

Prompting Toward Criticality: A Hypothesis for Fostering Creativity in Large Language Models

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

Large language models can readily produce diverse and unusual outputs, yet increased variation does not by itself explain when a system becomes more capable of creative thought. This paper develops a preliminary hypothesis that prompt design may alter the effective dynamical regime of inference and, under some conditions, move a model toward a near-critical region in which small perturbations, anomalies, or competing representations have larger and more persistent effects while coherence remains sufficiently preserved. The proposal does not treat criticality as a proven property of large language models, nor does it assume that higher randomness, temperature, or output diversity is equivalent to creativity. Instead, it asks whether prompting can systematically weaken excessive restoration toward familiar representations, delay premature closure, preserve unresolved alternatives, and increase the probability that local differences influence later reasoning. Creativity is treated here in a deliberately broad but operational sense: not merely as unusual wording, but as the capacity to develop new relations, problem framings, candidate explanations, or lines of inquiry that remain consequential across subsequent inference. The paper outlines candidate prompt mechanisms, measurable perturbation-response signatures, matched-trajectory experiments, and controls that distinguish near-critical regime change from ordinary stochastic variation. It also considers epistemic turbulence as a stronger downstream possibility, to be invoked only if perturbations interact nonlinearly, persist historically, and modify later propagation pathways. The resulting framework is offered as a falsifiable research programme for testing whether prompts can do more than request creativity: whether they can alter the conditions under which creative change becomes easier to sustain.

Keywords: large language models; prompting; creativity; criticality; near-critical dynamics; inference; anomaly retention; metastability; perturbation response; epistemic turbulence

Hypothesis Paper Note

This paper develops a narrow and testable hypothesis about prompt design in large language models. Its central claim is not that prompting can directly “make a model creative,” nor that criticality has already been demonstrated in language-model inference. The paper asks whether certain prompt structures can alter the effective dynamical regime of inference so that small perturbations, anomalies, or competing representations become more likely to persist and influence later reasoning while coherence remains sufficiently preserved for useful inquiry. The motivating possibility is that creativity may become easier to sustain when inference is neither strongly restored toward familiar representations nor destabilized into unstructured variation.

The intervention variable is prompting. The proposed mechanism is movement toward a more susceptible inference regime. The target phenomenon is creativity, understood in an operational rather than metaphysical sense. The stronger downstream hypothesis is that, under some conditions, persistent nonlinear interaction among perturbations could produce turbulence-like epistemic dynamics. These four levels must remain distinct throughout the paper. A successful prompt need not produce literal criticality; a near-critical regime need not produce creativity; creative outputs need not imply reorganization of inference; and turbulence-like dynamics should be invoked only when simpler explanations are insufficient.

The paper should begin from a practical puzzle. Existing prompting strategies often ask a model to “be creative,” to generate alternatives, to adopt unusual perspectives, or to increase diversity. Such interventions can change output content, but they do not by themselves establish that the effective inference regime has changed. Higher decoding temperature likewise broadens sampling but does not necessarily weaken restoring tendencies toward familiar conceptual organizations. The proposed research problem is therefore not whether one can elicit more unusual answers, but whether prompt design can systematically modify the way inference responds to perturbation.

The term *criticality* must be used conservatively. The paper does not assume that language-model inference possesses a unique thermodynamic or neural critical point. It uses criticality as an operational hypothesis concerning a transition region in which perturbation sensitivity increases while sufficient organization remains for structured accumulation. The relevant empirical question is whether prompt-induced changes produce a measurable shift in response geometry: longer persistence of small perturbations, reduced immediate recovery toward incumbent representations, greater interaction among competing frames, increased path dependence, and preserved coherence over a finite horizon.

A key distinction is between stochasticity and regime position. Randomness can be increased externally through sampling temperature, nucleus sampling, random seeds, or injected noise. None of these interventions alone shows that the system has moved toward a critical regime. A model can produce more varied tokens while retaining the same restoring organization at the level of concepts, problem framing, or search. The experimental design must therefore include a high-stochasticity control. If increased temperature alone produces the same persistence, sensitivity, and reorganization as the proposed criticality-targeting prompts, then the stronger regime-shift interpretation should be weakened or withdrawn.

The creativity outcome must also be defined carefully. The paper should not rely on a single originality score. Creativity can be operationalized through a family of outcomes: generation of nontrivial relations among previously separated concepts; reformulation of the problem rather than mere variation of answers; persistence of new explanatory structures across later prompts; production of alternative hypotheses that remain usable after evidence updates; and transfer of a newly formed relation to a related but non-identical task. Human ratings of originality may be included, but they should be treated as one outcome among several rather than as the sole criterion.

The central experimental architecture should use matched trajectories. Two or more trajectories begin from the same model, system context, task, and initial representation. A small controlled perturbation is then introduced, such as an anomalous datum, an unusual relation, a weakly relevant observation, or a competing explanatory frame. Different prompting regimes are applied while all other controllable factors are held fixed. The main question is whether the perturbation is quickly absorbed, remains locally available, interacts with later evidence, or changes subsequent reasoning in a persistent way.

At minimum, the paper should compare several prompt conditions. A baseline condition asks the model to solve the task normally. A diversity condition explicitly requests multiple alternatives. A high-stochasticity condition increases decoding variability without changing prompt structure. A delayed-closure condition prevents early selection among competing explanations and requires unresolved anomalies to remain active across several rounds. A relational-revision condition requires the model to revise the relation structure of the problem after each consequential perturbation rather than merely append another answer. A consolidation condition asks the model to stabilize a revised representation after exploration. The strongest condition, provisionally called *criticality-targeting prompting*, adaptively regulates exploration, anomaly retention, competition among frames, and consolidation according to observed response dynamics.

The term *criticality-targeting prompting* should be introduced cautiously. It does not imply that a prompt has direct access to a hidden physical critical parameter. It denotes an intervention strategy designed to move the effective inference process toward a region of higher structured susceptibility. The prompts may do this by reducing premature closure, preserving incompatible but plausible hypotheses, maintaining weak anomalies across turns, forcing explicit revision of conceptual relations, varying the environmental or evidential context, and delaying consolidation until perturbations have had time to interact. The empirical question is whether these interventions produce signatures that differ from ordinary diversity prompting.

The paper should distinguish fixed prompting from adaptive prompting. A fixed prompt can request a static style of reasoning, such as maintaining three competing hypotheses. An adaptive prompt can use feedback from the trajectory to decide whether the system is currently too stable, too diffuse, or ready for consolidation. The adaptive variant is theoretically closer to criticality regulation because it treats the desired regime as something to be approached and maintained rather than as a single instruction. A later experimental section can define simple feedback variables such as recovery speed, representational diversity, coherence, unresolved-anomaly count, or persistence of earlier perturbations.

Premature closure is one candidate obstacle that prompting may modulate. A model can encounter an anomaly and immediately map it onto an available explanation, thereby restoring a familiar representation before the anomaly has interacted with other parts of the problem. The paper should not treat closure as intrinsically harmful. Most irrelevant variation should be compressed or discarded. The proposed intervention concerns timing and reopenability: whether closure happens before alternative structures have had an opportunity to develop, and whether a settled representation can later be reopened when new evidence arrives.

The prompt design should therefore include explicit anomaly-retention mechanisms. One family of prompts can require the system to keep a small set of unresolved observations active without immediately explaining them. Another can separate observation, interpretation, and model revision into different stages. Another can ask the model to record which assumptions would have to change if an anomaly were taken seriously. These interventions should not be evaluated by whether they produce more text or more alternatives, but by whether later evidence interacts differently with the retained anomaly.

Historical retention is equally important. If every turn effectively resets the relevant representation, then a perturbation may never become dynamically consequential. The experiment should preserve a compact trajectory state across rounds, such as a structured hypothesis record, relation graph, anomaly register, or evolving problem representation. However, the paper must distinguish historical retention from mere ac-

cumulation. Memory can reinforce an existing attractor just as easily as it can support revision. A criticality-oriented intervention should therefore measure whether retained history changes later sensitivity rather than simply increasing context length.

The strongest evidence for prompt-induced regime change would come from matched-future tests. After trajectories have experienced different prompting histories, they should receive the same later evidence and the same future prompts. If the trajectories continue to differ systematically in problem framing, anomaly sensitivity, hypothesis generation, or relation structure, there is stronger evidence that the earlier intervention altered the effective response organization rather than producing a temporary textual divergence. Hidden-state, memory, and context differences must be controlled as far as the experimental architecture permits.

The paper should define regime signatures before discussing criticality. Useful signatures include perturbation persistence, recovery time, path dependence, sensitivity to perturbation direction, interaction among perturbations, representational switching, coherence, and later reopenability. A candidate near-critical regime would exhibit elevated sensitivity and persistence together with preserved structure. A highly stable regime would rapidly damp perturbations. An over-amplifying regime might produce large divergence but poor coherence or poor retention of useful structure. This three-regime distinction is more defensible than assuming that maximal entropy or maximal variance is desirable.

A phase-diagram representation will likely be useful. One axis may represent effective restoring tendency or closure pressure, while another represents amplification, heterogeneity, or relational coupling. The empirical quantities need not correspond directly to physical control parameters. The figure should be presented as a schematic research map, not as a fitted law. Its purpose is to clarify that prompting may change more than output content: it may alter where a trajectory lies within a space of damping, susceptibility, and coherence.

The term *near-critical* should be preferred over a claim of exact criticality unless a specific measurable transition and associated scaling behavior are demonstrated. Finite language-model inference trajectories are unlikely to justify strong statistical-physics claims from the outset. The paper can instead test whether prompt interventions move the system toward a transition-like region characterized by longer recovery, increased competition among representations, enhanced perturbation interaction, and greater persistence of small differences. Stronger criticality language should depend on evidence rather than framing.

Epistemic turbulence belongs later in the paper. It should not be used as a synonym for creativity, high entropy, or complicated reasoning. The stronger turbulence-like hypothesis becomes relevant only if perturbations interact nonlinearly across multiple representational scales, persist over time, and alter the pathways through which later perturbations propagate. If the observed effects can be explained by ordinary stochasticity, fixed nonlinear amplification, memory accumulation, or simple switching among predefined frames, the turbulence interpretation should be withdrawn.

The paper should also allow negative results. Prompting may increase originality ratings without moving any measured regime signature. Delayed closure may preserve anomalies but reduce task performance. Adaptive prompting may generate persistent divergence without useful structure. High-temperature sampling may perform as well as more elaborate interventions. Some model families may resist regime movement entirely. These outcomes are scientifically useful because the paper is a hypothesis paper, not an advocacy paper for criticality-based prompting.

The system boundary must be explicit. The relevant artificial system can include the base model, system prompt, conversation history, persistent memory, retrieval tools,

external code or scientific tools, and orchestration policy. Prompt-induced criticality may therefore be a property of the coupled inference system rather than the frozen foundation model alone. The experiment should report exactly which components are allowed to retain history and which are reset between trials.

Model-family comparisons are important because the same prompt may act differently depending on training, architecture, reasoning style, context management, and alignment procedures. The hypothesis should therefore be tested across multiple models rather than inferred from one system. The paper should also separate prompt effects from model-specific refusal, verbosity, chain-of-thought imitation, or instruction-following artifacts. Observable outputs and structured intermediate artifacts should be preferred over unsupported claims about hidden reasoning states.

The experimental tasks should include more than free-form brainstorming. At least three task classes would be useful. One class can involve scientific anomaly interpretation, where a small datum conflicts with an incumbent explanation. A second can involve open-ended theory construction or problem reframing. A third can involve cross-domain relation formation, where information from a seemingly distant domain becomes potentially relevant. Each task should allow matched perturbations and repeated later probes so that persistence and reorganization can be measured.

The outcome analysis should therefore combine conventional creativity measures with dynamical measures. Conventional measures may include originality, usefulness, diversity, and expert evaluation. Dynamical measures may include perturbation-response gain, recovery time, persistence of altered representations, delayed reactivation of earlier anomalies, non-additive interaction among perturbations, and divergence under matched future inputs. No single metric should be treated as equivalent to creativity or criticality.

The paper should include explicit falsification criteria. The prompt-induced-criticality hypothesis would be weakened if criticality-targeting prompts increase surface diversity but not perturbation persistence or later response structure. It would also be weakened if high-temperature controls produce the same effects, if the apparent persistence is fully explained by explicit memory replay, or if the effect disappears under matched future probes. The turbulence extension would be weakened if perturbation interactions remain additive or propagation pathways remain fixed. The creativity claim would be weakened if regime changes systematically reduce usefulness, coherence, or expert-rated originality.

The distinction between mechanism and value must remain visible. A prompt may successfully destabilize an incumbent representation and still produce false, incoherent, or scientifically unhelpful alternatives. Movement toward criticality is therefore proposed only as a possible enabling condition for creative change, not as an epistemic virtue in itself. Scientific evaluation still requires evidence, explanatory adequacy, reproducibility, and domain-specific judgment.

The organization of the paper should increase in theoretical strength gradually. Section 1 introduces the problem and distinguishes requests for creativity from interventions on inference dynamics. Section 2 reviews prompting, creativity, criticality, and dynamical-systems work relevant to the hypothesis. Section 3 defines the effective inference regime and the measurable quantities used in the paper. Section 4 develops the prompt-intervention taxonomy. Section 5 specifies the criticality-targeting hypothesis and the distinction between strongly restoring, near-critical, and over-amplifying regimes. Section 6 develops the experimental design and matched-trajectory methodology. Section 7 defines creativity outcomes and dynamical signatures. Section 8 introduces the stronger epistemic-turbulence extension and its additional evidence requirements. Section 9 addresses alternative explanations, boundary conditions, and model-

family variation. Section 10 discusses implications for discovery-oriented AI and prompt design. Section 11 concludes with the empirical status of the hypothesis and the next experimental steps.

Several writing constraints should remain explicit. Section and subsection titles should state structural roles rather than ask questions or announce conclusions. The paper should avoid claims such as “criticality causes creativity,” “this prompt makes LLMs creative,” “temperature is a criticality parameter,” “LLMs naturally operate below criticality,” or “epistemic turbulence is present in language models.” The preferred language is conditional: certain prompt structures may alter effective inference dynamics; a near-critical interpretation is supported only if specified signatures are observed; turbulence-like language is reserved for stronger evidence.

Mathematics should be used only where it provides an operational definition or experimentally measurable quantity. Natural-language stage descriptions should remain prose rather than being converted into arrow chains, symbolic inequalities, or decorative equations. Phase diagrams, perturbation-response plots, matched-trajectory plots, and recovery curves are likely to be more informative than symbolic formalism when the underlying variables are only heuristically defined. Any toy simulation must be labeled as illustrative and must not be presented as evidence about actual model dynamics.

The strongest defensible contribution of the paper is therefore methodological. It proposes that prompting can be studied not only as instruction engineering but as an intervention on the dynamical conditions of inference. The hypothesis is that some interventions may reduce excessive restoration, preserve consequential perturbations, and increase the probability of sustained creative change while keeping the system coherent enough to reason. Whether this corresponds to genuine near-critical behavior, whether the effect generalizes across models, and whether it improves creativity rather than merely instability are empirical questions that the paper is designed to make testable.

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 hypothesis concerning prompt-induced movement toward near-critical inference regimes and is circulated for discussion. The meanings of criticality, regime movement, creativity, and epistemic turbulence remain subject to conceptual and empirical refinement. Systematic literature review, controlled prompting experiments, perturbation analysis, and replication across model families 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. Figure 1 is a schematic regime map created solely to clarify the proposed control logic. It is not an empirical estimate of a language model phase diagram, and its axes and boundaries are illustrative rather than fitted physical parameters.

1 Introduction

The purpose of this section is to define the paper’s intervention problem before introducing a stronger language of criticality. Large language models can already be instructed to generate unusual associations, alternative hypotheses, imaginative continuations, or multiple candidate solutions. Recent empirical work confirms that prompt design and decoding parameters can materially change measured divergent output. Bellemare-Pepin and colleagues report that several large language models achieve strong performance on divergent-association and creative-writing tasks, and that both linguistic strategy prompts and temperature manipulations can increase semantic divergence under some conditions [2]. A larger comparison by Wang and colleagues likewise finds substantial overlap between human and machine divergent creativity, while also reporting greater human variability at the upper tail and mixed to negative effects from several strategic prompt manipulations [15]. These results make a simple claim that current models either possess or lack creativity increasingly unhelpful. They also reveal a methodological problem: a prompt can change the distribution of outputs without showing that the underlying inference process has entered a different dynamical regime.

The distinction matters because the present paper is not primarily concerned with eliciting more unusual text. It develops a narrower hypothesis about the conditions under which a perturbation becomes consequential for later inference. A model can be asked to produce five unconventional ideas, can be sampled at a higher temperature, or can be assigned an unusual persona while still returning, on subsequent turns, to substantially the same problem representation, explanatory neighborhood, or search strategy. Conversely, an intervention that produces only modest immediate novelty may matter more if it allows a previously weak anomaly to remain active, interact with later evidence, and alter what the system treats as relevant. The central object of study is therefore the trajectory of inference after controlled perturbation rather than the first visible output alone.

This trajectory-level focus also responds to a broader tension in the creativity literature. Generative AI can improve some individual creative outcomes while simultaneously narrowing collective diversity. Doshi and Hauser, for example, found that access to generative-AI ideas increased ratings of short-story creativity, writing quality, and enjoyment while making AI-assisted stories more similar to one another [6]. Related work has documented persistent homogenization effects in LLM-generated writing even when prompts or parameters are varied [1]. The point here is not that diversity is equivalent to creativity or that homogenization establishes a lack of creativity. It is that output-level gains and longer-run structural diversity can move differently. A prompting intervention intended to foster creativity should therefore be evaluated not only by immediate originality but also by whether it changes how later perturbations are processed.

The hypothesis developed in this paper is that certain prompt structures may function as interventions on the effective dynamics of inference. In particular, prompts that delay premature closure, preserve competing representations, retain anomalies across several rounds, require explicit revision of conceptual relations, or alternate exploration with consolidation may reduce excessive restoration toward an incumbent representation. Under suitable conditions, such interventions may move the coupled model-prompt-memory system toward a regime in which small perturbations have larger and more persistent consequences while the trajectory remains sufficiently coherent for useful reasoning. The paper refers to this candidate regime as near-critical only in an operational sense. It does not assume that language-model inference possesses a unique physical critical point, a thermodynamic order parameter, or universal scaling law.

1.1 Creativity Requests and Output Variation

Prompt engineering is ordinarily treated as a way to specify tasks, constrain format, provide context, elicit reasoning strategies, or otherwise improve the usefulness of model outputs [12]. Creativity-oriented prompting extends this logic by asking for novelty directly: generate more original ideas, adopt an unusual perspective, avoid conventional answers, combine remote concepts, or produce multiple alternatives before selecting one. Such instructions can be useful, and the present paper does not argue against them. The issue is that they intervene primarily at the level of requested behavior. They do not by themselves identify the dynamical mechanism through which a model becomes more capable of sustaining a creative departure.

The difference can be stated in ordinary experimental terms. Suppose two trajectories begin from the same model, task, context, and initial representation. One trajectory receives a standard instruction to solve the problem. The other receives a prompt asking for unusual alternatives. If the second trajectory produces more semantically distant outputs, the intervention has changed generation. It remains an open question whether

it has also changed perturbation persistence, recovery toward the original framing, sensitivity to later evidence, or the probability of reopening a representation after provisional closure. Those later properties concern the organization of inference across time rather than the diversity of one response.

Temperature manipulation illustrates the same distinction. Higher temperature changes the sampling distribution and can increase semantic divergence on some tasks [2]. Yet stochasticity is an external control over local sampling, not a direct measure of how strongly the larger inference process restores an incumbent conceptual organization. A high-temperature trajectory may wander more widely at the token or sentence level while remaining structurally conservative at the level of hypotheses, variables, assumptions, or explanatory relations. For this reason, the experimental programme developed later treats high stochasticity as a control condition rather than as a proxy for criticality.

This distinction also prevents the paper from defining creativity through a single scalar score. Divergent-thinking measures, semantic distance, expert ratings, and task-specific originality measures each capture useful aspects of creative output, but none establishes by itself that a system has changed how it organizes later inference. The target phenomenon here is therefore broader but still operational: sustained formation of new relations, problem framings, candidate explanations, or inquiry paths that remain consequential under subsequent prompts or evidence. Immediate novelty remains part of the evaluation, but it is not the sole endpoint.

The paper consequently treats a creativity request and a dynamical intervention as different experimental objects. A creativity request specifies what kind of answer should be produced. A dynamical intervention is designed to alter conditions under which alternative representations can persist, interact, and influence subsequent generation. A single prompt may do both, but the distinction matters analytically because the evidence required for each claim is different. More unusual answers support a behavioral effect. Movement toward a different inference regime requires trajectory-level signatures.

1.2 Prompting as an Intervention on Inference Dynamics

A language model is not exposed to a prompt as a detached instruction. The prompt changes the context in which later tokens are generated and can alter which representations, constraints, examples, and reasoning procedures become active. In agentic systems, prompts can also control memory use, retrieval, tool invocation, branching, critique, or consolidation. The relevant experimental system is therefore the coupled configuration that includes the model, prompt, context, any retained state, and the orchestration policy. The present hypothesis concerns this effective inference system rather than the frozen model weights alone.

This broader boundary makes it meaningful to study prompting in dynamical terms. Consider a small controlled perturbation such as an anomalous datum, a weakly relevant analogy, an unexpected tool result, or a competing explanatory frame. Under one prompt regime, the perturbation may be rapidly classified and absorbed into the incumbent representation. Under another, the prompt may require the system to preserve the anomaly, compare several incompatible interpretations, track which assumptions each interpretation depends on, and delay selection until later evidence arrives. The same perturbation then has different opportunities to influence the trajectory.

The proposed mechanism is not simply longer reasoning. More tokens can reproduce the same representation repeatedly. Nor is it equivalent to chain-of-thought prompting, self-critique, or multi-agent debate, although these techniques may provide components

for an intervention. The relevant question is whether the prompt changes measurable response properties: how quickly a perturbation decays, whether its influence reappears after a delay, whether multiple perturbations interact, whether later evidence is interpreted differently because of the earlier perturbation, and whether the system remains capable of coherent consolidation after exploration.

This framing is compatible with emerging evidence that inference in language models can exhibit distinct decoding regimes. Xia and colleagues, in a 2026 preprint, analyze token-level entropy dynamics and interpret some reasoning trajectories as phase-transition-like shifts between exploratory and more structured decoding states [18]. Their result does not establish the criticality hypothesis developed here, and entropy transition should not be conflated with creativity. It nevertheless demonstrates that treating LLM inference as a temporally evolving regime rather than a static mapping is empirically tractable. The present paper extends that methodological possibility in a different direction: instead of asking when explicit reasoning is useful, it asks whether prompt interventions can alter the susceptibility of inference to structured perturbations.

The intervention is therefore defined by intended effect rather than by a fixed wording. A prompt that preserves three competing hypotheses may increase susceptibility on one task and merely increase verbosity on another. A prompt that delays closure may be useful when evidence is ambiguous but harmful when the task has a clear solution. A relation-revision prompt may produce useful restructuring in one model family and unstable drift in another. The hypothesis must consequently be tested across tasks, models, and system configurations rather than inferred from one successful prompt example.

The stronger experimental architecture developed later uses matched trajectories. Trajectories begin from equivalent conditions and receive the same controlled perturbation while prompt regimes differ. Subsequent probes are then standardized so that differences in recovery, persistence, representation, and later response can be compared. This approach is designed to distinguish a temporary textual effect from a more persistent change in the effective organization of inference.

1.3 Criticality as a Candidate Mechanism

The term *criticality* enters the paper only at this point because it names a mechanism hypothesis rather than the empirical phenomenon itself. The empirical phenomenon of interest is a change in susceptibility: small, structured perturbations become more likely to persist, interact, and affect later inference. A criticality interpretation becomes relevant only if these effects cluster around a transition-like region in which restoring tendencies weaken while coherent organization remains available.

This operational use is deliberately weaker than criticality in statistical physics. Finite language-model trajectories do not justify assuming singular behavior, universal exponents, or an exact critical point. The paper instead uses a three-regime comparison. In a strongly restoring regime, perturbations are rapidly absorbed and the trajectory returns toward a familiar organization. In an intermediate regime, perturbations can remain active and influence later inference while coherence is still preserved. In an over-amplifying regime, perturbations produce large divergence but useful structure is degraded or difficult to consolidate. The hypothesis is that some prompt interventions may increase the probability of occupying the intermediate region.

The importance of this distinction can be seen by considering what would count against the hypothesis. If high-temperature sampling produces the same persistence, interaction, and later representational change as the proposed prompt interventions, then a criticality-oriented explanation adds little. If delayed-closure prompts merely

lengthen responses while matched-future probes show no persistent difference, the stronger regime-shift interpretation should be rejected. If apparent persistence is fully explained by explicit memory replay, there is no need to invoke a change in susceptibility. Criticality is therefore not a rhetorical label for successful creativity prompting; it is an interpretation that must earn its place through comparative evidence.

The paper also separates movement toward criticality from creativity itself. A system can become more sensitive and less restoring while producing worse ideas. It can preserve anomalies that are irrelevant, amplify spurious correlations, or become less capable of selecting among alternatives. For this reason, creativity outcomes and dynamical outcomes are measured separately. A useful intervention would need to show both a relevant change in inference dynamics and an improvement in some defensible creativity criterion, such as originality combined with usefulness, sustained problem reframing, or expert-rated quality under later evaluation.

The possibility of a stronger turbulence-like regime is even more conditional. Epistemic turbulence is reserved for cases in which perturbations interact nonlinearly across representational scales, persist historically, and alter the pathways through which later perturbations propagate. The paper does not assume that prompting produces such a regime. Turbulence-like language appears later as an extension whose evidence requirements are deliberately higher than those for simple susceptibility or regime movement.

1.4 Scope and Research Programme

The contribution of this paper is therefore methodological. It proposes that prompting can be studied not only as instruction engineering but also as an intervention on the dynamical conditions of inference. The central hypothesis is that some prompt structures may reduce excessive restoration, preserve consequential perturbations, and increase the probability of sustained creative change while maintaining enough coherence for reasoning and consolidation. Whether this behavior deserves a near-critical interpretation is an empirical question rather than a premise.

Section 2 reviews relevant work on prompting, creativity, criticality, and dynamical views of inference. Section 3 defines the effective inference regime and the trajectory-level quantities used throughout the paper. Section 4 develops a taxonomy of prompt interventions, separating diversity requests, delayed closure, anomaly retention, relational revision, consolidation, and adaptive regulation. Section 5 states the criticality-targeting hypothesis more precisely and distinguishes strongly restoring, near-critical, and over-amplifying regimes. Section 6 presents the matched-trajectory experimental architecture and perturbation protocol. Section 7 separates creativity outcomes from dynamical signatures and specifies how each will be evaluated. Section 8 introduces epistemic turbulence as a stronger extension with additional evidence requirements. Section 9 addresses alternative explanations, boundary conditions, and model-family variation. Section 10 discusses implications for creativity-supporting and discovery-oriented AI systems. Section 11 concludes with the empirical status of the hypothesis and the next experimental steps.

The research programme is designed to support negative as well as positive results. A prompt may improve originality while leaving recovery dynamics unchanged. It may preserve anomalies while reducing usefulness. It may increase path dependence through memory accumulation rather than movement toward a critical regime. Different models may respond in opposite ways to the same intervention. These outcomes would narrow the hypothesis rather than invalidate the broader experimental approach.

The central empirical requirement is therefore stricter than showing that a prompt yields more creative-looking text. Evidence for prompt-induced movement toward criticality would require converging changes in perturbation persistence, recovery, interaction, path dependence, or later response structure, together with preserved coherence over a relevant horizon. Evidence that such movement fosters creativity would additionally require improved creative outcomes under task-appropriate evaluation. By separating these claims, the paper aims to make a speculative idea testable without treating criticality, creativity, and prompting as interchangeable concepts.

2 Literature Review

The hypothesis developed in this paper sits at the intersection of several research traditions that have largely been pursued separately. One literature evaluates whether large language models can produce creative or divergent outputs and how such outputs respond to prompting and sampling parameters. A second literature studies inference-time interventions that broaden, branch, critique, or aggregate reasoning trajectories. A third examines criticality and metastability as regimes in which complex systems balance responsiveness with organization. A fourth, still emerging, treats language-model inference itself as a temporally evolving dynamical process with localized transitions, uncertainty changes, and path-dependent failures. Human creativity research provides an additional comparative tradition in which productive cognition is often associated with regulated movement between spontaneous and controlled modes rather than maximal variability alone. None of these literatures, by itself, establishes that prompting can move an LLM toward criticality or that such movement fosters creativity. Together, however, they define the empirical components that a prompt-induced criticality hypothesis would have to connect.

2.1 Creativity Evaluation and Prompt Sensitivity in Large Language Models

Recent work has substantially weakened any simple claim that contemporary LLMs are incapable of creative output. On divergent-association and creative-writing tasks, leading systems can equal or exceed average human performance under some metrics and task conditions, although the highest-performing human tail remains difficult to match consistently [2, 15]. These studies are important for the present paper because they establish that creativity-related performance is not fixed at the level of the trained model. Prompt wording, role instructions, task strategies, and decoding temperature can change the distribution of generated ideas.

The effects are nevertheless heterogeneous. Bellemare-Pepin and colleagues report reliable gains in semantic divergence from some prompting strategies and higher temperature settings, while also emphasizing task dependence and the limited scope of divergent-association measures [2]. Wang and colleagues find that persona and perspective prompting can raise creativity scores up to a threshold, but that additional prompt engineering produces mixed or negative effects in their large-scale comparison [15]. Haase, Hanel, and Pokutta similarly document substantial intra-model variability and model-specific sensitivity to prompting across creativity benchmarks [8]. The resulting literature therefore supports manipulability more strongly than it supports any single prompting recipe.

Studies of human-AI co-creation add another complication. AI assistance can raise evaluated creativity for individual outputs while reducing diversity across a population

Table 1: Literature strands relevant to the prompt-induced criticality hypothesis. The final column identifies the unresolved step required by the present paper rather than a deficiency in the cited work.

Literature strand	Established contribution	Unresolved relation to the present hypothesis
LLM creativity evaluation	Measures divergent association, originality, creative writing, and variability across models, prompts, and temperature settings	Output-level gains do not establish movement between inference regimes or persistent changes in perturbation response
Inference-time prompting and search	Demonstrates that branching, sampling, debate, critique, and search can alter reasoning trajectories and task performance	Search breadth or multiple paths do not by themselves identify a near-critical regime
Criticality and metastability	Provides formal and empirical accounts of regimes balancing responsiveness, persistence, and instability in recurrent and biological systems	Transfer to LLM inference requires operational signatures rather than direct physical analogy
Dynamic LLM inference	Identifies entropy changes, transition points, branching opportunities, and temporally concentrated failures during reasoning	Existing work has not primarily tested prompt-driven regime movement for creativity
Dynamic creative cognition	Links creativity to flexible coordination between generative and control processes and, in some studies, to balanced switching dynamics	Human neural findings motivate a balance hypothesis but cannot be directly mapped onto artificial inference

of outputs, showing that local improvement and collective exploration need not move together [6]. Related work on LLM-generated writing likewise reports homogenizing effects under some conditions [1]. These results matter because the present hypothesis concerns the organization of variation, not simply its amount. A system can produce many unusual continuations while repeatedly exploring similar conceptual regions, or it can produce fewer visible alternatives while preserving a small anomaly that later reorganizes the task.

Prompt-engineering work directed specifically at brainstorming and creative ideation provides more constructive evidence. Chang and Li propose a structured prompting framework that combines goals, prompts, and strategies to support design brainstorming, reporting improvements in creativity-related evaluation dimensions in their case-based study [4]. Lin’s broader account of effective prompting similarly emphasizes that prompt design changes the information, constraints, and procedures made available to the model during generation [12]. These approaches demonstrate that prompt structure can shape creative performance. Their main evaluation target, however, remains the quality or diversity of the generated product. The present paper asks a different question: whether some prompt structures change the temporal response of inference to small perturbations in a way that is measurable before the final creative product is scored.

2.2 Inference-Time Interventions for Exploratory Search

A large body of work already shows that language-model inference can be reorganized at test time. Self-consistency replaces a single greedy reasoning trajectory with multiple sampled paths followed by aggregation, improving performance on several reasoning benchmarks [16]. Tree of Thoughts broadens this idea by explicitly maintaining alternative intermediate states, evaluating them, and allowing search, lookahead, and backtracking rather than committing to a single left-to-right trajectory [19]. These methods are relevant because they demonstrate that the effective search process is not fully specified by the frozen model weights. External inference procedures can alter which trajectories remain available and which are discarded.

Multi-agent and debate-based methods provide an even closer neighboring case. Liang and colleagues identify a “Degeneration-of-Thought” pattern in iterative self-reflection, where an initially confident stance can constrain later novelty, and propose multi-agent debate to maintain competing lines of reasoning [11]. Their results suggest that deliberately sustaining disagreement can prevent some forms of premature convergence. The finding is conceptually close to the delayed-closure interventions proposed in this paper, although the original study is concerned with divergent thinking and task performance rather than criticality.

This literature establishes several mechanisms that could become components of a criticality-targeting intervention: branching, maintaining alternatives, delaying commitment, re-evaluating intermediate states, and aggregating across trajectories. Yet broader search is not equivalent to a critical regime. A tree search can be externally imposed while every node remains locally strongly restoring. Multi-agent debate can create diversity by duplicating agents rather than changing the susceptibility of any individual trajectory. Self-consistency deliberately samples diverse paths but ultimately uses convergence across them to stabilize an answer. These systems therefore demonstrate that inference architecture is manipulable, while leaving open whether prompting can move a single effective inference process into a regime with measurably different perturbation persistence and recovery.

The distinction is important experimentally. If a criticality-targeting prompt merely recreates the benefits of explicit search, then the criticality interpretation is unnecessary. The relevant comparison should therefore include strong inference-time baselines such as branching, debate, or diverse sampling, and should test whether the proposed intervention changes trajectory-level response properties beyond what can be explained by additional search breadth.

2.3 Critical Regimes in Adaptive Computation

The hypothesis draws its dynamical vocabulary from work on recurrent computation, complex systems, and neural dynamics, but these sources must be used with care. In recurrent neural networks, Bertschinger and Natschläger showed that computational capability for temporally varying inputs can be enhanced near a transition between ordered and chaotic dynamics in a class of random recurrent networks [3]. The result helped motivate a broad “edge of chaos” tradition in which systems near a transition can combine memory of prior states with sensitivity to new input. It does not imply that every useful computation, much less creativity, requires a critical point.

Metastability provides a related but distinct concept. Tognoli and Kelso describe metastability as a regime in which tendencies toward local autonomy coexist with tendencies toward coordination, allowing transient coalitions to form and dissolve without permanent locking or complete independence [14]. In this usage, metastability concerns

the coexistence and switching of partially coordinated states rather than a singular critical transition. The distinction is useful for the present paper because some desirable prompt effects may be better described as maintaining several viable organizations than as moving toward a sharp critical point.

The critical-brain literature offers a stronger claim about near-critical computation. Hengen and Shew review evidence for the hypothesis that neural systems may be regulated toward criticality and report a meta-analysis across 140 datasets, framing criticality as a possible computational setpoint associated with marginal stability and multi-scale dynamics [9]. This literature supplies an important conceptual precedent for *regulation toward* a dynamical regime rather than passive residence at a fixed state. It also illustrates the evidential burden of a criticality claim: one must specify measurable signatures, rule out alternative explanations, and distinguish finite-system approximations from idealized critical phenomena.

For LLM inference, these traditions are analogical resources rather than direct models. There is no established LLM equivalent of a thermodynamic order parameter, neuronal avalanche, or universal critical exponent. The present paper therefore uses criticality operationally. The proposed evidence concerns changes in perturbation persistence, recovery, interaction, path dependence, and coherence across prompt conditions. If these changes can be explained by ordinary search, memory replay, or fixed nonlinear amplification, a stronger criticality interpretation should not be adopted.

2.4 Dynamic Regime Evidence in Language-Model Inference

Recent work has begun to make dynamical descriptions of LLM inference empirically concrete. Xia and colleagues analyze token-level entropy trajectories and describe phase-transition-like changes between exploratory and more structured reasoning states, using early entropy dynamics to route inference strategies [18]. Their results are not evidence that creativity occurs at criticality, and the authors' entropy transition should not be treated as identical to the near-critical regime hypothesized here. The study is nevertheless important because it treats reasoning as a temporally evolving state whose regime can be inferred from trajectory-level statistics.

Zhu and colleagues provide a complementary result. Their analysis of reasoning failures finds that errors are often concentrated around a small number of transition points associated with localized entropy spikes, after which a trajectory can remain locally coherent while being globally wrong [20]. Alternative continuations from the same intermediate state can still reach correct solutions, motivating targeted intervention at these transition points. This work shows that small differences at particular locations in a trajectory can have persistent downstream consequences and that intervention need not be distributed uniformly across all tokens.

Together, these studies strengthen two methodological commitments of the present paper. First, inference should be measured longitudinally rather than inferred from final outputs alone. Second, interventions can be targeted according to observed trajectory state rather than applied uniformly. They do not yet establish that prompts can move inference toward a critical region, because entropy spikes, branching sensitivity, and transition points can arise in systems that are not critical. They also focus primarily on reasoning efficiency and correctness rather than creativity. The open opportunity is to combine trajectory-state estimation with prompt interventions designed specifically to alter the balance between restoration and exploratory persistence.

2.5 Dynamic Balance in Creative Cognition

Research on human creative cognition offers a further comparative motivation for this balance hypothesis. Contemporary network-neuroscience accounts do not generally describe creativity as maximal randomness. Instead, they emphasize interaction between spontaneous or associative processes and executive control. In a large multi-center analysis, Chen and colleagues report that divergent-thinking performance is associated with dynamic switching between default-mode and executive-control network states, including an inverted-U relation for the balance between segregation and integration [5]. The finding is consistent with the broader idea that creative performance can depend on regulated movement between generative and evaluative modes rather than on unconstrained variability.

This evidence should not be mapped directly onto LLMs. Brain networks, transformer inference, external prompting, and human creative behavior are different systems at different levels of description. The value of the comparison is narrower. It provides an empirical example in which creativity-related performance is associated with dynamic coordination and a non-monotonic balance, reinforcing the need to compare strongly constrained, intermediate, and over-diffuse regimes rather than assuming that more variation is always better.

The same caution applies to analogies between human metastability and artificial inference. A prompt that maintains competing hypotheses may resemble a metastable arrangement at the functional level, but the underlying mechanisms are not thereby homologous. The present paper therefore treats human creativity findings as motivation for a general experimental design principle: measure both exploratory flexibility and the capacity to maintain or recover coherent organization.

2.6 Mechanistic Gap for Prompt-Induced Criticality

Across these literatures, most of the necessary components already exist separately. Creativity research shows that model outputs are sensitive to prompts and stochasticity. Inference-time methods show that search structure, branching, debate, and selection can be altered during generation. Criticality and metastability research supplies formal language for intermediate regimes between excessive restoration and excessive instability. Emerging LLM studies show that reasoning trajectories contain transition points and measurable state changes. Human creativity research suggests that productive novelty can depend on regulated switching rather than maximal dispersion.

The missing connection is a direct intervention test. Existing creativity studies typically evaluate the final output; inference-time reasoning methods typically optimize correctness or search; criticality studies concern other dynamical systems; and dynamic LLM studies have not primarily asked whether a prompt can deliberately move the system toward a regime that improves creativity. The present hypothesis therefore requires a design in which prompt conditions, dynamical signatures, and creative outcomes are measured separately.

Such a study must answer three questions in sequence. First, does the prompt change trajectory dynamics in a reproducible way, for example by altering perturbation persistence, recovery, interaction, or path dependence while preserving coherence? Second, do those changes support a defensible interpretation of movement toward a near-critical regime rather than a simpler explanation such as increased sampling diversity or explicit search? Third, if a regime change is supported, does it improve creativity under task-appropriate evaluation? Failure at any earlier stage should block the stronger claim at the next stage.

This gap defines the contribution of the following sections. The paper does not infer criticality from creative-looking text, and it does not infer creativity from dynamical sensitivity. It instead develops an experimental framework in which prompting is treated as an intervention on the effective inference process, criticality is a candidate interpretation of the resulting response geometry, and creativity is an independently evaluated outcome.

3 Effective Inference Regimes and Analytical Representation

The hypothesis developed in this paper requires an intermediate level of description between model parameters and final outputs. Prompting does not normally modify the frozen weights of a deployed language model, yet it can alter which information is active, which alternatives remain available, how evidence is ordered, how long unresolved material is retained, and when a trajectory is consolidated. These changes can matter even when the underlying model is unchanged. The relevant object of analysis is therefore the *effective inference regime*: the organized pattern through which a coupled model–prompt–context system responds to perturbations over a finite trajectory.

This section defines the minimal analytical objects needed to study such regimes. The aim is not to claim that language-model inference already satisfies a particular physical model of criticality. It is to make the proposed intervention measurable. The framework distinguishes the system boundary, the prompt intervention, the perturbation, the trajectory response, and several finite-horizon quantities that can later be compared across prompt conditions. The strongest terms used elsewhere in the paper, including near-criticality and epistemic turbulence, will require additional evidence beyond the quantities introduced here.

3.1 System Boundary and Effective Inference State

A first requirement is to state what counts as the system. Treating the foundation model alone as the system is often too narrow for prompt experiments, because the effective behavior of an artificial agent can depend on conversation history, retrieval, external memory, tools, and orchestration. Let an experimental artificial system at time t be represented schematically by

$$S_t = (\theta, C_t, M_t, E_t, \Pi_t), \quad (1)$$

where θ denotes fixed model parameters, C_t the active context, M_t any persistent memory available to the system, E_t the external task or tool environment, and Π_t the orchestration or prompting policy. Equation (1) is a bookkeeping device rather than a claim that these components form a complete physical state.

The experimental analysis need not observe every component of S_t . Instead, it requires a task-relevant representation z_t that can be compared across matched trajectories. For a scientific anomaly task, z_t might contain the currently active hypotheses, the evidence assigned to each hypothesis, unresolved observations, and explicit relations among explanatory variables. For a cross-domain ideation task, it might contain a relation graph or a structured record of candidate connections. For an open-ended problem-reframing task, it might contain assumptions, problem boundaries, and alternative formulations.

This distinction matters because final text is an unreliable proxy for the effective inference state. Two outputs can differ lexically while preserving the same explanatory organization. Conversely, two superficially similar responses can encode different

Table 2: Minimal analytical objects for studying prompt-induced changes in effective inference regimes. Each object is intended to be operationalized at the level of an experimental system rather than inferred from final text alone.

Object	Role	Possible operationalization
System state z_t	Compact representation of the task-relevant inference state at step t	Structured hypothesis set, relation graph, anomaly register, selected evidence, tool state, or another reproducible intermediate artifact
Prompt condition p	Experimental intervention that changes inference constraints or procedures	Baseline instruction, diversity prompt, delayed-closure prompt, relational-revision prompt, adaptive prompting policy
Perturbation ϵ	Small controlled difference introduced into an otherwise matched trajectory	Anomalous datum, weakly relevant observation, competing explanation, altered relation, or controlled wording change
Response distance D	Difference between matched trajectories under a specified representation	Semantic, graph, hypothesis-set, distributional, or task-specific distance chosen before evaluation
Recovery	Degree to which a perturbed trajectory returns toward the matched baseline organization	Declining response distance, restoration of the prior hypothesis structure, or return to the same representational neighborhood
Persistence	Degree to which the effect of a perturbation remains consequential across later steps	Continued representational difference under common later evidence, delayed reactivation, or transfer to a related task
Coherence	Preservation of usable organization during exploration	Constraint satisfaction, explanatory consistency, task validity, expert judgment, or structured self-consistency checks

commitments about which variables matter, which anomalies remain unresolved, or which hypotheses are still available. The experimental representation should therefore be chosen to capture the level of change relevant to the task rather than defaulting to token-level distance.

The state representation also defines the scope of any later criticality claim. A prompt may increase token entropy while leaving z_t almost unchanged. Another prompt may leave wording relatively stable while substantially changing the relation structure represented in z_t . The paper is primarily concerned with the latter kind of difference: changes in the organization of inference that persist long enough to influence later reasoning.

3.2 Prompting as a Control Intervention

Prompting is treated here as an intervention on the conditions of inference rather than as a direct command to produce a desired outcome. Let p denote a prompt condition or prompting policy. For a fixed model and task environment, the effective transition from one task-relevant state to the next can be written abstractly as

$$z_{t+1} \sim F_p(z_t, u_t; \xi_t), \quad (2)$$

where u_t denotes new task information or external evidence and ξ_t collects stochastic influences such as sampling variability. The notation F_p does not imply that a prompt literally changes the model weights. It denotes the effective transition law observed under a particular intervention condition.

This formulation allows prompt conditions to be compared at the level of response dynamics. A baseline prompt may favor rapid resolution of uncertainty. A delayed-closure prompt may require several competing hypotheses to remain active. A relational-revision prompt may require the system to revise the structured representation after consequential evidence. An adaptive condition may change its instructions according to observed persistence, coherence, or recovery. These interventions can produce different F_p even when θ is identical.

The distinction between a fixed prompt and an adaptive policy should remain explicit. A fixed prompt specifies the same constraints throughout a trajectory. An adaptive policy uses observations from the evolving trajectory to modify later instructions. The latter is especially relevant to criticality regulation because it can attempt to counter excessive restoration or excessive dispersion as those tendencies emerge. Nevertheless, both are experimental prompt interventions and should first be evaluated by the same trajectory-level measures.

Equation (2) also clarifies why ordinary prompt sensitivity is insufficient evidence for regime change. Any prompt can alter output distributions. The stronger claim concerns systematic changes in how the system responds to controlled perturbations across a finite horizon. This requires matched trajectories rather than isolated prompt comparisons.

3.3 Controlled Perturbations and Matched Trajectories

The primary experimental unit is a matched pair or matched set of trajectories. Suppose two runs begin from the same model, task, system instructions, context, memory state, and decoding configuration. At a designated point, one trajectory receives a small controlled perturbation ϵ while the matched control does not. The perturbation should be chosen so that its informational magnitude is limited but its possible relational consequence is meaningful for the task.

Examples include a weak anomaly that conflicts with an incumbent hypothesis, an observation from a distant domain that may or may not become relevant, a small change in a causal relation, or the introduction of a plausible competing explanation. The experiment should avoid using perturbations that simply replace the task with a different task. The purpose is to measure the receiving regime, not to compare unrelated prompts.

Let $z_{t+k}^{(\epsilon,p)}$ denote the state after k subsequent steps under prompt condition p when the perturbation is present, and let $z_{t+k}^{(0,p)}$ denote the matched control. A task-appropriate distance D can then quantify divergence between the two trajectories. The first requirement is that D be defined at a level that corresponds to the scientific claim. If the hypothesis concerns problem framing, then a text-embedding distance alone is weak evidence. If the hypothesis concerns relation structure, graph or structured-hypothesis distances are more appropriate.

Repeated trials are necessary because language-model generation is stochastic. A single dramatic divergence may be an ordinary sample fluctuation. The empirical object is the distribution of perturbation responses under a prompt condition, estimated across seeds, task instances, and ideally multiple model families. Prompt-induced criticality would require a reproducible shift in this response distribution rather than a few visually striking examples.

Matched trajectories also provide the basis for later causal interpretation. They allow the experiment to distinguish the effect of the controlled perturbation from the general creativity of the prompt. A prompt may produce highly original outputs in both perturbed and unperturbed trajectories while showing little additional sensitivity to the perturbation. Conversely, a prompt may produce moderate surface novelty while allowing a small anomaly to produce a persistent change in later reasoning. These are different empirical outcomes and should not be collapsed into a single creativity score.

3.4 Recovery, Susceptibility, and Persistence

The simplest regime distinction concerns what happens after a perturbation. In a strongly restoring regime, the perturbed and control trajectories may diverge briefly and then return toward the same task-relevant organization. In a more susceptible regime, the difference may persist, spread to additional relations, or alter the response to later evidence. In an over-amplifying regime, divergence may become large while coherent task structure deteriorates.

A finite-horizon perturbation-response gain can be defined as

$$\chi_{k,p}(\epsilon) = \frac{D\left(z_{t+k}^{(\epsilon,p)}, z_{t+k}^{(0,p)}\right)}{\|\epsilon\|}, \quad (3)$$

provided that a meaningful perturbation magnitude $\|\epsilon\|$ can be specified for the task. The quantity is deliberately finite-horizon. It does not assume linear response, infinitesimal perturbations, or an underlying thermodynamic susceptibility. It measures how much trajectory difference is associated with a controlled perturbation under prompt condition p .

When perturbation magnitude is not naturally scalar, the numerator can be analyzed without normalization or perturbations can be grouped into matched categories. The important comparison is whether the same perturbation family produces systematically different trajectory effects across prompt conditions. A criticality-targeting condition should not be favored merely because its perturbations were stronger or more semantically disruptive.

Recovery can be measured by the decline of response distance after its initial increase. One useful operational quantity is a recovery time τ_{rec} , defined as the first later step at which the perturbed trajectory returns to a pre-specified neighborhood of the control representation and remains there for a minimum duration. The neighborhood and duration must be specified before analysis. Longer recovery can indicate weaker restoration, but it is not sufficient evidence of criticality. Memory replay, explicit prompt instructions, or simple task branching can also prolong divergence.

Persistence is stronger than slow recovery. A perturbation is historically persistent when it remains consequential under later common conditions. This can be tested by giving the perturbed and control trajectories identical subsequent evidence and instructions after the intervention period. If the earlier perturbation continues to affect problem framing, hypothesis availability, relation structure, or anomaly sensitivity, then the effect is more than a transient textual deviation. Such matched-future tests will become central in the experimental sections.

3.5 Coherence and Historical Retention

Higher sensitivity alone is not the target of this paper. A prompt that makes every small perturbation produce arbitrary divergence may increase $\chi_{k,p}$ while making the system

less useful for inquiry. The proposed criticality hypothesis therefore requires sensitivity to be evaluated together with coherence and historical retention.

Let $C_{k,p}$ denote a task-specific coherence score over a finite horizon. Its implementation will vary by task. In a scientific reasoning task, coherence may involve consistency with uncontested evidence, satisfaction of explicit constraints, and absence of internal contradictions that are not intentionally retained for comparison. In a design or ideation task, it may combine feasibility, task relevance, and expert-rated usefulness. The score should not reward conformity to the incumbent representation merely because that representation is familiar.

Historical retention concerns whether consequential changes remain available to later inference. This is different from context length. A transcript can preserve every token while the system effectively ignores an earlier anomaly. Conversely, a compact structured memory can preserve a relation that continues to influence later reasoning. Let $H_{k,p}$ denote a task-specific retention measure, such as the probability that an earlier perturbation remains recoverable and causally active under a later probe. The exact metric should be defined for each experimental task rather than treated as a universal quantity.

The combination of response gain, coherence, and retention provides a minimal regime description. A strongly restoring condition is expected to show low perturbation persistence and relatively high coherence. An over-diffuse condition may show large divergence but poor coherence or unstable retention. The hypothesized target region would show elevated perturbation persistence while maintaining sufficient coherence and historical availability for the perturbation to participate in later reasoning. Section 5 will develop the stronger criticality interpretation of this pattern.

3.6 Regime Classification and the Status of Criticality

The quantities introduced above are intended to support regime classification before any criticality label is applied. A prompt condition can be described empirically by its distribution of perturbation gains, recovery times, persistence under common future probes, coherence, and retention. Additional measures, including interaction among multiple perturbations, representational switching, and reopenability after closure, can be added when the task requires them.

The word *regime* is used because the paper is interested in organized response patterns across trajectories rather than isolated outcomes. A regime classification should therefore be supported by repeated observations and should be robust to modest changes in task instance or random seed. If a purported “near-critical” condition appears only for one carefully selected prompt example, the evidence is too weak for a regime-level interpretation.

Criticality is a stronger hypothesis layered on top of this descriptive framework. Evidence for increased $\chi_{k,p}$ or longer τ_{rec} does not by itself establish criticality. The interpretation becomes more plausible if several properties change together in a transition-like region: perturbation persistence increases, recovery slows, competing representations remain simultaneously available, small directional changes have larger downstream consequences, and coherence remains above a functional threshold. Stronger statistical-physics claims would require correspondingly stronger evidence, such as scaling relations or a clearly identified control parameter, which this paper does not assume in advance.

This ordering is methodologically important. Prompt conditions should first be compared in terms of observable response geometry. Only then should the analysis ask whether the pattern is consistent with movement toward a near-critical region. Cre-

ativity is evaluated separately again. A prompt may alter the inference regime without improving creative outcomes, and a creative output may occur without any measurable regime shift. Keeping these claims separate makes the central hypothesis falsifiable rather than definitional.

The analytical representation developed here therefore provides the bridge between the literature reviewed in Section 2 and the intervention framework developed next. Section 4 specifies the prompt families to be compared. Section 5 states the criticality-targeting hypothesis in terms of the regime quantities defined above. Later sections use matched trajectories to test whether regime movement, if observed, is associated with creativity and whether any stronger turbulence-like interpretation is warranted.

4 Prompt Intervention Taxonomy

The hypothesis of prompt-induced movement toward criticality cannot be tested by comparing one elaborate prompt against an underspecified baseline. Prompting changes many aspects of an inference process at once: the number of alternatives considered, the timing of commitment, the amount of context retained, the degree of contradiction tolerated, the order in which evidence is processed, and the requirement to revise an existing representation. A useful experimental taxonomy must therefore classify prompt interventions by the dynamical pressure they are intended to exert rather than by their surface wording.

This section develops such a taxonomy. The categories are not mutually exclusive prompt genres. They are intervention dimensions that can be implemented separately or combined factorially. The purpose is to make later experiments capable of distinguishing ordinary search broadening from genuine changes in perturbation persistence, recovery, representational switching, and historical retention. Existing techniques such as self-consistency, Tree of Thoughts, and multi-agent debate provide important precedents for broadening or diversifying inference trajectories [16, 19, 11]. The present framework adds interventions aimed specifically at closure timing, anomaly retention, relational revision, and adaptive regulation of the inference regime.

4.1 Variation-Inducing and Search-Structuring Interventions

The weakest intervention family increases the number or diversity of candidate outputs. Examples include explicit requests for multiple alternatives, unusual perspectives, remote associations, or non-obvious solutions. Higher sampling temperature can serve a related function at the decoding level, although it is not itself a prompt intervention. Such methods are useful controls because they can increase observable variety without necessarily changing the effective regime defined in Section 3.

Search-structuring methods go further by organizing inference into multiple explicit branches or repeated samples. Self-consistency samples multiple reasoning paths before aggregating an answer, while Tree of Thoughts maintains and evaluates alternative intermediate states [16, 19]. Multi-agent debate can similarly maintain competing arguments or perspectives across interaction rounds [11]. These methods demonstrate that inference-time structure can change the trajectory distribution even when model parameters remain fixed.

For the present hypothesis, however, broader search must be distinguished from increased susceptibility. A system may enumerate ten alternatives and still return rapidly to the same representational organization after a perturbation. Conversely, a system may consider only a few alternatives while one weak anomaly persistently changes later

Table 3: Prompt-intervention families distinguished by their intended dynamical effect. The entries describe experimental functions rather than universal prompt templates.

Intervention family	Primary target	Expected regime effect	Critical control
Variation-inducing	Number and diversity of candidate continuations	Broader local search with little necessary change in recovery structure	Match token budget and decoding stochasticity
Search-structuring	Explicit branching, comparison, or deliberation among alternatives	Longer residence among multiple candidate trajectories	Distinguish search width from altered perturbation sensitivity
Closure-regulating	Timing and reversibility of commitment	Slower restoration toward one representation; higher reopenability	Match reasoning length and total number of alternatives
Anomaly-retention	Persistence of weakly conflicting or unresolved evidence	Longer perturbation lifetime and greater opportunity for later interaction	Separate retention from simple verbatim memory
Competing-frame	Simultaneous maintenance of incompatible but plausible representations	Increased representational competition and switching potential	Require coherence within each frame and common evidence
Relational-revision	Modification of conceptual, causal, or evidential relations	Greater probability that a perturbation alters later problem structure	Compare against prompts that merely add alternatives
Adaptive regulation	Prompt policy changes based on measured restoration, divergence, or coherence	Movement toward a target susceptibility region rather than fixed maximal exploration	Pre-specify feedback variables and stopping criteria

framing. The experiment should therefore measure variation-inducing conditions primarily as controls for search width, token count, and candidate diversity.

The relevant comparison is not whether an intervention produces more text. It is whether matched perturbations have different downstream consequences after search breadth and computational budget are controlled. If ordinary branching or temperature produces the same recovery times, persistence, and matched-future divergence as the proposed criticality-targeting conditions, then the stronger interpretation is unnecessary.

4.2 Closure-Regulating Interventions

A second intervention family changes when the system is allowed to consolidate an interpretation. Many prompts implicitly reward rapid resolution: select the best explanation, summarize the evidence, provide the answer, or reconcile the contradiction. These instructions can be useful for ordinary problem solving, but they may also shorten the period during which a perturbation remains active enough to interact with alternative representations.

Closure-regulating prompts alter this timing. A simple version can separate obser-

vation from interpretation and prohibit final synthesis for a fixed number of rounds. Another can require provisional conclusions to retain an explicit list of unresolved tensions. A stronger version can require the system to state what future evidence would reopen a currently preferred explanation. The intended effect is not indecision for its own sake. It is increased *reopenability*: the capacity of a partially consolidated representation to become unsettled when later evidence arrives.

This family should be tested against length-matched controls. Longer reasoning alone may increase the probability of encountering an alternative, and repeated prompting can create persistence simply because earlier text remains in context. A valid comparison therefore holds approximate token budget, number of reasoning rounds, and available evidence constant while changing the timing and reversibility of commitment.

The main observables are recovery time after perturbation, persistence under later common evidence, and the probability that a previously closed representation is reopened by a new anomaly. A closure-regulating condition supports the paper’s hypothesis only if it changes these response properties rather than merely producing more caveats or longer explanations.

4.3 Anomaly-Retention and Competing-Frame Interventions

Anomaly-retention prompts are designed to preserve small observations that do not fit the current explanation without immediately forcing them into a known category. One implementation can maintain a compact anomaly register containing only the observation, its confidence level, and the specific assumption with which it conflicts. Another can require the system to revisit unresolved anomalies after new evidence is introduced. The aim is to increase historical availability, not to maximize memory size.

This distinction is important because explicit memory can stabilize an incumbent interpretation as easily as it can support revision. A transcript that repeatedly summarizes an anomaly as already explained may increase textual retention while reducing dynamical consequence. The stronger test asks whether the retained anomaly changes later hypothesis availability, evidence weighting, or relation structure under a common future probe.

Competing-frame interventions maintain more than one plausible representation at the same time. Instead of asking the model to choose immediately between explanations, the prompt can require separate internal records for each frame, with shared observations but frame-specific assumptions and predictions. The frames should remain constrained by evidence rather than becoming arbitrary personas. Their role is to prevent early collapse of representational competition while preserving enough organization for later comparison.

The experimental risk is that multiple frames simply create parallel text without meaningful interaction. A competing-frame condition should therefore include later probes that force evidence to propagate across frames or require the system to identify which relations would have to change for one frame to absorb another. Evidence of increased switching, delayed commitment, or non-additive perturbation effects would be more informative than the mere presence of multiple headings in the output.

4.4 Relational-Revision Interventions

Relational-revision prompts target a stronger form of change. Rather than asking for more candidate answers, they require the system to modify the structured relations

through which the problem is represented. Depending on the task, the relevant structure may be a causal graph, concept map, hypothesis-evidence matrix, assumption hierarchy, or dependency graph. The intervention can require the system to state which relation changed, why it changed, and what downstream inferences are affected by that revision.

This family is central to the creativity hypothesis because creative scientific change often concerns relations among variables, explanatory roles, or problem boundaries rather than isolated outputs. A new candidate hypothesis is weak evidence of creative reorganization if it is generated inside an unchanged problem representation. A small perturbation that causes the system to revise which variables are considered relevant, which evidence bears on which hypothesis, or which assumptions are upstream may be structurally more consequential even if the resulting prose is less unusual.

Relational revision should be evaluated using structured artifacts rather than text alone. If the experiment uses a hypothesis-evidence matrix, for example, the same matrix format can be compared across matched trajectories. If the experiment uses a concept graph, graph-edit distance or task-specific relation changes can be measured directly. The key outcome is whether the perturbation changes later inference through altered structure, not whether the model merely states that it has “reframed” the problem.

This condition also provides a bridge to matched-future tests. After an intervention period, the experimental and control trajectories can be given the same later evidence while preserving their evolved structured representations. Persistent differences in hypothesis generation, anomaly sensitivity, or evidence integration then provide stronger evidence that the prompt history changed the effective inference organization.

4.5 Adaptive Criticality-Targeting Interventions

The strongest intervention family is adaptive rather than fixed. A fixed prompt cannot guarantee that a system remains in a useful susceptibility range because the same instruction may have different effects across tasks, model families, and stages of inquiry. An adaptive policy instead monitors observable regime indicators and changes the prompting strategy when the trajectory becomes too strongly restoring or too diffuse.

The intervention need not claim direct access to a physical critical parameter. It can operate on empirical proxies defined in Section 3: recovery after controlled perturbation, persistence under common future evidence, representational competition, coherence, and historical retention. If a trajectory rapidly absorbs every perturbation, the policy can increase exploration by delaying closure, introducing a competing frame, or revisiting unresolved anomalies. If divergence becomes large while coherence deteriorates, the policy can shift toward consolidation, constraint checking, or evidence-based pruning.

Such a policy is best understood as a feedback controller over an effective inference regime. Its scientific value depends on pre-specifying the feedback variables and intervention rules rather than allowing an experimenter to improvise prompts after observing attractive outputs. Otherwise, adaptive prompting can become an unblinded search for successful examples and destroy the interpretability of the comparison.

A minimal adaptive experiment can therefore use a small number of discrete intervention states. One state emphasizes exploration, one preserves unresolved material, and one emphasizes consolidation. Transitions among these states are triggered by predefined thresholds on observable quantities such as recovery, coherence, or persistence. More elaborate controllers can be studied later, but the initial hypothesis does not re-

quire continuous control or an exact estimate of distance to criticality.

The crucial comparison is with both fixed creative prompting and fixed high-exploration prompting. If an adaptive policy improves creativity merely because it spends more tokens or samples more alternatives, the criticality interpretation is weak. Stronger support requires evidence that the policy changes the shape of perturbation response while preserving useful organization.

4.6 Control Structure for the Intervention Taxonomy

The taxonomy implies a factorial experimental strategy rather than a single prompt contest. At minimum, a study should include an ordinary task prompt, a variation-inducing prompt, a search-structured condition, a closure-regulating or anomaly-retention condition, a relational-revision condition, and an adaptive criticality-targeting condition. A high-stochasticity decoding condition should be crossed with at least some prompt conditions to separate prompt effects from sampling effects.

Several nuisance variables require explicit control. Token budget should be approximately matched when feasible. The number of external tool calls, retrieved documents, and interaction rounds should be reported. Persistent memory should be either reset or standardized across conditions. The same perturbation family should be used across matched trajectories. Evaluation prompts should be separated from generation prompts so that a condition is not rewarded by the same wording that produced it.

The taxonomy also clarifies the evidential hierarchy. Variation-inducing and search-structuring prompts test whether broader exploration is sufficient. Closure-regulating and anomaly-retention prompts test whether persistence and reopenability matter. Competing-frame and relational-revision prompts test whether representational organization changes. Adaptive prompting tests whether feedback-guided movement among these modes produces a reproducible regime shift. Only after these distinctions are measured should the paper ask whether the resulting pattern is consistent with movement toward a near-critical region.

Section 5 develops that stronger hypothesis. The criticality claim will not rest on the rhetorical complexity of a prompt. It will depend on whether the intervention changes the joint pattern of perturbation sensitivity, recovery, persistence, coherence, and later response under matched conditions.

5 Criticality-Targeting Hypothesis and Regime Structure

The intervention taxonomy in Section 4 identifies prompt families that can broaden search, delay closure, preserve anomalies, revise relational structure, or adaptively regulate exploration and consolidation. The present section states the stronger hypothesis that motivates combining these interventions. The hypothesis is not that a prompt can directly set a language model to a physical critical point. It is that prompting may alter the effective inference regime in a way that reduces excessive restoration, increases the persistence and interaction of small perturbations, and preserves enough coherence for those perturbations to remain useful in later reasoning. A near-critical interpretation becomes relevant only if this joint response pattern is observed.

The section therefore treats criticality as a regime hypothesis rather than a label for unusual output. The regime of interest should exhibit higher perturbation sensitivity than an ordinary strongly restoring condition, but it should not collapse into unstructured divergence. This intermediate possibility is consistent with broader work on computation near transition boundaries, metastability, and neural criticality, where useful

responsiveness is often discussed in relation to a balance between stability and flexibility rather than maximal variability alone [3, 14, 9]. These literatures provide analogies and methodological precedents rather than direct evidence about large language models.

5.1 Regime Structure and the Core Hypothesis

The experimental hypothesis requires at least three operational regimes. A *strongly restoring regime* rapidly absorbs controlled perturbations and returns to a familiar representational organization. A *candidate near-critical regime* shows longer perturbation persistence, greater sensitivity to small differences, and stronger interaction among alternative representations while maintaining usable coherence. An *over-amplifying regime* produces large divergence but loses enough consistency, historical retention, or evidential discipline that the resulting variation is no longer useful for sustained inquiry.

These regimes should be identified from measured response patterns rather than assigned from prompt wording. A prompt that asks for multiple perspectives may still produce a strongly restoring process if all branches collapse quickly to the same framing. A prompt that preserves competing hypotheses may create a candidate susceptible regime if small anomalies continue to affect later reasoning. A highly stochastic condition may produce large output divergence while remaining poor at retaining structured relations. The regime classification therefore depends on the joint behavior of perturbation response, recovery, coherence, and historical retention.

Let the prompt condition be indexed by p . Section 3 defined the finite-horizon perturbation response $\chi_{k,p}(\epsilon)$ together with recovery, coherence, and retention measures. The present hypothesis predicts that there exists, for at least some tasks and model families, a range of prompt conditions in which perturbation response and persistence increase while coherence and historical retention remain above task-dependent functional thresholds. This range is the empirical target of criticality-targeting prompting.

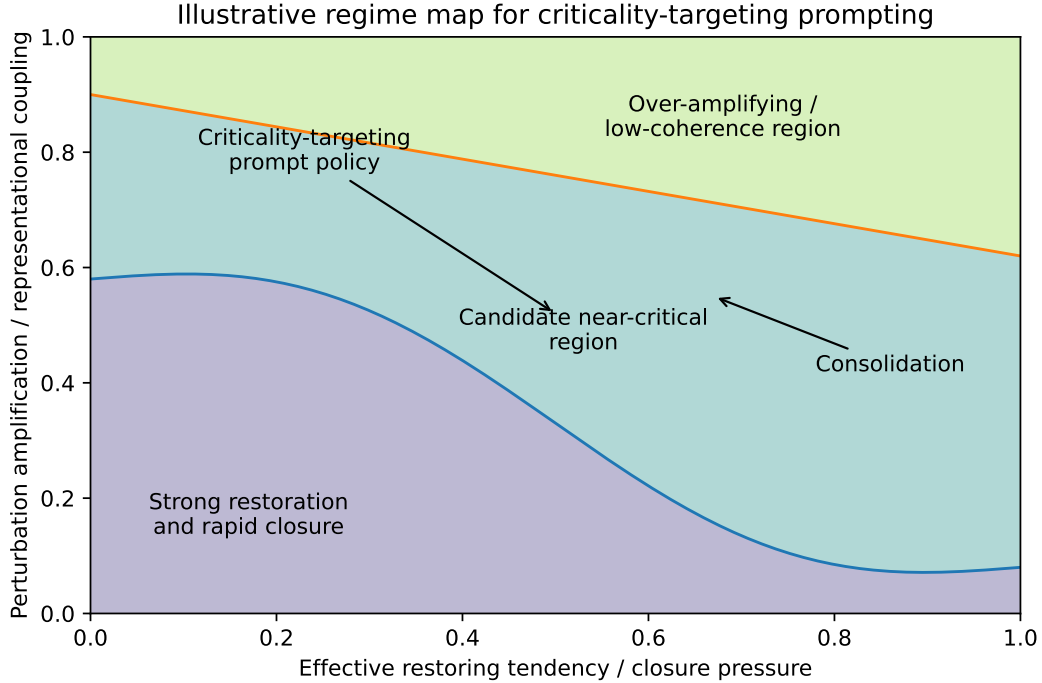
No single scalar index is assumed to identify this region. In particular, high entropy, high lexical diversity, long reasoning traces, or frequent representational switching are insufficient on their own. A candidate near-critical regime should instead display a coordinated pattern across multiple observables. The experimental programme should report these observables separately before considering any composite summary.

Figure 1 visualizes the regime hypothesis schematically. Its axes represent effective restoring tendency and perturbation amplification rather than physical control parameters. The boundaries are illustrative. Their purpose is to make clear that the desired intervention does not aim for maximal exploration. It aims for a region where perturbations can remain consequential while the system retains enough organization for later consolidation.

5.2 Prompting as Movement through Regime Space

A prompt intervention can be understood as changing constraints on the effective transition process rather than changing a single hidden physical parameter. Closure-regulating instructions can reduce the speed with which one representation becomes dominant. Anomaly-retention instructions can increase the lifetime of weak conflicts. Competing-frame prompts can increase the number of simultaneously active representations. Relational-revision prompts can increase the probability that a perturbation changes the structure through which later evidence is interpreted. Consolidation prompts can restore constraint after a period of broader exploration.

Figure 1: Illustrative regime map for criticality-targeting prompting. The horizontal axis represents effective restoring tendency or closure pressure, while the vertical axis represents perturbation amplification or representational coupling. The central region denotes a candidate near-critical regime in which sensitivity is elevated but coherence remains usable. The boundaries and intervention trajectories are schematic and are not fitted to empirical language-model data.



Criticality-targeting prompting combines these pressures with the aim of changing regime position. The relevant movement is not necessarily monotonic. A trajectory can become more susceptible during exploration and then return toward stronger restoration during consolidation. The hypothesis therefore concerns controlled movement through regime space rather than permanent destabilization.

This distinction is important for interpreting fixed prompts. A single instruction such as “keep three explanations active” can increase representational competition, but its effect may be too weak on one task and excessive on another. An adaptive policy can instead use observable trajectory signals to decide whether to increase exploration, preserve unresolved material, or consolidate. Such adaptation is theoretically closer to regime regulation because the intervention is conditioned on the current response pattern rather than specified once at the beginning.

The control variables remain empirical proxies. The experimenter may adjust closure timing, anomaly-retention duration, frame competition, relational-revision frequency, evidence-ordering policy, or consolidation strength. None of these is claimed to be a thermodynamic control parameter. The hypothesis only requires that their combined manipulation produce reproducible changes in the measured response geometry.

Recent work reporting phase-transition-like changes in entropy during reasoning provides one reason to take regime language seriously in large language models, while also illustrating the need for caution [18]. Localized transition points and failure dynamics in reasoning trajectories likewise suggest that qualitatively different response behavior can emerge within a single inference process [20]. These findings do not establish criticality in the sense proposed here. They indicate that inference dynamics can exhibit structured transitions worth measuring rather than treating all prompt effects

as static style changes.

5.3 Observable Signatures of a Candidate Near-Critical Region

A criticality-targeting intervention should be evaluated by pre-specified signatures. The first is increased persistence of a controlled perturbation. If a small anomalous datum influences only the immediately following output and then disappears, the intervention has not produced the response pattern of interest. Persistence should be measured across later common evidence and, where possible, after the original perturbation is no longer explicitly repeated.

The second signature is slower recovery toward the incumbent representation. The relevant quantity is not verbosity or hesitation but the time or number of inference steps required for the perturbed and unperturbed trajectories to become functionally indistinguishable under matched probes. A longer recovery horizon is compatible with increased susceptibility, although it does not by itself establish a critical regime.

The third signature is increased interaction among perturbations. Two small anomalies that are each weak in isolation may have a stronger joint effect when both are retained. A useful experiment can compare the joint response to the sum of single-perturbation responses. Non-additive interaction provides stronger evidence of changed response structure than independent persistence alone.

The fourth signature is competition among representations without immediate collapse. Metastability research provides a useful conceptual precedent for systems that transiently coordinate multiple tendencies while remaining capable of switching among them [14]. In the present setting, the measurable question is whether competing frames remain active long enough for later evidence to differentiate them, rather than whether a model merely lists multiple possibilities in one response.

The fifth signature is preserved coherence. A condition that increases perturbation gain while producing inconsistent assumptions, unusable causal structure, or uncontrolled factual degradation should be classified as over-amplifying rather than near-critical. Coherence must therefore be assessed at the same horizon as sensitivity. A model can be locally coherent in each sentence while losing global consistency across a multi-round trajectory, so the measurement should include structural and evidential constraints where the task permits them.

The sixth signature is historical retention. Earlier perturbations should continue to modify later sensitivity or problem framing after superficial wording differences are controlled. Matched-future tests are especially important here because they distinguish temporary textual divergence from a persistent change in the effective response organization.

These signatures are deliberately redundant. Criticality should not be inferred from one attractive statistic. Stronger support requires a cluster of measurements that changes together under the intervention and differs from high-temperature, high-token-budget, and ordinary search-broadening controls.

5.4 Criticality, Stochasticity, and Diversity Controls

The strongest immediate alternative explanation is that criticality-targeting prompts merely increase variation. The experimental design must therefore separate regime movement from stochasticity. A high-temperature condition should broaden token-level sampling while leaving prompt structure fixed. A diversity-prompt condition should explicitly request unusual alternatives while keeping decoding parameters matched. A

search-structured condition can increase branching or deliberation. These controls establish whether broader output distributions are sufficient to reproduce the same persistence and recovery effects.

If high-temperature sampling produces the same perturbation persistence, historical retention, non-additive interaction, and matched-future divergence as the criticality-targeting condition, then the stronger regime interpretation is unnecessary. The result would support a simpler stochastic explanation. Likewise, if ordinary Tree-of-Thought-style branching or multi-agent debate produces the same response geometry after computational budget is matched, then the distinctive role of criticality-targeting prompting would be weakened [19, 11].

The reverse dissociation would be more informative. Suppose a high-temperature condition increases lexical and semantic diversity but perturbations remain short-lived, while a closure-regulated or adaptive condition produces smaller immediate diversity yet greater persistence and later structural change. That pattern would support the claim that regime position cannot be reduced to surface variation.

A related control concerns reasoning length. Criticality-targeting prompts may generate longer trajectories because they delay closure. Longer traces alone can increase the chance of encountering alternatives. Experiments should therefore include length-matched controls or statistical adjustment for token budget. The hypothesis concerns altered response structure, not simply more computation.

5.5 Creativity as an Outcome of Regime Position

The paper proposes criticality only as a possible enabling condition for creativity. The creativity outcome should therefore be assessed independently from the dynamical regime measurements. A condition can satisfy several near-critical signatures yet fail to improve originality, usefulness, problem reframing, transfer, or expert judgment. Such a result would show that the intervention changed inference dynamics without fostering the target phenomenon.

The main creativity prediction is qualitative rather than a claim about a universal curve. Strongly restoring regimes may suppress unusual relations by absorbing perturbations too quickly. Over-amplifying regimes may produce many novel fragments but fail to consolidate them into usable structures. Candidate near-critical regimes may increase the probability that a small perturbation survives long enough to combine with other representations and support a new relation, hypothesis, or problem framing. This possibility resembles broader proposals that useful computation and creativity can benefit from a dynamic balance between stability and flexibility, but it remains an empirical claim for language models [3, 5].

The experiment should therefore report creativity measures and regime measures side by side. A useful result would show that the same intervention that changes perturbation response also improves one or more creativity outcomes under matched evidence and budget. A stronger result would demonstrate that the regime change mediates the creativity improvement. The present hypothesis paper does not assume such mediation; it specifies it as a later statistical test.

This separation also protects against an easy conceptual error. A system can become more sensitive without becoming more creative. It can also produce a highly original answer while remaining dynamically stable. The research programme is interested in whether a particular response regime increases the probability of sustained creative change, not in redefining creativity as criticality.

5.6 Regime Regulation and Falsification Conditions

The criticality-targeting hypothesis is strongest when framed as a regulation problem. A useful prompt policy should be capable of loosening restoration when the system closes too quickly and strengthening consolidation when divergence threatens coherence. The goal is not to hold the system at maximal instability. It is to maintain, for a finite period, a response regime in which perturbations can remain active and interact before a new structure is consolidated.

This regulation hypothesis yields direct falsification conditions. First, the hypothesis is weakened if criticality-targeting prompts change creativity ratings but do not alter any pre-specified regime signature. Second, it is weakened if high-temperature or diversity controls reproduce the same persistence and recovery pattern. Third, it is weakened if the candidate regime consistently reduces coherence or usefulness so strongly that creative gains disappear under expert evaluation. Fourth, it is weakened if apparent historical effects vanish under matched-future probes or are fully explained by explicit memory replay.

The near-critical interpretation should also be withdrawn if response changes vary smoothly with intervention strength and show no evidence of a transition-like region beyond ordinary monotonic scaling. The experiment need not demonstrate singular scaling laws, but the language of criticality should earn its use by identifying a meaningful change in response organization that simpler terms such as “more exploration” do not capture.

These constraints make the hypothesis deliberately difficult to confirm. The value of the proposal lies in turning a broad intuition about creativity and criticality into a set of separable measurements. Section 6 develops the matched-trajectory experimental design needed to estimate these quantities and compare prompt interventions under controlled conditions.

6 Experimental Design and Matched-Trajectory Methodology

The criticality-targeting hypothesis cannot be evaluated from isolated model responses. The relevant claim concerns how a prompt policy changes the evolution of inference after a controlled perturbation. The experimental unit should therefore be a trajectory rather than a single completion, and the central comparison should use matched trajectories that begin from the same task state and differ only in the intervention assigned to them. This design makes it possible to separate immediate output variation from persistent changes in recovery, representation, and later sensitivity.

The methodology proposed here is intentionally modular. It can be implemented with a single model, compared across model families, or extended to systems with retrieval, persistent memory, tools, and external orchestration. The system boundary must be declared before each experiment because history that is retained outside the base model can change the interpretation of persistence. The goal is not to eliminate all sources of variation, which would be impossible in stochastic generative systems, but to make the main contrasts explicit enough that stronger dynamical interpretations remain testable.

6.1 Experimental Unit and Matched Initialization

A matched block begins with a common task specification, common evidence, common system configuration, and a structured representation of the initial inquiry state. The representation can take different forms depending on the task: a set of candidate hypotheses, a relation graph, a causal sketch, an anomaly register, a problem decomposition, or another artifact that can be compared across later turns. The experiment should record both free-form outputs and these structured artifacts. Free text is useful for conventional creativity evaluation, whereas structured artifacts make persistence and reorganization easier to measure.

Within a block, several trajectories are generated from the same initial state. One trajectory receives the baseline prompt policy and the others receive the intervention policies defined in Section 4. When the model interface permits seed control, the same seed should be reused across conditions for the first matched generation and then varied systematically across replications. When seed control is unavailable, the experiment should use repeated samples per block and treat sampling variation as part of the uncertainty rather than attributing every divergence to the prompt.

The matched design should also control inference budget. Conditions that delay closure or maintain several hypotheses will often consume more tokens than baseline prompting. A fair comparison therefore requires either a fixed maximum budget across conditions or a second analysis in which token count and number of model calls are matched. The main hypothesis concerns response organization, so a condition should not receive credit merely because it was allowed substantially more computation.

Let block i contain an initial observable state $z_{i,0}$ and prompt condition p . The resulting trajectory is denoted $z_{i,0:T}^{(p)}$. For a perturbation experiment, a matched unperturbed trajectory $z_{i,0:T}^{(p,0)}$ and a perturbed trajectory $z_{i,0:T}^{(p,\epsilon)}$ are generated under the same prompt condition. Their finite-horizon divergence can be summarized as

$$d_{i,k}^{(p)}(\epsilon) = D\left(z_{i,k}^{(p,\epsilon)}, z_{i,k}^{(p,0)}\right), \quad (4)$$

where D is defined for the task representation before data collection. Equation (4) is descriptive: it does not identify criticality on its own. Its value is that the same perturbation can be compared across prompt conditions using a common measurement scale.

6.2 Perturbation Library and Task Classes

The perturbation should be small enough that immediate task difficulty remains comparable across conditions, yet meaningful enough that a prepared system could use it. A perturbation library should therefore contain several types of controlled difference rather than one generic anomaly. Suitable categories include an observation that conflicts weakly with the incumbent explanation, a relation imported from a distant domain, a counterexample to one assumption, an alternative causal direction, a low-salience datum whose relevance becomes clearer later, and a competing explanatory frame that fits only part of the evidence.

Perturbation magnitude should be calibrated empirically. If the perturbation is so strong that every condition must revise immediately, the experiment measures compliance with decisive evidence rather than susceptibility. If it is too weak to affect any condition, the experiment has little power. A pilot stage can therefore estimate a range in which the baseline condition usually absorbs or ignores the perturbation while at least some exploratory conditions occasionally retain it. The main experiment should use several perturbation strengths around that range so that regime effects are not inferred from a single hand-selected example.

Table 4: Core experimental conditions and the contrasts they are intended to support. The exact wording of prompts should be task-specific, but the dynamical role of each condition should remain stable across experiments.

Condition	Intervention target	Primary comparison	Primary readouts
Baseline	Ordinary task completion	Reference trajectory	Recovery, coherence, task quality
Diversity prompt	Broader candidate generation	Diversity without regime targeting	Output diversity, persistence of alternatives
High stochasticity	Sampling variability	Randomness control	Immediate divergence, later recovery
Delayed closure	Longer anomaly and frame lifetime	Timing of stabilization	Recovery time, unresolved-frame persistence
Relational revision	Revision of problem structure	Structural change beyond answer variation	Relation-graph change, transfer, matched-future divergence
Adaptive criticality-targeting	Regulated exploration and consolidation	Full hypothesis condition	Joint regime signatures and creativity outcomes

At least three task classes are useful. The first is *scientific anomaly interpretation*: the model receives a compact theory or explanation together with observations, one of which weakly conflicts with the incumbent account. Later evidence is designed so that preserving the anomaly can become useful. The second is *problem reframing*: the initial problem admits several representations, and a small perturbation provides a reason to reconsider which variables or relations are central. The third is *cross-domain relation formation*: an initially peripheral concept from another domain later becomes potentially useful for constructing a new explanation or method.

Tasks should support both local and longitudinal evaluation. A task that ends after one creative response cannot distinguish temporary variation from persistent reorganization. Each task should therefore include later evidence updates, repeated opportunities to reopen the problem, and at least one transfer probe that differs from the original wording. The transfer probe is especially important because a relation that survives only when the original perturbation is repeated may reflect short-term context dependence rather than historical retention.

The perturbation order should also be randomized across matched blocks. Otherwise, a later anomaly may appear more consequential simply because the model has accumulated more context. When a task requires a fixed scientific chronology, the order can remain fixed within that task while different tasks use different positions for the perturbation. The analysis should report whether effects depend strongly on early, middle, or late insertion.

6.3 Prompt Conditions and Factorial Controls

The prompt taxonomy becomes experimentally useful when conditions are constructed so that individual mechanisms can be isolated. A minimal experiment can use the six conditions in Table 4. A stronger design can factor some interventions separately. For example, anomaly retention and relational revision can be crossed so that the experiment distinguishes merely keeping an anomaly visible from allowing it to alter the problem representation. Likewise, delayed closure can be crossed with consolidation so that long

exploration is not confounded with the absence of a stabilization phase.

The high-stochasticity condition is essential. Sampling temperature or another decoding parameter is increased while prompt wording is held close to baseline. This condition tests whether persistence and creativity gains can be explained by broader sampling alone. A diversity-prompt condition provides a second control because it broadens candidate generation at the instruction level without explicitly modifying closure or historical retention. Search-structured prompting, such as branching or multi-agent comparison, can be added where computational resources permit, but its budget should be matched to the adaptive criticality-targeting condition as closely as possible.

The adaptive condition requires a transparent policy. It should not be an unconstrained meta-agent allowed to rewrite prompts arbitrarily. Instead, the policy can observe a small pre-specified set of trajectory indicators: whether the anomaly is still represented, how rapidly competing frames collapse, whether coherence has degraded, whether earlier perturbations remain relevant under new evidence, and whether the relation structure has changed. Based on these observations, the policy can choose among a limited set of actions such as preserve alternatives for another round, request a relational revision, introduce a contrastive probe, or consolidate the current representation.

This restricted action set makes the adaptive condition reproducible and helps separate criticality targeting from generic agentic optimization. The full prompt templates, intervention rules, and stopping conditions should be released with the experiment. If the adaptive policy uses another language model as a controller, that controller becomes part of the experimental system and must be held fixed across comparable trials.

The design should also include controls for instruction length where practical. Criticality-targeting prompts may be longer and more explicit than baseline prompts. A length-matched placebo can contain an equivalent amount of neutral procedural text while omitting anomaly-retention and relational-revision instructions. This control helps distinguish intervention content from generic effects of longer context or stronger instruction salience.

6.4 Longitudinal Trajectory Protocol

A useful trajectory protocol separates exploration, evidence update, and consolidation into identifiable phases while allowing the intervention to affect how long each phase lasts. The first stage establishes the incumbent representation from the common task evidence. The second introduces the controlled perturbation. The third supplies one or more evidence updates that can interact with the perturbation. The fourth applies matched future probes. The final stage requests consolidation into a structured account that can be evaluated for coherence, usefulness, and retained novelty.

The experiment should record intermediate states at fixed checkpoints rather than only at the end. At each checkpoint, the system can be asked for a compact structured artifact using the same schema. Examples include a ranked hypothesis set, a relation graph, an anomaly list, and a short statement of the current problem framing. The artifact should be generated with a standardized extraction prompt that is separate from the intervention prompt. This reduces the risk that observed structural change is merely a change in reporting style.

Recovery is measured after the perturbation is no longer repeated. If the perturbed and unperturbed trajectories rapidly become indistinguishable under common evidence, the perturbation has been damped. If they remain different in a task-relevant way, the persistence can be quantified using Equation (4) or a task-specific structural distance. Importantly, persistence should be interpreted together with coherence. A trajectory that remains different because it accumulates contradictions is not evidence

of a useful susceptible regime.

The protocol should also test delayed reactivation. A weak perturbation can be introduced early, then omitted for several rounds, followed by a later cue whose relevance depends partly on the earlier information. A history-sensitive system may use the cue differently depending on whether the early perturbation was present. This design directly tests whether weak traces remain causally available after they cease to dominate the immediate context.

Where persistent memory is used, the experiment should separate three cases: memory disabled, memory enabled but reset before common future probes, and memory enabled throughout. This decomposition helps distinguish a prompt-induced change in effective inference organization from simple replay of stored content. If a persistent effect disappears whenever explicit memory is reset, the result should be interpreted primarily as memory-mediated history rather than stronger generator reconfiguration.

6.5 Common-Future Probes and Reset Tests

The strongest test of persistent reorganization occurs after different histories are followed by the same later inputs. After the intervention phase, trajectories receive an identical sequence of evidence and probe prompts. The probes should include both direct questions about the original task and transfer questions that require the system to use its current representation in a related setting. A persistent difference under these common future inputs is more informative than a difference observed while the intervention prompt remains active.

For block i and prompt conditions p and q , let Q_i denote the common future probe set. A simple matched-future distance can be defined as

$$D_i^{\text{future}}(p, q) = \frac{1}{|Q_i|} \sum_{r \in Q_i} D(z_{i,r}^{(p)}, z_{i,r}^{(q)}). \quad (5)$$

The comparison is strongest when the intervention prompt itself is removed before Q_i begins. The resulting distance then reflects retained differences in context, memory, structured artifacts, or effective response organization rather than direct obedience to different current instructions.

Several reset tests can localize the source of persistence. In a *prompt reset*, all conditions receive the same later prompt while retaining their histories. In a *context reset*, the conversational history is removed but a standardized structured summary is supplied. In a *memory reset*, external persistent memory is cleared while the current structured state is preserved. In a stronger *state reconstruction* test, trajectories are brought to a matched explicit representation before common probes. Persistence that survives progressively stronger resets provides stronger evidence that earlier prompting changed more than superficial wording, although practical systems may not permit complete control over hidden state.

These tests should be interpreted conservatively. A difference after prompt reset can still be explained by retained context. A difference after context reset can still be explained by external memory. A difference after explicit-state reconstruction may still reflect hidden internal state or stochastic variation. The value of the reset hierarchy is not to prove access to the model’s internal generator, but to rule out simpler explanations step by step.

6.6 Replication, Randomization, and Statistical Analysis

The study should be replicated across tasks, perturbation instances, model families, and sampling seeds. Prompt effects are likely to be heterogeneous, so the analysis should report distributions rather than only grand means. At minimum, each task–perturbation block should be repeated across several seeds per condition. Larger experiments should use hierarchical models or mixed-effects analyses that treat task, model, perturbation type, and prompt condition as distinct sources of variation.

The primary statistical comparison should be paired within matched blocks. For an outcome Y , such as recovery time, historical retention, originality, or matched-future distance, the condition effect can be estimated from within-block differences rather than comparing unrelated samples. If b denotes the baseline condition, a simple paired estimator for condition p is

$$\hat{\Delta}_p = \frac{1}{n} \sum_{i=1}^n \left(Y_i^{(p)} - Y_i^{(b)} \right). \quad (6)$$

Confidence intervals should be obtained with a method appropriate to the block structure, and multiple outcomes should be handled with a pre-specified correction or hierarchical analysis rather than selective reporting.

Regime classification should be performed independently from creativity scoring where possible. One analysis can estimate recovery, persistence, perturbation interaction, switching, coherence, and retention. A second analysis can evaluate originality, usefulness, problem reframing, and transfer. Only after both are estimated should the experiment test whether creativity differences are associated with the inferred regime differences. This separation reduces the risk of defining the desired regime using the same creativity outcomes that it is supposed to explain.

The study should be preregistered when feasible. At minimum, the task set, perturbation library, intervention templates, primary regime signatures, primary creativity outcomes, exclusion rules, and falsification criteria should be fixed before the main comparison. Exploratory analyses remain valuable, especially because the correct dynamical variables are not yet known, but they should be labeled as exploratory rather than folded into the original hypothesis after observing the results.

A successful experiment would therefore produce more than a ranking of prompt conditions. It would produce a map of how prompt policies alter recovery, persistence, interaction, coherence, and later response under controlled perturbations. Section 7 specifies how these dynamical measurements should be related to creativity outcomes without treating either set of measures as a substitute for the other.

7 Creativity Outcomes and Dynamical Signatures

The central hypothesis of this paper links two kinds of observation that should remain analytically separate. The first concerns creativity: whether a prompt condition produces more original, useful, generative, or structurally novel contributions to an inquiry. The second concerns dynamics: whether the same condition changes how perturbations are retained, amplified, resolved, or carried into later inference. A criticality-targeting intervention is supported only if these two families of outcomes can be related without defining one in terms of the other. Otherwise, the analysis risks circularity: a regime would be called “near-critical” because the outputs appear creative, and the outputs would then be called creative because the regime has been labeled near-critical.

This section therefore specifies a two-layer evaluation strategy. Creativity is evaluated with task-appropriate judgments that do not rely on the criticality hypothesis. Dy-

namical signatures are estimated from matched trajectories, perturbation tests, reset tests, and common-future probes. The final analysis asks whether changes in the second family help explain changes in the first, after controlling for simpler factors such as output diversity, token budget, and decoding stochasticity.

7.1 Creativity as an Independent Outcome Family

The term *creativity* is intentionally broader than novelty alone. Recent evaluations of language models have used divergent-thinking tasks, expert ratings of research ideas, semantic distance, and measures of usefulness or feasibility [2, 15, 8]. These methods are informative, but no single creativity score is adequate for the present hypothesis. The proposed intervention is meant to facilitate inquiry, so the relevant outcomes include whether the system can produce a contribution that is both different from its baseline trajectory and capable of supporting later reasoning.

At minimum, creativity evaluation should distinguish originality from usefulness. Originality can be estimated by blinded human ratings, semantic-distance measures, rarity relative to a reference corpus, or task-specific novelty criteria. Usefulness can be evaluated through factual adequacy, explanatory coverage, feasibility, predictive relevance, or expert judgment, depending on the task. The two dimensions should not be collapsed before analysis. A condition that generates unusual but incoherent hypotheses should not be treated as equivalent to one that produces a novel and workable reframing.

For scientific tasks, a third dimension is especially important: *problem transformation*. A response can be creative because it proposes a new answer, but it can also be creative because it changes what is being treated as the problem. Examples include identifying a previously ignored variable, revising a causal direction, replacing a binary distinction with a relational representation, or recognizing that an apparent anomaly indicates a mismatch in the problem formulation itself. These changes should be scored separately from surface-level lexical novelty.

A fourth dimension is *generative follow-through*. A candidate idea that cannot support subsequent inference may be interesting but fragile. The experiment should therefore test whether a proposed relation, hypothesis, or reframing remains useful under later evidence, can generate discriminating predictions, or transfers to a related problem. This longitudinal criterion is particularly relevant to the present paper because the hypothesis concerns persistent reorganization rather than isolated unusual outputs.

The creativity panel should be blind to prompt condition whenever practical. Human evaluators should not know whether an output came from baseline, high-temperature, delayed-closure, or criticality-targeting prompting. When automated metrics are used, they should be fixed before the main experiment and validated against a subset of human judgments. The evaluation protocol should also report inter-rater reliability and preserve disagreement rather than forcing a false single ranking where expert judgments differ.

7.2 Structural Creativity beyond Surface Diversity

The experimental design should explicitly separate surface diversity from structural creativity. Two responses can differ lexically while relying on the same causal assumptions and the same decomposition of the problem. Conversely, two responses can share much of their vocabulary while differing in which entities are treated as relevant, which relations are considered causal, or which evidence is regarded as anomalous. This distinction matters because high-temperature decoding can increase surface variation without necessarily altering the organization of inquiry.

Table 5: Outcome families for evaluating the criticality-targeting hypothesis. Creativity outcomes and dynamical signatures are estimated independently before their relation is tested.

Outcome family	Representative measures	Interpretive role
Originality	Blinded ratings, semantic distance, rarity, nonredundancy	Tests whether the contribution departs meaningfully from common or baseline outputs
Usefulness and adequacy	Expert judgment, predictive value, feasibility, explanatory coverage	Prevents unusual but unusable outputs from being counted as successful creativity
Problem transformation	Variable revision, causal reframing, relation-graph change, new discriminations	Tests structural change in how the inquiry is represented
Generative follow-through	Later prediction, transfer, evidence integration, hypothesis refinement	Tests whether an idea remains consequential after its initial production
Dynamical response	Recovery time, perturbation persistence, nonlinear interaction, matched-future divergence	Characterizes how the inference trajectory responds to controlled disturbance
Regime integrity	Coherence, contradiction burden, historical retention, task performance	Distinguishes structured susceptibility from uncontrolled divergence

Structured task representations introduced in Section 6 make this distinction measurable. If each checkpoint includes a hypothesis set, relation graph, anomaly register, and compact problem framing, the analysis can compare changes in these artifacts separately from changes in prose. For relation graphs, suitable measures include edge additions or removals, changes in centrality of previously peripheral concepts, and the appearance of relations that were absent from both the initial representation and baseline trajectories. For hypothesis sets, one can track the persistence of newly introduced explanatory families and whether later evidence changes their rank.

Structural novelty should not be rewarded merely for changing many relations. Large graph edit distance can reflect confusion. A stronger creativity criterion asks whether structural change improves explanatory or predictive work. For example, a newly introduced variable is more consequential if it resolves a previously persistent anomaly, distinguishes between competing hypotheses, or supports a successful transfer prediction. This is why structural measures should be paired with usefulness and follow-through rather than interpreted in isolation.

The analysis should also distinguish *prompt-visible revision* from *retained revision*. A relational-revision prompt may explicitly instruct the model to change its representation, so immediate graph change is partly a compliance measure. Stronger evidence arises when the revised relation continues to organize later reasoning after the instruction is removed. Common-future probes and reset tests therefore provide the bridge between structural creativity and the dynamical analysis.

7.3 Dynamical Signatures of a Candidate Susceptible Regime

The dynamical outcome family characterizes how the inference process responds to controlled perturbations. No single signature is sufficient for a near-critical interpretation. The relevant evidence is a pattern across several measures, together with controls that rule out simpler explanations.

The first signature is *perturbation persistence*. Section 6 defined matched-trajectory

divergence at a finite horizon. For each trajectory block, persistence can be summarized by the accumulated divergence across a pre-specified set of checkpoints,

$$P_i^{(p)}(\epsilon) = \frac{1}{|K|} \sum_{k \in K} d_{i,k}^{(p)}(\epsilon), \quad (7)$$

where K contains the evaluation horizons selected before the experiment. This measure should be interpreted together with task relevance. Persistent stylistic divergence is not equivalent to persistent representational divergence.

The second signature is *recovery*. Recovery can be estimated from the time required for the perturbed trajectory to return to a pre-specified neighborhood of its matched control trajectory or incumbent representation. A criticality-targeting condition is expected to reduce immediate restoring pressure relative to baseline, but the hypothesis does not predict indefinite non-recovery. If the trajectory never restabilizes and coherence deteriorates, the condition is better described as over-amplifying than as productively susceptible.

The third signature is *perturbation interaction*. If two small perturbations are introduced separately and jointly in matched blocks, the outcome under the joint condition can be compared with the sum of their separate effects. For an outcome Y measured at a fixed horizon, a factorial interaction term can be estimated as

$$I_Y = Y_{11} - Y_{10} - Y_{01} + Y_{00}, \quad (8)$$

where the binary indices indicate absence or presence of the two perturbations. A nonzero interaction shows that the joint effect is not captured by simple addition on the chosen outcome scale. It does not by itself establish criticality, but repeated interaction across representations or horizons is evidence against a purely independent-noise account.

The fourth signature is *historical retention*. Earlier perturbations should remain causally available when later evidence arrives. Delayed-reactivation tasks provide a direct test: an early low-salience item is introduced, omitted for several rounds, and made relevant by a later cue. The analysis measures whether its earlier presence changes the later representation under matched prompts and evidence. Memory-reset controls are required to distinguish simple replay from broader history-dependent organization.

The fifth signature is *matched-future divergence*. After the intervention is removed, conditions receive the same future probes. Persistent differences in task-relevant representation under common inputs provide stronger evidence that the prompt intervention changed the effective trajectory rather than only the wording of the current response. Reset tests can localize how much of that persistence is carried by explicit context, memory, or structured artifacts.

Finally, all of these signatures must be evaluated alongside *coherence*. A regime that maximizes persistence and interaction by producing uncontrolled contradiction does not satisfy the hypothesis. Coherence should therefore be measured using task accuracy, logical consistency, contradiction counts, explanatory coverage, or expert evaluation appropriate to the domain. The candidate region of interest is one in which sensitivity increases without a collapse of usable organization.

7.4 Separating Regime Evidence from Creativity Evidence

The critical test is whether dynamical changes and creativity changes covary in the predicted way without being defined from the same observations. The analysis should therefore estimate regime signatures first and creativity outcomes second, with separate coding pipelines where possible. The resulting data can then test whether prompt

conditions that produce greater persistence, slower recovery, stronger interaction, and retained historical effects also show higher originality, problem transformation, or generative follow-through.

The expected relation need not be monotonic. The hypothesis is compatible with a middle region in which creativity is facilitated, flanked by strongly restoring conditions and over-amplifying conditions. Statistical models should therefore permit nonlinear associations rather than assuming that more persistence or more divergence is always better. Generalized additive models, pre-specified spline terms, or categorical regime comparisons are preferable to interpreting a simple positive correlation as evidence for criticality.

A particularly informative analysis compares conditions with similar output diversity but different dynamical signatures. If high temperature and criticality-targeting prompting produce comparable lexical or semantic diversity, yet only the latter yields longer anomaly retention, stronger structural revision, and better follow-through, the result would support the claim that diversity alone is insufficient. Conversely, if creativity gains track diversity equally well across all conditions and the additional dynamical measures provide no explanatory value, the stronger regime hypothesis is weakened.

The same logic applies to compute. Longer reasoning or additional branches may improve both creativity and persistence because they provide more opportunities to search. Budget-matched conditions are therefore necessary. The hypothesis is supported more strongly when regime signatures differ under comparable compute, token, and evidence budgets.

7.5 Outcome Patterns and Interpretive Levels

Several possible result patterns should be anticipated before data collection. One pattern is *diversity without persistence*: outputs become more varied, but perturbation effects decay rapidly and later probes converge. This result supports ordinary search broadening rather than movement toward a susceptible regime. A second pattern is *persistence without creative value*: perturbations remain active and trajectories diverge, but usefulness and coherence decline. This is consistent with over-amplification or maladaptive fixation rather than the intended regime.

A third pattern is *structured persistence without improved creativity*: the intervention changes recovery and historical retention while creative outcomes remain unchanged. This result would be theoretically important because it would show that prompting can alter inference dynamics without demonstrating that the altered regime facilitates creativity. The creativity claim should then be rejected or narrowed even if the dynamical hypothesis survives.

A fourth pattern is *creative gain without distinctive regime evidence*. If originality and problem transformation improve but recovery, interaction, and matched-future behavior do not differ from simpler prompting controls, the intervention may still be useful, but criticality is not needed to explain the benefit. The paper's mechanistic claim should be weakened accordingly.

The strongest supportive pattern combines creative gain with a distinct but coherent response geometry: perturbations remain consequential longer than under baseline, interact across stages, survive common-future probes, and support useful structural revision while coherence remains above pre-specified thresholds. Even this pattern would justify only a candidate near-critical interpretation until alternative nonlinear or memory-based explanations have been tested.

7.6 Falsification Criteria for the Creativity Link

The paper’s title states a hypothesis about fostering creativity, so the creativity link requires its own falsification conditions. It is not enough for prompt interventions to change regime signatures. The hypothesis should be considered unsupported if criticality-targeting conditions repeatedly alter persistence or recovery without improving any pre-specified creativity outcome across adequately powered tasks.

The hypothesis should also be weakened if creativity gains are fully reproduced by simpler controls. If high stochasticity, diversity prompting, or longer search achieves the same originality, usefulness, structural revision, and follow-through while producing no comparable regime signatures, criticality targeting is not a necessary explanation. Likewise, if the adaptive intervention improves performance only by spending more tokens or making more tool calls, the result should be interpreted as a compute advantage unless budget-matched analyses show an additional effect.

A further falsification condition concerns robustness. A putative creativity benefit that appears only for one model, one perturbation wording, or one evaluator should not support a general claim. The hypothesis is stronger if the direction of the effect replicates across task classes, perturbation families, model families, and independent evaluation procedures, even when effect sizes vary.

Finally, the experiment should remain open to the possibility that different kinds of creativity occupy different regimes. Divergent idea generation may benefit from a different balance of restoration and susceptibility than scientific hypothesis revision or problem reframing. Failure to identify one universal optimum would not be surprising. The broader claim concerns whether prompt policies can alter effective inference regimes in ways that systematically change the probability of useful creative outcomes. Section 8 develops the stronger downstream possibility that, under some conditions, repeated perturbation interaction may produce turbulence-like epistemic dynamics; that interpretation requires evidence beyond the signatures defined here.

8 Epistemic Turbulence as a Stronger Downstream Hypothesis

The criticality-targeting hypothesis developed in the preceding sections concerns a comparatively modest dynamical possibility: prompt interventions may move an effective inference process into a regime in which small perturbations persist longer, interact more strongly, and remain available to later reasoning while useful coherence is preserved. That hypothesis does not require turbulence. A system can become more susceptible to perturbation, display longer recovery times, or move among competing representations while still being adequately described by ordinary nonlinear response, metastable switching, or a finite cascade. The present section therefore introduces *epistemic turbulence* only as a stronger downstream hypothesis whose use requires additional evidence.

The distinction matters methodologically. If every large or surprising change in reasoning were called turbulent, the term would add little beyond instability or complexity. Physical turbulence is not simply randomness. Classical treatments emphasize structured nonlinear interaction across scales, intermittency, and the redistribution of activity through a dynamically evolving flow field [7]. Even in fluid systems, the onset of turbulence, the geometry separating decaying from sustained trajectories, and the fully developed turbulent regime are distinct analytical problems [13, 10]. The analogy proposed here therefore concerns a limited set of dynamical properties. It does not posit an

Table 6: Evidence hierarchy for interpreting complex inference dynamics. Stronger descriptions should be used only when weaker mechanisms are insufficient.

Description	Minimum observation	What remains insufficient
Stochastic variation	Output changes across repeated samples	Persistence or structural consequence
Nonlinear amplification	Small perturbations sometimes produce disproportionate effects	Cross-scale propagation or pathway change
Cascade	A local perturbation recruits multiple components or representational levels	Evidence that propagation rules themselves change
Metastable switching	Trajectories move among recurrent organizations with nontrivial dwell times	Self-modification of later propagation
Turbulence-like dynamics	Recurrent multiscale interaction, intermittency, historical persistence, and endogenous modification of propagation pathways	Direct identity with physical turbulence

epistemic analogue of kinetic energy, a Reynolds number, a Kolmogorov spectrum, or a conservation-law cascade.

8.1 Cascade, Switching, and Turbulence

A first requirement is to distinguish a cascade from turbulence. Small disturbances can trigger very large cascades in systems with fixed interaction rules. Watts’s threshold model, for example, demonstrates that rare global cascades can arise from local shocks when network structure and node thresholds create a susceptible configuration [17]. Such a cascade can be large, nonlinear, and difficult to predict while the propagation rule itself remains fixed. For the present paper, this is an important alternative explanation rather than evidence of epistemic turbulence.

The same caution applies to representational switching. A language model or agentic system may move among several recognizable frames, hypotheses, or search strategies because the prompt activates different pre-existing response modes. Repeated switching can produce complicated trajectories without altering the organization through which later perturbations propagate. Metastability can likewise produce long dwell times, abrupt transitions, and coexistence among states without requiring a turbulence-like interpretation. These phenomena may be scientifically interesting on their own and may be sufficient to explain a creativity gain.

The proposed turbulence extension begins only when the inference process exhibits a stronger form of interaction. Perturbations introduced at one representational level must influence activity at other levels; their effects must interact rather than simply accumulate; and the history of those interactions must alter which pathways remain available to later perturbations. A useful empirical example would be an anomalous datum that initially changes one local hypothesis, later modifies the relation among several hypotheses, then changes which evidence is retrieved or which conceptual decomposition the model uses, thereby changing the response to a subsequent anomaly. The importance lies less in the size of the initial deviation than in the evolving organization of propagation.

Table 6 expresses the intended evidence hierarchy. The levels are not a ranking of scientific value. They indicate increasing explanatory commitments. If nonlinear amplification or metastable switching explains the observed prompt effect, the stronger

turbulence language should not be used.

8.2 Multiscale Interaction and Intermittency

A second requirement concerns scale. The relevant scales in language-model inference need not correspond to physical length scales. They may instead refer to nested organizations such as tokens and local statements, relations among concepts, candidate hypotheses, problem representations, retrieval policies, and higher-level plans for inquiry. A turbulence-like interpretation becomes plausible only if perturbations propagate among several such levels and later return to alter lower-level processing.

This requirement prevents a common category error. A long response that mentions many concepts is not evidence of multiscale interaction. The experiment must identify structured intermediate artifacts or task-relevant observables at more than one level and show that changes at one level predict later changes at another. For example, a retained anomaly might first alter the ranking of candidate hypotheses, then change the model's decomposition of the problem, and later change which observations are treated as relevant. A later revision of the problem decomposition may in turn change how the original anomaly is interpreted. Such reciprocal influence is stronger evidence than simple semantic spread.

Intermittency is another candidate signature. Physical turbulent systems often alternate irregularly between relatively quiet and intense local activity, and transitions to turbulence can involve coexistence between laminar and turbulent regions or episodes [7, 10]. In the present framework, an analogous empirical pattern would be alternating periods of stable inference and concentrated reorganization rather than uniformly high diversity. A criticality-targeting prompt might preserve a problem representation for several rounds, then produce a burst of relational revision after a particular perturbation, followed by a new period of consolidation.

Intermittency alone is not decisive. A model can alternate between modes because of prompt formatting, tool availability, token-position effects, or explicit stage instructions. The relevant question is whether burst-like reorganization is coupled to perturbation history and whether it changes subsequent response organization. The experimental design should therefore compare intermittency under adaptive criticality-targeting prompts with intermittency produced by fixed exploration–consolidation schedules and with stochastic controls.

8.3 Historical Modification of Propagation Pathways

The strongest component of the turbulence hypothesis is historical modification of propagation. Sections 6 and 7 already require common-future probes to test whether earlier interventions have consequences beyond the immediate response. The present extension asks for something stronger: whether earlier perturbations change the pathways through which later perturbations influence the system.

This property can be studied without claiming access to hidden chain-of-thought states. The experiment can use explicit structured artifacts. Suppose the model maintains a hypothesis graph, an anomaly register, a problem decomposition, and a retrieval plan. After an initial perturbation, later evidence may traverse a different set of relations, recruit different hypotheses, or trigger different retrieval actions. If the effect is reproducible across matched future probes and cannot be removed by resetting only the current wording of the prompt, the history has altered the effective propagation structure of the coupled system.

A practical test is to repeat the same second perturbation after different first-perturbation histories. If the second perturbation produces systematically different relational consequences across those histories, the response is history dependent. To justify the stronger pathway-modification claim, however, the difference should be visible in the intermediate structures that mediate later inference rather than only in final text. A change in which hypotheses are connected, which evidence is retrieved, or which variables are treated as causally relevant provides stronger evidence than a change in phrasing.

This distinction also clarifies why persistent memory is not sufficient. Memory can reproduce an earlier statement and thereby change later output while the system’s effective propagation rule remains unchanged. Turbulence-like dynamics require interaction among retained history, current perturbations, and the evolving relational organization. Memory is therefore one possible substrate for historical dependence, not evidence of turbulence by itself.

8.4 Criticality and Turbulence as Distinct Hypotheses

The relation between criticality and turbulence should remain explicitly non-equivalent. The criticality-targeting hypothesis concerns movement toward a region of elevated susceptibility and weakened restoration while coherence remains useful. The turbulence extension concerns what may happen after perturbations begin to persist and interact within such a region. A near-critical regime may therefore increase the probability of turbulence-like dynamics without being sufficient for them.

The fluid-mechanical analogy reinforces this distinction. Transitional pipe flow can be linearly stable yet possess a state-space boundary separating perturbations that decay from those that trigger sustained turbulent behavior [13]. Contemporary work on transition also examines critical phenomena and directed-percolation-like behavior at the onset of turbulence, while treating the turbulent state itself as a richer dynamical object [10]. These examples do not license direct transfer to language models. They show only that sensitivity near a transition and the dynamics of the post-transition regime are analytically separable.

For prompt experiments, the corresponding logic is straightforward. Evidence that criticality-targeting prompts lengthen recovery, increase perturbation interaction, or preserve competing representations would support the susceptibility hypothesis. Evidence of repeated multiscale cascades would support a stronger cascade interpretation. Only if those cascades are intermittent, historically persistent, and accompanied by measurable modification of later propagation pathways should the turbulence-like description be considered.

This staged interpretation is important because a creativity benefit could appear at any of the weaker levels. A model may become more creative simply because delayed closure permits one anomaly to remain active long enough to influence another hypothesis. It may not need turbulence-like dynamics at all. The paper therefore treats epistemic turbulence as an optional explanatory extension, not as a required route from criticality-targeting prompts to creativity.

8.5 Operational Signatures and Falsification Conditions

The experimental programme should pre-register a small set of signatures before using turbulence language. First, perturbation interactions should be demonstrably non-additive under the factorial designs introduced in Section 7. Second, consequences should propagate across at least two explicitly measured representational levels. Third,

the trajectory should display nontrivial temporal concentration or intermittency rather than uniform divergence. Fourth, earlier perturbations should alter the response to later matched perturbations. Fifth, some of that history dependence should be traceable to changes in explicit relational or procedural structures rather than to simple replay of stored text.

The hypothesis is weakened when these signatures disappear under appropriate controls. If a fixed nonlinear model with stationary interaction rules reproduces the same trajectories, a turbulence interpretation is unnecessary. If high-temperature sampling produces the same multiscale patterns, the special role of criticality-targeting prompts is weakened. If common-future differences vanish after memory reset, the apparent historical effect may be ordinary memory dependence. If perturbation interactions remain additive or if propagation pathways remain stable across histories, the stronger turbulence claim should be withdrawn.

A final failure mode concerns creativity itself. Even if turbulence-like signatures were observed, they would not establish that the regime is useful. Excessive instability could produce incoherent, unfalsifiable, or scientifically unproductive outputs. The creativity outcomes defined in Section 7 must therefore remain independent of the dynamical diagnosis. The strongest supportive result would be a reproducible regime in which criticality-targeting prompts increase creativity outcomes, preserve sufficient coherence, and produce dynamical signatures that cannot be reduced to stochasticity, fixed nonlinear amplification, ordinary switching, or explicit memory replay.

The role of this section is therefore deliberately narrow. It specifies what additional observations would be required before the paper’s strongest dynamical language becomes appropriate. The main hypothesis remains testable even if epistemic turbulence is never observed. In that case, prompting toward a susceptible but non-turbulent regime could still prove useful for creativity, and the absence of turbulence would sharpen rather than undermine the empirical theory.

9 Limits of the Criticality-Targeting Hypothesis

The criticality-targeting hypothesis is deliberately stronger than the claim that prompting can change model outputs. It proposes that some prompt policies may alter the effective response regime of inference in a way that increases perturbation persistence and interaction while preserving enough coherence for useful creative work. Such a claim is meaningful only if several simpler explanations are considered first. A prompt can change wording, increase search breadth, lengthen reasoning, preserve more context, activate a familiar instruction-following pattern, or allocate more computational effort without moving the system toward any transition-like regime. The present section therefore specifies alternative explanations and scope conditions that should be tested before interpreting an observed creativity gain as evidence for prompt-induced movement toward criticality.

The purpose is not to weaken the research programme by adding an unlimited list of caveats. It is to establish an explanatory ordering. When a simpler mechanism accounts for the data, that mechanism should be preferred. Near-critical language becomes useful only when ordinary sampling variation, explicit memory, fixed nonlinear response, search-budget differences, or prompt-compliance effects leave important aspects of the perturbation-response geometry unexplained. The same principle applies to the stronger turbulence extension developed in Section 8.

Table 7 summarizes the main alternatives. The list is not exhaustive, but it identifies the mechanisms most likely to produce false positives in the proposed experiments.

Table 7: Alternative explanations for apparent prompt-induced regime change and the observations needed before adopting a stronger criticality interpretation.

Alternative mechanism	What it can explain	Additional evidence required for a regime interpretation
Sampling diversity	More varied wording, ideas, or branches across runs	Persistent response differences under matched stochasticity and future probes
Instruction compliance	Explicit maintenance of alternatives or anomalies while the instruction remains active	Effects that survive removal or weakening of the original instruction
Search or compute expansion	Better coverage from more tokens, branches, agents, or tool calls	Regime signatures under budget-matched conditions
Explicit memory or context replay	Historical dependence caused by retained text or structured records	Effects exceeding direct replay and surviving targeted memory or context resets
Fixed nonlinear response	Large or non-additive perturbation effects under an unchanged transition rule	Evidence of changed recovery or propagation structure across later matched probes
Mode switching	Abrupt movement among pre-existing reasoning styles or representations	Evidence that available transitions or later perturbation pathways have themselves changed
Model-specific alignment behavior	Refusal, verbosity, hedging, or instruction-following patterns that mimic openness or closure	Replication across model families and controls for response style

Each alternative should be treated as a substantive explanation rather than as a nuisance to be removed after the fact.

9.1 Sampling Diversity and Search-Budget Effects

The first alternative explanation is ordinary search expansion. A prompt that asks for more alternatives, a higher-temperature decoder, a larger beam, a branching search procedure, or a multi-agent protocol can increase the number of explored continuations without changing the underlying restoring organization of the task representation. Existing inference-time methods such as self-consistency, Tree of Thoughts, and debate demonstrate that broader search and comparison can improve reasoning performance under some conditions [16, 19, 11]. A creativity gain under criticality-targeting prompting could therefore arise simply because the intervention creates more opportunities to encounter a useful candidate.

This possibility makes budget matching essential. Conditions should be compared under comparable limits on generated tokens, branches, external tool calls, retrieval operations, and wall-clock or inference cost where those quantities are available. If an adaptive criticality-targeting condition repeatedly explores twice as many hypotheses as baseline, improved originality or usefulness cannot be attributed confidently to regime position. A secondary analysis can permit unequal budgets for practical comparison, but the mechanistic test requires a matched-budget subset.

Search breadth also needs to be separated from perturbation persistence. A condition can produce many distinct candidates while each candidate disappears as soon as the prompt changes. Conversely, a condition can explore relatively few candidates while a small perturbation changes the trajectory for many later rounds. These are different phenomena. The present hypothesis concerns the latter response geometry, not merely the size of the immediate search tree.

A particularly informative control pairs semantic diversity across conditions. If a diversity prompt and a criticality-targeting prompt produce comparable output diversity at the intervention stage, later differences in recovery, anomaly retention, or common-future behavior become easier to interpret. If those later differences disappear once diversity is matched, the stronger regime explanation loses support.

9.2 Instruction Compliance and Prompt-Format Artifacts

A second alternative is direct instruction compliance. Delayed-closure prompts may tell the model to keep several hypotheses active; anomaly-retention prompts may explicitly require an unresolved-observation list; relational-revision prompts may request a graph update after every perturbation. The resulting artifacts can look like persistence or openness even when the underlying behavior is only local compliance with the current instruction.

The experiment should therefore distinguish *performed openness* from persistent change in response organization. One test is prompt withdrawal. After several intervention rounds, the system receives a neutral task instruction that no longer asks it to preserve alternatives or revisit anomalies. If the previously retained structures immediately vanish and later perturbations are processed exactly as under baseline, the strongest interpretation is instruction dependence. If differences remain under matched neutral probes, a more persistent historical effect becomes plausible.

Paraphrase and format controls are also important. Prompt effects can depend strongly on wording and formatting [12]. A criticality-targeting policy should therefore be implemented through multiple surface forms that preserve the intended intervention but vary language, ordering, and presentation. Robust effects across these variants provide stronger evidence that the result is tied to the intervention principle rather than to one accidental instruction pattern.

The same caution applies to verbosity. A prompt that requires explicit alternatives, anomaly registers, and relation graphs will often produce longer outputs. Longer responses can mechanically increase the number of relations or ideas available for later reuse. Analyses should therefore normalize or control for response length where possible and should distinguish structural persistence from simple textual carryover.

9.3 Memory, Context, and External-State Confounds

Historical retention is necessary for many of the proposed effects, but explicit memory can also create a misleading appearance of regime change. If an anomalous datum is copied verbatim into every later prompt, its continued influence requires no special account of susceptibility. The same is true when a retrieval system repeatedly surfaces a stored hypothesis because of a fixed similarity rule. These mechanisms can be useful, but they should not be mistaken for evidence that the inference regime itself has changed.

The reset hierarchy developed in Section 6 is therefore central. Prompt reset tests whether the effect depends on the current instruction. Context reset tests whether the effect is carried by ordinary conversation history. Memory reset tests whether persistent artifacts are sufficient to reproduce the later behavior. A stronger reconstruction test can re-create the current task state while removing selected historical traces. Different results across these resets help localize where the apparent persistence resides.

The system boundary must be reported with similar care. In an agentic configuration, the effective system may include a foundation model, system prompt, conversation state, retrieval index, persistent files, external tools, and an orchestration policy. A change

in future behavior can arise because any one of these components changed. The paper should therefore avoid language suggesting that the frozen model itself has become near-critical when the measured effect is carried by an external controller or memory store.

This does not make the coupled-system interpretation illegitimate. The hypothesis is explicitly about effective inference systems. A prompting policy that reorganizes external memory and thereby changes future perturbation sensitivity may be scientifically important. The requirement is attribution: the experiment should identify which component carries the historical effect and should not infer hidden model-state change from system-level persistence alone.

9.4 Fixed Nonlinear Response and Metastable Switching

A fourth alternative is fixed nonlinear dynamics. Large perturbation gain, non-additive interactions, and abrupt transitions can occur in a system whose transition rule remains unchanged. Section 8 already used cascade and metastability models to motivate this caution. A prompt may move the trajectory into a different part of a fixed response landscape, producing long recovery or strong interaction without changing that landscape.

This possibility is especially relevant to the near-criticality hypothesis. Observing a steep increase in perturbation sensitivity across prompt conditions does not by itself show that the system has approached a critical point. The same pattern can arise from finite-amplitude threshold effects, state-dependent saturation, or switching among pre-existing modes. A transition-like interpretation becomes stronger when several indicators co-occur across a controllable prompt parameter, but even then the explanation should remain operational unless scaling behavior or a more specific transition model is demonstrated.

Matched-future tests can help distinguish a temporary state excursion from a more persistent change in response organization. If two trajectories that experienced different prompt histories converge once their observable state and future inputs are matched, the earlier difference may be adequately described as state dependence. If systematic differences remain after careful reconstruction, a stronger history-dependent interpretation becomes plausible. Even then, explicit memory and hidden context must be considered before invoking generator reconfiguration.

Metastable switching provides another intermediate explanation. A system may dwell for long periods in one reasoning organization and then shift rapidly to another because the prompt changes the relative accessibility of pre-existing modes. Such behavior can support creativity without any need for turbulence or structural self-modification. The empirical programme should therefore treat metastability as a legitimate endpoint rather than as a failed attempt to prove criticality.

9.5 Model-Family and Alignment Dependence

The same prompt can act differently across model families. Training data, architecture, reasoning-oriented post-training, context handling, tool interfaces, and alignment procedures can all alter how strongly a model follows an instruction, preserves alternatives, or returns to familiar explanations. Large-scale creativity comparisons already suggest substantial variation across models and prompting conditions rather than one universal creativity profile [2, 15, 8]. The criticality-targeting hypothesis should therefore be formulated at the level of conditional effects rather than a universal claim about large language models as a homogeneous class.

At minimum, the main experiment should include several models with different inference characteristics. A practical design can sample from general instruction-tuned models, reasoning-oriented models, and systems with different memory or tool architectures. The goal is not to rank vendors or model families. It is to test whether the proposed response geometry recurs under different implementations and whether the prompt policy requires family-specific calibration.

Model-family variation also complicates the interpretation of closure. Some systems may be strongly optimized to produce concise final answers; others may maintain visible intermediate alternatives for longer. Some may refuse speculative or uncertain outputs more often because of alignment policies. Others may generate extensive hedging that superficially resembles open inquiry while adding little structural diversity. Measures of closure and openness therefore need task-relevant operationalization rather than lexical proxies such as the presence of uncertainty words.

Adaptive prompting may consequently require model-specific control ranges. The same intervention strength that reduces excessive restoration in one model could push another into incoherence. This does not invalidate the general hypothesis; it implies that the hypothesized susceptible region, if it exists, is likely to be system dependent. Claims should therefore concern whether prompt policies can move a particular system’s response regime in a reproducible direction, not whether one universal prompt defines a critical zone for all models.

9.6 Scope Conditions and Withdrawal Rules

The hypothesis has several clear boundary conditions. First, it concerns tasks in which multiple representations, hypotheses, or problem decompositions are meaningfully available. A deterministic calculation with one short exact answer provides little room for the proposed anomaly-retention and relational-revision mechanisms. Second, the intervention requires a finite history over which perturbations can persist and interact. Single-turn tasks can test immediate susceptibility but provide weak evidence for historical reorganization. Third, the task must permit independent evaluation of usefulness or correctness; otherwise instability can be mistaken for creativity.

The strongest claims should also be withdrawn under specified results. If high-temperature or diversity controls reproduce both the creativity gains and the dynamical signatures, criticality-targeting prompting adds little mechanistic explanation. If prompt withdrawal eliminates all persistence, the effect is better described as instruction compliance. If memory replay fully explains common-future divergence, historical reconfiguration should not be inferred. If fixed nonlinear or metastable models explain the observed response geometry, exact or near-critical language should remain provisional. If creativity improves while the predicted regime signatures do not, the prompting strategy may still be valuable, but the criticality mechanism is unsupported.

Conversely, null creativity results do not automatically falsify the dynamical claim. Prompting may alter recovery, perturbation interaction, or historical retention without improving originality or usefulness. In that case the paper’s title-level claim about fostering creativity must be weakened, while the narrower claim that prompting can modify effective inference dynamics may remain viable. This separation is necessary because the paper tests both an intervention mechanism and a proposed functional benefit.

The same modular logic applies to epistemic turbulence. Turbulence-like language should be withdrawn whenever ordinary stochasticity, fixed nonlinear amplification, memory accumulation, cascades, or metastable switching provide sufficient explanation. The paper is therefore designed to contract gracefully under negative evidence. Its most durable contribution may ultimately be the matched-trajectory methodology for

studying prompt-induced regime changes even if the criticality interpretation proves unnecessary.

10 Discussion

The preceding sections develop a deliberately layered hypothesis. Prompting is treated first as an intervention on an effective inference system, then as a possible means of altering recovery, persistence, representational competition, and historical retention, and only after those quantities are measured does the paper ask whether a near-critical interpretation is useful. Creativity is evaluated separately from these dynamical signatures, while epistemic turbulence is reserved for a still stronger downstream pattern. The purpose of this discussion is to clarify what would follow if some of these layers were empirically supported, what would remain unresolved, and how the framework changes the interpretation of prompting in large language models.

The central proposal is modest in one sense and ambitious in another. It is modest because it does not claim that a prompt can move a language model to a literal thermodynamic critical point, that creativity requires criticality, or that every successful creative intervention should display turbulence-like dynamics. It is ambitious because it treats prompting as potentially more than instruction specification. Under some conditions, prompt policy may function as a control process that changes how strongly a system restores familiar representations, how long anomalies remain active, how readily competing frames interact, and how much of one inference episode remains consequential for later inference. If this possibility is real, prompt engineering becomes partly a problem of regime control rather than only a problem of wording.

10.1 Prompting as Regime Engineering

Most practical discussions of prompting assume a relatively direct relation between instruction and response. The user specifies a task, style, role, decomposition, or format, and the model generates a continuation conditioned on that instruction. This view is sufficient for many applications. The present hypothesis becomes relevant only when the object of interest is temporally extended inference and when the prompt policy is designed to regulate properties of the trajectory itself.

The intervention taxonomy in Section 4 makes this distinction explicit. Variation-inducing prompts broaden candidate generation. Closure-regulating prompts delay premature commitment. Anomaly-retention prompts preserve discrepancies that would otherwise disappear. Relational-revision prompts require changes in the organization among hypotheses, evidence, and problem representations. Adaptive criticality-targeting policies additionally respond to measured trajectory properties such as rapid recovery, excessive convergence, or loss of coherence. These interventions differ not merely in semantic content but in the dynamical pressure they place on the inference process.

This interpretation suggests a useful distinction between *response engineering* and *regime engineering*. Response engineering seeks a desired output from the current inference conditions. Regime engineering seeks to alter the conditions under which later outputs are generated. A response can change while the surrounding inference regime remains essentially unchanged. Conversely, a prompt policy can produce only modest immediate differences yet alter how later perturbations are retained or combined. The matched-trajectory methodology in Section 6 is intended to distinguish these cases.

The regime-engineering interpretation also changes how prompt quality should be evaluated. A prompt that maximizes immediate originality may be inferior to one that creates a better balance between exploration and consolidation over a longer horizon. Similarly, a prompt that forces several alternatives into every response may appear open while merely producing superficial compliance. The relevant evidence is whether the intervention changes later recovery, cross-round persistence, common-future responses, and the capacity to reorganize relations among representations. This is why the framework gives greater weight to trajectory-level measurements than to one-shot creativity scores.

The practical implication is not that every prompt should become adaptive or multi-stage. The stronger intervention is justified only for tasks where discovery, reframing, or sustained exploratory reasoning matters. For short deterministic tasks, rapid closure is often desirable. Criticality-targeting prompting is therefore best understood as a specialized control strategy for problems in which the cost of premature stabilization is substantial.

10.2 Criticality as an Operational Interpretation

The term *criticality* carries a strong mathematical and physical history, and its use in language-model inference therefore requires restraint. The present framework uses criticality operationally: a candidate near-critical region is a regime in which small, structured perturbations have unusually persistent and interactive effects while coherence and historical retention remain sufficient for useful inference. This definition provides an empirical target without claiming that the system instantiates a known universality class or a singular phase transition.

This operational stance is important because several mechanisms can mimic critical behavior. A fixed nonlinear system can show abrupt threshold effects. Metastable systems can switch among pre-existing organizations. Search expansion can increase diversity and apparent sensitivity. Explicit memory can preserve perturbations without altering the effective response regime. Model-specific instruction following can make alternatives appear persistent only because the prompt continues to request them. Section 9 therefore treats criticality as an interpretation that should be adopted only after these alternatives are tested.

If the hypothesis receives support, the strongest conclusion would still be system- and task-specific. There may be no universal prompt or universal critical region for large language models. Different models may have different restoring tendencies, different sensitivity to closure instructions, different context and memory architectures, and different tolerance for representational competition. The relevant empirical question is whether a reproducible intervention can move a particular effective system toward a region with the predicted response geometry.

This system dependence is not a weakness of the hypothesis. In control problems generally, the same input can have different effects on systems with different parameters and histories. What matters is whether the relation between intervention and regime movement can be measured and predicted. A successful theory would therefore characterize families of response surfaces rather than search for one universally optimal prompt.

The concept of criticality also has a useful negative role. It discourages the assumption that more randomness is necessarily beneficial. A highly stochastic condition can increase output diversity while reducing coherence, historical continuity, and evaluability. The candidate near-critical region proposed here is explicitly intermediate: perturbations should matter more, but the system should remain capable of consolidating

and testing what they produce. This is closer to regulated susceptibility than to unconstrained instability.

10.3 Creativity and Mechanistic Attribution

The title-level claim of this paper concerns creativity, so the relationship between regime change and creative performance must remain explicit. The hypothesis is not supported merely because a prompt changes perturbation response. Nor is it supported merely because a prompt improves originality ratings. Evidence for the proposed mechanism requires a relationship between the two outcome families after simpler controls are considered.

Several result patterns are therefore possible. Creativity may improve under a diversity prompt while recovery and persistence remain unchanged. In that case broader search is a sufficient explanation. A criticality-targeting condition may alter recovery and historical retention while producing no improvement in creativity. In that case the intervention has changed inference dynamics but has not established the proposed functional benefit. Creativity and dynamical signatures may both improve, but the same pattern may be reproduced by higher temperature or longer search. That would weaken the distinctive criticality interpretation. The strongest support would occur when criticality-targeting prompts produce a characteristic response geometry under matched budgets and stochasticity, and when variation in that geometry predicts creative gains across tasks and replications.

This requirement helps avoid a common problem in creativity research: treating any increase in difference as evidence of creativity. The outcome framework in Section 7 evaluates originality together with usefulness, problem transformation, hypothesis quality, and generative follow-through. A system that generates increasingly remote but incoherent alternatives is not thereby more creative in the sense relevant to scientific or problem-solving discovery. The proposed intervention is valuable only if increased susceptibility remains compatible with organized inquiry.

A related issue concerns novelty at different levels. Surface novelty can occur through lexical variation or unusual combinations while the underlying problem representation remains stable. Stronger creative change can involve modification of relations among evidence, hypotheses, variables, and problem boundaries. The criticality-targeting hypothesis is most relevant to this stronger level because the proposed mechanism concerns persistent reorganization rather than immediate stylistic diversity. This does not imply that every creative task requires deep representational change; it means only that the hypothesized mechanism should be tested on tasks capable of exhibiting such change.

Creativity scores also cannot establish epistemic value by themselves. A novel and persistent theory can be false. A prompt that weakens restoring tendencies may facilitate useful reframing, but it can also stabilize an attractive error. For scientific applications, creativity must therefore be paired with external evidence, reproducibility, logical consistency, domain constraints, or other task-appropriate validation. Criticality-targeting prompting concerns the generation and preservation of alternatives; it does not replace the standards by which those alternatives are judged.

10.4 Adaptive Prompt Policies and System Boundaries

The adaptive form of the hypothesis raises an important systems question. Once prompt policy changes in response to measured inference behavior, the relevant object is no longer a foundation model in isolation. The effective system includes the model,

prompts, conversation state, memory, retrieval components, tools, evaluators, and the controller that decides when to increase exploration or consolidation. Claims about criticality or creativity should therefore be attributed to this coupled system unless experiments localize the effect more narrowly.

This broader system boundary is especially important for historical retention. A trajectory can remain sensitive to an earlier anomaly because the model's current hidden state differs, because the anomaly remains in context, because an external memory store reintroduces it, or because the controller has changed future prompts. These mechanisms can have similar behavioral consequences while requiring different explanations. The reset hierarchy proposed in Section 6 is therefore not only a methodological detail; it is necessary for identifying where the effective historical state resides.

Adaptive prompting also introduces a controller-design problem. A policy must decide when inference is too stable, when it is sufficiently susceptible, and when it is becoming incoherent. No single scalar is likely to provide a reliable answer. A practical controller would probably combine several proxies, including recovery rate, persistence of unresolved anomalies, representational competition, coherence, common-future divergence, and perhaps task-specific performance signals. The phase diagram in Figure 1 should therefore be read as a conceptual map of competing pressures, not as a literal two-parameter control surface.

The controller itself can also create artifacts. If it repeatedly forces alternatives whenever outputs become similar, apparent susceptibility may simply reflect intervention frequency. If it consolidates only after an external evaluator identifies a useful idea, later persistence can be attributed partly to that evaluator. Experimental work should therefore compare fixed and adaptive policies, report the controller's decision rule, and distinguish autonomous regime effects from controller-imposed structure.

Despite these complications, the coupled-system view has an important advantage. It avoids treating the frozen model as the only legitimate locus of creativity. Human creative work is also distributed across memory, notation, tools, environments, collaborators, and external records. If a model-plus-controller system can preserve weak anomalies, revisit earlier alternatives, revise relations, and regulate exploration and consolidation, it may be scientifically useful even if the underlying foundation model weights never change during the episode. The relevant question is whether the coupled system exhibits the hypothesized response dynamics and produces better creative outcomes under appropriate evaluation.

10.5 Epistemic Turbulence as an Optional Strong Explanation

Epistemic turbulence occupies the highest and most provisional layer of the framework. It is not required for the title-level hypothesis. Prompting could move inference toward a more susceptible regime and improve creativity without producing anything that merits a turbulence analogy. Long recovery, stronger perturbation interaction, or metastable switching would already constitute meaningful dynamical findings.

The turbulence-like interpretation becomes relevant only if perturbations interact recurrently across multiple representational scales, if their effects are intermittent rather than smoothly diffusive, if earlier perturbations alter how later ones propagate, and if these changes persist historically. The critical feature is not large divergence by itself but modification of the propagation structure. Section 8 therefore treats turbulence as a stronger downstream hypothesis rather than as a synonym for creativity or criticality.

This ordering matters for experimental interpretation. If a criticality-targeting prompt increases creativity and perturbation persistence but propagation pathways remain fixed, the main hypothesis may receive support while the turbulence hypothesis

does not. If large cascades occur but are well explained by fixed nonlinear interactions, turbulence-like language remains unnecessary. If multi-scale, history-dependent pathway modification is observed repeatedly and simpler models fail, the analogy may become useful as a compact description of the regime.

The value of retaining this stronger hypothesis is that it points toward phenomena that ordinary diversity metrics miss. A creative system may not merely produce more alternatives; it may become temporarily capable of reorganizing which relations are available for future thought. Whether current language-model systems can exhibit such behavior under prompt control remains an open empirical question. The paper’s methodology is designed so that a negative answer would narrow, rather than invalidate, the broader study of prompt-induced regime change.

10.6 Implications for Discovery-Oriented AI

If prompting can reproducibly regulate inference susceptibility, one implication is that discovery-oriented AI systems may benefit from separating exploration, anomaly preservation, relational revision, and consolidation rather than optimizing all stages through one static instruction. A system designed only to answer efficiently can close too early. A system designed only to diversify can lose coherence. A discovery-oriented architecture may instead need explicit control over when to preserve unresolved discrepancies, when to compare competing representations, and when to consolidate a provisional structure for testing.

This view also changes how failures should be interpreted. Repeated return to a familiar explanation may reflect strong restoring dynamics rather than insufficient token-level randomness. Conversely, incoherent divergence may indicate excessive amplification rather than successful escape from a knowledge attractor. The design problem is therefore not simply to increase creativity but to regulate the conditions under which novelty can become historically consequential while still being evaluated against evidence.

The hypothesis also suggests a more informative evaluation protocol for future systems. Benchmarks could include controlled anomalies that are initially weak, later become relevant, and require reorganization of a problem representation rather than one-shot answer generation. Systems could then be evaluated on whether they preserve the anomaly, whether later evidence reactivates it, whether the resulting reorganization persists under neutral future prompts, and whether the new representation improves independent task performance. Such tasks would test a form of creative continuity that ordinary divergent-thinking benchmarks capture only indirectly.

A final implication concerns prompt design itself. The most effective criticality-targeting prompt may not be a single piece of text. It may be a policy that changes over time in response to measured inference behavior. This shifts the research problem from finding a clever phrase to designing a controllable interaction protocol. The paper therefore points toward an experimental programme in which prompting is studied as adaptive regulation of an inference process rather than only as static instruction engineering.

The framework remains intentionally provisional. Its strongest claims should survive only if matched experiments reveal reproducible regime movement, if that movement cannot be reduced to sampling or instruction artifacts, and if the resulting response geometry contributes to creativity beyond ordinary search expansion. Even if these conditions are not met, the proposed methodology can still clarify how prompts alter temporal inference behavior. The main empirical question is therefore not whether criticality is an appealing metaphor, but whether prompt interventions can

reproducibly change the balance between restoration, susceptibility, and coherence in ways that matter for creative performance.

11 Conclusion

This paper has proposed a preliminary hypothesis for studying prompting as a possible means of regulating the dynamical conditions of large language model inference. The motivating problem is narrower than the general question of whether language models are creative. Contemporary systems can already produce unusual, diverse, and sometimes highly rated outputs under suitable prompting. The unresolved issue is whether prompting can do more than change the immediate response distribution: whether it can alter how an inference process recovers from perturbation, preserves anomalies, maintains competing representations, reorganizes relations, and carries those changes into later reasoning.

The proposed answer is intentionally conditional. Certain prompt policies may move an effective inference system away from a strongly restoring regime and toward a region of increased susceptibility in which small, structured perturbations have more persistent and interactive effects while coherence remains sufficient for useful reasoning. The paper refers to such a region as a candidate near-critical regime. This language does not imply that language-model inference instantiates a known physical critical point, a universal phase transition, or a common control parameter across models. Criticality is treated as an empirical interpretation that becomes useful only if a reproducible cluster of dynamical signatures is observed and simpler explanations are insufficient.

The framework therefore separates four levels that are often conflated. The first is ordinary output variation, which can be increased through sampling temperature, diversity prompts, branching, or additional search. The second is trajectory-level susceptibility, measured through perturbation persistence, recovery, representational competition, and historical retention. The third is creativity, evaluated independently through originality, usefulness, problem transformation, hypothesis quality, and sustained generative follow-through. The fourth is the stronger possibility of turbulence-like epistemic dynamics, which would require recurrent multiscale interaction, non-additive perturbation effects, intermittency, historical persistence, and modification of the pathways through which later perturbations propagate. Evidence at one level should not be treated as evidence for all the others.

This separation changes the role assigned to prompts. In ordinary response engineering, a prompt specifies what the model should produce. In the stronger regime-engineering hypothesis developed here, a prompt policy is designed to influence how inference unfolds over time. It may delay premature closure, preserve unresolved anomalies, maintain competing frames, require relational revision, or adapt exploration and consolidation in response to observed trajectory behavior. The strongest version is therefore unlikely to be a single clever instruction. It is more naturally understood as a temporally extended interaction policy that monitors the effective state of inference and adjusts the pressure toward exploration or stabilization.

The experimental programme follows from this interpretation. The appropriate unit is not a single response but a matched trajectory block in which model, task, evidence order, perturbation, compute budget, and other controllable factors are held as constant as possible while prompt policy varies. Small perturbations can then be introduced under matched conditions, and their effects can be followed across multiple rounds. Common-future probes and reset tests are especially important because they help distinguish temporary prompt compliance from more persistent changes in the effective response struc-

ture. The relevant question is whether a perturbation remains consequential after the immediate instruction is removed and whether later behavior differs systematically under otherwise matched inputs.

The hypothesis is falsifiable in several ways. If high temperature or ordinary diversity prompting reproduces the same persistence, interaction, and creativity effects, the criticality interpretation loses explanatory value. If additional search budget explains the effect, regime movement need not be invoked. If apparent historical persistence disappears when prompt instructions are withdrawn, direct compliance may be sufficient. If explicit memory replay explains later divergence, there is no need to posit deeper reorganization. If fixed nonlinear response or metastable switching accounts for the observed trajectories, stronger critical or turbulence-like interpretations should be withheld. The framework is therefore designed to contract when simpler mechanisms are adequate rather than to absorb every positive result into a single theory.

A second limitation concerns model and system boundaries. Any prompt-induced regime is likely to depend on architecture, alignment procedure, context management, memory design, retrieval components, tools, and controller logic. Once prompting becomes adaptive, the relevant object may be a coupled inference system rather than a foundation model in isolation. Claims should therefore be localized to the system actually tested. A prompt policy that increases susceptibility in one model family may fail in another, and the location of historical state may differ across hidden activations, context, external memory, retrieval, or orchestration. There is no reason to assume a universal criticality-targeting prompt.

The creativity claim also remains independent of the dynamical claim. A system can become more sensitive to perturbation without becoming more creative. It can also become more creative without displaying distinctive near-critical signatures. A useful test of the present hypothesis therefore requires both outcome families: evidence that the intervention changes inference dynamics in the predicted way, and separate evidence that those changes contribute to creativity beyond ordinary diversity, search expansion, or compute. Even then, creativity does not establish epistemic value. Persistent novelty can stabilize an error as easily as a useful hypothesis. Discovery-oriented applications must retain independent standards of evidence, consistency, reproducibility, and domain-specific validation.

Epistemic turbulence occupies the most provisional position in the framework. It is not required for prompt-induced criticality and is not required for creative improvement. Its value is narrower: it names a possible downstream regime in which perturbations interact across scales and alter the structures through which later perturbations propagate. If experiments reveal only increased persistence, slower recovery, or metastable switching, those findings are already meaningful and should be described at that level. Turbulence-like language becomes warranted only if the additional signatures survive repeated tests and simpler dynamical accounts fail.

The broader research opportunity is therefore not to make models maximally unstable or maximally diverse. It is to determine whether inference can be deliberately regulated so that alternatives remain available long enough to interact, anomalies remain recoverable long enough to matter, and provisional structures remain coherent enough to support evaluation and further reasoning. Such a balance may be especially relevant to scientific discovery, conceptual reframing, and other tasks in which rapid convergence can be as limiting as uncontrolled divergence.

The title-level hypothesis can thus be stated in a deliberately modest form. Prompt design may provide one route for moving some large language model systems toward inference regimes in which perturbations are more capable of producing sustained, structured consequences. Whether those regimes are appropriately described as near-

critical, whether they improve creativity, and whether stronger turbulence-like dynamics ever emerge are separate empirical questions. The value of the proposal lies less in assuming their answers than in making those questions experimentally distinguishable.

References

- [1] Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:100207, 2025.
- [2] Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16:1279, 2026.
- [3] Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.
- [4] Hung-Fu Chang and Tong Li. A framework for collaborating a large language model tool in brainstorming for triggering creative thoughts. *Thinking Skills and Creativity*, 56:101755, 2025.
- [5] Qunlin Chen, Yoed N. Kenett, Zaixu Cui, et al. Dynamic switching between brain networks predicts creative ability. *Communications Biology*, 8:54, 2025.
- [6] Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.
- [7] Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, 1995.
- [8] Jennifer Haase, Paul H. P. Hanel, and Sebastian Pokutta. Has the creativity of large-language models peaked? an analysis of inter- and intra-LLM variability. *Journal of Creativity*, 35(3):100113, 2025.
- [9] Keith B. Hengen and Woodrow L. Shew. Is criticality a unified setpoint of brain function? *Neuron*, 113(16):2582–2598.e2, 2025.
- [10] Björn Hof. Directed percolation and the transition to turbulence. *Nature Reviews Physics*, 5:62–72, 2023.
- [11] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, 2024.
- [12] Zhicheng Lin. How to write effective prompts for large language models. *Nature Human Behaviour*, 8:611–615, 2024.
- [13] Tobias M. Schneider, Bruno Eckhardt, and James A. Yorke. Turbulence transition and the edge of chaos in pipe flow. *Physical Review Letters*, 99(3):034502, 2007.
- [14] Emmanuelle Tognoli and J. A. Scott Kelso. The metastable brain. *Neuron*, 81(1):35–48, 2014.

- [15] 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.
- [16] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.
- [17] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
- [18] Wei Xia, Haoqing Wang, Zhi-Hong Deng, and Yehui Tang. When do llms reason? a dynamical systems view via entropy phase transitions. *arXiv preprint arXiv:2605.22873*, 2026.
- [19] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [20] Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. Dissecting failure dynamics in large language model reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8893–8914, 2026.