eventual restabilization. This description does not imply that every episode contains a sharp critical point, or even that the stages occur in a fixed order. It instead emphasizes that the capacity to leave a stable representation and the capacity to form a new stable representation are complementary rather than opposed.

The strongest empirical version of this hypothesis predicts adaptive modulation. Systems exposed to repeated unresolved anomalies should become more reopenable, broaden retrieval, lengthen search, or reduce restoring pressure in the relevant representational region, while systems receiving consistent confirming evidence should consolidate. A fixed strategy that always maximizes exploration or always minimizes uncertainty would provide a weaker form of regulation.

The edge-of-the-known formulation can now be stated precisely enough for the next stage of the paper. The relevant boundary is not a line separating “known” from “unknown.” It is a regime in which an incumbent organization remains present but no longer absorbs every consequential perturbation with the same restoring strength. Metastability can keep alternatives available; near-critical language can describe elevated transition sensitivity when empirically justified; and structured susceptibility identifies the broader functional requirement that amplification remain compatible with coherence and retention. Section 7 considers a stronger possibility: under some conditions, captured perturbations may begin to interact across scales and modify the very pathways through which later perturbations propagate.

## 7 From Perturbation to Epistemic Turbulence

The preceding sections developed a sequence of progressively stronger claims. A generative system can produce novel outputs while remaining within an established representational organization. It can also receive anomalies, preserve them, reopen a problem, or move into a regime of elevated susceptibility without thereby undergoing a deeper reorganization. The present section introduces a stronger hypothesis only after these distinctions are in place. The term *epistemic turbulence* is used for a provisional dynamical regime in which perturbations do more than produce isolated divergence: they interact nonlinearly across epistemic scales, are retained through history, recur intermittently, and modify the pathways through which later perturbations propagate.

The analogy is deliberately limited. Fluid turbulence is a physical phenomenon governed by the Navier–Stokes equations, with well-developed concepts of inertial transfer, intermittency, dissipation, and multiscale structure [6]. Nothing in this paper implies an epistemic conservation law, an analogue of kinetic energy, or a direct mapping from intellectual activity to fluid mechanics. The useful comparison is structural. In both cases, the analytical problem concerns how local disturbances can become coupled across scales, how apparently stable regimes can coexist with bursts of large reorganization, and how the medium of propagation matters for what a disturbance eventually becomes. Research on the onset of turbulence also illustrates why transition phenomena cannot be reduced to “more noise”: susceptibility, finite-amplitude disturbances, spatiotemporal coexistence, and regime structure all matter [1, 10]. The present argument borrows that caution rather than those equations.

### 7.1 Cascades before Turbulence

The first boundary to preserve is between a cascade and a turbulence-like regime. A small disturbance can trigger a large cascade without producing any persistent change in the propagation medium. Threshold models on networks demonstrate this point clearly: a local shock can recruit neighboring nodes and eventually produce a global cascade when network structure and local thresholds place the system in a susceptible regime [18]. The existence of a large cascade therefore establishes nonlinear amplification under a particular structure, but it does not establish that the structure itself has been altered by the cascade.

An epistemic cascade can be described across levels of organization. Let  $R_t^{(\ell)}$  denote relational organization at epistemic scale  $\ell$ , where a scale may correspond to an observation, concept, hypothesis, theory, research programme, or institutional practice. An initial perturbation may first alter a local relation; that change may recruit effects at broader levels, and those broader effects may later influence lower-level interpretation. A laboratory anomaly, for example, can alter the interpretation of a variable; the altered variable can destabilize a hypothesis; the revised hypothesis can redirect experiment design; and a changed research programme can later alter what counts as a salient local observation. The propagation is therefore neither necessarily upward nor one-directional.

To distinguish ordinary accumulation from nonlinear cascade, one can compare the joint effect of perturbations with the effects produced separately. Let  $\mathcal{T}_k(E)$  denote a trajectory after  $k$  steps under perturbation set  $E$ , and let  $D$  be a task-appropriate distance. A simple interaction term is

$$I_k(e_1, e_2) = D(\mathcal{T}_k(\{e_1, e_2\}), \mathcal{T}_k(\emptyset)) - D(\mathcal{T}_k(\{e_1\}), \mathcal{T}_k(\emptyset)) - D(\mathcal{T}_k(\{e_2\}), \mathcal{T}_k(\emptyset)). \quad (18)$$

Table 4: Dynamical distinctions relevant to the epistemic-turbulence hypothesis. The rows are ordered by increasing explanatory commitment rather than by epistemic value.

| Phenomenon              | Minimal signature                                                                      | Additional evidence required for a stronger interpretation                        |
|-------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Stochastic variation    | Different outputs under matched conditions                                             | Persistent dependence on perturbation history rather than sampling alone          |
| Nonlinear amplification | Perturbations interact or grow non-additively                                          | Propagation across multiple representational scales                               |
| Cascade                 | Local change recruits wider relational structure                                       | Recurrence, intermittency, and history-dependent propagation                      |
| Metastable switching    | Movement among pre-existing recurrent organizations                                    | Evidence that propagation pathways themselves are modified                        |
| Chaotic divergence      | Sensitive divergence from nearby trajectories                                          | Coherent multiscale organization and retained structural consequence              |
| Epistemic turbulence    | Recurrent multiscale nonlinear interaction with intermittency and pathway modification | Evidence that this stronger model outperforms simpler fixed-dynamics explanations |

Large nonzero  $I_k$  indicates that perturbations interact rather than merely add. Positive values can indicate amplification, while negative values can indicate cancellation, competition, or saturation. This kind of measurement is more informative than simply counting unusual outputs, because it asks whether the generative consequences of one perturbation depend on the presence of another.

A single nonlinear cascade still remains weaker than the turbulence hypothesis developed below. The stronger claim requires repeated interaction, multiscale participation, historical persistence, and endogenous change in how later disturbances travel. Table 4 summarizes the evidence hierarchy used in this section.

## 7.2 Multiscale Interaction and Intermittency

The second requirement is multiscale interaction. In the present context, “scale” does not denote a fixed physical length. It denotes a level at which epistemic organization can be meaningfully measured. An observation may interact with a concept, a concept with a hypothesis, a hypothesis with a theory, and a theory with a research programme or institution. These levels can be represented differently across empirical settings, and no universal hierarchy is assumed.

Let  $\Phi_{\ell m}(t)$  denote an empirically defined transfer measure from organization at scale  $\ell$  to organization at scale  $m$ . Depending on the application,  $\Phi$  could be estimated through changes in concept graphs, retrieval structure, hypothesis distributions, experimental choices, citation paths, tool use, or latent-state representations. The notation should not be read as an energy flux. Its function is only to make explicit that a disturbance becomes more structurally consequential when its effects are detectable across levels that would otherwise be treated as separate.

Multiscale propagation also permits feedback. A revised theory can change which observation is treated as anomalous; a changed retrieval policy can alter which hypotheses are accessible; a newly selected experiment can generate data that further modifies the theory. The stronger case is therefore recursive: a change at one level affects another level, and the resulting reorganization later modifies the interpretation or propagation rule operating at the original level. Such feedback is already stronger than linear diffu-

sion because later local dynamics depend on prior higher-level reorganization.

Intermittency provides a second structural feature. Turbulent systems are not characterized by homogeneous disorder; bursts of intense activity can coexist with comparatively quiescent periods [6]. The epistemic analogue, if it exists, would likewise involve uneven temporal organization. A research process can remain locally stable for many steps, undergo a burst of rapid cross-level reinterpretation, and then settle again. The hypothesis therefore predicts distributions with clustered restructuring events rather than constant maximal divergence.

This distinction matters for artificial systems. A model that produces highly diverse answers on every step may simply be noisy. A system exhibiting long intervals of coherent continuation punctuated by historically triggered bursts of relational reorganization presents a different dynamical profile. The relevant empirical question is not whether outputs look surprising, but whether the timing and scale of reorganization show structured dependence on retained perturbations and current regime position.

An operational intermittency statistic could compare local structural-change magnitudes  $a_t$  across windows. Heavy-tailed or clustered  $a_t$  values may be compatible with intermittency, but they are not sufficient evidence by themselves. Similar distributions can arise from heterogeneous prompts, external task switching, or fixed nonlinearities. Intermittency becomes theoretically relevant only when it co-occurs with multiscale coupling and history-dependent modification of future propagation.

### 7.3 Historical Modification of Propagation Pathways

The strongest part of the turbulence analogy concerns not amplitude but the medium through which later disturbances travel. In a fixed network, a cascade can be large while the network remains unchanged. A stronger generative process occurs when earlier perturbations modify the effective relations, retrieval paths, interpretive rules, memory structures, or orchestration policies that determine later propagation.

Let  $\mathcal{K}_t$  denote an effective perturbation-propagation operator: a compact representation of how currently active relational changes influence later ones under the present state, retained history, and regime conditions. A fixed-medium explanation treats this propagation structure as stationary, or as changing independently of the perturbation sequence under study. The stronger historical hypothesis instead requires the propagation structure itself to change because of prior perturbations. In empirical work, that claim should be supported by matched later perturbations that travel differently after controlled differences in prior history.

This condition gives historical retention a precise role. Suppose a perturbation  $e_1$  changes retrieval so that a later perturbation  $e_2$  is routed through a different conceptual neighborhood. Then the effective consequence of  $e_2$  depends on the prior history even when  $e_2$  itself is held fixed. A strong test is therefore a matched-future-input design: two systems receive different earlier perturbation histories and are later exposed to the same probes. If their perturbation-response surfaces remain systematically different after current-state and explicit-memory controls, there is evidence that the effective propagation operator has changed.

For artificial systems, the substrate of  $\mathcal{K}_t$  need not be model weights. It can include persistent memory, retrieval indexes, concept graphs, tool-selection policies, search strategies, agent-to-agent communication structure, or rules for reopening resolved questions. This system-level view avoids the false conclusion that a fixed-weight language model cannot participate in historical reconfiguration. It also prevents the opposite mistake of treating any context accumulation as generator change. The empirical requirement is functional: prior perturbations must alter how later perturbations

propagate under matched conditions.

The same logic clarifies why anomaly preservation matters. If every anomaly is quickly normalized,  $\mathcal{K}_t$  may remain effectively unchanged and later disturbances will continue to be routed through the same explanatory organization. If anomalies remain available long enough to interact, their joint effects can modify both representation and pathway. In this sense, premature closure is one candidate damping mechanism that can prevent a cascade from becoming a self-modifying regime.

## 7.4 Epistemic Turbulence as a Controlled Analogy

The paper uses the term *epistemic turbulence* only for a conjunction of features. Within an observation window  $W$ , let

$$\mathbf{T}_W = (M_W, N_W, I_W, P_W, H_W) \quad (19)$$

denote a provisional signature vector, where  $M_W$  represents multiscale participation,  $N_W$  nonlinear perturbation interaction,  $I_W$  intermittency,  $P_W$  endogenous pathway modification, and  $H_W$  historical persistence. The notation does not define a universal metric. Each component must be operationalized for the system under study.

The turbulence label is warranted only if the conjunction explains observations that weaker models do not. High  $M_W$  with low  $P_W$  can describe a broad cascade through a fixed network. High  $N_W$  with low  $H_W$  can describe transient nonlinear amplification. High  $I_W$  without history dependence can result from externally varying tasks. Sensitive divergence without preserved coherence can reflect chaos or decoding noise. The stronger interpretation becomes useful only when multiscale interaction repeatedly modifies the relational pathways that sustain later interaction.

This restriction is essential because the metaphor otherwise becomes unfalsifiable. Any complicated sequence of ideas can be redescribed retrospectively as turbulent. The purpose of the formal distinctions above is to block that move. If a stationary nonlinear model predicts the observed transitions, that model should be preferred. If explicit memory replay explains historical persistence, no additional turbulence mechanism is required. If apparent cross-scale effects disappear after controlling for changing prompts, external tools, or human interventions, the endogenous-turbulence claim should be withdrawn.

The analogy also has explicit physical limits. No epistemic Reynolds number is proposed. No Kolmogorov scaling law is expected. No energy cascade or conserved quantity is assumed. Physical turbulence often concerns transfer across spatial or spectral scales under constraints specific to fluid dynamics [6]. Epistemic systems involve semantic, social, computational, and institutional relations whose scales can be heterogeneous and whose propagation can be bidirectional. The analogy therefore operates at the level of dynamical organization rather than governing equation.

Transition-to-turbulence research nevertheless offers one methodological lesson that transfers more safely: onset can depend on regime position, finite disturbances, co-existence between different dynamical organizations, and spatiotemporal proliferation rather than on a single scalar amount of noise [1, 10]. This reinforces the central argument of the present paper. If an artificial discovery system ever exhibits turbulence-like epistemic dynamics, the relevant mechanism is unlikely to be “increase randomness.” It would involve a susceptible generative organization in which retained perturbations can proliferate, interact, and alter their own future channels.

## 7.5 Empirical Signatures and Falsification Conditions

The epistemic-turbulence hypothesis yields stronger empirical requirements than a standard creativity benchmark. One would need longitudinal trajectories rather than isolated outputs, matched perturbations rather than uncontrolled prompt changes, and measurements at multiple representational levels. At minimum, five classes of evidence should be sought.

First, *multiscale propagation*: a perturbation introduced at one level should produce measurable changes at several other levels, such as concept relations, hypothesis distributions, retrieval paths, experiment choices, or research-plan structure. Second, *non-additive interaction*: paired or sequential perturbations should produce consequences not predictable from the sum of isolated effects. Third, *intermittency*: reorganization should occur in structured bursts rather than as uniform noise. Fourth, *historical dependence*: the same later probe should produce different perturbation responses after different retained histories. Fifth, *pathway modification*: evidence should show that the effective propagation structure itself has changed.

These signatures can be tested in human, artificial, or coupled human–AI systems. For an artificial research agent, one could hold the foundation model fixed while manipulating memory persistence, retrieval diversity, anomaly-retention policy, search depth, and orchestration rules. Small, semantically controlled anomalies could be inserted at matched points. The experiment would then track whether the anomalies disappear, produce local branching, trigger wider cascades, or alter the system’s future response to matched probes. The turbulence-like interpretation should be accepted only when the final pattern requires self-modifying multiscale dynamics rather than simpler explanations.

Several outcomes would weaken the hypothesis. If high novelty and discovery-level reorganization are fully predicted by ordinary stochastic search, then perturbation capture and turbulence-like structure add little. If large cascades occur but later propagation pathways remain unchanged, a cascade model is sufficient. If apparent pathway change vanishes after controlling for explicit memory contents, memory-mediated steering is the better explanation. If a stationary nonlinear model reproduces the same intermittency and history dependence, the claim of endogenous propagation reconfiguration should be reduced. If systems with strong damping nevertheless achieve robust discovery across matched tasks, the role assigned to susceptibility should be reconsidered.

The value of the turbulence hypothesis therefore lies partly in its vulnerability. It is not needed to explain every unusual idea, every scientific revision, or every non-deterministic trajectory. It is a candidate model for a narrower class of episodes in which preserved perturbations begin to interact across scales and, through that interaction, change the conditions under which later perturbations are interpreted and propagated. Section 8 applies this distinction to the recurrent observation that artificial systems can display considerable local novelty while repeatedly returning to established knowledge organizations.

## 8 Return to the Known in Artificial Generative Systems

The preceding sections developed a sequence of distinctions between novelty, attractor-like recovery, perturbation capture, damping, structured susceptibility, and turbulence-like reorganization. This section applies those distinctions to artificial generative systems. Its purpose is not to claim that current systems are intrinsically confined to existing knowledge. The narrower question is why systems capable of producing substantial local novelty may nevertheless exhibit recurrent return toward established representa-

Table 5: Candidate mechanisms that can produce return-to-known behavior in artificial generative systems. The mechanisms operate at different system layers and therefore require different controls before a stronger attractor interpretation is warranted.

| <b>Mechanism layer</b> | <b>Candidate restoring process</b>                                                                                | <b>Empirical discriminator</b>                                                                     |
|------------------------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| Learned generator      | High-probability continuation and familiar conceptual neighborhoods dominate repeated inference                   | Hold retrieval, memory, and orchestration fixed while testing recovery after matched perturbations |
| Retrieval              | Similar queries repeatedly retrieve similar documents, concepts, or exemplars                                     | Randomize or diversify retrieval while keeping the base model and prompt matched                   |
| Persistent memory      | Experience-following and replay stabilize previously successful or salient trajectories                           | Manipulate memory addition, deletion, and similarity while controlling current-state information   |
| Orchestration          | Planning templates, tool policies, stopping rules, or evaluation criteria repeatedly restore a preferred workflow | Replace or perturb orchestration policies while preserving model, memory, and task                 |
| Closure process        | Plausible explanation reduces anomaly persistence before alternative representations develop                      | Manipulate explanation timing, reopening policy, and anomaly-retention duration                    |
| Coupled system         | Humans, tools, databases, and external evaluators stabilize or destabilize the effective trajectory               | Match system boundaries and localize where persistent recovery or reconfiguration occurs           |

tional, explanatory, or procedural regions. The answer may depend on several mechanisms operating at different layers: learned generative priors, retrieval and memory, orchestration policies, closure behavior, and the absence or weakness of persistent pathway modification. These mechanisms should be separated empirically rather than collapsed into a single claim about “AI creativity.”

A useful starting point is the apparent tension in recent evidence. Large language models can generate research ideas that expert evaluators judge highly novel, and large-scale creativity comparisons show substantial competence on divergent-generation tasks [15, 17]. Yet other studies identify limits in anomaly-driven scientific discovery, diversity across repeated generation, and autonomous iterative dynamics [5, 9]. These observations are not inconsistent. High local novelty can coexist with strong recovery toward familiar explanatory or representational organization. The relevant dynamical question is therefore not whether an artificial system can leave a local neighborhood at all, but whether departures can become historically retained, relationally amplified, and capable of changing the system’s later response surface.

## 8.1 Generative Priors and High-Probability Recovery

A trained generative model encodes regularities from a large historical corpus. At inference time, these regularities shape the conditional distribution over later tokens, concepts, plans, and explanations. This fact alone does not imply an attractor. Section 3

emphasized that an attractor claim requires recovery after perturbation rather than merely high frequency. Nevertheless, learned probability structure provides one plausible source of restoring force. A local intervention can make an output unusual while subsequent generation again favors conceptual neighborhoods with dense statistical support.

Let  $z_t$  denote an effective representational state and let  $\mathcal{K}$  denote a reference region corresponding to an established explanatory organization. A minimal return statistic can be written as

$$\mathcal{R}_k(\epsilon) = P\left(z_{t+k}^{(\epsilon)} \in \mathcal{N}(\mathcal{K}) \mid z_t^{(\epsilon)} = z_t + \epsilon\right), \quad (20)$$

where  $\mathcal{N}(\mathcal{K})$  is an operational neighborhood rather than a metaphysical set of “known ideas.” The return-to-known hypothesis predicts that, for some perturbation classes and some systems,  $\mathcal{R}_k$  remains high even when immediate outputs diverge substantially from baseline.

This formulation separates two phenomena that are easily confused. The first is *novel sampling*: a trajectory temporarily enters a low-density region. The second is *structural escape*: subsequent inference continues to reorganize around relations that were weak or absent in the incumbent representation. A model can succeed at the first while failing at the second. The empirical question is not whether rare states are reachable, but whether they alter the distribution of later transitions.

The autonomous language–image loop results of Hintze, Proschinger Åström, and Schossau provide a useful example of why repeated dynamics matter [9]. Across many iterative text–image–text cycles, trajectories converged toward a small family of generic motifs despite varied initial prompts and temperature settings. Their setting is cross-modal rather than scientific, and the observed convergence should not be generalized directly to all reasoning systems. Still, it demonstrates that generative diversity at one step can coexist with strong iterative convergence over longer horizons. This is precisely the type of phenomenon for which an attractor-like analysis becomes more informative than a one-shot creativity score.

## 8.2 Retrieval and Memory as Stabilizing Mechanisms

The base model is only one component of contemporary artificial agents. Retrieval systems, persistent memory, notes, tool outputs, and long-horizon planning can all increase historical dependence. These additions are often treated as mechanisms for escaping the limitations of a static model. They can indeed widen the accessible state space. Yet they can also create new restoring forces.

A retrieval system that maps similar queries to similar documents can repeatedly return the agent to the same conceptual neighborhood. A memory system can preserve a successful procedure and make later behavior more path dependent. Such persistence may improve reliability while simultaneously reducing openness to alternatives. Recent work on LLM-agent memory reports an *experience-following* effect in which retrieved experiences similar to the current task strongly shape subsequent outputs, together with risks of error propagation and misaligned experience replay [19]. These results are important for the present argument because they show that historical retention is not equivalent to generative reorganization. Memory can preserve a perturbation, but it can also deepen an incumbent basin.

For this reason, a discovery-oriented system should not be evaluated by memory capacity alone. The more relevant question is how memory changes the perturbation-response profile. Suppose  $M_t$  denotes persistent memory and  $Q_t$  the retrieval operator.

Then an effective transition can be written schematically as

$$z_{t+1} = F_{\theta}(z_t, Q_t(z_t, M_t), u_t), \quad (21)$$

where  $u_t$  represents current external input. Historical change in  $M_t$  can increase either flexibility or restoration. If new memories diversify retrieval and expose weakly represented relations, susceptibility may increase. If memory repeatedly retrieves the same successful explanation, restoration may strengthen.

The relevant empirical comparison therefore manipulates not only whether memory exists but how it is written, indexed, deleted, and reopened. A perturbation that survives in memory but is never retrieved remains historically stored but generatively inactive. Conversely, a small memory item that repeatedly reorganizes search or tool use can become dynamically consequential. The distinction between stored persistence and relational persistence developed earlier is therefore especially important for artificial agents.

### 8.3 Anomaly Normalization and Closure Dynamics

Scientific discovery often depends on the treatment of anomalous evidence. Section 4 argued that an anomaly is not an intrinsic property of an observation but a relation between evidence and an existing representation. Section 5 then introduced epistemic damping and premature closure. These mechanisms provide a second route by which artificial systems may return to the known: an anomaly can be detected yet rapidly translated into an available explanation.

The scientific-discovery experiment of Ding and Li is relevant here because it separates background competence from anomaly-driven revision [5]. In their molecular-genetics task, the tested generative system could perform parts of the discovery process within known representations but did not reliably use anomalous experimental outcomes to initiate the kind of hypothesis revision observed in the human comparison. The study concerns one model, one task family, and a particular prompting protocol, so it does not justify a universal claim about AI. Its value for the present framework is narrower: it demonstrates how a system can possess substantial domain knowledge while still differing in the temporal treatment of anomaly.

One possible mechanism is rapid normalization. Let  $a_t$  denote an anomaly signal and let  $c_t$  denote the strength of an available closure or explanatory mapping. A schematic damping relation is

$$a_{t+1} = \lambda a_t - \gamma c_t a_t + \eta_t, \quad (22)$$

where  $0 < \lambda \leq 1$  represents persistence in the absence of closure,  $\gamma$  the effective damping induced by explanation, and  $\eta_t$  new evidence or contradiction. The equation is illustrative rather than cognitive or computationally literal. It expresses a testable idea: an explanatory completion can reduce the persistence of anomaly before alternative representations have had time to develop.

This mechanism also clarifies the paradox of extensive knowledge. A system with access to a very large explanatory repertoire may be able to assimilate unusual observations rapidly. That capability is often beneficial. It reduces redundant rediscovery and prevents every surprising input from destabilizing inquiry. Yet the same capacity may raise the threshold at which an observation remains unresolved long enough to reorganize the representation itself. The problem is therefore not “knowing too much” in any simple sense. It is whether explanatory abundance interacts with closure policies in a way that systematically shortens anomaly lifetime.

The relevant design variable is reopening. A system can produce a provisional explanation while preserving a record of unresolved residuals, counterevidence, or alternative framings. Such a system may achieve both coherence and historical openness. By contrast, if explanation removes the anomaly from later retrieval and planning, local competence can coexist with strong structural damping. This distinction is central to discovery-oriented design because maximal openness would be as dysfunctional as maximal closure.

#### 8.4 Historical Reconfiguration across System Boundaries

Claims that an artificial system “returns to the known” depend strongly on where the system boundary is drawn. A single LLM call has little persistent history. A research agent with long-term memory, retrieval, tools, notebooks, external databases, and human collaborators can possess extensive path dependence even if its foundation-model weights remain fixed. Conversely, an apparently autonomous agent may derive much of its stability from a human-designed orchestration layer that repeatedly selects familiar tasks, evaluates outputs with conventional criteria, or terminates exploration once a plausible answer is produced.

The effective generative system should therefore be represented broadly as

$$\mathcal{G}_t^{\text{AI}} = (F_\theta, M_t, Q_t, T_t, O_t, E_t), \quad (23)$$

where  $F_\theta$  is the foundation model,  $M_t$  memory,  $Q_t$  retrieval,  $T_t$  tools,  $O_t$  orchestration, and  $E_t$  the coupled environment. Return-to-known behavior can arise from any combination of these components. Likewise, persistent reconfiguration can occur outside the weights, for example through changed retrieval structure, altered memory organization, revised tool-selection policy, or newly preserved anomaly states.

This system-level formulation prevents two symmetrical errors. The first is to attribute all restoring dynamics to the pretrained model. The second is to treat any persistence in an agentic system as evidence that the generator itself has been reorganized. Stronger claims require localization. If a trajectory returns to a familiar region because retrieval repeatedly supplies the same documents, the relevant attractor may reside primarily in the retrieval layer. If the same recovery persists after retrieval is randomized, model-level structure becomes a stronger candidate. If a perturbation changes later tool use under identical probes while current memory contents are controlled, orchestration or latent state may be implicated.

The same boundary issue matters for human comparison. Human scientific inquiry is not an isolated biological generator. It includes notebooks, colleagues, instruments, institutions, literature, archives, and environments. Comparing a solitary model call to this distributed system would systematically exaggerate the difference. A fair comparison asks which coupled systems can preserve weak traces, reopen closed questions, diversify retrieval, and modify future perturbation pathways under matched resource constraints.

#### 8.5 Operational Criteria for Return Dynamics

The return-to-known hypothesis should be treated as a family of measurable claims rather than a single metaphor. At least four levels can be distinguished. *Output recovery* occurs when surface responses return to familiar forms after perturbation. *Representational recovery* occurs when concept relations, variable sets, or explanatory frames return to baseline. *Procedural recovery* occurs when search, tool use, or experimental planning returns to an incumbent workflow. *Generative recovery* is stronger: even after

a transient excursion, the system’s response to matched future probes remains effectively unchanged.

These levels can be tested through paired trajectories. Let two matched systems receive the same initial task, with one receiving a controlled perturbation  $\epsilon$ . After a recovery interval, both receive a common probe distribution  $Q$ . A return measure can be defined as

$$\mathcal{D}_{\text{return}} = \mathbb{E}_{q \sim Q} \left[ D \left( P_{t+k}^{(\epsilon)}(\cdot | q), P_{t+k}^{(0)}(\cdot | q) \right) \right], \quad (24)$$

with stronger recovery corresponding to smaller  $\mathcal{D}_{\text{return}}$  after the transient perturbation has dissipated. The same design can be applied at the level of retrieval paths, concept graphs, tool sequences, or hypothesis distributions rather than raw text.

The hypothesis is weakened if artificial systems routinely preserve small anomalies, reorganize later search around them, and retain altered response profiles under matched future inputs. It is also weakened if observed convergence disappears once retrieval, memory, or orchestration are controlled. Conversely, evidence for strong return dynamics would consist of systematic recovery across multiple levels despite substantial immediate novelty, together with identifiable restoring mechanisms.

The resulting picture is deliberately conditional. Current artificial systems may return to the known because of learned priors, retrieval concentration, experience-following memory, closure policies, or external orchestration. Different systems may exhibit different combinations, and future systems may alter these mechanisms substantially. The significance of the return-to-known problem is therefore not that it marks an immutable boundary between human and artificial discovery. It identifies a dynamical obstacle that any discovery-oriented system must sometimes overcome: local deviation must become sufficiently persistent and relationally consequential to change what later inquiry does with the same world. Section 9 turns from this diagnostic problem to the design conditions under which artificial systems might better support such transitions.

## 9 Conditions for Discovery-Oriented Artificial Systems

The preceding diagnosis does not imply that artificial systems should be made maximally unstable, maximally stochastic, or maximally resistant to closure. Scientific inquiry depends on the opposite capacities as well: reliable retrieval, cumulative memory, reproducible procedures, and the ability to consolidate explanations that survive criticism. The design problem is therefore not to remove restoring structure, but to make restoration selectively revisable. A discovery-oriented artificial system should be able to preserve enough coherence to accumulate knowledge while also retaining some disturbances that current representations do not yet accommodate. In the terminology developed above, the relevant target is *historical openness under constraint*: the system remains organized, but selected anomalies can persist long enough to alter later search, representation, and response.

This section develops that target as a set of provisional design conditions and experimental contrasts. The conditions are not presented as a recipe for scientific discovery. They specify mechanisms that would have to be distinguished empirically if one wants to test whether an artificial research system can move from local novelty toward persistent reorganization. The same mechanism can be beneficial in one regime and harmful in another. Anomaly preservation can protect a discovery opportunity or preserve noise; memory can enable delayed reactivation or deepen an error basin; relational revision can widen inquiry or dissolve useful structure; increased susceptibility can permit transition or destroy coherence. The central requirement is therefore regulation across time

Table 6: Provisional design conditions for discovery-oriented artificial systems and the failure modes they are intended to address. The conditions are hypotheses for experimental manipulation rather than requirements known to be sufficient for scientific discovery.

| Design condition          | Targeted dynamical problem                                                        | Operational indicator                                                                                 |
|---------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| Anomaly preservation      | Plausible explanation removes unresolved residuals too quickly                    | Residual anomaly state remains retrievable and affects later hypothesis generation                    |
| Reopenable closure        | Early completion prevents alternative representations from developing             | Closed explanations can be reopened by later evidence without reconstructing the problem from scratch |
| Weak-trace retention      | Low-salience observations disappear before their relevance becomes representable  | Earlier low-priority traces are selectively reactivated under later relational conditions             |
| Heterogeneous encounter   | Search remains concentrated in familiar conceptual neighborhoods                  | Later representations use relations or evidence classes absent from the incumbent search path         |
| Relational revision       | New evidence is added while the same concept graph or problem frame remains fixed | Concept, variable, evidence, or dependency structure changes persistently                             |
| Susceptibility regulation | The system is either over-restoring or incoherently exploratory                   | Perturbation gain increases while coherence and historical retention remain above task thresholds     |
| Consolidation             | Reorganization produces temporary divergence but no usable new regime             | New structure survives matched future probes and supports coherent later inquiry                      |

rather than maximalization of any single quantity.

## 9.1 Historical Openness under Provisional Closure

A discovery-oriented system needs a way to conclude provisionally without converting every conclusion into irreversible epistemic closure. Section 5 introduced reopenability as a property distinct from simple answer production. This distinction can be made architectural. Rather than representing an explanation as a terminal state, the system can preserve an associated record of unresolved residuals, alternative framings, evidential tensions, and conditions under which the explanation should be reconsidered. The resulting closure is functional: it permits action and cumulative work while preserving a route back to the problem if later evidence changes its status.

Let  $h_t$  denote a current hypothesis or explanation, and let  $r_t$  denote a residual state containing observations, relations, or counterevidence not fully absorbed by that hypothesis. A completion-oriented architecture tends to eliminate the residual when an explanation is accepted. A reopenable architecture instead preserves a structured residual after provisional closure. That residual records what remains unexplained, weakly explained, or dependent on contested assumptions. The point is not to preserve every discrepancy indefinitely. Residual information should be compressed, ranked, and periodically reevaluated so that explanatory completion does not erase the historical conditions that could later justify reopening the problem.

This design principle follows directly from the anomaly literature discussed in Section 4. Anomalous evidence can be ignored, rejected, held in abeyance, reinterpreted, accommodated, or used to support deeper change [3, 4]. A discovery-oriented artificial

system should therefore make its treatment of anomaly explicit enough to be manipulated and measured. If an observation is classified as noise, the system should retain the grounds for that classification; if it is accommodated by an auxiliary assumption, that assumption should remain available for later revision; if it is placed in abeyance, later evidence should be able to retrieve it.

A useful experimental contrast would compare immediate explanatory closure with delayed or reopenable closure under identical evidence. Both conditions can be permitted to produce a working answer. The difference lies in what happens later, when a second anomaly or a cross-domain relation appears. If the reopenable system reactivates the earlier residual and reorganizes the problem while the conventional system reconstructs the anomaly only after explicit prompting, there is evidence that closure policy affects historical availability rather than merely current verbosity.

This proposal is deliberately weaker than asking a model to “remain uncertain.” Generic uncertainty language can be added to an answer without changing later dynamics. The relevant variable is causal persistence: whether the unresolved structure survives in a form that changes subsequent retrieval, hypothesis generation, experiment selection, or problem decomposition. Discovery-oriented closure is therefore not stylistic hedging. It is a property of the system’s longitudinal state.

## 9.2 Weak Traces and Delayed Reactivation

Some potentially consequential observations are not recognizable as important when they first appear. Section 4 distinguished perturbation entry, recognition, and capture in time. That distinction has a direct design implication: memory systems optimized only for current relevance can systematically discard precisely the material whose relevance depends on future relations.

The problem can be stated as a retention policy under uncertain future value. Let  $e_t$  be an encountered item and let  $s_t(e_t)$  be its current salience score. A purely salience-thresholded memory writes the item only when

$$s_t(e_t) > \theta_s. \quad (25)$$

Such a rule is efficient, but it creates a structural blind spot. If the future importance of  $e_t$  depends on a relation  $R_{t+j}$  that does not yet exist, then current salience is an incomplete proxy for future generative consequence. A discovery-oriented memory may therefore require a low-cost channel for retaining weak traces whose immediate relevance is uncertain.

This does not imply storing all encountered information. A more plausible architecture would use heterogeneous retention times and representations. High-confidence evidence can enter durable structured memory; low-salience anomalies can enter compressed episodic storage; discarded material can leave statistical traces in indexes or novelty maps; unresolved relations can be periodically resampled. The design objective is not exhaustive memory but delayed causal availability.

The experience-following effects reported in agent-memory research illustrate why this issue is two-sided [19]. Retrieved experiences can strongly shape later behavior, which means memory is capable of creating substantial historical dependence. The same mechanism can reinforce a successful but narrow trajectory or propagate an earlier error. Discovery-oriented memory should therefore be evaluated by the diversity and conditionality of its reactivation, not by persistence alone. A weak trace that is never retrieved is inert; a trace retrieved on every superficially similar task can become an additional restoring force.

One experimental design is a delayed-reactivation task. A system encounters a low-salience observation  $e_1$  that has no immediate instrumental value. Many steps later it receives  $e_2$ , which makes a relation between  $e_1$  and the current problem newly informative. The primary outcome is not explicit recall of  $e_1$ . It is whether the earlier trace changes later hypothesis space, variable selection, or search policy under conditions in which the connection was not specified in advance. A control condition can provide the same current evidence without the earlier encounter. The difference measures whether weak historical traces can become generatively active after their significance changes.

This criterion also sharpens the comparison between human and artificial research systems. Humans forget extensively; they do not preserve a complete record of experience. The relevant comparison is therefore not biological memory versus perfect machine storage. It is whether a coupled system can preserve enough low-confidence structure to support later reactivation while still filtering most irrelevant variation. That is a problem of memory governance rather than memory size.

### 9.3 Heterogeneous Encounter and Relational Revision

Discovery can require more than retaining anomalies within a fixed representation. It can require encounters that expose the system to relations its incumbent search policy would rarely produce. Increasing sampling temperature is one way to diversify local outputs, but Section 6 argued that stochasticity is not equivalent to regime change. A stronger intervention alters the distribution of what the system can encounter and how those encounters are allowed to modify its relational organization.

For an artificial research system, heterogeneous encounter can be introduced through multiple sources: deliberately diverse retrieval, cross-domain corpora, tool-mediated observations, adversarial evidence search, alternative model families, simulated experiments, human collaborators, or agents operating under different representational constraints. The relevant property is not source count but relational heterogeneity. Ten retrievals from the same conceptual neighborhood may add little. One observation from a different measurement regime can expose a hidden assumption in the incumbent representation.

Let  $\mathcal{E}_t$  denote the distribution of encountered evidence or candidate relations. A conventional search policy may adapt  $\mathcal{E}_t$  toward regions expected to improve the current objective. Discovery-oriented exploration sometimes requires a countervailing term that samples outside this high-utility neighborhood:

$$\mathcal{E}_{t+1} = (1 - \alpha_t) \mathcal{E}_t^{\text{exploit}} + \alpha_t \mathcal{E}_t^{\text{hetero}}, \quad (26)$$

where  $\alpha_t$  is not fixed. When the current representation explains new evidence well, exploitation can dominate. When unresolved residuals accumulate, relational diversity can increase. The equation is schematic, but it makes the design problem explicit: exploration should respond to epistemic state rather than remain a constant randomness parameter.

Exposure alone remains insufficient. The stronger condition is relational revision. Suppose the system maintains a structured representation

$$\mathcal{M}_t = (C_t, V_t, E_t^{\text{rel}}, \Pi_t), \quad (27)$$

where  $C_t$  denotes concepts,  $V_t$  variables,  $E_t^{\text{rel}}$  evidential or dependency relations, and  $\Pi_t$  interpretation or inference policies. A new observation has produced stronger reorganization when it changes this structure rather than being appended as another item. Such

change can include splitting a concept, introducing a previously absent variable, reversing an assumed dependency, changing evidential status, or altering which experiment would discriminate among hypotheses.

This distinction makes the idea-generation results discussed earlier easier to interpret. A system can generate a research proposal that expert raters find novel while leaving its underlying representation unchanged [15]. Conversely, a modest change in representation can later generate many ordinary-looking but scientifically important consequences. Discovery-oriented evaluation should therefore track structural revision directly instead of inferring it from the novelty of prose.

## 9.4 Susceptibility Regulation and Critical-Zone Interventions

The preceding mechanisms create a further control problem. Preserving anomalies, widening encounter, and allowing relational revision can increase susceptibility, but excessive susceptibility can produce incoherent search. Section 6 therefore defined a structured-susceptibility regime in which perturbation gain rises while coherence and historical retention remain above functional thresholds. A discovery-oriented system should ideally regulate its position relative to such a regime rather than maximize entropy or exploration.

Let  $\hat{\mu}_t$  denote an estimated regime state based on measurable quantities such as anomaly persistence, representational recovery time, hypothesis diversity, relational change, coherence, and task performance. A regulation policy can use this estimate together with unresolved anomaly, current coherence, and retained history to decide whether to widen exploration, preserve tension, or consolidate. The policy itself should be specified by the experiment rather than represented by an otherwise unconstrained abstract function.

A minimal experiment can compare at least four conditions: standard completion, increased decoding stochasticity, delayed-closure prompting, and adaptive susceptibility regulation. The critical prediction is not that the regulated condition generates the most unusual text. It is that controlled perturbations produce greater persistent change in later representation and search under matched future probes, without a corresponding collapse in task coherence. If temperature alone yields the same effect, the critical-zone hypothesis is unnecessary. If delayed closure increases diversity but not historical persistence, the intervention changes exploration rather than regime structure. If adaptive regulation improves both persistence and consolidation, the stronger dynamical interpretation gains support.

The same design should include a return phase. A useful discovery system must be capable of leaving a stable organization and later establishing another. Permanent destabilization is not discovery. The regulation problem is therefore cyclic. A research system may begin from a relatively restored organization, accumulate unresolved problems, become more susceptible to revision, reorganize its representation, and then consolidate a new working structure. The sequence is descriptive rather than universal: stages may overlap, repeat, or be externally scaffolded. The sequence is a hypothesis about useful control structure, not a claim that all discoveries follow these stages.

## 9.5 Consolidation and Common-Future-Input Tests

The strongest design objective is not sustained divergence but durable change in the effective generator. A system that produces a radical hypothesis in one episode and then immediately returns to its previous response surface has generated novelty without strong historical consequence. Conversely, a perturbation that changes later re-

trieval, hypothesis formation, or experiment choice under matched conditions provides stronger evidence of reconfiguration.

Consolidation therefore needs to be treated as a distinct phase. Once a candidate representation has survived exploratory challenge, the system can compress it into a form that supports stable inference while retaining provenance, residual anomalies, and conditions for reopening. Consolidation can include revised concept graphs, durable memory changes, updated retrieval indexes, changed tool policies, or externally stored research artifacts. Weight updates are one possible mechanism but are not required for effective generator change.

A common-future-input test provides a direct experimental criterion. Let two matched systems receive histories  $H^{(a)}$  and  $H^{(b)}$ , with  $H^{(b)}$  containing a controlled sequence of perturbations and relational revisions. Both systems are then exposed to the same future probe distribution  $Q$ . Define

$$\Delta_G \left( H^{(a)}, H^{(b)}; Q \right) = \mathbb{E}_{q \sim Q} \left[ D_Y \left( P^{(a)}(\cdot | q), P^{(b)}(\cdot | q) \right) \right]. \quad (28)$$

A persistent nonzero value is not sufficient evidence for generator reconfiguration because hidden memory content or current state may still differ. Stronger tests progressively control those factors and ask whether the historical intervention changed a stable transition rule, retrieval policy, representation, or orchestration process.

The test should be extended beyond textual outputs. The future probes can measure anomaly sensitivity, evidence ranking, concept decomposition, hypothesis distributions, experiment selection, retrieval paths, or tool sequences. If the system has genuinely reorganized, the historical difference should appear across a coherent family of later behaviors rather than only in one wording choice.

Consolidation also introduces an epistemic constraint that the dynamical framework cannot solve by itself. A persistent new generator can be wrong. A hallucinated causal structure, a conspiracy framework, or a systematically biased model can satisfy dynamical criteria for reorganization. Discovery-oriented systems therefore require independent validation through evidence, reproducibility, predictive performance, robustness, and where relevant human or institutional review. The present framework concerns the conditions under which reorganization can occur and persist; it does not replace standards of justification.

## 9.6 Experimental Programme and Falsification Structure

The proposed conditions can be tested through a factorial programme rather than a single creativity benchmark. A minimal study can begin from matched system configurations and manipulate six factors: decoding stochasticity, anomaly-retention policy, memory treatment of low-salience traces, encounter heterogeneity, relational-revision requirements, and susceptibility regulation. The tasks should include both domains with well-defined evidence and open-ended research settings, because a mechanism that appears useful only when evaluation is subjective may reflect stylistic divergence rather than discovery dynamics.

Primary outcomes should be longitudinal. Immediate originality can remain a secondary measure, but the central quantities are recovery time after perturbation, persistence of altered representations, delayed reactivation of weak traces, breadth of relational cascade, reopening after prior closure, common-future-input divergence, and the quality of subsequent consolidation. The Ding and Li discovery setting provides one example of why anomaly-driven revision should be measured separately from background competence [5]; the memory results of Xiong et al. show why persistent history

must be separated from beneficial reconfiguration [19]; and the convergence observed in autonomous generation loops motivates measurements across repeated dynamics rather than isolated outputs [9].

The hierarchy of claims should remain explicit. If increased stochasticity alone produces durable representational change, then the stronger susceptibility-regulation account is unnecessary for that setting. If delayed closure improves later revision but weak-trace retention adds nothing, memory-based historical availability can be dropped from the explanation. If fixed nonlinear dynamics explain all cascade behavior, epistemic turbulence should be withdrawn. If common-future-input differences disappear when explicit memory and retrieval are matched, there is weak evidence for generator reconfiguration. Each stronger concept is justified only by residual phenomena that simpler mechanisms fail to explain.

This structure turns discovery-oriented design into a sequence of discriminable questions rather than a single objective called “creativity.” Can unresolved evidence survive plausible explanation? Can low-salience traces become relevant later? Can heterogeneous encounters revise the representation rather than merely expand the output set? Can susceptibility increase without loss of coherence? Can a reorganized state later consolidate? Can the historical intervention change the response surface under matched future conditions? A system need not satisfy every condition in every task. The purpose of the framework is to make clear which transition is being tested when an artificial system appears to move beyond the known.

## **10 Discussion**

The preceding analysis reframes the boundary between generative novelty and scientific discovery as a problem of dynamical organization rather than output rarity alone. The central proposal is deliberately modest. Contemporary artificial systems can generate unusual, useful, and expert-rated ideas, and nothing in the present framework implies that biological agents possess an exclusive capacity for theoretical innovation. The narrower claim is that novelty, anomaly response, persistent reorganization, and discovery belong to different analytical levels. A system can vary extensively while its effective organization remains stable; it can also undergo a comparatively small revision that later alters what it notices, retrieves, tests, and regards as explanatory. This section draws out the implications of that distinction, clarifies the status of the stronger dynamical concepts introduced in Sections 6–7, and identifies the limits of a discovery account based on attractors and perturbation dynamics.

### **10.1 Discovery beyond Maximal Novelty**

The first implication concerns the target of evaluation. Much contemporary work on machine creativity necessarily begins with outputs: originality ratings, semantic distance, diversity, usefulness, or expert judgments of proposed ideas. These measures are valuable because discovery has an observable expressive surface. They nevertheless underdetermine whether the system that produced the output has changed in a way that matters for later inquiry. The distinction introduced in Section 2 can therefore be restated as a hierarchy of increasingly demanding claims. A novel output is the weakest claim. Representational revision requires evidence that the problem has been organized differently. Persistent reorganization additionally requires that the change survive into later inquiry. Epistemically warranted discovery is stronger still because the reorganization must also survive domain-appropriate evidential evaluation. This ordering concerns

structural and epistemic commitment rather than scientific worth.

This hierarchy helps reconcile apparently conflicting empirical results. A language model can propose ideas that human experts judge highly novel [15] while still showing recovery toward familiar explanatory regions in repeated dynamics [9]. Likewise, a system can perform well on broad ideational tasks while encountering difficulty when discovery requires preserving an anomalous result long enough to revise an incumbent hypothesis space [5]. These observations need not be interpreted as contradictions. They may simply concern different levels of change.

The practical consequence is that evaluation should not ask only whether an output lies far from a comparison set. It should also ask what the output changes. Does it alter the variables that are subsequently considered relevant? Does it change how evidence is partitioned? Does it create new discriminating experiments? Does it modify later retrieval and hypothesis generation? Does the system continue to behave differently under matched future probes? These questions shift attention from the originality of a product to the historical consequences of a perturbation.

The distinction is also important for human science. Scientific history contains many cases in which later-significant revisions were initially modest, technically local, or only retrospectively recognized as reorganizing. Conversely, highly imaginative proposals can remain peripheral to later inquiry. The present framework therefore does not oppose machine novelty to an idealized human discovery process. It proposes a common analytical vocabulary for asking when any research system moves from exploration inside an established organization to alteration of the organization itself.

## **10.2 Structured Susceptibility and Regulated Openness**

The second implication concerns the relation between stability and discovery. If strong restoring dynamics absorb every perturbation, a system can become highly competent within an established representation while remaining difficult to redirect. If restoring structure is weakened indiscriminately, however, the result may be incoherence rather than discovery. Section 6 therefore proposed structured susceptibility as a regime in which perturbation gain is elevated while enough coherence and historical retention remain for consequences to accumulate.

This position is intentionally weaker than a literal edge-of-chaos thesis. The language of metastability and near-criticality is useful because it directs attention to regime-dependent sensitivity, competing organizations, slowing recovery, and transition boundaries [14, 12, 8]. It does not establish that scientific discovery occupies a unique physical critical point. In finite, heterogeneous research systems, the relevant transition region may be broad, multidimensional, and task dependent. Non-normal amplification, finite-amplitude basin crossing, memory-mediated history, or external scaffolding can also generate strong perturbation effects away from any conventional critical point.

The more defensible conclusion is therefore regulatory. Discovery-oriented systems may benefit from mechanisms that adjust restoration and openness in response to epistemic conditions. Stable periods support accumulation, compression, and reliable inference. Persistent residuals, unresolved contradictions, or repeated failure can justify wider retrieval, delayed closure, alternative decompositions, or increased model diversity. Successful reorganization can then be followed by consolidation. The relevant control problem is not to maximize openness but to coordinate exploration and stabilization over time.

This interpretation also changes how stochasticity should be understood. Sampling temperature, random seeds, and heterogeneous generation policies can widen local ex-

ploration, but they do not determine where the effective system lies in a regime space. A model can be highly stochastic while repeatedly reconstructing familiar explanations. Conversely, a low-amplitude anomaly can have a large effect if it enters a relation on which the current representation strongly depends. The design target is therefore not variability by itself but conditional susceptibility to structured perturbation.

A useful future direction is to estimate this susceptibility directly. Rather than infer regime position from entropy or diversity alone, experiments can measure perturbation-response curves, recovery times, reopening probabilities, persistence of revised relations, and the degree to which earlier perturbations change later sensitivity. If a system can increase these quantities when unresolved anomalies accumulate and reduce them after successful consolidation, the notion of criticality regulation becomes empirically meaningful without requiring a literal thermodynamic analogue.

### 10.3 Return Dynamics as a System-Level Property

The third implication concerns the phrase “return to the known.” It is tempting to attribute recovery toward familiar explanations to the foundation model alone. Section 8 argued that this inference is generally too coarse. The effective research system can include a learned generator, retrieval engine, persistent memory, external databases, tools, prompts, agent policies, human interventions, and institutional constraints. Each component can add restoring forces or create alternative routes of exploration.

A return-to-known effect should therefore be localized rather than assumed. Repeated convergence can arise from the learned prior of the model, but it can also result from retrieval that repeatedly surfaces canonical sources, memory that replays earlier successful strategies, benchmarks that reward conventional decompositions, or orchestration rules that terminate exploration once a plausible explanation appears. Conversely, an apparently adventurous model can be embedded in a strongly conservative system if downstream filters discard unfamiliar hypotheses. The natural unit of analysis is the coupled generative arrangement under study.

This system-level perspective is especially important when comparing artificial and human research. A working scientist is not a stateless biological model. Scientific reasoning is distributed across notebooks, instruments, archives, collaborators, disciplinary vocabularies, software, institutions, and accumulated material practices. Artificial systems increasingly possess analogous external supports. A fair comparison should therefore match functional boundaries as closely as possible rather than contrast a socially embedded human with a single-turn language model.

The same point complicates claims about memory. Persistent memory can reduce forgetting and thereby increase historical dependence, but historical dependence is not equivalent to productive reorganization. Experience-following and error propagation in agent memory show that preserved history can narrow later search as readily as it can enable discovery [19]. The important variable is how stored history is reactivated, revised, weighted, and connected to current anomalies. Memory can deepen an attractor as well as weaken one.

This suggests a broader design principle: discovery-oriented systems should expose their restoring structure to measurement. If a system repeatedly returns to the same explanatory neighborhood, one should be able to test whether the restoration is produced by model priors, retrieval, memory, evaluation policy, or some interaction among them. Such decomposition is more informative than assigning a global creativity score to the system and is more actionable for engineering interventions.

## 10.4 Interpretive Status of Epistemic Turbulence

The turbulence analogy is the strongest and most provisional component of the framework. Its value depends on disciplined use. Physical turbulence involves mathematically specific properties that have no established epistemic counterpart, including energy transfer across scales and dynamics governed by fluid equations [6]. The present paper therefore does not propose an epistemic Reynolds number, a conservation-law analogue, or an equation-level identification between thought and fluid flow.

The analogy is retained because it directs attention to a narrower structural possibility: perturbations may interact nonlinearly across scales while altering the pathways through which later perturbations propagate. A local anomaly can change a conceptual relation; the revised relation can change a hypothesis family; the changed hypothesis family can alter retrieval, experimentation, and later interpretation of local evidence. If this process becomes recurrent, intermittent, historically persistent, and self-modifying, a simple one-off cascade description may become insufficient.

Even then, turbulence should remain a defeasible interpretation. Fixed nonlinear dynamics can produce large amplification without pathway modification. Metastable systems can switch among pre-existing organizations. Chaotic divergence can generate strong trajectory separation without meaningful historical retention. Memory accumulation can create path dependence without multiscale interaction. Watts-style cascades demonstrate how small disturbances can have large network effects under fixed propagation rules [18]; such effects are already powerful and should be preferred whenever they explain the observations.

The epistemic-turbulence hypothesis earns additional explanatory work only if the propagation structure itself changes in a history-dependent manner. Operationally, this means demonstrating not merely that a perturbation spreads, but that its spread modifies later coupling, salience, or transition probabilities. The strongest evidence would come from controlled histories followed by matched perturbations that propagate differently because of those histories. If simpler models reproduce the observed dynamics, the turbulence interpretation should be withdrawn.

This conservative stance is not a weakness of the framework. It creates a hierarchy in which stronger language corresponds to stronger empirical obligations. Attractor recovery can be tested without turbulence. Structured susceptibility can be tested without criticality. Generator reconfiguration can be studied without claiming discovery. Each concept can survive or fail independently, allowing the research programme to contract toward simpler explanations when evidence requires it.

## 10.5 Dynamical Reorganization and Epistemic Evaluation

The most important limitation is that dynamical transformation is not epistemic success. Escaping an incumbent attractor can produce a more adequate theory, but it can also produce a hallucinated causal structure, a self-reinforcing error, an ideological system, or an unproductive research programme. Novelty, persistence, and internal coherence do not supply truth conditions.

Dynamical reorganization and epistemic evaluation should therefore be treated as separate dimensions. Greater departure from an incumbent organization does not in general imply greater epistemic value. Epistemic evaluation requires standards that depend on domain and practice: evidential support, predictive success, explanatory adequacy, reproducibility, robustness, calibration, causal discrimination, and appropriate forms of expert or institutional scrutiny.

This separation is particularly important for AI systems because mechanisms designed to weaken closure can also increase hallucination and speculative coherence.

A system encouraged to preserve anomalies and generate alternative representations may become better at leaving familiar explanatory regions while becoming worse at distinguishing productive alternatives from unsupported ones. Discovery-oriented design therefore requires a second layer that evaluates candidate reorganizations against evidence rather than rewarding structural departure itself.

The relation between these layers should also be temporal. Immediate rejection of weakly supported alternatives can recreate premature closure, while indefinite suspension prevents consolidation. A plausible architecture would allow hypotheses to remain provisionally active under explicit uncertainty, seek discriminating evidence, and then strengthen, revise, or discard them according to empirical performance. This is one reason the framework emphasizes reopening and consolidation rather than permanent destabilization.

This distinction also limits the philosophical reading of “the edge of the known.” The edge is not automatically the location of truth. It is the region where current representational resources become less secure and where reorganization may become possible. Scientific discovery requires that some reorganizations later survive independent epistemic tests. The dynamical framework describes conditions of transformation; it does not decide which transformed structure deserves acceptance.

## **10.6 Research Programme and Comparative Implications**

Taken together, the framework suggests a shift from static creativity benchmarks toward longitudinal perturbation experiments. The central comparison is no longer simply whether humans or artificial systems generate more original answers. It is whether a research system can preserve unresolved evidence, reopen prior closure, retain weak traces, amplify selected perturbations across levels, revise propagation pathways, and later consolidate a new organization that remains effective under matched probes.

This programme allows human and artificial systems to be compared without presupposing a metaphysical boundary between them. Human inquiry may exhibit rich historical dependence because biological memory, material environments, social interaction, and institutional practices continually reshape the effective generator. Artificial systems can increasingly acquire analogous forms of historical embedding through persistent memory, tool use, external archives, multi-agent interaction, and long-running research environments. Whether these mechanisms produce similar or different dynamical regimes is an empirical question.

At the same time, current artificial systems provide unusually useful experimental objects. Their prompts, memories, retrieval policies, temperatures, tools, and interaction histories can be manipulated in ways that are difficult or impossible in human research. The critical-zone prompting proposal in Section 9 should be understood in this methodological spirit. If delayed closure, explicit anomaly preservation, heterogeneous retrieval, and relational-revision requirements produce persistent changes under matched future probes, the intervention would support a stronger regime-based account. If ordinary stochastic sampling yields the same results, the additional machinery should be rejected.

Several alternative explanations should remain active throughout this programme. Apparent discovery can reflect contamination from training data, retrieval of an existing solution, stylistic novelty, evaluator bias, hidden external scaffolding, or simple accumulation of context. Apparent attractor recovery can reflect task wording, decoding choices, or evaluation artifacts rather than a stable knowledge organization. Apparent generator change can reflect explicit memory content rather than revised transition structure. Strong claims therefore require matched controls, withheld evidence, prove-

nance tracking, repeated trajectories, and where possible preregistered intervention criteria.

The broader contribution of the framework is consequently methodological rather than definitive. It supplies a sequence of distinctions that can be tested separately. These include novelty versus reorganization, anomaly versus capture, memory versus generative history, stochasticity versus susceptibility, cascade versus turbulence, and dynamical transition versus epistemic success. These distinctions make it possible to ask what kind of change has actually occurred when an artificial system appears to move beyond its prior knowledge organization.

The phrase *at the edge of the known* can therefore retain its dual role without collapsing into a single thesis. Epistemically, it marks situations in which available concepts and explanations become inadequate to organize new evidence. Dynamically, it names a possible region in which restoring tendencies weaken enough for selected perturbations to have persistent effects. Scientific discovery may sometimes involve both conditions, but their relation remains an empirical and philosophical problem rather than an established identity. The value of the proposed framework lies in making that relation more precise, more measurable, and more open to falsification.

## 11 Conclusion

This paper has examined a narrow problem at the intersection of artificial intelligence, creativity research, and the philosophy of scientific discovery: how a generative system can move from producing novelty within an established organization of knowledge to changing the relations through which later inquiry is conducted. The distinction matters because contemporary artificial systems can generate unusual, diverse, and sometimes highly rated ideas while still leaving open a different question: whether those outputs alter what the system later treats as a relevant variable, a meaningful anomaly, an admissible explanation, or a promising direction of search. The paper has therefore treated novelty as an insufficient proxy for discovery and has focused instead on the dynamics by which a perturbation can become historically consequential for subsequent generation.

The proposed framework begins with a hierarchy of changes. Output novelty can occur while the effective representational and generative organization remains stable. Relational reorganization requires a stronger change in the connections among concepts, evidence, questions, or hypotheses. Representational change concerns the structure through which a problem is formulated. Generative reorganization is stronger still: it concerns a persistent alteration in the effective transition structure that shapes later inquiry. Scientific discovery can involve any of these levels, but it cannot be inferred from them mechanically because discovery also requires epistemic evaluation. A system can move far from an incumbent representation and still move toward error.

The language of *knowledge attractors* was introduced to describe one possible source of stability in this hierarchy. The relevant claim is not that frequent outputs or dominant concepts are automatically attractors. An attractor interpretation becomes informative only when a system displays recovery after perturbation, recurrent return to a region of conceptual organization, or other evidence of restoring dynamics under controlled conditions. This makes the central AI question more precise. A generative system may be capable of producing rare outputs while retaining strong restoring tendencies at a higher level of organization. What looks like departure in a single response can therefore coexist with recovery in later reasoning, retrieval, or problem framing.

Contingency enters this picture through perturbation rather than through random-

ness alone. Random sampling can widen immediate variation, but the downstream consequence of a perturbation depends on the relational configuration into which it enters. The same event can be absorbed under one organization and become destabilizing under another. For this reason, the paper distinguished surprise from anomaly, anomaly from capture, and capture from historical consequence. A perturbation is generatively important only when some effect of it survives long enough to influence later inquiry. That survival can be explicit, as when an anomaly is deliberately preserved, or latent, as when a weak trace remains available for later reactivation before its significance is fully understood.

The concept of *epistemic damping* was introduced to describe the opposite process. Generative systems require mechanisms that suppress irrelevant variation, compress evidence, and restore coherent representations. Damping is therefore necessary rather than intrinsically defective. The potential problem is premature closure: stabilization that occurs before alternative representations, delayed evidence, or unresolved anomalies have had sufficient opportunity to develop. This temporal formulation avoids treating openness as universally desirable. A discovery-oriented system must both resist some restoring forces and retain enough organization to evaluate and consolidate what emerges.

This requirement motivated the notion of *structured susceptibility*. The phrase *the edge of the known* names, at the dynamical level, a hypothesized regime in which perturbations can acquire greater persistence and gain while coherence and historical retention remain sufficient for cumulative inquiry. Metastability and near-criticality were introduced as possible resources for describing such regimes, but the paper has not claimed that discovery occurs at a unique physical critical point. Increased decoding temperature is likewise not equivalent to movement toward such a regime. Stochasticity changes variation; regime position concerns the structure of restoration, coupling, memory, and transition.

Only after these distinctions were established did the paper introduce *epistemic turbulence*. The term designates the strongest and most provisional part of the framework. A large cascade is not yet turbulence. Neither are ordinary metastable switching, chaotic divergence, or repeated memory effects. The turbulence-like interpretation becomes relevant only if perturbations interact across levels, exhibit intermittent and historically persistent propagation, and alter the pathways through which later perturbations themselves travel. In that stronger case, the propagation medium is no longer fixed: prior history changes subsequent coupling. No fluid equation, conservation law, or literal physical equivalence is assumed. The concept should be abandoned whenever simpler dynamical explanations are sufficient.

Applied to artificial generative systems, this framework suggests that a possible “return to the known” should be decomposed rather than attributed to a single property of a language model. Restoring tendencies can arise from learned model priors, retrieval policies, persistent memory, evaluation procedures, system prompts, tool use, or orchestration. Memory can preserve anomalies but can also deepen existing attractors through experience-following and repeated retrieval. External tools can widen the encounter space while simultaneously narrowing what is subsequently considered relevant. The proper object of comparison is therefore often a coupled system consisting of model, memory, tools, environment, and interaction history rather than an isolated model call.

The design implications are correspondingly conditional. A discovery-oriented artificial system may require mechanisms for preserving unresolved anomalies, retaining weak traces whose importance is not yet known, reopening prior closure, encountering heterogeneous evidence, revising relations among concepts, and regulating the balance between exploration and consolidation. These mechanisms do not guarantee discov-

ery. They specify candidate conditions under which perturbations may remain available long enough to test whether an incumbent knowledge organization should change. The stronger claim that such conditions can induce a turbulence-like epistemic regime remains experimental.

The empirical programme therefore places trajectory-level tests above static creativity scores. Matched perturbations can test restoring structure, while controlled histories can test whether later probes are processed differently because of earlier events. Common-future-input designs can distinguish temporary state differences from persistent changes in effective generation. Critical-zone interventions can compare stochastic diversity with delayed closure, anomaly preservation, relational revision, and adaptive susceptibility regulation. These experiments can falsify the framework layer by layer. If simple sampling explains the observed novelty, stronger dynamical concepts are unnecessary. If fixed nonlinear dynamics explain amplification, generator reconfiguration need not be invoked. If memory content explains path dependence, epistemic turbulence should not be claimed.

The broader implication is methodological. Evaluating artificial scientific discovery requires attention not only to what a system produces, but also to what kinds of disturbance it can preserve, how it returns after disturbance, which relations can be revised, and whether earlier encounters change the conditions of later inquiry. This shifts the question from whether an AI can generate a surprising idea to whether a knowledge-generating system can remain coherent while allowing selected surprises to become structurally consequential.

The framework remains preliminary. Knowledge attractors may prove useful in some experimental settings and misleading in others. Near-critical language may add explanatory value in one class of systems and collapse into ordinary nonlinear response in another. Epistemic turbulence may turn out to be unnecessary. These possibilities are compatible with the central purpose of the paper, which is to separate levels of change that are often conflated and to state stronger hypotheses in forms that permit empirical withdrawal.

At the edge of the known, then, the relevant problem is not simply how to produce something different. It is how a system can preserve enough stability to inquire while remaining capable of revising the structures that make some questions, anomalies, and explanations available in the first place. Whether present or future artificial systems can reliably inhabit and regulate such conditions is still open. That question, rather than a general claim about the presence or absence of machine creativity, is the principal research problem left by this paper.

## References

- [1] Marc Avila, Dwight Barkley, and Björn Hof. Transition to turbulence in pipe flow. *Annual Review of Fluid Mechanics*, 55:575–602, 2023.
- [2] Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2 edition, 2004.
- [3] Clark A. Chinn and William F. Brewer. The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63(1):1–49, 1993.
- [4] Clark A. Chinn and William F. Brewer. An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, 35(6):623–654, 1998.

- [5] Amy Wenxuan Ding and Shibo Li. Generative AI lacks the human creativity to achieve scientific discovery from scratch. *Scientific Reports*, 15:9587, 2025.
- [6] Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, 1995.
- [7] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.
- [8] Keith B. Hengen and Woodrow L. Shew. Is criticality a unified setpoint of brain function? *Neuron*, 113(16):2582–2598.e2, 2025.
- [9] Arend Hintze, Frida Proschinger Åström, and Jory Schossau. Autonomous language-image generation loops converge to generic visual motifs. *Patterns*, 7(1):101451, 2026.
- [10] Björn Hof. Directed percolation and the transition to turbulence. *Nature Reviews Physics*, 5:62–72, 2023.
- [11] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982.
- [12] J. A. Scott Kelso. Multistability and metastability: Understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):906–918, 2012.
- [13] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 4 edition, 2012.
- [14] Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.
- [15] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 2025.
- [16] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Chapman and Hall/CRC, 3 edition, 2024.
- [17] Dawei Wang, Difang Huang, Haipeng Shen, and Brian Uzzi. A large-scale comparison of divergent creativity in humans and large language models. *Nature Human Behaviour*, 10:531–540, 2026.
- [18] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5766–5771, 2002.
- [19] Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–645. Association for Computational Linguistics, 2026.