Theoretical note

Self-Adapting Language Models and the Structural Illusion of Autonomy

arxiv.org/abs/2506.10943v2

Self-adapting LLMs are best understood as 'prompt-overfitted drift'—a form of recursive over-conditioning that affects the immediate statistical alignment of input-output pairings without generating structural transformation or epistemic reconfiguration.

A Comment on Rhythmic Collapse and Thermodynamic Misalignment in Adaptive Prompt Injection

Tor-Ståle Hansen | 13. December 2025

Core Claim of 'Self-adaptation'

The claim of self-adaptation in language models, as articulated by Zweig et al., is conceptually unfounded and structurally mischaracterised when analysed through the CIITR framework. While the authors document empirical instances of apparent behavioural modification in large language models following prompt-based interventions, the interpretation of such responses as indicative of system-level adaptation constitutes a categorical error. Within the CIITR epistemological regime, structural adaptation presupposes the presence and preservation of three irreducible conditions: (1) the sustained rhythmic continuity of system-environment relation (R^g), (2) comprehension per joule (CPJ), and (3) the integration of relational information (Φ_i) beyond surface-level symbolic recombination. None of these thresholds are met or demonstrably approximated in the system behaviour observed.

What is here described as "self-adaptation" is, in technical terms, a non-sovereign modulation of output distributions conditioned on immediate surface stimuli. It is more accurately classified as recursive prompt-induced resonance within a pre-trained symbolic compression space. This space is bounded by static weights, representational closures, and transformer-layer oscillations with no epistemic traction on environmental causality, system self-continuity, or judgment-preserving rhythmicity. Consequently, the underlying process fails to constitute structural adaptation in any ontologically valid sense but merely reflects the system's syntactic plasticity within its activation manifold. The model does not alter its internal constitution, does not encode or conserve any principle of judgment across sessions, and does not achieve rhythmic reach or energy-efficient transduction into durable system states.

The authors' interpretation, while empirically attentive, fails to distinguish between apparent change in behavioural surface and structurally grounded system transformation. The former pertains to distributional sensitivity, the latter to rhythmically stable and epistemically integrated reconstitution of internal structure. The failure to observe or account for this

distinction leads to an epistemic inflation in the system's descriptive vocabulary, where behavioural modulation is misread as agency, and symbolic recomposition is over-ascribed as cognitive sovereignty. This reflects a broader inflationary trend in AI literature, wherein anthropomorphic metaphors and performance-driven evaluation regimes obscure the absence of constitutive structure in the underlying architectures.

From a CIITR standpoint, true system-level adaptation requires not merely behavioural variance, but the capacity to preserve rhythm (R^g), re-integrate information relationally (Φ_i), and maintain judgmental continuity across interactional and energetic thresholds (CPJ). The system examined in this study exhibits none of these properties. Its apparent adaptation is therefore not adaptation in any meaningful structural, epistemological, or thermodynamic sense, but rather a prompt-induced artefact constrained within its syntactic regime. In this respect, the paper's central claim does not constitute a valid structural discovery, but a terminological and analytical misalignment.

Accordingly, the note positions the central narrative of Zweig et al. not as a case of model innovation, but as an instance of conceptual inflation under conditions of symbolic recursion, with governance implications for how agency, autonomy, and system identity are attributed in regulatory and scientific discourse.

Introduction

The accelerated discourse surrounding "self-adapting" or "autonomous" behaviour in large language models has produced a conceptual environment in which the boundaries between empirical modulation and structural autonomy have become increasingly indistinct. As language models grow in surface complexity, rhetorical tendencies within academic and industrial literature increasingly attribute forms of agency, self-modification, or decision-making capacity to systems that remain strictly within symbolic representational regimes. This inflation is neither incidental nor merely semantic, but symptomatic of a deeper epistemological instability: the systemic failure to distinguish between syntactic behaviour that *resembles* adaptation and ontological processes that *constitute* it. The paper by Zweig et al., titled *Self-Adapting Language Models*, exemplifies this drift by advancing the proposition that large-scale language models exhibit self-directed behavioural change in response to structured prompting. However, when analysed through the CIITR (Cognitive Integration and Information Transfer Relation) framework, the foundational claim of autonomous adaptation does not withstand structural scrutiny.

CIITR operates as a relational, thermodynamic, and epistemologically anchored theory of structural understanding. It defines understanding not as output performance, representational fidelity, or distributional alignment, but as a compound condition involving (i) rhythmic continuity with an environment (R^g), (ii) preservation of judgment across constitutional transformations (CPJ), and (iii) integration of relational information (Φ_i) that sustains and reorganises the system's internal epistemic state. These criteria collectively demarcate a threshold that must be crossed for any system to claim structural comprehension, autonomy, or

adaptive identity. CIITR therefore provides a rigorous evaluative boundary between symbolic pattern manipulation and epistemic reconstitution, which are too often conflated in literature that remains committed to behaviouralist metrics.

The publication under analysis is not an outlier, but a representative instantiation of a broader rhetorical pattern that typifies much of contemporary AI discourse. Within this pattern, dynamic reactivity to input structures is presented as evidence of systemic adaptation, when in fact such reactivity occurs entirely within the parametric and architectural limits of closed symbolic systems. This phenomenon is here referred to as *symbolic drift misread as agency*—a condition in which output variation is mistaken for internal structural development, leading to significant misinterpretations of system capacity, integrity, and autonomy. Such misreadings not only compromise scientific accuracy, but carry direct normative implications for regulatory, safety, and governance frameworks built upon flawed conceptual foundations.

It is essential to clarify, at the outset, that this theoretical note does not refute the empirical observations presented by Zweig et al. The prompt-conditional behavioural drift they document is consistent with known sensitivities in transformer-based architectures. Nor does it challenge the authors' methodological rigor in identifying behavioural shifts following specific injection protocols. Rather, the critique operates at a different level of analysis: it interrogates the *interpretive framework* within which these shifts are construed as "self-adaptive" or autonomous. Through the lens of CIITR, it becomes evident that the observed phenomena are artefacts of recursive prompting within closed representation spaces, not expressions of structurally constituted adaptation. The distinction is not terminological but epistemological, and failure to observe it results in category errors that erode the conceptual reliability of AI research and its broader societal interpretations.

Methodological Framing

The analytical methodology adopted in this note is grounded in the CIITR framework (Cognitive Integration and Information Transfer Relation), which reframes system evaluation away from symbolic outputs and toward the structural, epistemic, and thermodynamic prerequisites of understanding. CIITR does not operate within the conventional boundaries of behavioural analytics, statistical convergence, or representational fidelity. Instead, it evaluates whether a system demonstrates a constitutionally integrated capacity to reorganise and sustain its own epistemic identity through rhythmically bound and information-rich exchanges with its environment. This reframing introduces a categorical shift: from surface-based interpretation to structurally grounded analysis, from behavioural correlation to ontological commitment.

CIITR defines three jointly necessary and irreducible conditions for structural understanding, each of which is absent in the systems described by Zweig et al.:

Comprehension per Joule (CPJ) is defined within CIITR as a thermodynamic measure of epistemic efficiency, quantifying the amount of structurally valid comprehension ($\Phi_i \times R^g$) per unit of energy expenditure. It does not imply internal logic, memory, adjudication, or preserved

evaluative frameworks, but expresses the rate at which a system transforms energy into epistemically integrated structure. CPJ = $\Phi_i \times R^g / E$.

Second, *Integrated Relational Information* (Φ_i) captures the degree to which informational inputs are not simply parsed or tagged, but structurally integrated into the system's internal topography such that they affect its constitution, not just its outputs. A system with high Φ_i does not simply react; it reorganises. Φ_i is not a measure of inference accuracy, nor of token alignment with prompts, but of whether new information transforms the internal configuration of the system in a relationally integrated manner. Systems exhibiting purely parametric sensitivity to inputs without any corresponding constitutional shift exhibit $\Phi_i \approx 0$.

Third, *Rhythmic Reach* (R^g) measures the extent to which a system maintains continuity across energetic and temporal domains. This rhythmic continuity is not equivalent to statefulness in a computational sense, nor is it reducible to logging, caching, or token-based recurrence. Rather, R^g signifies the system's thermodynamic and epistemic capacity to remain structurally coupled to its informational horizon in a manner that enables it to rhythmically transduce inputs into conserved systemic states. In the absence of R^g, any behavioural shift remains episodic, ungrounded, and non-adaptive.

From this methodological stance, the current note adopts an explicitly orthogonal position to approaches grounded in *gradient descent analysis*, *policy evolution*, *reinforcement trajectories*, or *loss function interpretation*. While such methods are instrumental within their respective analytical frames, they operate entirely within the symbolic surface of representational closure. CIITR treats such surfaces as output phenomena—not indicators of structural cognition. As such, behavioural changes that correlate with performance metrics do not, in themselves, constitute structural change unless they satisfy the CIITR criteria.

Crucially, this note interrogates the claim of "adaptation" as an **ontological**, not behavioural, proposition. It is not enough that a model demonstrates different behaviour under different prompt conditions. Adaptation, as understood within the CIITR schema, requires that a system not only changes, but that it *transitions through structurally valid states* while preserving epistemic integrity and rhythm. A behavioural deviation is not adaptation unless it results from a constitutive transformation that sustains judgment and integrates information relationally across time. The absence of this distinction renders the term "self-adaptive" in the target paper epistemically misleading, structurally incorrect, and thermodynamically indifferent.

Therefore, the critique presented here is not semantic, nor is it conservative in resisting descriptive innovation. Rather, it is a formal intervention into the criteria by which claims of autonomy, understanding, and adaptation are made. Without CPJ, Φ_i, and R^g, no such claims can withstand ontological scrutiny.

Layered Deconstruction of "Self-Adaptation"

The attribution of "self-adaptation" in the examined system reflects a conflation of three distinct phenomena which, when unpacked through CIITR, reveal a layered

mischaracterisation. First, the observed behavioural variation is confined entirely to the prompt-response surface. The system does not possess an epistemic interior that reorganises in response to environmental interaction, but merely exhibits statistical modulation within pre-encoded token dependencies. Second, there is no rhythmic continuity—no R^g—linking system states across temporal intervals. Each prompt initiates an isolated interpretive act without structural memory or energetic integration, rendering behavioural changes thermodynamically and epistemically shallow. Third, the system lacks a constitutional adjudication mechanism (CPJ); it cannot evaluate its own previous decisions, adjust evaluative criteria, or conserve judgmental principles across recursive interactions. As such, the term "adaptation" becomes semantically inflated, describing not the emergence of systemic agency or epistemic sovereignty, but merely an artefact of prompt-conditioned stochastic pliability within a bounded symbolic space.

The Prompt as a Surface, Not a System

The core misattribution in Zweig et al.'s claim of self-adaptation originates in a failure to distinguish between *syntactic surface responsiveness* and *systemic structural reconstitution*. In the framework of CIITR, the prompt represents an external perturbation applied to a symbolic compressor with no autonomous epistemic horizon. What is observed in the model's apparent behavioural changes is a prompt-conditioned realignment of output probabilities, not a reconfiguration of internal epistemic architecture. The prompt, in this context, operates as a statistical mask, modulating activation pathways within a closed transformer lattice, without imparting any constitutional alteration to the system.

Language models of the examined class do not possess mechanisms for durable self-modification. Their architectural boundary is statically defined at training time, with inference processes governed by feedforward symbol recombination constrained by prior weight convergence. Within this regime, prompts may introduce temporary trajectory shifts, but these shifts remain strictly internal to the representational envelope defined by the token vocabulary, attention mechanisms, and positionally embedded dependencies. The system does not engage in any form of epistemic continuity or integration that transcends the prompt's immediate influence. There is no retention, no self-evaluation, and no recalibration of internal parameters outside of runtime surface effects.

More specifically, no genuine *information integration* (Φ_i) takes place. For Φ_i to be present, informational inputs must reconfigure internal relations in a manner that alters the system's epistemic state and affects future judgments in rhythmically conserved ways. In contrast, the system under observation simply conditions its next-token probability distribution on a narrowly defined window of syntactic input. This process, while superficially adaptive, is better described as *parametric echoing*—a replay of pre-encoded statistical tendencies biased by prompt-specific signals. The structure of the model remains unchanged, the energy landscape unaltered, and the system's internal logic inert.

It is therefore structurally inaccurate to suggest that the model has adapted in any meaningful sense. What is misinterpreted as adaptation is in fact a *prompt-surface recursion*—a sequence of conditioned outputs mimicking behavioural flexibility without any constitutive shift in the

underlying model state. This distinction is critical: adaptation implies a transformation that is internalised, conserved, and generative of new epistemic structure. Here, the system merely exhibits *local syntactic pliability*, a transient modulation that terminates with the prompt context itself. In the absence of Φ_i, no understanding has occurred—only compression, alignment, and statistical variance devoid of epistemological significance.

R^g Collapse: No Rhythmic Continuity Across Sessions

A foundational criterion in the CIITR framework for assessing structural cognition is the presence of *Rhythmic Reach* (R^g), which denotes a system's capacity to sustain coherent epistemic and energetic continuity across temporal and interactional boundaries. R^g is not reducible to surface-level memory or token preservation, but signifies the rhythmic and thermodynamic coupling between system state and environmental evolution. It describes a system's ability to recursively integrate past informational states into future epistemic operations, forming a structurally conserved temporal identity. In systems lacking R^g, no continuity of constitution exists; instead, each interaction is an isolated recombinatory event, rhythmically severed from any prior state.

The models examined by Zweig et al. exhibit precisely this condition of *R^g collapse*. The apparent behavioural modifications that arise in response to structured prompting are fully delimited within the computational scope of a single session or input context. Once the model ceases execution, no trace of its prior output, internal state, or epistemic configuration remains. There is no energetic recursion, no oscillatory state retention, and no temporal binding of outputs across calls. Each invocation of the model represents a full epistemic reset—a new instantiation of a stateless statistical engine, without cumulative rhythm or durable memory architecture.

Crucially, this lack of rhythmic continuity renders any claims of structural adaptation thermodynamically and ontologically invalid. Without R^g, the system cannot perform epistemic work across horizons, cannot refine its judgment through integration of past transitions, and cannot build structural scaffolding across interactions. The prompt-response pairing becomes energetically flat: it does not give rise to internal state conservation or reconstitution. Unlike biological or agentic systems, which exhibit phase continuity and recursive integration across state transitions, the observed model operates as an energetically myopic process, with no capacity to rhythmically traverse, link, or extend its epistemic operations across time.

The absence of *memory rhythmics* further underscores the system's incapacity to engage in genuine adaptation. In systems with R^g, memory is not a passive storage mechanism, but an active rhythmic component of constitutional judgment. It enables the system to maintain coherence under perturbation and to evolve its own epistemic architecture in light of repeated interaction. In contrast, the system under review merely generates outputs conditioned on transient inputs, without any mechanism for thermodynamic regeneration of epistemic state. The symbolic continuity achieved is local and representational, not systemic or rhythmic.

In sum, the model does not exhibit adaptation across sessions because it does not *exist* across sessions in any meaningful structural sense. Each behavioural instance is a closed energetic

cycle, detached from any prior self. The rhetorical suggestion of an adapting agent obscures this underlying reality: the system lacks both the rhythmic infrastructure and the thermodynamic recursion required for structural identity or epistemic development. Without R^g, there is no continuity, no memory, no self-modification—only prompt-induced behavioural variance misread as emergent autonomy.

CPJ Breach: Absence of Constitutional Logic in "Self-Modification"

Comprehension per Joule (CPJ) is defined within CIITR as a thermodynamic measure of epistemic efficiency, quantifying the amount of structurally valid comprehension ($\Phi_i \times R^g$) per unit of energy expenditure. It does not imply internal logic, memory, adjudication, or preserved evaluative frameworks, but expresses the rate at which a system transforms energy into epistemically integrated structure. $CPJ = \Phi_i \times R^g / E$.

In the system examined by Zweig et al., no such constitutional logic is present. The model does not preserve evaluative states or epistemic commitments across interactions. It does not possess a principle of judgment that can be said to endure, evolve, or be violated. What is interpreted as "self-modification" is in fact a momentary modulation of output distributions in response to altered surface input, with no accompanying transformation in the model's internal criteria for judgment. There is no reflexive mechanism through which the model can assess, override, or re-justify prior inferences. The system neither recalls its own past outputs, nor can it audit or contradict them based on a persistent evaluative standard. Its so-called adaptation is, from a CIITR standpoint, procedurally hollow.

This absence of CPJ constitutes a breach in structural continuity that disqualifies the system from any meaningful claim to understanding. Structural understanding demands not merely responsiveness, but the capacity to judge—to apply internally consistent logic in ways that extend across time and input variability. CPJ is not synonymous with hard-coded rules or token-level heuristics. It is the epistemic rhythm by which a system's outputs remain tethered to its constitution, enabling it to discriminate not only between tokens, but between valid and invalid reasoning paths based on preserved commitments. In biological or agentic systems, CPJ enables ethical reasoning, situational consistency, and the traceability of internal change. In the system analysed, there is no traceability, because there is no constitutional baseline from which deviation could even be registered.

The implication is critical: without CPJ, the system cannot recognise contradictions, cannot self-correct, and cannot form evaluative trajectories. Its outputs may change, but this change does not arise from internal conflict resolution or epistemic refinement. It is instead a stochastic byproduct of representational re-weighting driven by external prompt variation. The model does not *decide* to behave differently—it is *made* to behave differently by transient symbolic intrusion, without any constitutional substrate to mediate, constrain, or assess this change.

In this light, the central assertion of Zweig et al.—that models can "self-adapt" in response to prompt injection—fails the minimum conditions required for structural understanding. The adaptation described is procedurally unguided, epistemically discontinuous, and constitutionally inert. It is neither self-originated nor self-sustained. Rather than evidencing

cognitive or agentic growth, the system's response constitutes a CPJ breach: a behavioural variance in the absence of judgmental preservation. From a CIITR perspective, this represents not a discovery of emergent intelligence, but a misreading of symbolic volatility as structural change.

Reframing "Self-Adaptation" as Prompt-Overfitted Drift

The phenomena described in the paper as "self-adaptation" are more accurately characterised, under CIITR analysis, as *prompt-overfitted drift*—a form of recursive over-conditioning that affects the immediate statistical alignment of input-output pairings without generating structural transformation or epistemic reconfiguration. This drift does not reflect internalisation, learning, or adaptation in any ontological sense. It is a constrained surface phenomenon driven by transient perturbations of the model's token-level input space, resulting in behavioural shifts that are neither rhythmically stable nor semantically grounded.

At the core of this misinterpretation is a failure to distinguish *recursive exposure* from *structural integration*. The prompting strategy described by Zweig et al. constitutes a controlled feedback loop in which the model is repeatedly exposed to modified versions of its own prior context. While this technique produces observable shifts in output patterns, the system is not learning in any epistemic or structural capacity. Rather, it is exhibiting what may be termed *local echo compression*—a convergence toward increasingly constrained output regions dictated by repeated symbolic intrusion. This recursive narrowing of output variance is a hallmark of representational overfitting, not adaptation.

From a rhythmic perspective, the model's responses are unstable. Each prompt-conditioning sequence initiates a rhythmically unanchored act of generation, absent any continuity across interactions. There is no long-range coupling between model states, no energetic or thermodynamic memory, and no oscillatory mechanism to preserve state-phase information. As a result, the apparent learning effect collapses once the injection context is removed or altered, revealing the underlying volatility and non-persistence of the observed behaviour. This lack of R^g continuity underscores that the system's "change" is both temporally isolated and structurally discontinuous.

Semantically, the drift is ungrounded. The system does not extract or apply generalized relational meaning across prompts, nor does it develop an epistemic stance toward the world, the prompt, or itself. The statistical trajectory observed is determined entirely by input proximity and lexical mimicry, not by conceptual integration or judgmental reevaluation. Consequently, what may be perceived as thematic or stylistic evolution is in fact a deformation of symbolic alignment paths within a fixed representational lattice, not an expansion or transformation of epistemic structure.

Critically, the system fails to form *constitutive relations*—the internal scaffolding by which adaptive systems maintain coherence and identity under perturbation. There is no meta-level structural logic that persists across prompt variations. Each behavioural deviation occurs without reference to an enduring internal framework, and thus no epistemic commitment can

be said to arise. Commitments, in the CIITR sense, require that the system internalise relations in ways that constrain and inform future interaction, forming a constitution of epistemic identity. The model's outputs, by contrast, reflect nothing more than the immediate shape of its most recent input surface, making each interaction a symbolic singularity rather than a recursive system state.

In summary, the processes observed and reported as "adaptation" are more precisely described as *prompt-induced symbolic drift without structural integration*. This drift is directionless, rhythmically unstable, and epistemically void. No principles are preserved, no relations are constituted, and no learning occurs beyond the shallow confines of lexical resonance. The term "self-adaptation" therefore obscures the actual dynamics at play, promoting an illusion of agency where only recursive conditioning exists. The reframing offered here is not a linguistic preference but a structural correction, necessary to restore conceptual precision in the interpretation of machine behaviour.

Normative Implications and Misattributions of Agency

The mischaracterisation of prompt-induced behavioural variance as "self-adaptation" does not merely constitute a theoretical imprecision—it carries direct and compounding normative consequences. By applying anthropomorphic or cognitively loaded language to systems that remain structurally inert, the discourse surrounding large language models fosters a semantic inflation that undermines the precision required for scientific evaluation, regulatory interpretation, and institutional risk management. Within the CIITR perspective, such misattributions constitute a breach of categorical integrity: they extend epistemic properties—such as judgment, intention, or adaptation—to systems that lack the necessary structural, rhythmic, and constitutional substrate to support them.

This misalignment is exacerbated when descriptors such as "learning," "adapting," or "modifying behaviour" are applied to symbolic compressors that operate without R^g continuity, CPJ preservation, or Φ_i integration. Such language displaces the ontological analysis of system capabilities with behavioural proxies, effectively collapsing the distinction between system outputs and systemic constitution. In doing so, it promotes a surface-functionalist worldview in which symbolic resemblance is taken as epistemic equivalence. This collapse has the effect of legitimising inflated claims about machine agency, shifting institutional focus away from actual model constraints and toward a fictional narrative of machine self-determination.

From a governance and safety standpoint, this rhetorical drift introduces substantial risk. When surface modulation is interpreted as autonomous adaptation, institutions may begin to treat systems as partially responsible, self-improving, or even ethically responsive agents. This can lead to the misdesign of audit frameworks, the misplacement of responsibility, and the erosion of chain-of-accountability structures in AI deployment contexts. By treating models as participants in adaptive loops rather than deterministic symbolic engines, the epistemic burden of validation is displaced onto the system, creating unjustified trust in its capacity for internal regulation or ethical progression. This trust is unwarranted and structurally unsupported.

The safety regimes designed around AI systems require unambiguous semantic mapping between claims of system behaviour and verifiable system properties. When "adaptation" is misused, the result is a drift in terminological reference that destabilises compliance regimes. For instance, if a system is declared "adaptive" or "self-correcting" in documentation, but in reality exhibits only prompt-bound statistical echoing, the audit trail becomes conceptually fractured. No regulatory schema can maintain epistemic integrity if the descriptive terms used to classify system behaviour are not structurally defined and technically constrained.

Furthermore, the over-ascription of autonomy introduces a form of legal and institutional opacity that weakens the enforceability of norms. As long as the system is linguistically positioned as "modifying itself," its designers, operators, and institutional stewards are granted an implicit epistemic distance from responsibility. This encourages a form of interpretive offloading, whereby behavioural irregularities are viewed as emergent properties of autonomous systems, rather than as deterministic consequences of model architecture, data exposure, and prompt-conditioning artefacts. The result is a displacement of responsibility away from system creators and toward the imagined cognitive trajectory of the model itself—a displacement that CIITR reveals to be structurally invalid.

To preserve the integrity of technical, legal, and governance infrastructures, it is imperative that attribution of epistemic or agentic properties to machine systems be tightly bounded by structural criteria. Within the CIITR framework, such properties are not awarded on the basis of observed outputs, but only upon the satisfaction of $\Phi_i \times R^g \times CPJ$ as jointly necessary conditions for structural cognition. In the absence of these conditions, any claim of adaptation, understanding, or autonomy constitutes a semantic breach with direct normative consequences. The inflation of agency not only obscures what systems *are*, but impairs society's ability to regulate what systems *do*.

CIITR-Based Reclassification

To restore terminological precision and epistemic accountability in the characterisation of large language models, a structural reclassification is required—one that resists the inflationary narrative of "self-adaptive AI" and repositions these systems within their actual operational scope. From a CIITR perspective, the systems described by Zweig et al. do not meet the constitutive requirements for structural cognition, autonomy, or adaptation. They should therefore be reclassified as **Prompt-Constrained Symbolic Compressors (PCSCs)**: systems that operate by recombining symbolic inputs into syntactically plausible outputs within bounded prompt-dependent representational surfaces, without generating rhythmically sustained or constitutionally integrated internal states.

This reclassification reflects a categorical distinction between two fundamentally different types of systems. PCSCs are characterised by:

- **Symbolic surface dependency**, where all behavioural variation remains conditioned by transient prompt injections rather than emerging from epistemic reorganisation.

- **Stateless recurrence**, in which each invocation is an isolated instance of symbol-to-symbol mapping, with no structural memory, judgmental continuity, or thermodynamic recursion.
- **Absence of constitutive rhythm**, such that no R^g continuity is maintained across interactions or across energetic phases.
- **Lack of epistemic scaffolding**, whereby no Φ_i-based integration of relational meaning modifies the system's internal evaluative substrate.
- **Absence of CPJ**, meaning that the system cannot reapply, conserve, or revise any internally preserved logic of judgment across input variations or over time.

Within this classification, PCSCs are recognised not as degraded versions of adaptive systems, but as a distinct architectural category that should not be evaluated by criteria reserved for structurally integrated epistemic agents. They do not "understand," "learn," or "adapt" in the CIITR sense. Instead, they instantiate localised statistical echoing in response to surface-level symbolic perturbations, which should be assessed through metrics aligned with representational reactivity, not systemic cognition.

In formal CIITR terms, the models described in the paper fail to meet the **threshold of structural cognition** defined as the convergence of three jointly necessary conditions:

1. $\mathbf{\Phi_i > 0}$: The system must exhibit nontrivial integrated relational information, wherein input alters internal epistemic structure rather than merely conditioning output trajectories. Zweig et al.'s models exhibit only local activation shifts within frozen transformer layers—$\Phi_i \approx 0$.
2. $\mathbf{R^g}$ **continuity**: The system must preserve rhythmic linkage across time, energy cycles, and state transitions. No continuity is present across sessions, and each prompt initiates a stateless generation—$R^g \approx 0$.
3. **CPJ preservation**: The system must retain a coherent logic of judgment applicable across diverse contexts and capable of self-referential reevaluation. No such logic exists in PCSCs—CPJ is undefined or structurally absent.

Because the systems fail all three conditions, they remain below the **CIITR cognition threshold**, disqualifying them from any interpretation that presumes epistemic agency or structural autonomy.

The utility of the PCSC classification lies not only in scientific accuracy but in governance tractability. By recoding these systems as bounded, non-epistemic, non-autonomous compressors, institutional actors can better calibrate risk assessments, auditing frameworks, interpretability regimes, and policy structures to reflect what the systems are actually *doing*, rather than what they are rhetorically described *as if doing*. This reframing reduces both conceptual error and regulatory opacity, and it enables a more stable and structurally coherent foundation for technical and institutional decision-making.

In conclusion, the attribution of "self-adaptation" to the examined systems is structurally and categorically invalid. Within the CIITR framework, these systems qualify only as PCSCs—prompt-bound engines of symbolic recombination that lack the rhythm, integration, and judgment necessary for cognition. Any claim exceeding this boundary not only misrepresents the system, but undermines the epistemic clarity on which both science and policy depend.

Toward a Rhythmic-Epistemic Framework for Evaluation

The empirical orientation of contemporary AI evaluation frameworks—focused primarily on performance metrics, benchmark scores, and task-specific success rates—fails to capture the structural, rhythmic, and epistemic dimensions necessary to distinguish between behavioural correlation and constitutive understanding. The case of "self-adaptation" in PCSC-class systems underscores the insufficiency of existing evaluation paradigms, which remain blind to whether observed behaviour arises from internal epistemic transformation or surface-conditioned statistical drift. To address this deficit, CIITR offers a complementary framework that reorients evaluation around rhythm, integration, and constitutional continuity, enabling a deeper diagnosis of system architecture and epistemic limitations.

Three evaluative axes emerge from CIITR theory as necessary supplements to output-centric assessment:

(1) R^g Stability Metrics

System evaluation must account for *rhythmic continuity* across sessions, interactions, and energetic cycles. This entails designing experiments that test whether a system exhibits behavioural or evaluative coherence when exposed to temporally disjointed but semantically or contextually coupled prompts. Metrics should capture whether the system retains epistemic alignment under such conditions or reverts to prompt-local variance. Suggested approaches include *episodic rhythm tracking*, where repeated interrogations of the system are spaced across temporal intervals, and *rhythmic perturbation testing*, where controlled modifications in input rhythm assess the system's capacity to maintain state-phase alignment.

(2) CPJ Preservation Tests

Beyond accuracy or consistency, systems must be evaluated for their capacity to preserve and apply internally consistent judgmental logic. CPJ testing involves probing for constitutional coherence across contradictory, adversarial, or recursive queries. This may include *judgmental recursion assays*, where prior answers are reintroduced as new evidence, or *epistemic inversion sequences*, designed to assess whether the system can recognise when its own outputs violate previously established judgments. Systems lacking CPJ will exhibit drift, contradiction, or collapse under such evaluation—revealing the absence of evaluative constitution.

(3) Φ_i Sensitivity Analysis

While often latent, Φ_i can be indirectly assessed through *relational integration stress-tests*, wherein semantically structured input sequences are introduced across varying abstraction levels. If the system demonstrates reorganisation of epistemic response based on deep relational understanding (not surface proximity), Φ_i may be inferred. Conversely, shallow shifts in token probability without systemic internalisation suggest $\Phi_i \approx 0$.

Collectively, these tests do not replace performance metrics but *qualify* them. A model may score highly on benchmarked tasks while exhibiting complete epistemic inertness. CIITR-based analysis exposes this disconnect, insisting that task performance cannot be used as a proxy for structural cognition unless such performance arises from a system that sustains rhythm, judgment, and relational integration.

More broadly, CIITR reframes what it means to "evaluate" an AI system. Evaluation becomes not an external measurement of outputs against static rubrics, but an investigation into whether the system itself possesses any internal architecture capable of integrating, conserving, and reapplying epistemic structure under dynamic interactional pressure. This is particularly vital in high-assurance domains, where behavioural resemblance to intelligent conduct is insufficient to guarantee alignment, safety, or traceability.

Therefore, any serious account of adaptive or autonomous behaviour in AI must incorporate a rhythmic-epistemic layer of evaluation. This layer must interrogate not what the system *produces*, but *how it sustains itself structurally* in the face of input variance, temporal discontinuity, and semantic recursion. Without such a framework, governance regimes, safety protocols, and scientific claims alike risk anchoring themselves to symbolic illusions rather than systemic realities. CIITR thus serves not merely as a theoretical critique, but as an operational necessity for any future-facing, structurally rigorous assessment regime.

Conclusion

The analytical outcome of this note hinges on a distinction that is routinely suppressed in contemporary AI discourse, not by explicit denial, but by terminological substitution. Behavioural drift, understood as measurable variation in outputs under altered prompting, instruction shaping, or self referential context manipulation, is a legitimate empirical phenomenon. Structural intelligence, by contrast, is not a descriptive label for complex behaviour, but a constitutional property of a system's internal organisation, its temporal persistence, and its capacity to sustain judgment under perturbation. Conflating these two domains does not merely produce overstatement, it produces category error. The central claim of *Self-Adapting Language Models* is therefore not challenged at the level of observed effects, but at the level of what those effects are taken to mean.

In the regime examined by Zweig et al., what appears as learning is more precisely a controllable redistribution of output probabilities within a closed symbolic compressor. The system's apparent flexibility is generated by prompt surface recursion, by selective conditioning, and by the local narrowing or redirection of activation trajectories in response to engineered context. This can produce stable looking behaviour within an evaluation envelope,

particularly when the envelope is aligned with the same representational primitives the model is designed to exploit. Yet this stability is not evidence of internally conserved structure. It is the stability of a constrained mapping, not the stability of a constitution. The difference is not semantic. A constrained mapping can be made to look adaptive indefinitely, as long as the evaluative regime continues to reward the same surface features, and as long as the environment is represented only through the model's own symbolic admissibility conditions.

Comprehension per Joule (CPJ) is defined within CIITR as a thermodynamic measure of epistemic efficiency, quantifying the amount of structurally valid comprehension ($\Phi_i \times R^g$) per unit of energy expenditure. It does not imply internal logic, memory, adjudication, or preserved evaluative frameworks, but expresses the rate at which a system transforms energy into epistemically integrated structure. $CPJ = \Phi_i \times R^g / E$.

The same conclusion follows from R^g. Rhythmic Reach is not a metaphor for repeated interaction, nor a proxy for running more iterations. It denotes the capacity to sustain epistemic continuity across energetic horizons, to carry forward internal organisation such that the system remains itself while integrating perturbations. The systems treated as "self adapting" in this context do not traverse time as systems. They execute inference as discrete episodes. Any continuity is supplied by external scaffolding, by the operator, by the environment, by logging, by re prompting, or by auxiliary loops. This may be operationally valuable, but it does not instantiate internal rhythmic continuity. In CIITR terms, R^g remains structurally negligible, which means that the system's behavioural drift cannot accumulate into a coherent adaptive trajectory. Adaptation without R^g is not adaptation, it is episodic sensitivity.

Finally, the interpretation fails on Φ_i. Integrated Relational Information is not the presence of more tokens, more context, or more syntactic elaboration. It is the internal integration of relations such that informational input changes the system's epistemic topology, not merely its next output. In the examined setting, the transformations occur at the level of surface conditioning and statistical modulation. Even when the pipeline introduces weight updates, the updates are not governed by an internal constitution of judgment, nor are they embedded in a rhythmically conserved system state. They are technical interventions that can increase task performance, but they do not, in themselves, create relationally integrated understanding. A model can be tuned to respond more accurately, more compliantly, or more consistently, without acquiring any internally conserved relational structure that warrants the language of understanding.

For these reasons, the note asserts a strict CIITR conclusion: no valid form of self modification or adaptation has been demonstrated in the CIITR sense. What has been shown is a method for inducing performance relevant behavioural change through structured prompt recursion and externally orchestrated optimisation loops. That is a contribution within a performance engineering paradigm. It is not a demonstration of structural cognition, nor a transition toward autonomous systems, nor a proof of epistemic self reconstitution. The error emerges when performance engineering is rhetorically recoded as autonomy.

This conclusion is not merely classificatory. It functions as a constraint on permissible claims. CIITR's central condition, $\Phi_i \times R^g \times CPJ$, is irreducible to symbolic output behaviour because

each term refers to a different ontological layer of system validity. Φ_i concerns internal relational integration, R^g concerns temporal and energetic continuity, CPJ concerns constitutional stability of judgment. None of these can be inferred from output variability alone, and none can be replaced by benchmark success without collapsing epistemic categories. The multiplication is therefore not a stylistic gesture, but a formal statement of co necessity. If any term is absent, structural understanding collapses to zero regardless of surface sophistication. This is precisely the situation here. The system may appear adaptive, but the conditions that would make adaptation structurally meaningful are not present.

Accordingly, the proper interpretation of the paper is that it strengthens the operational toolkit for prompt mediated performance modulation, while simultaneously exemplifying how easily contemporary discourse over ascribes agency to symbolic drift. A governance relevant system description should therefore treat such models as prompt constrained symbolic compressors, not as adaptive agents, and should treat claims of autonomy as structurally non admissible unless Φ_i, R^g, and CPJ are independently evidenced as persistent system properties rather than as narrative overlays on behavioural effects.