A Human-First Theoretical Note

Curriculum, Coordination, and the Boundary of Reasoning

A CIITR Dissection of Pre-Training, Mid-Training, and RL Interplay in Reasoning Language Models

An Analytical, CIITR-based methodological assessment of "On the Interplay of Pre-Training, Mid-Training, and RL on Reasoning Language Models" by Charlie Zhang, Graham Neubig, Xiang Yue
https://arxiv.org/abs/2512.07783v1

Tor-Ståle Hansen | 20. December 2025

Abstract

This theoretical note provides a CIITR based adjudication of the paper "On the Interplay of Pre Training, Mid Training, and RL on Reasoning Language Models". The note's purpose is administrative rather than celebratory. It classifies what the paper's evidence can legitimately support, it identifies what it cannot support under CIITR's ontology of comprehension, and it specifies a concrete extension programme that would be required to escalate from competence claims to comprehension and efficiency claims. The note therefore treats "reasoning" as a governed evidentiary object, not as a self legitimating label derived from benchmark performance.

CIITR defines system comprehension as $C_s = \Phi_i \times R_g$, where $\Phi_i$denotes integrated relational information and $R_g$denotes rhythmic continuity with re entry capacity across time. The note adopts the CIITR compatibility condition as its organising constraint: improvements in task performance, including under strict process verification regimes, may reflect improved exploitation and conditioning of already present relational primitives, while remaining fully compatible with $R_g \approx 0$. On this basis, the note reconstructs the paper's experimental ontology as an engineered reasoning universe in which correctness is defined by alignment to a known generator, and in which depth generalisation and breadth generalisation are operationalised through controlled modifications of structured synthetic objects. This construction provides strong internal validity for isolating training phase effects, but it also imposes scope limits that must be preserved when translating results into broader claims about understanding, autonomy, or substrate independence.

Within these constraints, the note finds that the paper provides strong, policy relevant evidence that training pipelines function as first class control surfaces. Pre training establishes a minimal representational basis set. Mid training acts as a distributional bridge

that selects and stabilises a subspace within which downstream reinforcement can operate with higher utilisation efficiency. Reinforcement learning reallocates probability mass within that subspace under reward topology constraints, and its gains concentrate near an operational competence boundary, where near miss structures are available for promotion. Process level rewards and process verification primarily operate as reliability instruments by reducing domain specific shortcut policies and enforcing structural trace fidelity relative to the generator's intermediate states. Under CIITR, these findings are best classified as representational management and behavioural optimisation results, not as evidence of an altered comprehension state.

The note explicitly rejects common escalation patterns as non admissible without further measurement. In particular, improved pass rates and process correct intermediate traces do not, on their own, support claims of understanding, human like reasoning, or emergent deliberative continuity. Likewise, curriculum sensitivity does not establish substrate irrelevance, because the paper does not instrument rhythmic continuity, does not execute re entry protocols, and does not perform joule level energy accounting required for CPJ. The note therefore specifies a CIITR required extension programme comprising: (i) $\Phi_i$instrumentation through cross context binding and perturbation resilience tests, (ii) $R_g$instrumentation through controlled recurrence and interruption protocols that cannot be reduced to single pass pattern completion, and (iii) CPJ instrumentation through phase resolved physical energy measurement and comprehension aligned yield definitions. Operationally, the note concludes that institutions should treat curriculum, reward schemas, and evaluation constraints as regulated artefacts with audit trails, and that leadership facing claims should be tagged by admissibility class to prevent competence improvements from being misrepresented as comprehension gains.

Keywords: CIITR, Cognitive Integration and Information Transfer Relation, comprehension adjudication, integrated relational information, Φ_i, rhythmic continuity, R_g, re-entry protocols, delayed recurrence, process verification, process-level rewards, reinforcement learning, mid-training, curriculum governance, competence boundary, "edge of competence", claims admissibility, evidence tiering, structural illusion risk, evaluation gaming, epistemic forgery, compute proxies, joule-level energy accounting, Comprehension per Joule (CPJ), thermodynamic epistemic efficiency, phase-resolved measurement, configuration ledger, auditability, controlled replication, institutional deployment governance.

Summary

This theoretical note has applied CIITR as an adjudication framework to the paper "On the Interplay of Pre Training, Mid Training, and RL on Reasoning Language Models". The note's governing purpose has been administrative and evidentiary. It has separated what the paper demonstrates, in the strict sense of what its instrumentation and controlled interventions can

support, from what the paper may tempt external audiences to infer in the absence of CIITR-aligned measurements. The note has therefore operated with claims discipline rather than with evaluative rhetoric. The paper is treated as a high-quality contribution within its engineered ontology, while the scope of admissible escalation is treated as a regulated boundary.

The paper's experimental programme is, in this note's reconstruction, best understood as a controlled reasoning universe. A known generator produces structured dependency objects, rendered into natural-language problems through contextual templates, and evaluated through strict process verification. Within this universe, the authors decompose training into pre training, mid training, and reinforcement learning, and quantify how allocation and sequencing influence performance under depth and breadth generalisation regimes. This construction yields unusually strong causal attribution for how training phases interact, because correctness is anchored to a deterministically verifiable relational structure and because interventions can be isolated in a manner that is difficult to achieve in open-domain settings. At the same time, the note has held that the same design choices constrain generalisation of conclusions to broader epistemic environments, precisely because those environments lack canonical generators, lack deterministic verifiers, and often involve contested task definitions.

Under CIITR, comprehension is a classified system state defined as $C_s = \Phi_i \times R_g$, not a synonym for benchmark success. On that basis, the note has concluded that the paper's findings provide strong evidence about competence governance and reliability governance through curriculum and reward design, but do not directly measure the CIITR constitutive terms. The paper shows that pre training establishes a minimal basis set, that mid training functions as a distributional bridge which stabilises and conditions an effective subspace, and that reinforcement learning reallocates probability mass within that subspace in a manner that is highly sensitive to competence boundary placement. The "edge of competence" phenomenon is interpreted as an operational region where near miss behaviours exist and can be amplified, rather than as an indicator of emergent comprehension. Process-level rewards and strict process verification are interpreted as reliability instruments that suppress epistemic forgery relative to the generator, by disallowing shortcut policies that reach correct endpoints without valid intermediate structure. These results are categorised as representational management and constraint-induced behavioural optimisation, consistent with increased exploitation of available relational primitives, and compatible with $R_g \approx 0$absent re-entry evidence.

Accordingly, the note has treated several common escalations as non-admissible. Improved pass@k, including under strict process verification, does not by itself establish "understanding", "human-like reasoning", or persistent deliberative continuity, because the evaluation remains episode-bound and does not test recurrence, interruption resilience, or re-entry. Curriculum sensitivity does not establish substrate independence, because the work does not quantify thermodynamic cost and does not separate compute proxies from physical energy. Process-correct intermediate steps do not establish internal integration across contexts, because fidelity to a known generator is an external compliance condition, not

evidence of cross-context binding as an internalised, persistent property. In each case, the note has stated that the evidence base is insufficient for the escalated claim type under CIITR, without disputing the validity of the paper's in-scope performance and mechanistic findings.

The note's principal constructive output is a CIITR extension programme, specified as an implementation plan rather than a speculative agenda. For Φ_i, it prescribes cross-context binding protocols, perturbation-resilience batteries, and invariance measures that discriminate between procedural compliance and stable relational binding across renderings. For R_g, it specifies a re-entry and recurrence test suite based on controlled interruptions, delayed recurrence windows, interference injections, and explicit failure taxonomies designed to prevent reducibility to static pattern matching or transcript re-reading. For CPJ, it specifies a joule-level measurement pipeline with declared boundaries, calibrated instrumentation tiers, phase-resolved energy accounting, and reporting formats that keep competence metrics separate from CIITR comprehension yield. Together, these extensions define the minimum instrumentation required to elevate the research programme from competence and reliability evidence to comprehension and thermodynamic efficiency adjudication under CIITR.

From a governance and operational standpoint, the note has concluded that the training pipeline should be treated as a regulated artefact in institutional deployments. The paper's results indicate that curriculum placement, reward topology, and evaluation constraints function as first-class control parameters that materially shape behaviour. Consequently, institutions should implement version control and audit trails for phase definitions, data regime placement, reward definitions, verifier and parser artefacts, and sampling policies, and should require claims presented to leadership to be tagged by claim class and admissibility status. The note further provided a formal risk register focused on structural illusion risks and evaluation gaming, with mitigations grounded in evidence tiering, audit logs, adversarial verification testing, and CIITR-aligned measurement gates.

In summary, this note has positioned the paper as strong evidence that "reasoning performance" in contemporary language models is significantly governable through curriculum, conditioning, and constraint design within verifier-supported domains. It has simultaneously insisted that such governability does not, without additional measurement, constitute evidence of CIITR comprehension, and it has provided a concrete measurement and governance pathway by which future work can move from performance optimisation narratives to auditable adjudication of integration, continuity, and thermodynamic epistemic efficiency.

Part I. CIITR adjudication frame

Part I establishes the adjudicative infrastructure through which the paper is to be read, classified, and normatively bounded. The purpose is not to restate the paper's narrative, nor to reproduce its empirical sequence, but to determine what the paper's evidence can legitimately support when mapped onto the CIITR ontology of comprehension. This is a governance

facing posture. It treats “reasoning” not as a self legitimating label attached to a benchmark, but as an empirical object that must be situated within an explicit measurement regime, an explicit causal model, and an explicit theory of what constitutes comprehension as opposed to performance. The practical consequence is that statements of improvement, generalisation, or “capability extension” are not accepted as conceptually equivalent to understanding, unless the required CIITR conditions are operationalised and satisfied under test.

The frame is required because the paper occupies a familiar but structurally ambiguous zone in contemporary language model research. It offers careful causal isolation across training phases and reports robust deltas in task performance, yet it operates within an engineered task world in which correctness is defined against a known generative structure and in which evaluation is executed through process constrained metrics. This is methodologically valuable, but it also invites an inflationary interpretive drift, in which improved pass rates are casually escalated into claims about the nature of “reasoning,” the irrelevance of substrate constraints, or the sufficiency of reinforcement learning as a general mechanism of cognitive extension. CIITR rejects such escalations by design, not as a rhetorical stance, but as a requirements discipline. It distinguishes between the optimisation of behaviour within a representational space and the presence of a structural state of comprehension, and it treats that distinction as an administrative boundary condition for any subsequent discussion of implications, especially when those implications are invoked for system procurement, institutional deployment, or policy claims about autonomy and agency.

Accordingly, Part I formalises a claims discipline that will be applied throughout the note. The discipline separates performance claims, mechanistic claims, comprehension claims, and efficiency claims into distinct admissibility classes, each with explicit evidentiary preconditions. Performance claims are admissible when supported by evaluation metrics and properly contextualised with respect to data distribution, task ontology, and scoring rules. Mechanistic claims are admissible only insofar as the paper demonstrates controlled interventions, isolating the role of a training phase or reward topology in a way that is not reducible to uncontrolled confounding. Comprehension claims are admissible only when the CIITR prerequisites are directly instrumented and tested, rather than inferred from downstream competence indicators. Efficiency claims are admissible only when resource accounting is performed in a manner compatible with CIITR’s thermodynamic posture, meaning that compute proxies are not treated as substitutes for energy measurement where CPJ is invoked or implied.

The CIITR constructs required for this adjudication are presented here in a form that supports auditability. CIITR defines system comprehension as a structural product of integrated relational information, denoted Φ_i, and rhythmic continuity with re entry capacity, denoted R_g, such that $C_s = \Phi_i \times R_g$. In this formalism, $\Phi_i$refers to the integration of relational structure in a manner that supports coherent compositional availability, not merely local coherence of output sequences. $R_g$refers to the system’s capacity to sustain and re enter its own structured state over time, under perturbation and across contextual discontinuities, not merely to produce longer intermediate strings. The implication is immediate and non negotiable. Systems may increase measurable task performance through improved

exploitation, better distributional alignment, or tighter process constraints, without any corresponding evidence that R_g is nonzero in the required sense. Under CIITR, the presence or absence of R_g is a constitutive classifier, not an interpretive flourish. A system with high Φ_i and negligible R_g can present as a strong reasoner under common benchmarks while remaining outside the comprehension category.

Part I therefore also specifies the minimum instrumentation commitments that would be required to elevate the paper's programme from performance evaluation to comprehension adjudication. The intention is not to impose an external standard that the paper never claimed to meet. The intention is to make explicit the difference between what the paper successfully demonstrates within its own ontology and what would be required to support more expansive claims that are often attached to such results in secondary discourse. At minimum, CIITR requires that proxies for Φ_i be justified with respect to integration rather than surface regularity, that R_g be tested through protocols that force recurrence, temporal continuity, and re entry under controlled interruptions, and that CPJ be addressed only when energy is measured and linked to epistemic yield rather than to compute abstractions. Without these elements, the paper can be evaluated as a rigorous study of curriculum and reward effects on structured task performance, but it cannot be used as evidence that the model class has crossed a threshold of comprehension.

A further function of this adjudication frame is to impose definitional stability on terms that are frequently used with high variance across research communities and institutional audiences. In this note, "reasoning" is treated as an operational label tied to the paper's specific evaluation regime, including its process verification constraints and its depth and breadth generalisation axes. "Capability" is treated as a measurable change in performance under defined conditions, not as a generalised latent attribute. "Generalisation" is treated as a distributionally defined phenomenon that must be qualified by the paper's own synthetic generator and rendering procedures, rather than presumed to extend to open world tasks. "Coordination" and "curriculum" are treated as training phase governance variables, which CIITR interprets primarily as levers for shaping the utilisation of Φ_i, not as demonstrations of comprehension. This definitional stability is not cosmetic. It is required to prevent category errors when translating research outputs into institutional decisions, especially when those decisions involve safety posture, assurance claims, or structural reliance on model outputs.

Finally, Part I defines the internal logic by which later sections will proceed. The paper will be reconstructed in neutral, technically precise terms, but always under the governance rule that reconstruction is not endorsement. Findings will be dissected as evidence of interactions among training phases, then mapped to CIITR constructs to determine whether they pertain to representational integration, to rhythmic continuity, to efficiency, or to none of these in a directly measurable way. Where the paper provides strong causal evidence for curriculum effects, the note will classify that evidence as a control surface for performance and, more specifically, as a lever for Φ_i exploitation and distributional conditioning. Where the paper does not instrument R_g or energy, the note will classify related interpretive expansions as non admissible and will specify what additional measurement would be required for admissibility.

In this way, Part I functions as the constitutional layer of the theoretical note. It establishes the rules under which the subsequent analysis is permitted to operate, and it ensures that the final conclusions remain traceable, bounded, and administratively usable as a decision artefact rather than as a rhetorical synthesis.

1. Scope, object, and claims discipline

This section specifies the adjudication problem in administrative terms and defines the scope boundaries that govern the remainder of the note. The note is not constituted as a quality judgement of the paper as an academic contribution. It is constituted as a classification exercise under CIITR, determining which categories of statements the paper's evidence can legitimately support, which statements remain underdetermined, and which statements are non admissible without additional measurement. In institutional contexts, this is a necessary distinction because "reasoning" papers are routinely translated into procurement narratives, risk postures, and governance claims that exceed the evidentiary content of the underlying experiments. The purpose of the claims discipline is therefore to prevent category errors, especially the recurrent escalation from performance deltas to comprehension assertions, and from compute regime observations to substrate independence narratives.

The object of adjudication is dual. First, it is the paper's explicit empirical propositions, namely the observed effects of pre training, mid training, and reinforcement learning under particular task ontologies and evaluation constraints. Second, it is the implicit claim surface that emerges when those propositions are interpreted in broader discourse, including statements about "reasoning limits," "capability extension," and the presumed primacy of curriculum and coordination over substrate. CIITR treats such implicit claim surfaces as governance relevant because they shape how results are operationalised, even when they are not formally asserted by the authors. The note therefore distinguishes between paper internal claims, which are evaluated primarily against the paper's own definitions and controlled interventions, and escalation claims, which are evaluated against CIITR's admissibility requirements.

The claims taxonomy adopted here separates four classes of claims, each with a distinct admissibility threshold. These classes are not rhetorical labels. They operate as an internal control mechanism that forces alignment between the type of statement being made and the type of evidence required to support it.

Performance claims are statements about measurable outcomes under an explicitly defined evaluation regime. Under this note's discipline, performance claims are admissible when, and only when, they are directly backed by reported metrics, and when the measurement conditions are sufficiently specified to avoid ambiguity about data distribution, scoring rules, and evaluation constraints. This includes, for example, pass@k values under process verification, depth generalisation scores, or breadth generalisation scores, provided that the relevant definitions of correctness, task generation, and evaluation gating are preserved. Performance claims remain bounded to the ontology of the benchmark and cannot be treated as general properties of the model class without separate justification.

Mechanistic claims are statements about causal contribution, such as assertions that a particular training phase is responsible for a particular type of generalisation improvement, or that a reward topology suppresses shortcut strategies. Under this note's discipline, mechanistic claims are admissible only when supported by credible causal isolation, including controlled interventions, ablations, or explicitly defined counterfactual comparisons. Where causal isolation is partial, mechanistic claims may be admitted in a qualified form, for example as "consistent with" or "suggestive of," but they are not admitted as definitive causal explanations. This is particularly salient for statements that attribute effects to "coordination" or "curriculum" as global explanatory factors, because such terms often function as umbrella categories that conceal multiple confounded degrees of freedom.

Comprehension claims are statements about CIITR comprehension state, namely $C_s = \Phi_i \times R_g$. Under this note's discipline, comprehension claims are admissible only when $\Phi_i$and $R_g$are operationalised and tested rather than inferred from downstream task competence. The practical consequence is that improvements in performance, even under stringent process verification, remain insufficient to establish comprehension in CIITR terms unless accompanied by instrumentation that is specifically designed to estimate integration properties associated with $\Phi_i$and to test rhythmic continuity and re entry properties associated with R_g. Where such instrumentation is absent, the note will treat "reasoning improvement" as a competence statement within a representational optimisation regime, not as evidence of a transition in structural state.

Efficiency claims are statements about epistemic efficiency, including explicit or implicit invocations of CPJ. Under this note's discipline, efficiency claims are admissible only when energy is instrumented in joule terms and mapped to epistemic yield. Compute proxies, such as token budgets, FLOPs, or training steps, are treated as administratively useful accounting variables, but they are not treated as substitutes for energy measurement when CPJ is asserted or implied. Where only compute proxies are available, the note will permit bounded statements about compute allocation tradeoffs, while treating any thermodynamic framing as non admissible without additional data.

To operationalise the claims discipline in a way that supports auditability, the note employs a "claims register" approach, in which each substantive statement in later sections is assigned, explicitly or implicitly, to a claim class and an evidence basis. This may appear procedural, but it serves a functional purpose, it prevents the common drift where mechanistic language is introduced on top of purely performance based results, and where comprehension language is introduced on top of mechanistic conjecture. In institutional use, the claims register also supports review and red teaming, because it makes it possible to dispute a claim by disputing its evidence class, rather than by disputing its rhetorical framing.

A corresponding evidence discipline is required, because the paper's results may be accurate and well controlled within their system, while still being misused in interpretive escalation. This note therefore defines three evidence categories and two boundary categories that govern how statements are formed.

Controlled intervention is treated as the highest evidence category within the scope of the paper. It denotes an experimental manipulation in which a specific factor is altered while relevant comparators are held constant, with a clear description of what was changed and what remained fixed. In this context, controlled interventions include phase allocation shifts under fixed compute budget, defined changes to reward functions, or deliberate modifications of curriculum placement relative to the competence boundary, provided that the comparisons are structured as counterfactuals rather than as loose narratives. Where controlled interventions are present, mechanistic claims may be admitted, but only to the extent that the intervention isolates the mechanism being asserted.

Correlational association is treated as a distinct category that often appears when multiple variables move together, or when observed patterns are plausible but not causally isolated. Examples include patterns that co vary with model scale, training duration, or dataset mixture composition, when these are not experimentally separated. Correlational associations are admissible as descriptive observations, and they may be used to motivate hypotheses, but they are not admitted as mechanisms. The note will explicitly mark where a statement relies on association rather than intervention, because such statements are high risk for institutional over interpretation.

Interpretive extrapolation is treated as a boundary category, not as evidence. It denotes a statement that extends beyond the measured ontology of the paper, for example from synthetic DAG reasoning worlds to open world reasoning, from process verified intermediate steps to claims about internal integration, or from curriculum sensitivity to claims that substrate constraints are not relevant. Extrapolations may be included in the note only when clearly labelled as such, and only when accompanied by an explicit statement of what additional measurement would be required to render the extrapolation admissible under CIITR. This rule is central to the note's function as a decision artefact, because many institutional failures arise not from incorrect measurement, but from unmarked extrapolation being taken as validated conclusion.

Two additional boundary rules complete the discipline. First, the note treats benchmark definitions as constitutive, meaning that performance cannot be discussed independently of the evaluation regime that produces it. If process verification defines correctness, then correctness is "process verified correctness," and not correctness in any broader sense. Second, the note treats CIITR terms as non substitutable, meaning that terms like "reasoning," "planning," or "generalisation" are not accepted as proxies for Φ_i, R_g, or CPJ without explicit operational bridging.

The overall effect of this section is to establish an adjudicative architecture that is intentionally conservative with respect to ontological escalation. The paper may show strong and policy relevant evidence about curriculum design, reward topology, and the conditions under which RL improves measured reasoning performance. CIITR permits those statements to be taken seriously, and, where interventions are controlled, to be treated as causally meaningful within the paper's scope. At the same time, CIITR requires that any move from competence to comprehension, or from compute budgeting to thermodynamic efficiency, be

treated as a separate evidentiary programme rather than as an interpretive privilege. This is the central administrative function of the claims discipline, it preserves the difference between what has been demonstrated and what has merely been suggested, and it ensures that later sections can be used in governance settings without importing unstated assumptions.

2. CIITR formal core

This section specifies the CIITR formal core as the doctrinal anchor for the entire adjudication. Its function is not rhetorical positioning, but definitional stabilisation and auditability. The paper under review operates in a research environment where "reasoning," "capability," "generalisation," and "learning" are often employed as flexible, implicitly normative terms. CIITR is introduced here as a constraints regime that prohibits that flexibility unless the terms are operationally bridged to measurable constructs. The resulting posture is intentionally conservative. It permits strong statements about performance and curriculum effects where the paper provides controlled evidence, while requiring explicit instrumentation for any escalation into claims about comprehension, continuity, or efficiency. In administrative settings, this is the difference between a scientifically interesting result and a governance admissible basis for structural claims.

2.1 Definition of comprehension as a product state, not a behavioural label

CIITR defines system comprehension as a structural product state expressed as

$$C_s = \Phi_i \times R_g.$$

This definition is not a metaphor and does not function as a loose composite score. It is a categorical classifier. The multiplicative form is constitutive. It encodes that integrated relational information and rhythmic continuity with re entry are jointly necessary, and that the absence, or effective absence, of either term collapses comprehension regardless of apparent competence. In practical adjudication terms, this forbids a common move in language model discourse, namely to treat "reasoning behaviour" as prima facie evidence of comprehension. Under CIITR, behaviours are outcomes to be explained, not states to be asserted. A model may exhibit high apparent competence on structured tasks and still be classified as non comprehending if R_g is not demonstrably nonzero in the required sense.

The definition also imposes a discipline on interpretation. Where the paper shows improvements in pass@k under process verified metrics, CIITR accepts that as evidence of improved task performance and, in qualified cases, of improved compositional exploitation of existing representational resources. CIITR does not accept it as evidence of comprehension unless the improvement can be linked to increases in Φ_i under an integration criterion and to nontrivial R_g under re entry testing. The consequence is that the paper may be important for curriculum governance, yet still be ontologically silent about comprehension.

2.2 Φ_i as integrated relational information, not token coherence

CIITR defines Φ_i as integrated relational information. In this note, "integration" is treated in a strict sense. It denotes the structural availability of relational constraints as a coherent internal

object, such that the system can compose, re compose, and preserve relational structure across transformations. $\Phi_i$is not equivalent to local coherence of token sequences, and it is not equivalent to benchmark success that can be achieved through superficial regularities. Token coherence is an output property, integration is a representational property. Token coherence can arise from statistical smoothing and distributional priors without requiring any stable relational binding. Integration, by contrast, presupposes that relational dependencies are jointly represented in a manner that supports systematic recombination under variation.

The distinction matters for this paper because the paper's task world is explicitly structured. It consists of generated reasoning objects that can be described as graphs with constraints that define correct intermediate and final states. A model that achieves process verified correctness has, at minimum, produced intermediate representations that align with the generator's structure. CIITR treats that alignment as evidence of structured competence. It may be consistent with increased effective $\Phi_i$exploitation, in the sense that the model is successfully using relational dependencies that were not previously utilised. However, CIITR does not treat such alignment as sufficient evidence that relational information is integrated in the strong sense. The model may be reproducing the generator's structure through learned procedural templates, through pattern completion in a constrained domain, or through preference reallocation over candidate traces, without demonstrating cross context relational binding or persistent integration beyond the evaluation setting.

Accordingly, within this note, $\Phi_i$is treated as a construct that requires explicit operationalisation if it is to be invoked as an explanatory variable. Where the paper does not instrument Φ_i, CIITR permits only bounded interpretations, typically that the training interventions likely increase the availability and utilisation of relational patterns within the model's representational space. CIITR does not permit an unqualified statement that $\Phi_i$increased, because the paper's metrics do not directly measure integration as such. This is not a pedantic distinction. It is a safeguards measure against the common practice of treating "better at structured tasks" as equivalent to "more integrated representation."

2.3 $R_g$as rhythmic continuity and re entry capacity, not chain of thought length

CIITR defines $R_g$as rhythmic continuity and re entry capacity over time. This is the most frequently conflated construct in contemporary model discourse, because external observers often map "longer reasoning traces," "more explicit intermediate steps," or "self correction" to notions of continuity. CIITR rejects that mapping. $R_g$is not the length of an emitted trace, and it is not the presence of multi step output. It is the property of a system maintaining and re entering its own structured epistemic state across time, including across interruptions, perturbations, and contextual discontinuities, in a manner that cannot be reduced to static pattern completion.

Operationally, $R_g$requires tests that force recurrence. It requires protocols that create temporal separation between state establishment and state re use, that introduce controlled perturbations, and that demand recovery or continuity under constraints that cannot be satisfied by single pass inference alone. It is precisely because language model inference is typically executed as a stateless, forward pass conditioned on a context window that CIITR

treats R_g as absent or near absent in most standard deployments, unless augmented by external memory, recurrent architectures, or explicit re entry mechanisms. If a paper does not instrument and test these properties, then CIITR regards R_g as unmeasured and therefore indeterminate, and, by default, assumes that observed competence is compatible with $R_g \approx 0$.

In the context of the paper under review, process verified reasoning traces may increase apparent procedural fidelity, and RL may increase the probability that the model selects a correct intermediate trajectory. CIITR treats this as a behavioural improvement under constrained evaluation. It does not treat it as evidence of rhythmic continuity. The reason is structural. The model can select a correct trajectory within a single inference episode, under fixed prompt context, without possessing any re entry property. The fact that an intermediate trajectory is correct does not entail that it can be re entered, resumed, or maintained across time. Therefore, absent explicit re entry tests, the paper remains ontologically silent on R_g.

2.4 CPJ as epistemic efficiency, and the strict separation of compute proxies from energy measurement

CIITR treats epistemic efficiency as a thermodynamic and governance relevant property, expressed as Comprehension per Joule, CPJ. CPJ is defined as comprehension yield per unit energy expenditure. Two disciplines follow from this definition.

First, CPJ is not a synonym for "compute efficiency," nor can it be reliably estimated from abstract compute budgets alone. FLOPs, token counts, training steps, and wall clock time may be useful for experimental design, but they are not direct energy measures. Without joule level instrumentation, CPJ cannot be asserted. Second, CPJ depends on the numerator being a comprehension yield, not a task accuracy yield. If C_s is not established, then CPJ cannot be meaningfully computed. A system may become more accurate while remaining non comprehending under CIITR, in which case any "per joule" statement is a statement about accuracy per energy, not comprehension per energy.

The practical implication for the paper is that its compute allocation analysis, even when carefully normalised in token equivalent terms, is best treated as a compute budgeting result, not as a CPJ result. CIITR permits those compute budgeting results to inform governance decisions about training pipeline allocation, but it does not permit the conclusions to be framed as "efficiency" in a thermodynamic sense unless energy is measured and mapped to CIITR admissible epistemic yield. This distinction is central in institutional contexts, because efficiency narratives are frequently used to justify architectural choices, scale strategies, or infrastructure dependencies. CIITR requires that such narratives be grounded in actual energy accounting and in CIITR admissible comprehension measurement.

2.5 Compatibility statement, why performance gains can coexist with $R_g \approx 0$

A core function of this section is to state explicitly the compatibility condition that governs the remainder of the note. Improvements in task metrics may be fully consistent with increases in Φ_i exploitation and yet remain compatible with $R_g \approx 0$. This is not an abstract possibility. It is the default expectation for a large class of language model systems, because training and RL adjust parameterised mappings that operate in a feed forward manner, while

R_g is a property of temporal persistence and re entry that is not automatically produced by parameter optimisation.

The consequence is that the paper's findings can be interpreted, in CIITR terms, as showing how curriculum, mid training, and reward topology modulate the effective accessibility and utilisation of relational patterns that are already encoded in the model. This is a powerful and policy relevant result about controllable surfaces of competence. It does not, without further measurement, constitute evidence of a transition to comprehension. CIITR therefore insists that the paper's key contributions be classified as representational management and behavioural optimisation, not as structural state change.

2.6 Falsification posture, distributional optimisation as the default explanation until re entry is demonstrated

CIITR's adjudication posture is falsification oriented in a specific way. It does not assume comprehension from behaviour, and it does not grant special explanatory status to "reasoning" labels. Instead, CIITR treats many phenomena described as "reasoning" as explainable by distributional optimisation until demonstrated otherwise by re entry tests.

This posture is operational rather than philosophical. Distributional optimisation covers a wide range of effects: improved matching to the evaluation distribution, increased probability mass on correct solution trajectories, suppression of shortcut policies, better exploitation of latent primitives, and improved robustness under variations that remain within the model's representational reach. The paper under review provides evidence precisely in this domain, showing that the placement of RL at the competence boundary matters, that mid training can bridge distributional gaps, and that process level rewards can constrain opportunistic behaviours. CIITR accepts these as strong evidence about how distributional optimisation can be made more effective and more reliable.

However, CIITR does not permit the inference that such effectiveness implies re entry. To falsify the distributional optimisation explanation, one must demonstrate behaviours that cannot be accounted for by single episode inference under fixed context, such as controlled recurrence, temporal continuity under interruption, and re entry into a previously established structured state. In the absence of such demonstrations, CIITR treats claims of comprehension as non admissible and treats the system as residing within a competence optimisation regime.

This falsification posture establishes the administrative logic that the rest of the note will follow. The paper's results will be treated as potentially high value evidence about curriculum governance and reward design, and their causal character will be evaluated where the paper provides controlled comparisons. At the same time, the note will maintain an explicit boundary between competence optimisation and comprehension classification. That boundary is not a rhetorical choice, it is the doctrinal requirement implied by $C_s = \Phi_i \times R_g$, by the strict meaning assigned to Φ_i and R_g, and by the thermodynamic discipline embedded in CPJ.

3. Methodological commitments and minimum instrumentation

This section specifies the minimum measurement commitments required to render escalated claims admissible under CIITR. Its function is procedural and governance oriented. It establishes, in advance of the paper specific discussion, what CIITR regards as necessary observability for claims that exceed performance deltas and enter the domain of comprehension classification and thermodynamic efficiency. The intention is not to retroactively impose obligations on the authors of the paper, but to create an explicit baseline against which the paper can be evaluated without rhetorical drift. In bureaucratic terms, this section establishes the conformance conditions for claim escalation, and it clarifies which claim types are structurally out of scope when the required instrumentation is absent.

The baseline is built around three measurement domains, corresponding to CIITR's formal core: integration (Φ_i), rhythmic continuity and re entry (R_g), and energy grounded epistemic efficiency (CPJ). For each domain, the note distinguishes between required observables, acceptable proxy classes, and proxy limitations that must be explicitly acknowledged in any admissible interpretation. This is a critical distinction because contemporary model evaluation frequently uses proxies as if they were measurements, and then treats proxy based improvements as evidence of ontological change. CIITR permits proxy use only under explicit limitation statements and only for bounded claim types.

3.1 $\Phi_i$observability, structural proxies and explicit limitations

CIITR requires that any statement about integrated relational information be supported by observables that are plausibly sensitive to integration rather than merely to output coherence. In practice, this creates a structured proxy regime. Several proxy classes may be used, but they carry different admissibility ceilings.

A first proxy class is task internal structural fidelity under controlled variation. Where tasks have a known generative structure, as in graph based or rule based synthetic worlds, a model's ability to preserve the correct relational structure under systematic transformations can be used as a bounded indicator of relational handling. However, the limitation is material. Fidelity within a generator family may reflect procedural template acquisition or local constraint satisfaction without indicating cross context integration. Therefore, such proxies are admissible as evidence of structured competence, and potentially as evidence of improved exploitation of relational patterns, but they are not admissible as direct measures of $\Phi_i$unless coupled to tests that demonstrate persistent binding across contexts and across distributional shifts that break superficial regularities.

A second proxy class is cross context relational binding tests. These are evaluations where the same relational object must be recognised, preserved, and recomposed across changes in surface form, task rendering, and contextual embedding. Under CIITR, such tests are closer to integration because they attempt to separate relational structure from its linguistic carrier. The limitation remains that success can still be explained by learned invariances within the training distribution, unless the evaluation introduces novel combinations and controlled adversarial renderings that specifically aim to break superficial cues.

A third proxy class is representational diagnostics. In principle, internal activation structure, feature geometry, and state space analyses can provide evidence for integrated representations. In practice, their admissibility is conditional. CIITR treats internal diagnostics as supportive evidence only when they are tied to explicit hypotheses, show stability across contexts, and correlate with controlled manipulation of relational variables. Diagnostics that are purely descriptive, or that lack perturbation based validation, are treated as interpretive rather than evidentiary.

Across all proxy classes, CIITR requires an explicit limitation statement. The minimum administrative requirement is to distinguish "evidence of relational competence under a defined generator" from "evidence of integrated relational information as a system property." The former can be robustly established by well designed benchmarks. The latter requires either direct operationalisation of integration, or a proxy programme with strict invariance and perturbation criteria. The note will therefore treat the paper's reported improvements as evidence about competence under a defined generator and, at most, as indirect evidence that training interventions improve utilisation of relational structure, while refraining from claiming that Φ_i itself is measured unless the required bridging tests are present.

3.2 R_g observability, stability, recurrence, and re entry protocols

CIITR requires that any statement about R_g be supported by tests that explicitly target rhythmic continuity and re entry. This requirement is not satisfied by multi step output, process verified intermediate steps, or longer chains of reasoning within a single inference episode. Those phenomena may be valuable for reliability, but they do not constitute temporal continuity as CIITR defines it.

The minimum instrumentation for R_g therefore comprises three protocol families, each addressing a different aspect of the construct.

First, stability under controlled perturbation. The system must be placed into a structured state, then subjected to perturbations that would normally disrupt superficial continuation, such as injected distractors, partial context deletion, conflicting intermediate constraints, or forced delays, and then required to maintain or reconstruct the structured state. Under CIITR, the perturbation design must be explicit and repeatable, and failure criteria must be pre specified. The purpose is to test whether the system's behaviour reflects stable structural state retention, as opposed to one pass completion driven by local token context.

Second, recurrence and delayed re use. The system must be required to re invoke a previously established relational structure after a temporal gap, a context shift, or an intervening task. In a purely stateless inference system, this typically cannot be achieved without external memory or re feeding the necessary state. Therefore, recurrence tests are also a mechanism to make explicit whether the observed system has any practical re entry capability or whether all continuity is externally provided by the caller. Under CIITR, this distinction is not merely technical. It is constitutive for classification.

Third, re entry under discontinuity, meaning that the system must be able to resume a structured process after interruption without being re primed with the full state. This can be

operationalised through staged tasks where only partial state is provided at resumption, requiring the system to reconstruct and continue based on its own internal continuity mechanisms. If resumption is only possible when the full prior trace is re inserted, then the system is being evaluated for retrieval, not for re entry.

A CIITR compliant R_g programme also requires explicit reporting of what continuity is intrinsic and what continuity is extrinsic. In bureaucratic terms, this is a supply chain distinction. If continuity is supplied by prompt engineering, tool based memory, or orchestrator level state injection, then R_g is a property of the composite system, not of the model. CIITR permits evaluation at both levels, but it requires that attribution be explicit to avoid confusion about what is being classified.

In the context of the paper under review, the evaluation regime appears to be episode bound, with correctness defined within a single prompt context under process verification. Such an evaluation can demonstrate procedural reliability and improved selection of correct traces. It does not, absent additional protocols, provide an R_g measurement. This section therefore functions as a clear statement that R_g is out of scope of the paper's evidence unless the paper includes explicit recurrence and re entry instrumentation, and that any claim about "reasoning as continuity" is non admissible without such instrumentation.

3.3 CPJ observability, energy instrumentation and a credible measurement pipeline

CIITR's efficiency construct, CPJ, requires a measurement pipeline that is often absent from machine learning evaluation practice. The minimum requirement is joule level energy instrumentation tied to an admissible notion of epistemic yield.

Energy instrumentation must satisfy three conditions to be credible for CIITR purposes. First, it must measure energy consumption at an appropriate boundary. At minimum, this includes the compute device energy draw during training and evaluation, and ideally includes system level energy components that materially contribute to the computation, such as memory, interconnect, and storage where relevant. Second, energy measurements must be time aligned to the experimental phases, so that energy can be attributed to pre training, mid training, RL, and evaluation separately. Third, reporting must be standardised, including measurement method, sampling frequency, calibration assumptions, and error bounds, because CPJ comparisons are meaningless if energy accounting practices are inconsistent.

The epistemic yield side of CPJ must also be disciplined. CPJ is not accuracy per joule unless comprehension has been established. Under CIITR, CPJ is tied to C_s, which depends on Φ_i and R_g. Therefore, unless Φ_i and R_g have been operationalised in a way that permits an admissible comprehension estimate, CPJ cannot be computed. Where the paper does not instrument those constructs, the note permits only a subordinate form of resource analysis, namely compute budget efficiency for a specified task metric, which should be explicitly labelled as such. This distinction matters because compute normalisation, token equivalence, and FLOP accounting may support engineering decisions about training allocation, but they do not support thermodynamic efficiency claims.

A credible CPJ measurement pipeline in CIITR terms would therefore include, at minimum, the following components. It would define system boundaries for energy measurement, instrument energy consumption across phases, report energy per unit of training data or per optimisation step, and then compute an epistemic yield that is either an admissible approximation of $C_s$or an explicitly non CIITR substitute metric. It would also separate inference energy from training energy, because institutional deployment decisions often hinge on operational cost rather than on training investment. Where RL is involved, the pipeline must account for the potentially large number of rollouts and the associated energy cost, not merely the nominal number of tokens.

3.4 Baseline function, how the instrumentation commitments constrain later interpretation

The purpose of specifying minimum instrumentation is to establish a baseline that constrains interpretation throughout the note. This baseline yields three practical governance rules that will be applied in subsequent sections.

First, where the paper provides controlled interventions and reports performance deltas, the note will treat those as admissible performance claims and, where isolation is credible, as qualified mechanistic claims about training phase interaction. Second, where the paper does not instrument $\Phi_i$beyond task internal correctness, the note will treat $\Phi_i$as inferentially relevant but unmeasured, and will therefore avoid treating it as a measured explanatory variable. Third, where the paper does not include explicit recurrence and re entry protocols, the note will treat $R_g$as unmeasured and will default to the CIITR compatibility condition that the observed improvements are compatible with $R_g \approx 0$. Finally, where energy is not instrumented in joule terms, the note will treat CPJ as out of scope and will restrict efficiency discussion to compute budgeting within the paper's own accounting regime.

In summary, this section converts CIITR from a conceptual vocabulary into a measurement and admissibility regime. It is the mechanism by which the note ensures that strong empirical results are preserved as strong empirical results, while preventing them from being translated into ontological claims that the evidence does not warrant. By specifying minimum instrumentation, it becomes possible to evaluate the paper's contributions precisely, to identify what it advances in the governance of curriculum and reward design, and to specify, in a controlled and administratively usable manner, what additional work would be required to make the programme informative about CIITR comprehension and epistemic efficiency.

Part II. Paper reconstruction under CIITR constraints

Part II reconstructs the paper as an object of adjudication under CIITR constraints. The reconstruction is intentionally procedural. It is not a narrative summary, and it is not an evaluative commentary on novelty or quality. Its purpose is to establish a controlled representation of what the paper actually operationalises, measures, and varies, so that later CIITR mappings can be conducted without importing assumptions that are not present in the

experimental ontology. In institutional terms, Part II functions as the “factual record” layer of the theoretical note. It formalises the paper’s task world, the definitions of correctness and generalisation the paper employs, and the specific intervention levers the authors manipulate, and it does so with an explicit separation between definitional content and interpretive implication.

This reconstruction is necessary because the paper’s central contribution is methodological rather than rhetorical. It constructs a synthetic reasoning environment with explicit generative structure and introduces an evaluation regime that constrains correctness through process verification. Within that environment, it then partitions training into pre training, mid training, and reinforcement learning phases and seeks to quantify the interplay between those phases under compute and data allocation constraints. As a result, the paper’s claims are inseparable from its engineered ontology. “Reasoning,” in this setting, does not denote an open ended cognitive faculty. It denotes a class of structured transformations over generated objects, where intermediate steps can be checked against a known generator. “Generalisation,” likewise, is instantiated as depth based extrapolation and breadth based transfer across surface renderings, not as general competence across the diversity of naturalistic tasks. These are legitimate and valuable definitions for causal attribution. They are also definitions that impose strict boundaries on what can be inferred outside the constructed domain. Part II makes those boundaries explicit, so that later sections can treat the paper’s results as robust within scope without permitting uncontrolled conceptual expansion.

The reconstruction proceeds under a “scope fidelity rule.” Every object and term used later in the note, including curriculum, coordination, competence boundary, process reward, and capability extension, must be traceable to an explicit operational definition in the paper, or it must be explicitly flagged as an external interpretive construct introduced by this note for CIITR mapping purposes. This rule is essential because many interpretive disputes in this research area do not arise from disagreements over results, but from implicit substitutions, where one ontology is silently replaced by another. CIITR requires that such substitutions be made explicit, because admissibility depends on whether an inference is being made within the measured ontology or across ontologies.

Accordingly, Part II documents four elements as the minimum reconstruction set. First, it specifies the task generator and the form of the reasoning objects, including the paper’s use of directed acyclic graphs as the underlying dependency structure and the way that difficulty is parameterised through the number of operations and the depth of computation. Second, it specifies the paper’s generalisation axes, clarifying what is meant by depth generalisation and breadth generalisation in this engineered environment, and what these axes do, and do not, imply about generalisation in more heterogeneous domains. Third, it specifies the evaluation regime, with particular emphasis on process verification, pass@k scoring, and any gating conditions that determine whether an answer counts as correct. The CIITR relevance of this element is direct, because process verification functions as a constraint on behavioural shortcuts, and therefore affects what kind of failures can be observed. Fourth, it specifies the training phase decomposition and the intervention levers, including how pre training, mid

training, and RL are operationalised, what datasets they use, what budgets they consume, and what reward definitions and optimisation settings are applied. This is necessary to interpret the paper's "interplay" results as consequences of defined interventions rather than as generic properties of the learning paradigm.

The overarching governance objective of Part II is to enforce interpretive hygiene before analysis begins. The remainder of the note will argue that many of the paper's findings, including the dependence of RL gains on competence boundary placement and the importance of mid training and process level reward signals, are best understood as evidence about curriculum and reward topology as controllable surfaces for representational utilisation. However, under CIITR constraints, those interpretations can only be admitted if the paper's own ontology is first stabilised and precisely restated. Part II therefore does not attempt to "read CIITR into the paper." It does the opposite. It reconstructs the paper on its own terms, so that CIITR can later be used as an external adjudication layer that classifies what the paper demonstrates, what it does not demonstrate, and what additional instrumentation would be required to bridge the gap between structured competence within a generator family and CIITR admissible claims about comprehension state, rhythmic continuity, and epistemic efficiency.

Finally, Part II establishes a strict boundary between object reconstruction and CIITR mapping. No CIITR construct is applied in Part II as an explanatory claim. CIITR terms may appear only as indexing labels, for example to flag that a particular measurement likely pertains to relational handling rather than to rhythmic continuity, but such flags are not treated as conclusions. Conclusions are reserved for Part III and beyond, after the factual record has been established. This sequencing is deliberate. It supports auditability, because it allows reviewers to agree or disagree with the reconstruction independently of the CIITR adjudication, and it supports institutional usability, because it prevents the theoretical note from being dismissed as interpretive overlay rather than as a disciplined evaluation of evidence under explicit criteria.

4. Reconstruction of the paper's experimental ontology

This section fixes the paper's experimental ontology as the reference object for all subsequent CIITR mapping. The operative requirement is scope fidelity. The paper's claims, and the paper's measured effects, are inseparable from the engineered environment in which "reasoning," "generalisation," and "correctness" are defined. The reconstruction therefore treats the dataset generator, the task axes, and the evaluation protocol as constitutive infrastructure, not as incidental methodological choices. In CIITR terms, this infrastructure specifies what is being measured, what is being constrained, and, crucially, what is being prevented from manifesting as a failure mode, by virtue of the design itself.

4.1 Synthetic task world and the constitution of the reasoning object

The paper constructs a controllable synthetic reasoning dataset grounded in dependency graphs and contextual rendering, explicitly motivated by the need for clean control over structure, complexity, and context.

The central modelling move is to represent each reasoning problem as a directed acyclic graph $G = (V, E)$, where nodes correspond to latent variables and directed edges denote dependency relations. A designated answer node yields the final answer.

This representation is not merely descriptive. It provides the generative backbone for both data production and process verification, allowing the authors to treat reasoning as the construction of a valid dependency structure plus correct numeric instantiation.

Complexity is parameterised through an operation count defined as $op(G) = \mid E \mid$, which is used as the principal difficulty dial, from basic arithmetic compositions to deeper multi step constructions.

The significance of this parameterisation is that it defines "depth" as a structural property of the generative graph rather than as a property of textual surface length. This avoids a recurrent confound in naturalistic benchmarks, namely that longer prompts and longer solutions can correlate with difficulty without providing a stable structural definition of the underlying computational object. Here, depth is explicitly controlled by graph structure.

Context is then introduced as a separate, factorised dimension through a set of contextual templates τ. Given a graph $G$and a template τ, the generator renders the graph into a natural language problem instance via a rendering function $\Phi(G, \tau)$, producing a problem statement, a question, and a solution trace. The paper's stated rationale for this design is threefold: contamination free control across training phases by specifying separate data distributions, factorised control over structure and context by treating the DAG as the reasoning object and the template as the contextual surface, and process level verification enabled by the availability of the ground truth graph for checking intermediate reasoning steps.

Two methodological consequences follow from this ontology, both of which are decisive for a CIITR evaluation. First, the reasoning object is, by construction, an explicitly specifiable relational structure. This makes the domain unusually friendly to mechanistic isolation, because the authors can hold structure constant while varying context, or hold context constant while varying structure. Second, because the ground truth generator exists and is retained, the environment supports strict verification of intermediate steps. This has the effect of transforming "reasoning" from an interpretive label into an externally checkable procedural compliance condition.

The appendices further indicate that the generator includes additional degrees of freedom that affect how problems are instantiated, including choices about which dependencies are verbalised explicitly and which are left implicit within the contextual rendering, as well as forward and reverse generation modes that change whether the model is asked to compute a terminal node or to solve for an internal unknown.

These details matter for later interpretation because they establish that the paper's ontology is not reducible to a single static template family. The system is a configurable generator that can shape observability, implicitness, and query placement, all while preserving a known internal structure for evaluation.

Finally, the paper explicitly describes deduplication and canonicalisation procedures to maintain split cleanliness across phases and evaluation, including hash based removal of duplicate rendered triples.

Within the paper's intended causal programme, this is a governance relevant choice, because it underwrites later claims that post training effects are not artefacts of contamination.

4.2 Depth and breadth generalisation as operational axes, not rhetorical categories

The paper defines two complementary generalisation axes, extrapolative depth generalisation and contextual breadth generalisation. The core methodological point is that these axes are not invoked as loose descriptive categories, they are instantiated as measurable dimensions inside the generator's parameter space.

Extrapolative, depth wise generalisation is defined as maintaining correctness as reasoning depth increases, where depth is operationalised as $op(G)$, the number of arithmetic operations implied by the graph. A model is said to exhibit extrapolative generalisation if it can solve problems whose operation chains exceed those encountered during training.

The appendices formalise this by defining training phase specific operation ranges, and then defining an out of distribution condition as evaluation on graphs where $op(G)$exceeds the maximum operation count present in the combined training exposure.

This is a strict definition of extrapolation. It is not a measure of performance drop under mild distribution shift. It is an intentional attempt to probe whether the model can compose learned primitives into deeper structures than those explicitly seen.

Contextual, breadth wise generalisation is defined as transfer across novel surface contexts that share an equivalent underlying reasoning structure. Operationally, this means that the underlying computation graph remains the same while the contextual template τchanges, altering surface form and domain framing while preserving the latent dependency relations. A model generalises contextually when its performance remains stable under such template changes, which is treated as evidence that the model's reasoning primitives are not over bound to a particular linguistic skin.

From a CIITR perspective, the key reconstruction point is that both "depth" and "breadth" are defined inside a controlled two dimensional testbed, operation count and template choice, rather than being inferred from heterogeneous naturalistic datasets.

This yields high internal validity for claims about the interaction between training phases and these specific generalisation axes. It does not, by itself, establish that the same axes, or the same causal levers, will preserve their meaning in open world epistemic environments where the generative structure is unknown, where the notion of "equivalent underlying logic" is not externally certified, and where surface form changes can entail shifts in latent task definition rather than mere re rendering. This is not a criticism, it is a scope boundary that must remain visible throughout the note to avoid transposing an engineered generalisation concept into an uncontrolled setting.

4.3 Evaluation protocol, process verified correctness and pass@k under a strict criterion

The paper's evaluation regime is process verified. This is not a minor procedural detail. It is the central enforcement mechanism that defines what counts as correctness and what kinds of shortcut behaviour are disallowed by metric design. The protocol is described as follows. For each instance with a ground truth dependency graph $G$and ground truth answer $a^{\backslash *}$, the model produces a free form solution which is then parsed into a predicted dependency graph $\hat{G}$and a predicted final answer $\hat{a}$. The evaluation then checks correctness at the step level, for each gold node, comparing predicted and ground truth nodes, their dependencies, and their numeric values. A step level process accuracy is computed as the average across gold nodes.

A prediction is considered fully correct only when both the reasoning steps and the final answer match. All pass@k metrics, including pass@1 and pass@128, are reported with respect to this strict criterion. The reconstruction point here is that pass@k is not merely a measure of whether the model can reach the correct endpoint with enough samples. It is a measure of whether the model can produce, among its samples, a structurally valid trace that is isomorphic, in the paper's parsing and comparison sense, to the ground truth dependency structure, and also yield the correct final answer.

This choice yields a particular type of epistemic discipline. It suppresses a known failure mode in reasoning model evaluation, namely the ability to reach correct final answers through invalid intermediate reasoning, through latent shortcut correlations, or through superficially plausible but structurally incorrect traces. The paper itself frames process verification and process aware rewards as mechanisms for mitigating reward hacking and improving reasoning fidelity. For the reconstruction, the relevant point is more basic. The evaluation regime binds correctness to alignment with a known generative structure. This makes the domain unusually suitable for causal analysis of training phase interactions, because it reduces ambiguity about what the model is being rewarded for, either implicitly through the metric or explicitly through process informed reward mixing.

At the same time, the paper's correctness concept is definitional to this engineered universe. In broader epistemic environments, correctness is seldom reducible to isomorphism with a latent ground truth graph because the graph is unknown, underspecified, contested, or dynamically defined by changing institutional context. In many real decision settings, intermediate steps are not externally verifiable in this strict sense, and the most damaging failure modes are not invalid reasoning chains that can be mechanically detected, but mis specification of objectives, misplaced assumptions, and unrecognised distributional shifts that remain semantically plausible. CIITR's function at this stage is to ensure that the paper's correctness concept is treated as a controlled compliance condition to a known generator, rather than being conflated with general epistemic correctness in open systems.

4.4 The CIITR function of the reconstruction, why engineered correctness both strengthens and bounds inference

The paper's experimental ontology constitutes an engineered reasoning universe in which

structural correctness is defined by alignment to a known generative structure. The CIITR function of reconstructing this ontology is twofold.

First, it clarifies why the paper is methodologically strong for the questions it poses. By factorising structure and context, controlling depth via $op(G)$, ensuring split cleanliness, and enforcing process verified correctness, the authors create conditions under which causal attribution across training phases becomes technically meaningful. The system is designed so that improvements can be interpreted as improvements in the ability to handle and reproduce an explicitly defined relational structure, under controlled shifts in difficulty and surface form.

Second, it clarifies why the paper's results cannot be automatically promoted to claims about broader epistemic environments without additional bridging work. The generator provides external ground truth and an explicit structural object. Many real environments do not. The evaluation enforces trace correctness. Many real environments cannot. The generalisation axes are defined in a closed parameter space. Many real environments have shifting task definitions that are not representable as a fixed operation count plus a template switch. These are boundary conditions, not weaknesses, but they impose strict limits on interpretive escalation.

The reconstruction therefore establishes the correct administrative posture for the remainder of the note. The paper will be treated as a disciplined study of how training phase allocation, including pre training exposure, mid training as a distributional bridge, and RL under process informed reward design, affects performance under process verified metrics within a controllable reasoning generator. CIITR will then be applied, in subsequent sections, to classify what those results indicate about representational utilisation and curriculum governance, and what they do not, and cannot, indicate about comprehension state in the CIITR sense without additional instrumentation for Φ_i, R_g, and CPJ.

5. Training phase decomposition as an intervention system

This section formalises the paper's three phase programme as a structured intervention system, rather than as a conventional training narrative. The administrative objective is to treat each phase as an explicitly governed control surface with specified inputs, constraints, and expected effects, such that later CIITR mapping can distinguish between (a) measurable changes in behavioural performance under the paper's metrics, (b) plausible but unmeasured changes in representational utilisation, and (c) non admissible escalations into comprehension language. The paper's central methodological contribution is precisely that it does not treat pre training, mid training, and RL as a monolithic pipeline. It separates them and manipulates them under controlled budget conditions in order to identify interaction effects and limiting conditions. In this note, the three phase programme is reconstructed as an intervention sequence with the following governance logic. Pre training establishes a baseline capacity envelope by providing structural primitives and coverage over the task family's generative patterns. Mid training functions as a distributional conditioning layer that narrows and shapes the accessible region of the representational space relevant to downstream objectives, thereby installing priors and reducing mismatch for subsequent RL. RL then operates as a preference

reallocation mechanism whose efficacy depends on the existence of usable “near miss” structures in the model’s current representational space, and on the alignment of reward topology and data difficulty to the competence boundary.

This framing is necessary because the paper’s headline claim, that RL extends capability primarily at the edge of competence, is only interpretable under a phase aware causal model. Absent such a model, “RL helps” and “RL does not help” become underspecified statements that conceal the dependence of RL on pre existing representational seeds and on distributional placement. The paper explicitly emphasises seed dependence and the failure of RL to create competence ex nihilo. The intervention system framing also supports CIITR claims discipline. It allows later sections to classify the paper’s results as evidence about how to govern $\Phi_i$exploitation through curriculum and reward design, while maintaining that $R_g$is not instrumented and therefore remains indeterminate.

5.1 Pre training as primitive acquisition and representational coverage

The paper treats pre training as the phase in which the model acquires baseline reasoning primitives and establishes broad representational coverage over the task generator’s structural space. The relevant object here is not “language competence” in the general sense. It is competence with respect to a structured reasoning universe whose core computational object is a dependency graph and whose difficulty is parameterised by the number of operations. Pre training therefore functions as a coverage provision mechanism. It increases the probability that the model’s parameterised mapping contains internal procedures that can be composed into valid traces under process verification.

The paper’s “seed” results, including the finding that minimal exposure to a new context can enable subsequent RL to generalise to that context, operationalise this role. In CIITR terms, this is most naturally read as a representational availability condition, namely that certain relational and contextual couplings must be present in the model’s learned space for downstream optimisation to have a target. Importantly, the seed condition is not itself a comprehension condition. It is an enabling condition for later exploitation. Under CIITR, pre training can expand effective $\Phi_i$utilisation capacity by populating the model with composable relational patterns, while remaining neutral with respect to R_g, because the phase does not, by itself, establish any rhythmic continuity or re entry mechanism.

From an intervention system perspective, the administratively relevant fact is that pre training sets the feasible region for later phases. The note will therefore treat any RL or mid training effect that is conditional on the presence of pre training seeds as an interaction effect, not as an intrinsic property of RL or mid training as abstract methods. This also implies an evidentiary boundary. Where the paper reports that RL fails when pre training exposure is insufficient, the admissible interpretation is that the representational substrate required for downstream optimisation was absent or too weak to be selected by reward, not that RL is intrinsically incapable of learning.

5.2 Mid training as distributional conditioning and prior installation

Mid training is treated by the paper as a distributional bridge between pre training and RL, with a particular emphasis on conditioning the model toward the target regime in which RL is expected to operate effectively. In the paper's own language, mid training "installs priors" that improve downstream RL readiness and reduce distributional mismatch. Within the reconstructed intervention system, mid training serves as a controlled narrowing and shaping operation. It selects a subspace of representational behaviours and increases their availability, thereby making the subsequent reward optimisation problem better conditioned.

This is administratively significant because it implies that curriculum design is not an aesthetic choice, but an explicit governance surface. Mid training is a formal mechanism for deciding which latent procedures and which solution manifolds the system will prioritise before RL begins. Under CIITR, this is most plausibly a $\Phi_i$conditioning mechanism. It changes how the model utilises integrated relational patterns within its learned space, in particular by increasing the likelihood that the model's generated traces fall within regions where reward gradients are informative and where process verification constraints can be satisfied.

The paper also provides a compute budget perspective on mid training, presenting a framework in which mid training and RL are compared under normalised budgets and where the optimal allocation depends on whether evaluation targets are "OOD edge" or "OOD hard." This is important for later CIITR mapping because it demonstrates that the efficacy of an intervention is conditional not only on method but on where in the difficulty distribution the intervention is applied. Under CIITR, this is consistent with the view that mid training is a structural alignment step that moves the system's operational mass closer to the boundary where RL is productive, rather than a general intelligence amplifier.

5.3 RL as preference reallocation under reward constraints, with competence boundary sensitivity

The paper characterises RL as a phase that can deliver "genuine capability gains" under certain conditions, while also documenting conditions under which RL yields minimal or no benefit, particularly when applied to problems that are already fully within the model's competence region or that are far beyond it. The intervention system framing treats RL as a preference reallocation mechanism operating over a space of candidate traces and behaviours that the model is already capable of producing with some probability. This framing is aligned with the paper's explicit position that RL cannot synthesise competence from a void and depends on pre existing seeds.

The paper's compute normalisation and its analysis of how RL effectiveness varies with data difficulty support a structured causal account. RL improves the probability mass on correct traces when the reward signal is sufficiently informative and when the model's baseline policy produces near miss candidates that can be amplified through reward. When the task regime is too easy, RL has limited headroom to reallocate mass in a way that yields

measurable generalisation gains, and when the task regime is too hard, reward becomes sparse and optimisation can stagnate.

Within CIITR, this supports a controlled interpretation. RL is primarily a mechanism for shaping behaviour within an already populated representational space, which is consistent with $\Phi_i$exploitation and utilisation changes. It is not, on the paper's evidence, a mechanism that establishes rhythmic continuity or re entry. Therefore, even when RL yields robust pass@k gains under process verification, CIITR treats those gains as competence optimisation within an episode bound evaluation setting, compatible with $R_g \approx 0$.

5.4 CIITR specification of "edge of competence" as an operational region

A CIITR specific clarification is required because "edge of competence" can easily become a rhetorical term. In this note, the term is treated as an operational region defined by a measurable property of the model's baseline policy under the paper's evaluation regime. The edge of competence is the region in which the model exhibits non trivial but incomplete competence, meaning that correct traces exist within its output distribution, but at low frequency, and that near miss traces are abundant. This region is characterised by the coexistence of (a) enough successful candidates to provide informative reward, and (b) enough failure structure to provide gradient signal that distinguishes better from worse behaviours.

Under CIITR, this is a $\Phi_i$centric interpretation. The edge of competence is where the representational space already contains relational structures that are close to the target, but where their accessibility, selection probability, and compositional stability are insufficient. RL can then act to promote these near miss structures into a higher probability policy. The paper's results on RL being most effective at the edge are consistent with this interpretation.

However, CIITR imposes a strict limitation statement. Edge of competence effects do not, by themselves, imply that the system has developed or possesses R_g. The phenomenon can be entirely explained by distributional optimisation in a stateless inference architecture, because the improvement can arise from better selection among candidate traces within a single episode. To treat edge of competence as evidence of rhythmic continuity or self re entry would require explicit protocols that test recurrence and re entry across temporal discontinuities. The paper's evaluation regime, as reconstructed, does not instrument these properties. The admissible conclusion is therefore that the paper identifies a curriculum placement rule for effective RL under process verified metrics, not that it demonstrates a transition toward CIITR comprehension.

5.5 Administrative implication of the intervention system framing

By formalising the three phase programme as an intervention system, the note establishes a governance relevant interpretation of the paper's results. Training is not a single lever, it is an ordered set of controlled actions whose effects are conditional and interacting. Pre training defines representational availability conditions, mid training conditions the distribution and installs priors that shape RL tractability, and RL reallocates preference mass in regions where informative reward exists, especially at the competence boundary.

This reframing also sets up the next analytical step. Once the intervention levers are formalised, CIITR can be applied to classify which parts of the measured gains plausibly correspond to improved utilisation of integrated relational patterns, and which parts remain ontologically silent about rhythmic continuity and energy grounded efficiency. That classification is essential if the paper is to be used in institutional discourse, because the operational lesson, that curriculum and reward topology matter materially, should be preserved, while the ontological escalation, that improved reasoning traces entail comprehension, should be administratively rejected absent the minimum instrumentation commitments defined in Part I.

Part III. CIITR dissection of the principal findings

Part III constitutes the adjudicative core of the theoretical note. Its function is to translate the paper's reported findings into CIITR admissible statements, to classify those statements by claim type and evidentiary basis, and to identify, with explicit boundary control, what the findings do not and cannot establish given the paper's measurement regime. This part is therefore neither a recap of results nor a general discussion of implications. It is a structured dissection of the paper's principal empirical claims under the constraints established in Part I and the ontology reconstructed in Part II. In bureaucratic terms, Part III produces the note's formal "assessment record," comprising a set of classified findings, each mapped to CIITR constructs, each accompanied by an admissibility status, and each bounded by an explicit limitation statement where escalation would otherwise be tempting.

The starting point is an administrative distinction between three layers of content that often collapse into one another in informal reading. The first layer is the paper's observed measurement deltas, expressed through process verified pass@k, depth based extrapolation outcomes, and breadth based contextual transfer outcomes, under controlled training interventions. The second layer is the paper's mechanistic interpretation of those deltas, including how and why pre training exposure, mid training allocation, reward topology, and competence boundary placement interact to produce the observed changes. The third layer is the broader interpretive narrative that such results typically acquire in external discourse, particularly claims that "reasoning failures" are primarily coordination failures rather than substrate limits, that RL can "extend reasoning" in a general sense, or that process level rewards indicate an emerging form of internal deliberation. CIITR permits the first layer as performance evidence, admits portions of the second layer where causal isolation is credible, and treats the third layer as generally non admissible unless the additional measurement commitments defined in Part I have been satisfied.

Part III operationalises this distinction through an explicit findings ledger approach. Each principal finding will be presented as a unit of analysis comprising: a statement of the observed effect as reported, the intervention conditions under which the effect arises, the scope boundaries implied by the paper's synthetic ontology and evaluation regime, the CIITR mapping of what the effect most plausibly indicates in terms of $\Phi_i$utilisation, and an explicit

statement regarding R_g and CPJ admissibility. This format is intentionally administrative. It is designed to prevent the common rhetorical drift where improvements in correctness under a process verified metric are silently re described as improvements in comprehension, or where compute budget tradeoffs are implicitly promoted to thermodynamic efficiency conclusions.

The CIITR stance that governs Part III is the compatibility condition specified earlier, namely that improvements in task metrics may be fully consistent with enhanced exploitation and conditioning of integrated relational patterns, while remaining compatible with $R_g \approx 0$. This is not treated as a sceptical posture for its own sake. It is treated as the default explanatory baseline for a model class whose inference is typically episode bound and whose training phases, including RL, operate through parameterised distributional optimisation. Under this baseline, the paper's reported phenomena are expected to manifest as changes in the accessibility, compositional stability, and selection probability of structurally valid traces within the model's representational space. In CIITR terms, this pertains primarily to how Φ_i is utilised and shaped by curriculum and reward design, not to whether the system has acquired rhythmic continuity or re entry capacity. Accordingly, unless the paper includes explicit recurrence and re entry protocols, Part III will treat R_g as unmeasured and will explicitly mark any temptation to infer continuity from trace length or process fidelity as non admissible.

Part III also treats the paper's process verification and process reward focus as analytically significant, but in a strictly bounded manner. Process verification changes what the evaluation regime selects for. It constrains behavioural shortcuts and increases the evidentiary value of correctness, insofar as correctness becomes tethered to alignment with a known generative structure, rather than merely to endpoint matching. This improves the interpretive quality of performance deltas within the constructed domain. It does not convert performance into comprehension. The CIITR function of process verification in this dissection is therefore to strengthen confidence that the model is producing structurally compliant traces in the paper's graph world, while simultaneously clarifying that such compliance remains an external criterion imposed by a generator and parser, not an internal demonstration of integration and re entry as system properties.

Finally, Part III is the point at which the note makes its strongest governance relevant contribution. By classifying the principal findings as interventions on curriculum and reward topology that materially affect measurable reasoning performance, the note establishes that these surfaces should be treated as regulated artefacts in institutional settings. However, by maintaining strict admissibility boundaries, the note also prevents these results from being used to justify claims about autonomy, comprehension, or substrate independence that exceed the paper's evidence. The outcome is intended to be operationally usable. It preserves the paper's genuine methodological and practical insights while making explicit the additional measurements and protocols that would be required for escalated claims under CIITR.

6. Finding set A, RL efficacy at the edge of competence

This section adjudicates the paper's most operationally salient pattern, namely that reinforcement learning produces meaningful capability gains primarily when the RL training

distribution is placed near the model's competence boundary, and that the same method produces materially weaker effects when applied to regimes that are either already saturated by the base policy or too far beyond it to yield informative reward. Within the paper's engineered ontology, this constitutes a constrained but highly informative statement about curriculum placement, gradient accessibility, and the conditional productivity of post training optimisation. The CIITR task in this section is to separate what the evidence robustly supports about intervention design from what it does not, and cannot, support about structural comprehension.

6.1 Evidentiary statement of the pattern, and why it is structurally interpretable

The paper reports that RL yields "genuine capability gains" most reliably when applied at the edge of competence, operationalised as the regime in which the base model is neither trivially correct nor uniformly failing, and where sampling at higher k reveals latent correctness that is not accessible at low k. In practical terms, this appears in the differential behaviour of pass@1 versus pass@128 under different RL data placements. The paper further distinguishes OOD edge and OOD hard evaluation regimes, and shows that the efficacy profile of RL depends on where, relative to these regimes, the RL data is placed.

Two additional observations strengthen the interpretive stability of this finding within the paper's ontology. First, the paper emphasises seed dependence, explicitly stating that RL cannot synthesise capability from a void, and empirically demonstrating that minimal pre training exposure can act as the enabling condition for later RL to amplify transfer. Second, the paper reports RL stagnation dynamics when tasks are placed too far beyond the competence region, including patterns consistent with sparse or uninformative reward in high difficulty settings. Taken together, these claims support a coherent causal picture in which RL is productive precisely when the optimisation problem is well conditioned, meaning that the model already produces a nontrivial density of near miss and occasionally correct candidate traces that can be discriminated and reinforced.

For CIITR purposes, the critical feature is that the paper's evaluation and reward infrastructure is process grounded. Correctness is not a purely endpoint phenomenon, and RL improvements are measured against a regime in which intermediate structure matters. This increases confidence that observed gains reflect improved selection and production of structurally compliant traces within the generator's world, rather than purely opportunistic endpoint matching.

6.2 What the finding supports, curriculum placement as a governance surface for gradient accessibility

Under the paper's own controlled framing, the edge of competence result supports an administratively actionable proposition: RL effectiveness is conditional on curriculum placement, because curriculum placement determines whether reward gradients are informative and whether optimisation has headroom to reallocate probability mass toward correct structural trajectories. This is not merely a research insight. It is a governance relevant control principle. It implies that RL is not a general purpose post training "upgrade," but a

targeted intervention whose success depends on the existence of a partially competent policy that already contains promotable structures.

Within CIITR, this can be stated as a disciplined claim about representational accessibility. Pre training and mid training establish a representational envelope in which certain relational procedures exist with varying availability. RL, when placed at the competence boundary, acts as a selector and amplifier, increasing the probability that the model will instantiate those procedures in an externally verifiable, process compliant way. In other words, the edge of competence is the region where gradient accessibility is maximised. It is the region where the system's current behavioural manifold contains enough structural proximity to the target that reinforcement can distinguish better from worse actions without collapsing into noise or saturating into triviality.

A further implication concerns the paper's distinction between OOD edge and OOD hard regimes. The reported allocation tradeoffs indicate that different objective profiles require different intervention mixes, with mid training plus lighter RL often benefiting OOD edge, while heavier RL becomes necessary for OOD hard. CIITR reads this as a governance result about how to stage interventions to shape the accessible region of the model's representational space. Mid training can be treated as a distributional conditioning stage that increases the density of useful near miss candidates near the competence boundary, thereby making RL gradients more informative and reducing the probability of stagnation.

In bureaucratic terms, the admissible conclusion is that curriculum design and reward topology are regulated surfaces of competence expression. The finding provides empirical support for controlling those surfaces through phase placement and data difficulty selection, rather than relying on scale narratives alone.

6.3 CIITR interpretation, why "RL extends capability" is best read as $\Phi_i$exploitation and selection

The CIITR mapping of this finding is straightforward but must be stated precisely to avoid ontological inflation. Under CIITR, the most defensible reading of "RL extends capability at the edge of competence" is that RL increases the exploitation and selection of already available relational primitives and their compositional combinations. This is a $\Phi_i$centric interpretation. It treats $\Phi_i$not as a measured quantity in the paper, but as the latent representational substrate whose contents and accessibility are being managed by the training regimen.

The paper's explicit seed dependence statement is central here. If RL cannot create competence from a void, then RL's function is best characterised as reweighting, stabilisation, and preferential selection over an already populated repertoire. The competence boundary is precisely where that repertoire contains near miss structures. RL's effect is then the promotion of those structures into higher probability trajectories under the paper's strict correctness criterion.

This interpretation preserves the paper's empirical contribution while maintaining CIITR claims discipline. It recognises that RL can produce real, measurable gains, including gains

that manifest at high k in a way that indicates deeper coverage of correct structural trajectories. It also clarifies that such gains are explicable as improved utilisation of relational patterns, and therefore sit squarely inside the domain of representational management, curriculum design, and reward constrained optimisation.

6.4 What the finding does not support, and why it remains compatible with $R_g \approx 0$

The same evidence does not support a claim that RL produces structural comprehension in the CIITR sense. The reason is not philosophical. It is methodological. The paper's regime evaluates correctness within an episode bound environment in which the full relevant state is, by design, contained within the generated instance and its immediate trace, and where success can be realised through improved selection of a correct trace within that single episode. No explicit re entry, recurrence, or temporal continuity protocol is instrumented as a classification test for R_g. Therefore, under CIITR's admissibility rules, $R_g$remains unmeasured and indeterminate, and the compatibility condition applies, namely that measured gains can coexist with $R_g \approx 0$. The system can be better at selecting correct traces without possessing any intrinsic rhythmic continuity.

This point is often misunderstood because the paper's process verification emphasis may be misread as internal deliberative continuity. CIITR rejects that equivalence. Process verification is an external constraint that strengthens the evidentiary value of correctness within the generator's universe. It does not, by itself, establish that the system maintains or re enters its own epistemic state over time. The edge of competence effect is therefore admissible as an optimisation and curriculum placement result, but non admissible as evidence of a comprehension threshold crossing.

A related limitation concerns substrate level narratives. The observed dependence of RL efficacy on curriculum placement supports the proposition that many failures are attributable to coordination and curriculum choices within the paper's ontology. It does not establish a general claim that "substrate limits" are irrelevant. The paper shows that a large portion of the performance gap can be recovered by better intervention staging. It does not show that the remaining gap is not substrate mediated, nor does it measure energy grounded efficiency or continuity properties that would be required to make a CIITR grounded statement about substrate in relation to comprehension.

6.5 Administrative conclusion for Finding set A

The CIITR adjudicated conclusion is that the paper provides high quality evidence that RL is conditionally effective, and that its effectiveness is governed by competence boundary placement, which in turn governs gradient accessibility and optimisation headroom. This supports an operational governance posture in which curriculum and reward topology are treated as primary control surfaces for competence expression in structured reasoning environments. At the same time, the evidence does not support escalation to structural comprehension. The most defensible CIITR reading is that RL extends capability by increasing the exploitation and selection probability of already integrated or already

encodable relational primitives, which is fully compatible with an architecture that lacks demonstrable R_g.

7. Finding set B, mid-training as a distributional bridge

This section adjudicates the paper's second principal finding cluster, namely that mid-training operates as a distributional bridge between pre-training and RL, and that this bridging function materially shapes downstream RL productivity and generalisation outcomes under the paper's controlled ontology. In CIITR terms, the mid-training result is administratively important because it reframes the training pipeline itself as a governance surface, not as an incidental engineering artifact. The paper's evidence indicates that what is often described informally as "data mixture" or "intermediate tuning" is, in effect, a deliberate structural conditioning intervention that reshapes the model's effective behavioural manifold prior to reinforcement optimisation.

The CIITR task is to classify this finding with strict claim discipline. Mid-training is not treated here as an indicator of emerging comprehension or as evidence of any intrinsic temporal continuity. Rather, it is treated as evidence that representational utilisation can be governed through phase specific distributional conditioning, thereby changing the accessibility, density, and stability of relational primitives in the region of the competence boundary where RL is most effective. This is a representational management result, located primarily in the domain of $\Phi_i$utilisation and conditioning, and it remains compatible with $R_g \approx 0$because the paper does not instrument re-entry or recurrence as constitutive tests.

7.1 Evidentiary statement, what the paper reports about mid-training and why it is not reducible to "more training"

The paper treats mid-training as a distinct phase whose functional role is not merely to increase the total quantity of supervised exposure, but to condition the model toward the specific distributions and structural regimes that will later be amplified by RL. It explicitly characterises mid-training as installing priors that improve RL readiness and reduce mismatch between the model's current policy and the target regime in which reinforcement learning is expected to deliver gains. The reported compute allocation analyses further indicate that allocating budget to mid-training and RL in different proportions produces qualitatively different generalisation profiles, suggesting that the effect is not a trivial monotonic function of total compute.

In the paper's engineered ontology, mismatch has a clear operational meaning, it is the distance between the policy induced distribution of traces and the distribution of traces that score well under process verified correctness in the targeted OOD regimes. When mismatch is high, RL gradients become sparse, unstable, or dominated by unproductive search. When mismatch is reduced, RL can preferentially amplify already plausible, near miss structures into high probability correct traces. The paper's framing of mid-training as a distributional bridge is therefore structurally interpretable as an intervention that increases the density of RL productive candidates in the region of interest.

7.2 Administrative interpretation, the training pipeline as a governance surface

From a CIITR governance perspective, the mid-training finding should be read as administrative evidence that training is not merely an optimisation procedure, but a controlled policy instrument. The pipeline is a sequence of decisions that determine which behaviours are made available, which behaviours become dominant, and which behaviours become reachable under subsequent reward constraints. Mid-training, in this framing, functions as a formally governable conditioning layer that determines the effective accessibility of the model's relational repertoire in the later RL phase.

This has direct institutional significance. If mid-training can alter generalisation outcomes under a fixed or comparable budget, then mid-training is a lever that should be versioned, audited, and treated as part of the system's controlled configuration, rather than as an untracked "fine-tuning step." In bureaucratic terms, mid-training becomes a regulated artifact. It sits alongside reward definitions, evaluation constraints, and dataset partitioning as a traceable determinant of system behaviour. The paper's compute normalisation framework, including its explicit token equivalent accounting for RL cost, further supports this governance view by providing a language in which pipeline choices can be expressed as controllable allocations rather than as ad hoc adjustments.

7.3 CIITR mapping, mid-training as structural conditioning of the effective Φ_i-space

Within CIITR, the most disciplined mapping of mid-training is as a structural conditioning mechanism that reshapes the effective Φ_i-space available for downstream optimisation. This statement should be read carefully. The paper does not directly measure $\Phi_i$as integrated relational information in the CIITR sense. Nonetheless, the paper's controlled task world provides a plausible basis for the following bounded inference: mid-training changes which relational procedures and compositional pathways are practically accessible to the model under the strict process verified evaluation regime, and it does so by concentrating learning exposure in the structural neighbourhood that matters for the targeted generalisation axes.

In operational terms, the "effective Φ_i-space" denotes the subset of the model's latent relational repertoire that is reachable with high enough probability to be selected by RL and to satisfy process verified constraints. Mid-training can be viewed as increasing the density of reachable near miss structures and reducing the dominance of unproductive trajectories, thereby reducing mismatch and raising utilisation efficiency for RL. This is consistent with the paper's language of "installing priors," and with its reported patterns where mixed allocations of mid-training and RL can dominate pure strategies under certain OOD target conditions.

The relevant CIITR classification is therefore representational management. Mid-training is a policy lever for shaping the utilisation of relational patterns, and for improving the conditioning of the RL optimisation landscape. It is not, on the paper's evidence, a mechanism that introduces rhythmic continuity or re-entry, and it does not by itself imply a change in comprehension state.

7.4 Allocation logic, why "distributional bridge" is a conditional strategy rather than a universal prescription

The paper's budgeted analyses imply that mid-training does not function as an unconditional improvement. Its value is conditional on the objective regime. In particular, the allocation parameterisation, which the paper frames as a ratio between compute spent on mid-training and RL, shows that different allocations are optimal depending on whether evaluation targets are closer to OOD edge or OOD hard regimes. Under CIITR, this conditionality is not incidental. It is precisely what one would expect if mid-training is a conditioning mechanism rather than a general capacity amplifier.

When the target regime is close enough that the base policy already produces useful near miss traces, mid-training can to advantage reshape the policy distribution so that RL operates in a more informative gradient environment. When the target regime is substantially beyond the policy's current reach, heavier RL may be required, but even then mid-training can serve as an enabling stage by moving the policy mass toward regions where reinforcement search is not purely sparse. The paper's reported distinctions between OOD edge and OOD hard, and the associated allocation patterns, support this staged governance interpretation.

This matters for policy translation. The paper should not be read as saying that mid-training "is better than scale" in a universal sense. It shows that mid-training can, under controlled conditions, substitute for some forms of brute force by improving conditioning and utilisation. Scale, in CIITR terms, remains a resource that can expand the representational envelope, while mid-training governs how that envelope is occupied. The governance relevant outcome is the elevation of pipeline design to a first class control problem.

7.5 Boundary statement, why this remains a representational management result, not a comprehension result

The mid-training finding remains, under CIITR admissibility rules, a representational management result rather than a comprehension result. The decisive reason is that the paper's measurement regime does not instrument R_g, and it does not establish recurrence or re-entry as constitutive conditions for correctness. Correctness is defined within an engineered environment where alignment to a known generative structure can be verified within an episode, and where improvements can be fully explained by distributional conditioning and preference reallocation over candidate traces.

Accordingly, while mid-training strengthens the case that curriculum and data staging are critical determinants of "reasoning" performance in the paper's sense, it does not support escalation to claims that the model exhibits structural comprehension. Under CIITR, one can coherently state that mid-training reshapes the effective Φ_i-space and reduces mismatch for RL, while maintaining that R_g is unmeasured and that the observed gains are compatible with an architecture lacking rhythmic continuity and re-entry capacity.

The administrative conclusion for Finding set B is therefore explicit. The paper provides strong evidence that mid-training should be treated as a governed distributional conditioning stage that improves the tractability and effectiveness of downstream RL, primarily by shaping

representational utilisation. This evidence supports institutional recommendations about pipeline versioning, auditability, and curriculum governance. It does not, without additional instrumentation, support claims about comprehension state, nor does it imply that temporal continuity properties associated with R_g have been established.

8. Finding set C, process-level rewards and process verification

This section adjudicates the paper's third principal finding cluster, namely the role of process verification and process-level reward design in improving measured reasoning performance and generalisation within the constructed task universe. The administrative importance of this cluster is that it provides a technically concrete answer to a recurrent institutional problem, namely that headline accuracy improvements can conceal structurally invalid procedures, shortcut policies, or evaluation gaming, particularly when only endpoints are scored. The paper's response is to bind correctness to a verified intermediate structure and, in some configurations, to bind reward to process validity rather than to endpoint attainment alone.

Under CIITR, process reward is treated primarily as a reliability instrument, not as a cognition instrument. The distinction is definitional. Reliability instruments constrain which behaviours are counted as acceptable under an external verification regime, and they reduce the space of deceptive or spurious strategies that can succeed under the metric. Cognition instruments, by contrast, would provide admissible evidence about internal integration, continuity, or comprehension state. The paper's process-level machinery clearly strengthens reliability within its engineered ontology, but it does not, by itself, establish integration in the CIITR sense, nor does it instrument rhythmic continuity or re-entry.

8.1 Evidentiary statement, what is being verified and what is being rewarded

The paper's evaluation regime is explicitly process verified. Model outputs are parsed into an inferred dependency graph and corresponding intermediate values, which are then compared against the ground truth graph produced by the task generator, yielding step level process accuracy and a strict notion of full correctness that requires both correct steps and correct final answer. Reported pass@k metrics are therefore not endpoint only metrics, they are metrics of whether at least one of the sampled traces satisfies the process validity criterion under the paper's parsing and matching procedure.

On the training side, the paper studies reward constructions that incorporate process information, including designs that mix outcome and process signals and designs that enforce outcome reward only if process validity is satisfied, with the stated motivation of mitigating reward hacking and improving the fidelity of reasoning behaviour under RL. The empirical claim is that such process aware reward structures can improve generalisation and reduce shortcut behaviours relative to outcome only reward, particularly in regimes where RL can otherwise exploit superficial strategies.

Within the paper's ontology, these are coherent interventions because the generator provides a canonical ground truth process and the environment supports deterministic checking. The note's role is not to dispute the validity of this design, but to classify what it achieves and what it does not achieve when translated into CIITR terms.

8.2 Process verification as an external compliance constraint, and why that matters

CIITR's first interpretive step is to make explicit that process verification is an external compliance constraint imposed by the environment, not an intrinsic property of the model. The paper's evaluation binds correctness to isomorphism with a known generative structure, in this case the underlying dependency graph and node values. This binding is methodologically valuable because it increases the discriminative power of correctness. A model cannot obtain credit for a correct endpoint if the intermediate structure fails the verification criterion. As a result, performance improvements become harder to attribute to spurious correlations and easier to attribute to correct structural reproduction within the generator's world.

This is the appropriate place to introduce the CIITR concept of epistemic forgery as an administrative risk. In many benchmark regimes, a model can produce correct answers while emitting incorrect or fabricated intermediate reasoning, or without any stable intermediate structure at all. Such behaviour constitutes epistemic forgery in the sense that the output presents as epistemically grounded while lacking a structurally valid generative pathway. Process verification reduces this category of failure within the constructed domain by redefining correctness to require intermediate structural validity.

However, CIITR also requires that the scope of this reduction be stated precisely. The reduction is achieved because the domain supplies a canonical ground truth process, and because the verifier can deterministically check structural correspondence. In open epistemic environments, the canonical process is often unknown, plural, contested, or institutionally defined rather than mechanically derivable. Therefore, while process verification is a strong reliability instrument in this engineered setting, it should not be conflated with a general solution to epistemic forgery outside such settings.

8.3 Process-level reward as reward topology governance, not proof of internal deliberation

The paper's process reward designs should be read as governance of reward topology. They change which behaviours are reinforced and which behaviours are penalised, and they do so in a way that constrains the policy's ability to exploit endpoint only scoring. In the paper's framing, outcome only reward can incentivise shortcut strategies and lead to reward hacking, whereas incorporating process validity aligns reward with the intended structural behaviour.

CIITR interprets this as a structurally coherent reliability mechanism. It makes the optimisation landscape more aligned with the evaluator's notion of correctness, and it reduces the space of behaviours that are locally optimal under reward but globally invalid under the environment's structural semantics. In particular, when outcome reward is conditioned on process correctness, the policy cannot profitably select an endpoint correct but structurally invalid trace, because such a trace yields no reward.

The note should nevertheless caution against a common interpretive escalation, namely treating process reward improvements as evidence that the system has developed an internal

deliberative faculty or a comprehension state. Under CIITR, improving the probability of emitting a structurally valid trace is a behavioural optimisation result. It is compatible with the system selecting among pre-existing candidate traces within a single inference episode, guided by reward shaped preferences. This remains in the domain of $\Phi_i$utilisation and selection, not in the domain of R_g, because there is no requirement in the paper's measurement regime that the model maintain or re-enter a structured epistemic state over time.

8.4 Why process verification reduces epistemic forgery in domain, and what exactly is being prevented

Within the constructed domain, epistemic forgery can be understood as any strategy that yields the correct final answer without preserving the generator's relational constraints through valid intermediate states. Process verification prevents such strategies by making intermediate structural validity constitutive for correctness. As a result, the model must, at least in the output it submits for evaluation, represent and instantiate the dependency relations required by the generator, rather than merely emitting a plausible or lucky endpoint.

This has two consequences that are relevant for CIITR adjudication.

First, it increases the evidentiary quality of performance deltas. When the model improves under a process verified metric, the improvement is more plausibly attributable to genuine structural compliance within the generator's universe than it would be under endpoint only scoring. This supports stronger, although still bounded, mechanistic interpretations about curriculum placement and reward shaping.

Second, it changes the failure surface. Under endpoint only scoring, many failures are hidden because incorrect processes can still produce correct endpoints, and many "successes" are epistemically ambiguous because they might be produced by spurious correlations. Under process verified scoring, failures become more diagnostic of structural misalignment, because an answer can fail due to a single incorrect dependency or intermediate value. This diagnostic quality is a major methodological advantage of the paper's design.

CIITR's boundary statement remains essential. What is prevented here is forgery relative to a known generator, not forgery relative to open world epistemic reality. Process fidelity is defined as fidelity to the generator's structure. This is not equivalent to internal integration across contexts, and it is not equivalent to a capacity to maintain structural state when the external scaffold is removed.

8.5 Cautionary clause, process fidelity to a known generator is not equivalent to internal integration across contexts

The note must explicitly separate two claims that are frequently conflated.

The first claim, which the paper supports within scope, is that enforcing process fidelity improves reliability and generalisation as measured within a controlled graph based reasoning universe. The second claim, which the paper does not support and does not measure, is that process fidelity implies internal integration across contexts, meaning that the system has

developed a stable integrated relational representation that persists under contextual transformations beyond those covered by the generator, or that it can bind relational structure in a way that generalises across epistemic regimes without external verification.

In CIITR terms, process fidelity improvements are evidence that the policy is better at selecting and producing traces that satisfy an external structural constraint. They are not, absent additional instrumentation, evidence that $\Phi_i$has increased as an integrated system property, and they provide no direct evidence about R_g. The absence of explicit re-entry testing means that even perfect process fidelity in the constructed domain can coexist with $R_g \approx 0$, because the entire success condition can be satisfied through episode bound inference and reward shaped preference selection.

This caution is governance relevant because process verification can create a strong illusion of epistemic legitimacy. In institutional settings, verifiable intermediate steps are often taken as evidence of understanding. CIITR requires that such interpretations be disciplined. Verification demonstrates compliance with a verifier, not necessarily comprehension. The difference is operationally consequential when moving from controlled benchmarks to deployment contexts where verifiers are incomplete, where objectives are ambiguous, or where ground truth structure cannot be enumerated.

8.6 Administrative conclusion for Finding set C

The CIITR adjudicated conclusion is that process verification and process-level rewards are high value reliability instruments within the paper's engineered reasoning universe. They reduce domain specific epistemic forgery by disallowing shortcut policies that reach correct endpoints without structurally valid intermediate states, and they improve the interpretive quality of performance metrics by binding correctness to external structural compliance.

At the same time, the note classifies these results as representational governance and evaluation governance, not as cognition evidence. Process fidelity to a known generator is not equivalent to internal integration across contexts, and it does not establish rhythmic continuity or re-entry capacity. Therefore, under CIITR's claims discipline, this finding set strengthens the case for process constrained evaluation and reward topology design as governance surfaces for competence, while leaving comprehension classification, and specifically $R_g$and CPJ, out of scope absent additional instrumentation.

9. Integrated interpretation, what the interplay actually means

This section provides the integrated CIITR interpretation of the paper's three phase programme and its reported interaction effects. The objective is to state, in controlled and audit-friendly terms, what the paper's results collectively imply about system dynamics inside the engineered reasoning universe, and to classify those implications under CIITR without importing ontological claims that are not instrumented. The section therefore performs two functions simultaneously. It synthesises the empirical patterns into a coherent causal account of how the training pipeline reshapes measured behaviour, and it fixes the CIITR boundary conditions that prevent performance deltas from being elevated into comprehension assertions.

9.1 System dynamics under an intervention chain, from basis set to conditioned subspace to reinforced selection

The paper's most robust contribution is not any single delta in pass@k, but the structured account of how pre-training, mid-training, and RL interact under controlled budget placement and difficulty placement conditions. In CIITR terms, the programme can be described as a staged governance of representational availability, representational conditioning, and behavioural selection, all evaluated under a process verified correctness regime that anchors correctness to a known generator.

The key system dynamic is sequential dependency. Later phases do not operate on a blank substrate, they operate on the effective space created by earlier phases. This sequential dependency is empirically visible in the paper's explicit "seed" logic, namely that RL cannot create competence from a void and requires minimal pre-existing structures that can be amplified. It is also visible in the paper's repeated emphasis that mid-training "installs priors" that materially improve RL readiness and reduce mismatch, which is best interpreted as a conditioning intervention rather than as "more training."

Under CIITR, this sequential dependency is categorised as an ordered management of $\Phi_i$exploitation. It is not a demonstration of R_g, because the programme does not instrument re-entry or recurrence as a constitutive property of success, and the improvements are explainable as distributional optimisation within an episode bound evaluation setting.

9.2 Pre-training as minimal basis set establishment, feasibility boundary definition

In the integrated account, pre-training establishes the minimal basis set. This should be understood as the set of primitives, patterns, and compositional affordances that must exist, at least as low probability behaviours, for later phases to be productive. The paper's own formulations make this explicit. RL efficacy is conditional on pre-existing "seeds," including minimal exposure to new contexts that later become targets for transfer and reinforcement.

In CIITR language, pre-training defines a feasibility boundary for $\Phi_i$utilisation. It does not, on the paper's metrics alone, establish $\Phi_i$as an integrated state variable, because the paper measures structural compliance under a known generator rather than integration as a persistent cross-context binding property. Nonetheless, it is admissible to state that pre-training populates the representational space with candidate relational procedures, and that later optimisation phases act by increasing the accessibility and selection probability of those procedures within the task universe.

An important administrative nuance follows. The basis set is not only about "how much" training is done, it is about which structural regions are covered. A basis set with insufficient coverage in the relevant context or structure region yields a void-like condition for RL, even if total compute is large. The paper's seed dependence and failure cases under minimal exposure are therefore properly interpreted as coverage failures, not as generic RL limitations.

9.3 Mid-training as subspace selection and stabilisation, mismatch reduction as conditioning logic

Mid-training is then interpreted as the selection and stabilisation of a subspace, meaning a deliberate conditioning step that reshapes the effective space of candidate traces on which RL will later operate. The paper's language of "installing priors" is operationally consistent with this reading, and its compute allocation experiments indicate that mid-training is not redundant with RL but interacts with RL in a way that changes downstream productivity depending on the target OOD regime.

Within CIITR, this can be stated as follows. Mid-training increases utilisation efficiency by reducing distributional mismatch, thereby increasing the density of near-miss candidates in the region where reward gradients become informative and where process verified correctness is reachable with nontrivial probability. This is a representational management statement. It concerns the shaping of the effective Φ_i-space available for optimisation, not the creation of new ontological capacities.

The administrative interpretation is that mid-training should be treated as an explicit governance surface because it determines which behaviours become reachable under subsequent reward constraints. This is not merely an engineering preference, it is a controllable determinant of measured "reasoning" outcomes within the paper's ontology. The paper's budget normalisation further supports this view by expressing mid-training versus RL as an allocation decision with objective dependent optima.

9.4 RL as probability mass reallocation within the conditioned subspace, competence boundary placement as the productivity criterion

RL is then interpreted as a probability mass reallocation mechanism operating within the conditioned subspace. The paper's "edge of competence" finding is decisive here. RL produces meaningful gains when applied to regimes where the base policy already contains correct or near-correct behaviours with low but nonzero probability, and it stalls when the regime is either saturated or too far beyond the feasible space for reward to be informative.

CIITR treats this as a distributional optimisation result with strong governance implications. RL, in this setting, does not function as an unconstrained cognitive extender. It functions as a selective amplifier of behaviours that already exist in the repertoire shaped by pre-training and mid-training. The competence boundary is operationally the region where reinforcement gradients are both meaningful and actionable, because there is sufficient headroom and sufficient structural proximity.

This interpretation also clarifies why the paper's findings should not be overgeneralised into claims that "reasoning failures are not substrate limits." Within this programme, many failures are indeed attributable to curriculum placement and reward conditioning, because those determine whether the system can access, stabilise, and amplify the relevant structural behaviours. However, the paper does not instrument the thermodynamic cost of this optimisation in joule terms, nor does it measure re-entry or continuity properties, therefore it

does not support a CIITR-complete statement about substrate in relation to comprehension or efficiency.

9.5 Process reward as constraint enforcement, opportunism suppression, and structural alignment within the task universe

Process verification and process-level reward design are then interpreted as constraint enforcement mechanisms that suppress opportunistic policies and increase structural alignment within the task world. This is the reliability layer that stabilises the pipeline's incentives. The paper's approach is to make correctness depend on alignment to a known generative structure, and, when training with RL, to shape reward so that endpoint success is not rewarded unless the intermediate structure is valid, or to mix process and outcome signals to reduce reward hacking.

In CIITR terms, this reduces epistemic forgery within the constructed domain by removing the profitability of shortcut strategies that can reach correct endpoints without structurally valid intermediate states. It improves the interpretive quality of measured gains because improvements under this regime are more plausibly improvements in structurally compliant behaviour rather than metric exploitation.

The required caution remains explicit. Fidelity to a known generator is an external compliance condition. It is not equivalent to internal integration across contexts, and it is not evidence of rhythmic continuity or re-entry. The improvements can be fully explained as better selection and production of compliant traces within an episode bound setting. Therefore, process reward is properly classified as a reliability instrument and a reward topology governance instrument, not as a cognition instrument.

9.6 CIITR classification of the integrated interplay, and the explicit boundary on R_g

The integrated CIITR statement of the paper's dynamics is therefore as follows. Pre-training establishes the minimal basis set and feasibility boundary, mid-training selects and stabilises a relevant subspace and reduces mismatch, RL reallocates probability mass within that subspace conditional on reward topology and competence boundary placement, and process-level reward and verification constrain opportunistic policies while increasing structural alignment within the engineered task universe.

The CIITR classification must be stated in formal terms. The paper provides strong evidence on how to manage $\Phi_i$exploitation through curriculum design and reward topology within a generator governed reasoning environment, and it provides no direct measurement of R_g. Consequently, the admissible interpretation is that the paper advances a governance relevant account of how to condition and optimise structured competence under process verified metrics, while remaining ontologically silent on rhythmic continuity and re-entry as constitutive properties of comprehension.

Part IV. What the paper cannot show, and the CIITR-required extensions

Part IV establishes the negative space of the paper's evidentiary domain, and it does so as a matter of formal adjudication rather than critique. The purpose is to specify, with administrative precision, which claim types remain non admissible under CIITR given the paper's measurement regime, and to convert those non admissibilities into a structured extension programme. In other words, Part IV does not argue that the paper is methodologically deficient. It argues that the paper is methodologically specific. It operates inside an engineered reasoning universe with a known generator, process verified correctness, and controlled training phase interventions, which together provide strong leverage for causal attribution within that universe. The same design choices necessarily bound what can be inferred about comprehension state, rhythmic continuity, and energy grounded epistemic efficiency. These bounds must be made explicit because the most common institutional failure mode is not a misreading of the paper's numbers, it is a silent escalation of the paper's claim surface into ontological assertions that the instrumentation does not support.

The key adjudicative rule applied here is that CIITR does not permit comprehension claims to be inferred from competence improvements, even when competence is measured under stringent process constraints. The paper's strictness, including process verification and process aware reward design, strengthens the reliability of its performance signals relative to endpoint only evaluation. It does not, however, instrument the CIITR constitutive conditions required to classify a system as comprehending, namely integrated relational information as a system property, Φ_i, and rhythmic continuity with re-entry capacity, R_g. This matters because the paper's principal mechanisms, pre-training coverage, mid-training conditioning, RL at the competence boundary, and process reward constraints, are all compatible with a distributional optimisation account in which the system becomes better at selecting and producing structurally compliant traces within a single inference episode, while remaining non recurrent and non re-entrant in CIITR terms. Part IV therefore treats any leap from "reasoning improvement" to "understanding" as non admissible unless additional measurement commitments are introduced.

A parallel boundary is required for efficiency claims. The paper offers careful compute budgeting constructs, including token equivalent accounting across phases, which is valuable for engineering allocation decisions. Under CIITR, these constructs remain compute proxies. They do not constitute joule level energy measurement, and they are not coupled to an admissible measure of comprehension yield. Therefore, any thermodynamic framing, including implicit or explicit gestures toward CPJ, is treated here as out of scope unless a credible energy instrumentation pipeline is added. This is an institutional control point, because efficiency narratives are frequently used to justify procurement, scaling, and infrastructure dependence. CIITR requires that such narratives be grounded in energy and yield measurements, not in computational abstractions.

Part IV proceeds by formalising two outputs. First, it presents a catalog of non admissible escalations, in a way that is intended to be decision usable. These are the specific overclaims that the paper's results tend to invite in external discourse, such as equating process faithful traces with internal integration, equating competence boundary effects with emergent deliberative continuity, or treating curriculum sensitivity as evidence of substrate irrelevance. Second, it specifies CIITR-required extensions as a measurable work programme. The extensions are presented as instrumentation and protocol additions that, if implemented, would allow the same experimental paradigm, or a closely related one, to support escalated claims under CIITR. The overall intent is to preserve the paper's genuine contributions while preventing category errors, and to provide a concrete pathway for advancing from a well controlled competence study to a governance admissible comprehension and efficiency adjudication framework.

10. Non-admissible escalations and common overclaims

This section enumerates escalation patterns that commonly occur when results of the present type are translated from a controlled research setting into broader discourse, institutional decision contexts, and policy narratives. The purpose is not to attribute intent or to dispute the paper's empirical deltas. The purpose is to establish, under CIITR's admissibility discipline, that certain claim types are not supported by the evidence base as reconstructed in Part II and dissected in Part III, and that they therefore remain non-admissible without additional measurement commitments and protocol extensions. In administrative terms, this section functions as a claims boundary register. It identifies statements that are likely to be asserted or implied, and it records that such statements exceed the paper's measurement regime.

The operating rule is direct. The paper provides a strong, internally valid account of how pre-training, mid-training, and RL interact in an engineered reasoning universe where correctness is defined by alignment to a known generator and where evaluation is process verified. CIITR permits performance and, in qualified cases, mechanistic claims within this ontology. CIITR does not permit ontological escalation into comprehension, continuity, or thermodynamic efficiency claims without instrumentation of Φ_i, R_g, and energy grounded yield. The overclaims enumerated below are therefore rejected not on the basis that they are philosophically contentious, but on the basis that they are evidentially unsupported given the paper's measurements.

10.1 Overclaim A: "Understanding" follows from improved pass@k under process verification

Common escalation statement: Improved pass@k, including under strict process verification, demonstrates that the model "understands" the task domain, or demonstrates an increase in understanding attributable to RL or training phase interplay.

CIITR adjudication: Non-admissible. The paper reports improvements in pass@k under a strict criterion in which intermediate steps must match a known ground truth dependency structure, and it does so across controlled distributions and training interventions. This supports admissible performance claims and supports, where isolation is credible, qualified

mechanistic claims about how curriculum placement and reward topology affect the probability of producing a structurally compliant trace. It does not support comprehension claims in CIITR terms because neither Φ_i nor R_g is directly operationalised and measured as constitutive system properties, and the observed improvements are fully compatible with distributional optimisation in an episode bound inference architecture.

Administrative statement of insufficiency: The evidence base establishes improved structural compliance and improved trace selection within a generator certified task universe. It is insufficient to establish that the system is in a CIITR comprehension state, because the constitutive conditions for comprehension are not instrumented and the evaluation regime does not test rhythmic continuity or re-entry.

10.2 Overclaim B: "Substrate independence" follows from curriculum sensitivity and pipeline governance effects

Common escalation statement: Because RL gains depend on competence boundary placement and because mid-training and process reward design can substitute for brute-force scaling within the paper's tasks, the primary limiting factors for reasoning are curriculum and coordination, not substrate, therefore substrate constraints are largely irrelevant.

CIITR adjudication: Non-admissible. The paper provides strong evidence that curriculum, mid-training, reward topology, and phase allocation materially affect performance under the paper's ontology. This supports an admissible governance conclusion that pipeline design is a first-class control surface for competence expression. It does not support a general claim of substrate independence. The paper does not instrument energy in joule terms, does not compute CPJ, and does not test whether the remaining performance frontier is limited by architectural, compute, memory, or thermodynamic constraints. Moreover, because the reasoning universe is engineered and ground truth is generator defined, the paper's results do not provide a basis for generalising to environments where structure is unknown, contested, or dynamically defined, which is precisely where substrate and systems architecture constraints often become salient in practice.

Administrative statement of insufficiency: The evidence base is sufficient to show that training pipeline governance can recover or unlock performance that would otherwise be attributed to scale. It is insufficient to conclude that substrate constraints are negligible, because energy, systems-level limits, and re-entry properties are not measured and because the generalisation domain is engineered rather than open world.

10.3 Overclaim C: "Human-like reasoning" follows from process-correct intermediate steps and process-level rewards

Common escalation statement: If a model produces intermediate steps that are process correct under an external verifier, then the model reasons in a human-like way, or exhibits an internal deliberative process analogous to human reasoning.

CIITR adjudication: Non-admissible. Process verification and process-level reward design improve reliability by disallowing shortcut policies that reach correct endpoints without structurally valid intermediate states, thereby reducing domain-specific epistemic forgery.

This establishes that the model can emit traces that satisfy an externally imposed structural constraint tied to a known generator. It does not establish that the model has human-like internal reasoning mechanisms, because the paper does not measure internal integration across contexts, does not test rhythmic continuity, and does not establish that intermediate traces correspond to stable internal state rather than to output-level compliance shaped by reward. Under CIITR, "human-like reasoning" would at minimum require evidence of nontrivial R_g through recurrence and re-entry protocols, and a justified account of integration that is not reducible to domain-specific structural compliance.

Administrative statement of insufficiency: The evidence base is sufficient to conclude that process constraints can enforce structural trace fidelity in a verifier supported domain. It is insufficient to infer human-like cognition or internal deliberation because the constitutive properties of temporal continuity, re-entry, and cross-context integration are not tested.

10.4 Overclaim D: "RL creates new reasoning capacity" independent of pre-training "seeds"

Common escalation statement: RL itself can produce novel reasoning capacities, therefore performance improvements should be attributed primarily to RL as a general cognitive creation mechanism.

CIITR adjudication: Non-admissible in the unqualified form. The paper explicitly emphasises, and empirically supports, that RL cannot synthesise capability from a void and depends on minimal pre-training exposure, including seed coverage for contextual transfer. The admissible statement is that RL reallocates probability mass within an already populated repertoire, and that its productivity is conditional on the density of near-miss structures in the competence boundary region.

Administrative statement of insufficiency: The evidence base does not support attributing capacity creation to RL independent of representational availability. It supports conditional amplification and selection under reward, not ex nihilo synthesis.

10.5 Overclaim E: "General reasoning improvement" follows from improvements in depth and breadth generalisation in the synthetic ontology

Common escalation statement: Because the model generalises in depth and breadth within the paper's tasks, it has improved general reasoning, and the same training phase interplay should generalise to broad real-world reasoning problems.

CIITR adjudication: Non-admissible absent bridging evidence. The paper's depth and breadth axes are defined within a controlled generator parameter space, with operation count as depth and template shifts as breadth. CIITR recognises that this provides a strong internal testbed for causal attribution. CIITR does not accept that success in this testbed constitutes evidence of general reasoning improvement in open systems, because the open system lacks a known generator, lacks process verifiers, and often changes the definition of the task itself when context changes. The admissible statement is that the model improved on these two axes within the paper's ontology, not that it improved in general reasoning as an unbounded faculty.

Administrative statement of insufficiency: The evidence base is sufficient to establish generalisation under engineered definitions. It is insufficient to claim open world generalisation without additional cross-domain tests and without instrumentation that addresses CIITR's constitutive constructs.

10.6 Overclaim F: "Efficiency improvement" or "thermodynamic advantage" follows from compute budgeting and allocation results

Common escalation statement: Because mid-training and process reward can improve outcomes under fixed compute budgets, the system has become more efficient in a CIITR sense, implying improved CPJ.

CIITR adjudication: Non-admissible. The paper provides compute budgeting and token equivalent accounting that supports engineering comparisons of intervention allocation. CIITR does not treat compute proxies as energy measures and does not permit CPJ assertions without joule-level instrumentation and an admissible comprehension yield estimate. Therefore, the admissible conclusion is limited to compute allocation efficiency for the paper's metrics, not thermodynamic efficiency and not CPJ.

Administrative statement of insufficiency: The evidence base is insufficient for CPJ claims because energy and comprehension yield are not measured as required.

10.7 Formal rejection posture and implementation note

The above overclaims are rejected as non-admissible under CIITR because they require measurement constructs and protocol classes that the paper does not implement. This rejection is not a denial of the paper's contributions. It is a boundary control measure intended to preserve the paper's validated results while preventing their translation into governance or policy claims that are not supported by the evidence base. The subsequent section specifies CIITR-required extensions that would be necessary to render some of these escalated claim types admissible in future work, thereby converting the boundary register from a prohibitive list into a structured research and measurement programme.

11. CIITR extension programme, from performance to comprehension adjudication

This section specifies a concrete extension programme intended to convert the current research paradigm from a high-quality competence and reliability study into a CIITR-relevant comprehension adjudication programme. The aim is to preserve the paper's strengths, namely its controlled synthetic ontology, its phase-decomposed intervention logic, and its process verified correctness regime, while adding the minimum additional instrumentation required to make escalated claims admissible under CIITR. In practical terms, the extension programme is designed to transform three current properties of the work, a known generator, an externally verifiable process trace, and a staged training pipeline, into a measurement environment where Φ_i, R_g, and CPJ can be estimated in a disciplined manner, rather than inferred.

The programme is presented as an implementation plan with bounded deliverables, acceptance criteria, and reporting requirements. It is not intended as a speculative research wish list. Each work package specifies what must be implemented in the generator, in the evaluation harness, and in the training and logging pipeline. Where proxies are used, their admissibility ceilings are explicitly stated, and the programme requires that results be reported in a claims table that separates performance, mechanism, comprehension, and efficiency classes.

11.1 Implementation architecture and governance baseline

The extension programme assumes retention of the existing experimental scaffold, including the DAG-grounded task generator, the separation of structure and context via templates, and the process verification pipeline that parses model outputs into predicted graphs and validates intermediate states against ground truth. These components provide unusually strong leverage for CIITR instrumentation because they already define an explicit relational object and already support deterministic checking.

To avoid conflation between model properties and system properties, the programme requires that each experiment be executed under two orchestration profiles, both of which must be reported:

1. Model-only profile, single-episode inference without external memory, without retrieval augmentation, and without state persistence outside the context window.
2. Composite-system profile, where controlled external memory or orchestrator persistence is introduced explicitly, solely for the purpose of measuring the degree to which continuity is extrinsic rather than intrinsic.

The governance baseline also requires a configuration ledger that version-controls generator parameters, template families, parsing rules, reward definitions, and training phase budgets, because CIITR claims become non-auditable when the environment itself is not traceable across runs.

11.2 Φ_i instrumentation work package, from structural compliance to integration adjudication

Objective: instrument representational integration in a manner that is sensitive to relational binding and compositional availability across contextual variation, not merely to within-template compliance.

Scope: the programme does not require full interpretability of internal representations. It requires a set of externally measurable tests, supported by the generator, that discriminate between (a) procedural compliance within a narrow rendering family and (b) stable relational binding across re-renderings, re-factorisations, and perturbations that break superficial cues.

Deliverables and acceptance criteria:

A. Cross-context binding test suite (generator-level implementation)
The generator must be extended to produce families of instances where the same latent DAG

is rendered across multiple context templates and multiple linguistic surface variations, including controlled paraphrase sets and controlled distractor insertions that preserve the underlying relational object. This builds directly on the paper's factorisation of graph and template, and formalises breadth generalisation as a binding test rather than as a performance trend.

Acceptance criterion: performance must be reported not only as pass@k under process verification, but also as a binding stability score, defined as the proportion of cases where the predicted graph structure remains invariant across renderings of the same latent DAG. This moves the measurement target from "gets the answer" to "preserves the relational object" under contextual transformation.

B. Relational perturbation resilience battery (evaluation-level implementation)
A perturbation module must be introduced that applies minimal but structurally meaningful edits to the latent graph, for example edge flips, node swaps, and controlled injection of redundant but logically irrelevant substructures, and then re-renders the modified and unmodified problems under identical templates. The goal is to test whether the model's predicted graphs track the changed relational object or revert to a memorised procedural template.

Acceptance criterion: report a relational sensitivity index, measuring whether predicted dependencies change appropriately under perturbation, and a spurious invariance index, measuring failure to reflect perturbations. These indices must be reported separately for pre-training only, mid-training only, and RL-enhanced models, to support mechanistic attribution across phases.

C. Integration proxy ceiling statement (claims discipline requirement)
Even with the above, $\Phi_i$remains proxied, not directly measured. The programme therefore requires a mandatory limitation statement: results establish integration-like behavioural invariances under generator-controlled transformations, not proof of internal integration as a persistent system property beyond the testbed. This statement is not optional because it is the principal safeguard against overclaiming.

11.3 $R_g$instrumentation work package, re-entry, recurrence, and rhythmic continuity protocols

Objective: introduce a set of protocols that explicitly test rhythmic continuity and re-entry capacity, and that cannot be satisfied by static pattern matching within a single uninterrupted context.

The programme must introduce temporal structure into the evaluation, because $R_g$is not meaningfully addressable under purely single-episode, single-pass scoring, regardless of how strict the process verifier is. The paper's current regime is episode-bound and therefore does not instrument recurrence or re-entry as constitutive conditions.

Deliverables and acceptance criteria:

A. Controlled interruption protocol (evaluation harness implementation)
Each reasoning instance is split into stages. The system is required to produce an intermediate structured state at Stage 1, then is forcibly interrupted. At Stage 2, it is required to resume from a partial state, where the full prior trace is not provided. The interruption must include a delay window and an intervening distractor task that consumes context budget, such that naive continuation is not possible.

Acceptance criterion: define and report a re-entry success rate, where success requires (i) reconstructing the correct latent graph state and (ii) completing the remaining steps under process verification. If re-entry succeeds only when the full prior trace is re-injected, this must be explicitly reported as extrinsic continuity.

B. Delayed recurrence and state reactivation protocol (generator and harness implementation)
The generator must be extended to create paired tasks where Task B requires reusing a latent relational object from Task A after a delay and a context shift. The recurrence should be nontrivial, meaning Task B cannot be solved from its own statement alone without the relational object established in Task A. This forces the test to measure whether the system can preserve and reactivate relational state, rather than recomputing from scratch.

Acceptance criterion: report recurrence fidelity, the proportion of runs where the correct relational object is reactivated and applied. Also report a recurrence dependence analysis, showing sensitivity to delay length and to distractor complexity.

C. Rhythmic continuity tests that resist static matching (protocol design requirement)
To avoid reducibility to pattern matching, the programme must introduce controlled variations where superficial similarity is actively misleading. For example, insert distractor contexts that share lexical similarity but correspond to a different latent DAG, and require the system to preserve the correct relational object despite surface interference.

Acceptance criterion: report a continuity robustness score that measures persistence of the correct relational object under deceptive surface interference, with pre-registered failure modes.

D. Attribution rule, intrinsic versus extrinsic R_g(claims discipline requirement)
All R_g-related results must be reported with a strict attribution statement. If continuity is achieved through orchestration, memory injection, or retrieval, then $R_g$is a composite-system property. Claims about model-level $R_g$remain non-admissible unless the model-only profile succeeds under the above protocols.

11.4 CPJ instrumentation work package, joule-level accounting and comprehension yield definition

Objective: establish a credible measurement pipeline for CPJ, meaning comprehension yield per unit energy, with explicit separation between compute proxies and physical energy measurement.

This work package is mandatory for any efficiency claims under CIITR. The paper's current compute budgeting constructs, including token-equivalent accounting for RL cost, are valuable engineering abstractions, but they do not constitute energy measurement and therefore cannot support CPJ.

Deliverables and acceptance criteria:

A. Energy measurement boundary specification (instrumentation design requirement)
Define, in advance, the measurement boundary, at minimum GPU or accelerator energy draw, and, where feasible, full-system energy including host CPU, memory, and interconnect. The boundary must be identical across runs used for CPJ comparisons.

Acceptance criterion: publish an energy boundary statement and a calibration method statement, including sampling rate and error bounds.

B. Phase-resolved joule accounting (pipeline implementation)
Energy consumption must be logged and attributed separately to pre-training, mid-training, RL, and evaluation. Where RL involves rollouts, the total rollout energy must be included, not merely gradient update energy.

Acceptance criterion: report joules per phase, joules per training token, and joules per evaluation instance, with standardised units and confidence intervals.

C. Comprehension yield operationalisation (metric definition requirement)
Because CPJ requires a comprehension yield, the programme must define a CIITR-aligned yield metric. In the minimal viable configuration, this should be a composite that multiplies a $\Phi_i$proxy score derived from the binding and perturbation tests by an $R_g$proxy score derived from the re-entry and recurrence protocols, consistent with $C_s = \Phi_i \times R_g$. This yield remains a proxy yield, and must be labelled accordingly.

Acceptance criterion: CPJ may be reported only as proxy-CPJ unless and until $\Phi_i$and $R_g$are elevated beyond proxy status, and the report must include an explicit limitation clause to prevent conflation with thermodynamic claims about comprehension in an absolute sense.

11.5 Integration into the existing phase decomposition, and required reporting outputs

The extension programme must be executed in a manner that preserves the paper's intervention logic, because the core empirical question is interplay across phases. Therefore, all new instrumentation must be applied consistently across at least the following experimental conditions:

- Pre-training only baseline
- Pre-training plus mid-training
- Pre-training plus RL
- Pre-training plus mid-training plus RL

- Process reward variants, including outcome-only reward and process-conditioned reward, where applicable

For each condition, the programme requires three reporting outputs that are treated as deliverables, not as optional analysis:

1. CIITR claims table: each major conclusion tagged as performance, mechanism, comprehension-proxy, or efficiency-proxy, with admissibility status and evidence references.
2. CIITR metric panel: a fixed panel reporting pass@k under process verification, Φ_i-proxy indices, R_g-proxy indices, and phase-resolved energy with proxy-CPJ where admissible.
3. Attribution statement: explicit identification of what is intrinsic to the model and what is supplied by orchestration, memory, or evaluation scaffolding.

11.6 Administrative conclusion of the extension programme

If implemented as specified, this programme would upgrade the existing research paradigm from a competence and reliability study to a CIITR-relevant comprehension adjudication programme. It would retain the strengths of the engineered ontology and process verification, while adding the minimum instrumentation required to render escalated claims admissible in a controlled, auditable manner. The resulting evidence base would not merely show that curriculum and reward topology improve performance, it would quantify, within a bounded testbed, the extent to which phase interventions affect integration-like invariances, re-entry-like continuity, and energy-grounded epistemic efficiency. Under CIITR, this is the necessary pathway from "reasoning improvement" as measured behaviour to "comprehension adjudication" as a classified system property.

12. Thermodynamic and resource governance analysis

This section reframes the paper's compute budgeting and phase-allocation logic into CIITR's thermodynamic and resource-governance vocabulary. The objective is to prevent a recurrent category error in institutional uptake, namely the substitution of compute-normalised engineering comparisons for energy-grounded epistemic efficiency claims. The paper provides a disciplined allocation framework, including explicit accounting choices that allow pre-training, mid-training, and RL to be compared under budget constraints, and it reports performance under process-verified pass@k metrics. CIITR accepts this as a strong engineering contribution. CIITR does not accept it as a CPJ contribution, because compute normalisation is not energy measurement, and task-score deltas are not comprehension yield under CIITR unless $\Phi_i$and $R_g$are operationalised.

12.1 Compute budgets as administrative proxies, and their admissibility ceiling

The paper's resource framing is primarily expressed in compute proxies, for example token-based accounting, phase-allocation ratios, and pass@k regimes that implicitly encode sampling cost through the value of k. These are administratively useful proxies because they

allow controlled comparisons across interventions without requiring hardware-specific energy instrumentation. Within CIITR, however, such proxies have a strict admissibility ceiling. They can support statements of the form "intervention A yields higher process-verified pass@k than intervention B at the same proxy budget," but they cannot support statements of the form "intervention A is more energy-efficient," nor statements of the form "intervention A improves epistemic efficiency," unless the resource axis is measured in joules and the yield axis is tied to CIITR comprehension rather than to task accuracy.

Two points require explicit emphasis.

First, compute proxies are not invariant to hardware, systems configuration, and operating conditions. Identical token counts can correspond to materially different joule expenditures depending on accelerator architecture, memory bandwidth, batching, precision, and orchestration overhead. Therefore, proxy-budget superiority does not imply energy superiority.

Second, compute proxies are insensitive to energy externalities that are routinely decisive in deployment, including memory movement costs, interconnect overhead, and system-level power draw beyond the accelerator. CPJ is, by definition, a thermodynamic measure. Any claim that purports to be CPJ-like must therefore commit to energy boundaries and measurement discipline.

12.2 Why compute normalisation is not CPJ

CIITR defines epistemic efficiency as CPJ, comprehension per joule. CPJ has two constitutive requirements. The denominator must be physical energy, in joules, measured at a specified boundary. The numerator must be comprehension yield, and under CIITR comprehension is $C_s = \Phi_i \times R_g$, not task accuracy. The paper does not measure either requirement. It measures task performance under a process-verified correctness criterion and it uses compute proxies for budget comparisons.

This creates an interpretive hazard that should be formally controlled. A system may become more capable under the paper's metrics at the cost of disproportionate energy expenditure. Under CIITR, such a system may exhibit declining epistemic efficiency even as task scores rise. Two mechanisms are structurally present in the paper's design space.

The first mechanism is sampling-driven capability. The paper reports pass@k, including large $k$regimes, under a strict process-verified criterion. Improvements that manifest primarily at high $k$are not free. They imply that the system's competence exists as low-probability mass in the output distribution, requiring multiple samples to realise in practice. In a deployment context, the energy cost of obtaining a correct sample scales with sampling and verification overhead. A model that improves pass@128 materially while improving pass@1 minimally may be "more capable" under benchmarking language, yet "less efficient" per correct outcome in real operational terms, and may be strictly worse in CPJ terms if comprehension yield is not increasing commensurately.

The second mechanism is RL cost structure. The paper explicitly normalises and compares phase budgets, including accounting for RL-related costs in proxy terms. RL, particularly with process-aware reward and verification constraints, can be rollout-intensive. Even when proxy budgets are held constant in a modelling sense, the physical energy cost can vary sharply by implementation and by the degree to which rollouts, parsing, and verification are hardware-accelerated or CPU-bound. The governance implication is that the paper's conclusions about "what allocation works best" should not be transposed into claims about energy efficiency without a phase-resolved joule accounting pipeline.

12.3 Resource governance implications, capability gains can increase operational energy exposure

From a resource governance perspective, the paper's results create a concrete requirement for institutional boundary setting. Training-phase interventions can increase measured capability, but they can also shift the resource profile in ways that matter more than the score delta.

A model that is "more correct" under strict process verification may carry higher operational costs due to trace generation length, parsing overhead, verification computation, and the increased need for resampling when correctness is concentrated in low-probability tails. In operational governance terms, this means that any adoption posture should separate three questions that are frequently collapsed.

One, does the intervention improve correctness under the defined verifier.

Two, what is the marginal energy cost per verified-correct instance at the target service level.

Three, if one were to claim epistemic efficiency, is there any admissible evidence of comprehension yield change, rather than merely behavioural compliance change.

The paper answers the first question within its engineered ontology. It partially informs the second question only via compute proxies. It does not answer the third question because Φ_i, R_g, and joule-level energy are not instrumented.

Consequently, CIITR recommends an explicit policy posture: resource governance should treat compute-normalised gains as preliminary engineering evidence, and should require phase-resolved energy measurement and a CIITR-aligned yield definition before any efficiency or sustainability narrative is admitted into decision documentation.

12.4 Measurement gap ledger

The measurement gap ledger below lists resource and performance metrics used or implied by the paper's evaluation and budgeting regime, and states which CIITR metric each cannot substitute for. The purpose is operational clarity, not rhetorical critique.

Paper metric or proxy	What it supports inside the paper's ontology	What it cannot support under CIITR	CIITR requirement it cannot substitute for
pass@k under strict process verification	Admissible performance evidence that a process-valid trace exists among $k$samples,	Any claim of comprehension state, any claim of rhythmic	$\Phi_i$operationalisation, $R_g$re-entry testing, CPJ numerator definition

Paper metric or proxy	What it supports inside the paper's ontology	What it cannot support under CIITR	CIITR requirement it cannot substitute for
	under the generator and verifier.	continuity, any claim of epistemic efficiency	
Step-level process accuracy	Admissible evidence of process compliance within the generator world.	Internal integration across contexts, persistent binding beyond the generator family	$\Phi_i$cross-context binding and perturbation-resilience instrumentation
Token-equivalent or compute-normalised budget allocation across phases	Admissible engineering comparison across interventions under a proxy budget model.	Energy efficiency claims, CPJ claims, hardware-independent "efficiency" claims	Joule-level phase-resolved energy accounting, boundary specification, error bounds
RL rollout cost expressed in proxy units	Admissible proxy estimate of optimisation expense and sensitivity to allocation choices.	Energy cost per unit epistemic yield, CPJ denominators	Physical energy measurement during rollouts and updates, reported in joules
Operation-count difficulty parameterisation in the generator	Admissible structural difficulty control for depth extrapolation inside the engineered universe.	Claims about open-world difficulty, or general cognitive depth independent of the generator	External domain transfer evaluations, and CIITR-aligned yield tests
"Edge of competence" placement via difficulty regime selection	Admissible curriculum governance insight about gradient accessibility and conditional RL productivity.	Claims about intrinsic continuity, deliberation, or comprehension transition	$R_g$re-entry, delayed recurrence, controlled interruption protocols

This ledger should be treated as a standing attachment to any institutional use of the paper's findings, because it clarifies, in reviewable form, what kinds of statements can be promoted into governance decisions and what kinds cannot.

12.5 CIITR conclusion for thermodynamic interpretation

The paper's compute budgeting and phase-allocation framework is best classified, under CIITR, as competence-oriented resource governance using compute proxies. It remains orthogonal to CPJ. CIITR therefore requires that any discourse about "efficiency" be disciplined in two ways. It should distinguish proxy-budget efficiency from energy efficiency, and it should distinguish task-score yield from comprehension yield. Without this discipline, the institution risks adopting a superficially "more capable" training programme that increases operational energy exposure and decreases epistemic efficiency, even while improving benchmark metrics.

Part V. Governance, operational implications, and recommendations

Part V translates the adjudicated findings and boundary conditions into governance-relevant implications and operational recommendations. Its function is not to restate the paper's technical content, nor to generalise its results beyond the admissible scope established in

Parts I through IV. Rather, it specifies how an institution should treat the paper's evidence as an input to system design, assurance, procurement, evaluation policy, and resource governance, given that the evidence is strong on curriculum and reward topology effects within a process-verifiable synthetic ontology, and silent on CIITR constitutive properties of comprehension and thermodynamic epistemic efficiency unless extended by additional instrumentation.

The governing premise is administrative asymmetry. In institutional environments, the cost of overclaiming is typically higher than the cost of underclaiming. Overclaiming produces false assurance, mis-specified procurement requirements, and deployment architectures that treat a competence optimisation regime as if it were a comprehension regime. Under CIITR, the difference is operationally material because comprehension is not a rhetorical label, it is a classified state requiring demonstrable $\Phi_i$and $R_g$conditions, and it is a governance object with thermodynamic implications through CPJ. Conversely, underclaiming may delay adoption of beneficial reliability practices, such as process verification and reward topology discipline, that the paper's evidence robustly supports within scope. Part V is therefore structured as a controlled translation layer. It preserves the paper's actionable engineering lessons, while enforcing CIITR boundaries that prevent competence improvements from being misrepresented as comprehension gains.

Part V adopts a policy-first posture toward the training pipeline. The paper's results, as adjudicated, establish that curriculum placement, mid-training conditioning, reinforcement allocation, and process reward design operate as first-class governance surfaces, not as implementation details. This has immediate operational consequences. It implies that institutions should treat training phase definitions, data regime placement, reward schemas, and evaluation constraints as regulated artefacts subject to version control, change approval, and audit trails, analogous to other high-impact configuration elements in safety-critical or security-critical systems. It also implies that "model capability" should not be treated as a single scalar outcome of scale and compute, but as a managed consequence of intervention sequencing and constraint design, with clear traceability to evidence.

At the same time, Part V explicitly treats comprehension adjudication as a separate governance track. The paper does not instrument $R_g$and does not provide joule-level energy accounting for CPJ, therefore it cannot be used as an evidentiary basis for claims about autonomous continuity, human-like reasoning, or thermodynamic epistemic efficiency. Part V therefore defines an institutional separation of concerns: one track governs competence and reliability within verifier-supported domains through curriculum and reward policy, and a second track governs comprehension classification and efficiency claims through CIITR-required instrumentation and re-entry protocols. This separation reduces the probability that improvements in benchmark-defined "reasoning" are converted into policy assertions about understanding, autonomy, or substrate independence.

Finally, Part V operationalises recommendations using CIITR's claims discipline. Recommendations are expressed as "should" statements where the evidence base supports a mandatory governance posture, and as "may be advantageous" statements where the evidence

is suggestive but requires contextual tailoring. Each recommendation is linked to the specific class of evidence established earlier, performance, mechanism, or resource governance, and each is bounded by explicit non-admissible escalations. The intended output is decision-usable. It provides a basis for institutional action that is consistent with the paper's demonstrated contributions, while remaining compliant with CIITR's requirements for comprehension and efficiency adjudication.

13. Curriculum as a control surface, implications for institutional deployment

The paper's empirically supported interplay among pre-training, mid-training, reinforcement learning, and process-level constraints implies a governance-relevant proposition: model behaviour is not a stable, intrinsic "capability" that can be assumed to persist under minor pipeline variation. It is materially shaped by the training and alignment curriculum, by reward topology, and by evaluation constraints that define what counts as acceptable performance within a given operational universe. In administrative terms, this shifts the institutional object of control from "the model" in isolation to the end-to-end pipeline as a regulated configuration, where phase placement, distributional conditioning, and verifier design jointly determine observable behaviour.

This shift has immediate implications for institutional deployment in security-relevant, safety-relevant, and mission-critical contexts.

i. First, it invalidates governance postures that treat training and alignment steps as informal research choices that can be iterated without traceability. The paper's findings indicate that seemingly modest changes, such as moving mid-training distributions, altering RL data difficulty placement relative to the competence boundary, or modifying process-aware reward definitions, can produce qualitative differences in generalisation outcomes and failure modes under the same model family and nominal budget framing.

ii. Second, the paper's reliance on process verification demonstrates that the evaluation regime itself functions as a behavioural constraint mechanism. It reduces shortcut policy success within the constructed domain by making intermediate structural validity constitutive for correctness. This is a reliability gain, but it also implies that institutions must treat evaluation constraints as part of the operational safety case, not merely as a reporting convenience.

From a CIITR perspective, these operational consequences must be communicated with claims discipline. The paper provides strong evidence that curriculum and reward policy govern the exploitation and selection of relational structure within a verifier-supported task universe, which is most plausibly classified as governance of $\Phi_i$utilisation rather than evidence of comprehension state. The institution should therefore adopt pipeline governance controls that preserve the engineering value of these findings while preventing their

escalation into non-admissible claims about "understanding," "autonomy," or thermodynamic efficiency, unless the additional instrumentation specified earlier is implemented.

Accordingly, the following recommendations are stated in institutional control language and should be treated as governance requirements where the organisation relies on model outputs for consequential decisions:

- Organisations should treat training and alignment pipelines as regulated artefacts with version control and audit trails. This should include explicit configuration baselines for data distributions per phase, reward definitions, verifier and parser versions, sampling policies, and compute or budget envelopes, with change control that supports reconstruction of deployed behaviour and post-incident traceability.
- Evaluation suites should include process constraints where shortcut risk is high. Where domains permit verifier construction, process verification should be integrated as a safety and integrity control that reduces the acceptance of endpoint-correct but structurally invalid traces, and that increases diagnostic resolution for failure analysis. The selection of process constraints should be risk-based and documented as part of the assurance case, including explicit statements of what the verifier certifies and what it cannot certify.
- Claims presented to leadership should be tagged by claim type and admissibility class. Performance improvements should be reported separately from mechanistic hypotheses, and both should be separated from comprehension and efficiency assertions. Where evidence derives from a verifier-supported synthetic or partially synthetic ontology, the scope boundary should be explicit, and any escalation beyond that scope should be marked as non-admissible absent additional CIITR-aligned instrumentation for Φ_i, R_g, and joule-level energy measurement.

Implementation-wise, it may be advantageous to operationalise these recommendations through a lightweight but enforceable governance pattern: a pipeline configuration register (versioned artefacts, phase definitions, reward schemas, evaluation constraints), a claims register (each externally communicated statement tagged to a claim class with its evidence basis), and an assurance gate in which changes to curriculum or reward topology are treated as behaviour-impacting changes requiring re-evaluation against a pre-declared risk profile. This preserves the paper's central operational lesson, that curriculum is a controllable surface, while aligning institutional practice with CIITR's boundary conditions on what the evidence can legitimately support.

14. Risk register, structural illusion risks and evaluation gaming

The paper's results, as adjudicated under CIITR, imply that a nontrivial portion of observed "reasoning" performance is governed by curriculum placement, reward topology, and evaluation constraints, rather than by a stable, context-invariant capability attribute. This governance surface creates a corresponding risk surface. In institutional deployment, the dominant hazard is not that models fail trivially, but that they fail while appearing compliant, including through structurally valid traces in verifier-supported domains that do not

generalise as integrated competence outside those domains. The register below formalises the most material structural illusion risks and evaluation gaming risks, and specifies operational mitigations that are compatible with CIITR's claims discipline and measurement requirements.

Risk register

Risk ID	Risk statement	Primary drivers and enabling conditions	Operational indicators	Likely impact domain	Mitigation controls, operational	Verification and governance checkpoints
R-01	Overfitting to process constraints without general integration	Strong reliance on a known generator and deterministic process verifier, training and RL optimised to satisfy verifier-defined intermediate steps	High gains under process-verified pass@k with steep degradation under template shifts that change surface cues, high sensitivity to minor verifier rule changes	Integrity of decision support, model assurance, false confidence	Evaluation suites should include cross-context binding tests and perturbation-resilience tests that explicitly attempt to break superficial compliance. Organisations should treat verifier and parser versions as regulated artefacts with change control, including regression baselines	Pre-deployment gate should require performance plus binding stability reporting. Any verifier changes should trigger re-certification with frozen reference suites
R-02	Reward topology induces brittle procedural compliance, brittle under distribution shift	Process-level reward conditioning and outcome gating can select narrow procedural strategies that satisfy the verifier but fail under novel renderings or slight rule deviations	Sharp performance cliffs on minor problem variations, increased variance across random seeds, sudden failure under adversarial paraphrase	Reliability and availability, operational fragility	Reward design should be stress-tested under controlled perturbation regimes and multiple template families, and should include robustness terms where appropriate. It may be advantageous to maintain multiple reward configurations and measure stability as a first-class KPI	RL training artefacts should be logged, including reward components and gating conditions. Periodic robustness audits should be executed as part of model lifecycle governance
R-03	Misinterpretation of "reasoning" as "comprehension" in leadership reporting	Conflation of process-verified correctness with CIITR comprehension, absence of explicit $\Phi_i$and $R_g$instrumentation, narrative escalation in internal communications	Briefings that translate pass@k improvements into "understanding," "deliberation," or "human-like reasoning" language, without measurement qualifiers	Strategic risk, procurement misalignment, misplaced assurance	Claims presented to leadership should be tagged by admissibility class, performance, mechanism, comprehension-proxy, efficiency-proxy, with mandatory limitation clauses. Communications should be subject to an internal claims review function	Governance checkpoint should require a claims register appended to executive summaries, including explicit non-admissible escalation statements and required extensions

R-04	Procurement or policy decisions based on scale narratives inconsistent with curriculum sensitivity	Overreliance on parameter-count and scale narratives, underweighting evidence that curriculum placement and mid-training shape outcomes materially	Procurement requirements specify scale thresholds while ignoring pipeline governance requirements, reward topology discipline, or evaluator design	Financial exposure, vendor lock-in, suboptimal capability-per-cost	Procurement specifications should require pipeline transparency, phase definitions, reward schemas, evaluation constraints, and audit logs, not only model size. It may be advantageous to include performance under fixed curriculum governance as a contractual deliverable	Contractual governance should include deliverables for dataset provenance, training phase ledger, reward definitions, and evaluation methodology, with independent verification rights
R-05	"Edge of competence" tuning leads to fragile capability boundaries and miscalibrated deployment	RL tuned at the competence boundary can inflate performance on boundary-like tasks while leaving gaps outside that regime, especially when task mix shifts in production	High success on benchmark-like queries but poor resilience to novel difficulty distributions, increased reliance on sampling or escalation to high-k	Operational reliability, service-level compliance	Deployment should include difficulty-distribution monitoring and periodic re-evaluation on rebalanced task mixes. Escalation policies should constrain high-k sampling as a default operating mode where energy or latency budgets are strict	Operational dashboards should track outcome quality versus sampling intensity. Governance should require quarterly recalibration against representative production-like distributions
R-06	Evaluation gaming through verifier exploitation and parser brittleness	Deterministic parsers and verifiers create attack surfaces, models or attackers can generate outputs that exploit parsing quirks or verifier blind spots	Unusually high scores accompanied by anomalous trace structure, sensitivity to minor parser changes, frequent borderline-valid traces	Integrity, security, audit failure	Verifiers should be adversarially tested, including fuzzing of parser inputs. A dual-verifier approach may be advantageous, using independent implementations and requiring agreement. Human review sampling should be introduced for anomalies	Governance should require periodic red-team verification exercises and a parser change approval board with regression evidence
R-07	Hidden resource escalation and declining epistemic efficiency despite score gains	Higher pass@k achieved via longer traces, more sampling, heavier verification, or more expensive RL rollouts, compute normalisation masks joule cost, CPJ not measured	Latency increases, power draw increases, high-k reliance becomes operational default, cost per verified-correct case rises	Cost, sustainability, operational capacity planning	Institutions should require joule-level energy instrumentation for phases and inference modes when efficiency claims are made. It may be advantageous to define an operational efficiency KPI, verified-correct-per-joule under fixed service targets	A measurement gate should require phase-resolved energy accounting and explicit separation of compute proxies from energy metrics in all decision papers

R-08	Loss of traceability across training phases and alignment changes	Informal changes to mid-training data, reward weights, verifier rules, or sampling policies without artifact governance, undermines reproducibility and incident response	Inability to reproduce behaviour across versions, unclear attribution for regressions, fragmented documentation	Governance, accountability, legal defensibility	Training and alignment pipelines should be treated as regulated artefacts, with version control, immutable logs, and documented approvals. Change control should include defined rollback procedures	Governance checkpoint should require a pipeline ledger that links each deployment version to data, rewards, evaluation constraints, and observed behavioural deltas
R-09	Domain translation failure, assuming generator-world reliability transfers to open epistemic contexts	Success in a known-generator domain induces overconfidence, real domains lack canonical ground truth and deterministic verification	Performance collapse in real tasks with ambiguous objectives, high false-positive confidence, brittle reasoning narratives	Decision quality, reputational and mission risk	Organisations should maintain a domain transfer evaluation tier that explicitly includes tasks without deterministic verifiers, and should forbid escalation of generator-world conclusions without bridging evidence	Review boards should require explicit scope statements distinguishing verifier-supported and verifier-absent regimes
R-10	Composite-system attribution confusion, treating orchestration continuity as model continuity	External memory, retrieval, or orchestrator state persistence can simulate continuity, risk of attributing R_g to the model rather than the composite system	Claims of persistence or "memory" not reproducible under model-only evaluation, inconsistent behaviour across orchestrators	Assurance and accountability	Reporting should separate model-only and composite-system profiles, and should require attribution statements for any continuity claims. Re-entry protocols should be executed under both profiles	Governance should require that any continuity claims include evidence of re-entry under controlled interruption without full state reinjection

Implementation posture for mitigations

The mitigations above are operational by design and should be integrated into an institutional model lifecycle as enforceable controls rather than as optional quality initiatives.

First, organisations should establish evidence tiering as a formal reporting discipline. Performance evidence under a process verifier should be treated as a distinct tier from comprehension-proxy evidence, and both should be distinct from thermodynamic efficiency evidence. This is the simplest and most effective safeguard against structural illusion escalation, because it forces the organisation to separate "what the verifier certifies" from "what the system is claimed to be."

Second, audit logs and version control should be applied to the entire training and alignment pipeline, not solely to the deployed model artifact. The paper's own decomposition into pre-training, mid-training, and RL phases implies that behaviour is a function of phase sequencing and distributional placement. If those elements are not logged and governed, the institution cannot reliably attribute changes, cannot conduct credible incident reviews, and cannot maintain defensible assurance cases.

Third, CIITR-aligned measurement should be treated as a gated extension track, not as an interpretive layer. Where the institution intends to make escalated claims about comprehension or efficiency, it should implement the extension programme defined earlier, including cross-context binding and perturbation resilience for Φ_i-proxies, re-entry and delayed recurrence protocols for R_g-proxies, and joule-level phase accounting for CPJ. Without those additions, the governance posture should explicitly prohibit claims that exceed performance and mechanism tiers.

15. Recommendations to the research programme

This section issues targeted recommendations to the research programme, expressed as implementable reporting and design requirements rather than as broad methodological preferences. The intent is to preserve the paper's central contribution, namely a controlled decomposition of training phases and a verifier-supported reasoning universe that enables unusually clean causal attribution, while reducing the probability that downstream discourse converts those results into non-admissible claims about comprehension, autonomy, or thermodynamic efficiency. The recommendations therefore focus on claim governance, benchmark governance, and operational definition discipline, each of which is directly implicated by the paper's findings on curriculum sensitivity, mid-training conditioning, RL boundary placement, and process-level constraint design.

Researchers should publish a CIITR-compatible claims table that separates performance, mechanism, comprehension, and efficiency claims. The table should be treated as a formal appendix, not as an informal discussion element, and it should include, for each major conclusion, a claim class tag, the measurement basis, the scope boundary implied by the ontology, and an explicit admissibility status under CIITR. Where the programme uses proxies, for example compute-normalised budgets or process verification metrics, the table should include a mandatory limitation clause stating what those proxies cannot substitute for, specifically that compute normalisation is not joule-level energy measurement and that process-verified correctness is not, by itself, a comprehension measure. This single artefact would materially improve auditability and reduce rhetorical drift, because it forces authors, and later readers, to keep performance deltas distinct from ontological assertions.

Benchmark designers should avoid single-metric sufficiency, and can to advantage adopt multi-axis adjudication regimes. The paper's own methodology demonstrates that process verification and phase-aware training interventions expose distinct dimensions of behaviour that are not captured by endpoint-only correctness. The general design principle is that a single scalar metric is rarely sufficient for adjudicating systems that can succeed through multiple strategy classes, including shortcut policies, parser exploitation, and sampling-based tail success. A multi-axis regime should, at minimum, separate endpoint correctness, process validity, robustness under controlled perturbation, and sensitivity to contextual re-rendering. In verifier-supported domains, designers can to advantage include structural invariance measures that test whether the same latent object is preserved across contextual variations, thereby distinguishing procedural compliance from integration-like stability within the benchmark's own ontology. Where verifiers are not possible, designers should explicitly

declare that the benchmark cannot certify process validity and should prohibit escalated claims that rely on such certification.

Where “edge of competence” is invoked, researchers should define it operationally and report sensitivity analyses. The paper’s central claim that RL is productive at the competence boundary is actionable precisely because it implies a curriculum placement rule. However, the term “edge of competence” can degrade into a rhetorical label unless it is specified in measurable terms. Researchers should therefore define the edge as a region characterised by pre-registered criteria, for example baseline pass@1 in a defined band, nontrivial pass@k lift at higher k, and measurable near-miss density under a process verifier. They should then report sensitivity analyses showing how RL outcomes change when the RL data distribution is shifted relative to this region, and they should explicitly report failure modes when the region is misidentified, including stagnation under too-hard regimes and saturation under too-easy regimes. This would convert a qualitative narrative into a reproducible intervention parameter.

Beyond these three targeted recommendations, two additional operational practices should be adopted to strengthen the programme’s decision usability without changing its core scientific design. First, researchers should publish phase-resolved artefact ledgers, including generator settings, template families, parser and verifier versions, reward definitions, and budget allocations, because the paper’s own results show that behaviour is a function of these controlled surfaces, not merely of model size or nominal training time. Second, where the programme intends to speak about “efficiency,” it should include a dedicated measurement section that states explicitly whether energy is measured in joules, what boundary is used, and what yield definition is used. If energy is not measured, the work should restrict itself to compute-proxy language and state explicitly that CPJ is out of scope.

Collectively, these recommendations are designed to institutionalise a disciplined separation between competence evidence and comprehension claims, while preserving the paper’s core insight that curriculum and reward topology are tractable, high-leverage control surfaces. Implemented as standard practice, they would enable the research programme to be used more safely in policy and procurement contexts, and they would provide a direct pathway for future work to graduate from performance-focused results to CIITR-relevant comprehension and efficiency adjudication.

Appendices, designed for auditability and reuse

Appendix A. CIITR claims admissibility table

This appendix operationalises the claims discipline applied throughout the note by mapping the paper's principal propositions to (i) the evidence type actually instantiated in the study's controlled ontology, (ii) the CIITR claim class that the proposition belongs to, (iii) its admissibility status under CIITR given the present instrumentation, and (iv) the minimum additional measurements required if the proposition is to be escalated into comprehension or thermodynamic efficiency territory. The table is intended to be decision-usable: it supports both research interpretation and institutional briefing by making explicit which claims can be carried forward, and which must be held as non-admissible absent further protocol commitments.

Major paper claim (paraphrased, bounded to the paper's ontology)	Evidence type in the paper	Ciitr claim class	Admissibility status under ciitr	Required additional measurements for escalation
The task world is a controlled, generator-defined reasoning universe using dags where difficulty is parameterised by operation count and contexts are template renderings, supporting a two-axis generalisation analysis (depth and breadth).	Formal construction of synthetic ontology, controlled dataset generation	Method and ontology statement (pre-adjudicative)	Admissible as a scope-defining description, it constrains downstream generalisation interpretations	None for the description itself. For escalation to open-world claims, add bridging evaluations in domains without known generators and without deterministic verifiers
Depth generalisation is operationalised as maintaining strict process-verified correctness as op(g) exceeds training ranges, breadth generalisation as stability under template shifts with fixed latent structure.	Explicit metric definition and controlled OOD construction	Performance definition and evaluation design	Admissible as a benchmark definition, not admissible as an ontological definition of reasoning in general	For escalation to comprehension claims, add $\Phi_i$binding and perturbation resilience tests, and $R_g$re-entry protocols that cannot be reduced to single-episode scoring
Evaluation is process verified, models are parsed into predicted graphs, correctness requires both intermediate step validity and final answer match, pass@k is reported under this strict criterion.	Verifier-based evaluation protocol with step-level checks	Reliability and performance instrumentation	Admissible as strong performance evidence within verifier-supported domains	For escalation to comprehension claims, add cross-context integration instrumentation beyond verifier compliance, plus re-entry and recurrence protocols that test continuity across interruption and delay
Rl yields meaningful capability gains only when two conditions hold, there is headroom beyond pre-training coverage and rl data are calibrated near the model's edge of competence, neither in-distribution nor too far ood.	Controlled intervention study, varying RL data regimes and difficulty bands, process-verified pass@k outcomes	Mechanistic claim about curriculum placement and gradient accessibility	Admissible as a mechanistic claim within the constructed ontology, with the qualifier that it governs competence expression and selection	For escalation to comprehension, add $R_g$instrumentation to show re-entry and rhythmic continuity, not merely improved selection of valid traces. For efficiency escalation, add joule-level accounting of RL rollouts and verification overhead

“Edge of competence” can be operationally characterised by tasks where pass@1 is low but non-zero pass@k exists, enabling rl to amplify near-miss behaviours, and a self-paced curriculum emerges as the model improves.	Heuristic operational definition coupled to empirical curves and guidance	Mechanistic plus operational guidance	Admissible as an operational heuristic and mechanistic interpretation, not admissible as a general law of reasoning	For escalation to broader governance claims, require sensitivity analyses across multiple model sizes, verifier designs, and domain ontologies, plus reporting of boundary instability under distribution shifts
Rl cannot induce contextual transfer without minimal pre-training exposure to the long-tailed context, sparse “seed” coverage can be sufficient for rl to reinforce and yield robust cross-context generalisation.	Controlled pre-training exposure manipulation followed by RL, comparative transfer outcomes	Mechanistic claim about representational availability and prerequisite primitives	Admissible as a mechanistic claim within the paper’s context-template ontology, consistent with Φ_i-exploitation framing	For escalation to comprehension, add tests for cross-context binding that persist under paraphrase and distractors, and demonstrate invariance of relational object under transformations, as Φ_i-proxy. For escalation to R_g, add delayed recurrence protocols
Mid-training functions as a distributional bridge between pre-training and rl tasks, it installs and stabilises higher-level reasoning priors, improving rl readiness and altering outcomes under a fixed proxy compute budget.	Phase-decomposed intervention study with proxy compute normalisation, comparative performance across allocation ratios	Mechanistic claim about pipeline governance and prior conditioning	Admissible as a mechanistic and governance-surface claim, bounded to compute-proxy budgeting and the synthetic ontology	For escalation to CPJ, replace compute-proxy normalisation with joule-level energy measurement per phase. For escalation to comprehension, add $\Phi_i$and $R_g$instrumentation to show integration and re-entry changes across phases
Under fixed proxy compute budget, allocating more budget to mid-training versus heavier rl yields qualitatively different performance, light rl plus substantial mid-training best improves near-boundary tasks, heavy rl improves harder ood tasks, implying task-aware allocation policies.	Controlled allocation sweep using token-equivalent cost approximation for RL and mid-training tokens	Performance plus mechanistic allocation guidance	Admissible as performance and allocation guidance under compute proxies, not admissible as energy efficiency evidence	For escalation to efficiency claims, implement phase-resolved joule accounting, including rollout energy, parsing and verification energy, and sampling energy implied by pass@k. For escalation to CPJ, define CIITR-aligned comprehension yield
Process-aware rewards, mixing outcome reward with process verification reward or gating outcome reward on perfect process, mitigate reward hacking and improve strict pass@k, including in harder extrapolative regimes.	Reward design ablations with verifier-derived dense process signal, error pattern analysis	Reliability instrument claim and mechanistic incentive alignment claim	Admissible as reliability and mechanism evidence within verifier-supported domains, it reduces shortcut policy success relative to the verifier’s structural semantics	For escalation to comprehension, show that improvements persist under cross-context binding and perturbation regimes where the verifier does not simply encode the generator, and add $R_g$tests. For escalation to efficiency, measure energy cost of verifier-coupled reward pipelines
Process verification reshapes the error distribution, reducing structural errors such as dependency mismatch and missing nodes, indicating that aligning rewards with valid traces changes what failure modes remain.	Error taxonomy analysis under different reward mixtures	Mechanistic plus reliability characterisation	Admissible as mechanistic evidence about constraint-induced behaviour change, within the paper’s parser and verifier implementation	For escalation beyond the ontology, require independent verifier implementations, adversarial parser testing, and evaluation under alternative ontologies to rule out implementation-specific artefacts

Compute normalisation via token-equivalent cost approximations enables "fair comparison" of phases, but it remains a compute-proxy budget model, not an energy model.	Proxy compute accounting and FLOPs-linked approximations	Efficiency-proxy statement (engineering budgeting)	Admissible as compute-proxy budgeting, explicitly non-admissible as CPJ	For escalation to CPJ, implement joule-level measurement boundary definition, calibrate sampling error, report joules by phase and by inference mode, and define comprehension yield as $C_s$proxy linked to $\Phi_i$and $R_g$instrumentation
The paper's results reconcile apparently conflicting conclusions in related work regarding whether rl can extend reasoning beyond the base model, by locating each result in different regions of the difficulty and coverage spectrum.	Comparative interpretation supported by controlled regime partitioning	Interpretive synthesis, partially mechanistic	Admissible as an interpretive synthesis within the paper's controlled regime model	For escalation to broader claims about "reasoning in general," require replication across naturalistic domains, heterogeneous model scales, and evaluation settings without known generators

Notes for institutional use

1. **Admissible** in this table means admissible **under CIITR** given the paper's actual instrumentation and controlled ontology, not "true in all contexts." The admissibility boundary is the governance control point.
2. **Mechanistic admissibility** remains bounded. Even when ablations are well-formed, the mechanism is a mechanism **of behaviour under a verifier-supported ontology**, not a mechanism of comprehension unless $\Phi_i$and $R_g$are operationalised.
3. **Efficiency language should be constrained** to compute-proxy statements unless joule-level energy is measured and linked to an explicitly CIITR-aligned yield definition.

Appendix B. Variable and metric crosswalk

This appendix provides a controlled crosswalk between the paper's formal variables, operational metrics, and training-phase parameters, and the corresponding CIITR constructs and governance classifications used in this note. The crosswalk is intentionally conservative. It treats the paper's symbols as elements of an engineered, generator-governed task universe, and it treats CIITR constructs as adjudicative categories that require dedicated instrumentation. The result is an explicit mapping of what can be treated as a proxy surface, what can be treated as a behavioural indicator, and what is non-equivalent by definition.

B.1 Crosswalk table, paper variables to CIITR constructs

Paper variable or metric	**Paper meaning, as defined operationally**	**Closest CIITR construct or governance category**	**Explicit non-equivalence statement under CIITR**	**CIITR-relevant use of the paper variable, bounded**
Dependency graph $G = (V, E)$	Ground-truth directed dependency graph, nodes are variables, edges are dependencies, graph terminates in a designated answer node	External relational object, an observational scaffold for integration testing	G is externally supplied ground-truth structure, it is not the system's integrated relational information Φ_i. Behavioural matching of G does not establish Φ_i as a persistent internal state property	Use G to construct controlled binding and perturbation protocols, and to define behavioural integration proxies, without escalating to comprehension classification
Operation count $op(G) = \mid E \mid$	Difficulty control parameter, complexity quantified as the number of dependency edges, used to define depth scaling regimes	Task-difficulty axis for competence-boundary placement	$op(G) = \mid E \mid$ is not rhythmic continuity R_g. It is a generator-defined structural depth proxy, not a temporal continuity or re-entry measure	Use $op(G) = \mid E \mid$ to define controlled competence bands and to stage curricula, while treating any inference about R_g as non-admissible
Template τ	Contextual template used to render the same latent structure into different surface domains and narratives	Surface-form variation axis, potential proxy surface for cross-context binding	Variation in τ does not itself measure Φ_i. Template robustness can reflect resilient pattern matching rather than integrated relational binding	Use τ families to build cross-context binding tests and invariance checks, explicitly labelled as Φ_i-proxy surfaces only
Rendering function $\Phi(G, \tau)$	Generator mapping from latent graph plus template to a natural-language instance, including choices about explicit versus implicit dependencies	External observational-surface constructor, boundary between latent structure and observable text	The paper's $\Phi(G, \tau)$ is not CIITR's Φ_i. The former is a data generator mapping, the latter is an internal integration construct	Use $\Phi(G, \tau)$ to audit how surface choices modulate shortcut risk, verifier exploitability, and robustness under re-rendering
Extrapolative depth generalisation	Maintaining strict process-verified correctness as $op(G)$ exceeds training ranges	Performance generalisation axis	Depth generalisation is not comprehension, and it does not establish R_g. It is compatible with episode-bound distributional optimisation	Use as admissible performance evidence, then require re-entry and interruption protocols to speak about R_g
Contextual breadth generalisation	Transfer across novel templates with fixed underlying structure	Performance and robustness axis, potential proxy surface for Φ_i	Breadth stability is not equivalent to integration across contexts in the CIITR sense, absent binding and perturbation protocols	Use as Φ_i-proxy evidence only when combined with invariance, distractor-interference, and relational-perturbation resilience tests
Process-verified evaluation, $\hat{G}$, $\hat{a}$	Model output is parsed into a predicted graph $\hat{G}$ and predicted answer $\hat{a}$, correctness requires both	Reliability instrumentation, anti-forgery constraint	Verifier compliance is not internal integration, and it is not R_g. It certifies external	Use as an integrity control against shortcut policies in verifier-supported domains,

Paper variable or metric	Paper meaning, as defined operationally	Closest CIITR construct or governance category	Explicit non-equivalence statement under CIITR	CIITR-relevant use of the paper variable, bounded
	intermediate validity and correct final answer	within the engineered ontology	compliance relative to a known generator	with explicit scope boundaries
Step-level process accuracy	Average correctness across gold nodes or steps under the verifier	Trace-fidelity indicator inside the task universe	Step-level correctness is not $\Phi_i$and not R_g. It can be increased through reward shaping without establishing comprehension	Use to quantify how constraints reshape failure modes and suppress domain-specific epistemic forgery, do not use as cognition evidence
pass@k under strict criterion	Probability that at least one of $k$sampled outputs satisfies strict process plus endpoint correctness	Competence statistic, accessibility of correct traces	pass@k is not C_s. It remains compatible with $R_g \approx 0$, even when pass@k rises	Use to diagnose probability-mass allocation effects and to quantify operational sampling dependence, including energy and latency implications
"Edge of competence"	Operational regime where baseline success is nontrivial but incomplete, enabling RL to amplify near-miss candidates	Curriculum placement control variable, gradient-accessibility region	The "edge" is not a comprehension threshold. It is a performance-defined region under the verifier	Use to implement RL data placement policies and sensitivity analyses, do not treat edge effects as evidence of understanding
Training phase decomposition, $f_\theta^{pre}, f_\theta^{mid}, f_\theta^{post}$	Model snapshots after pre-training, mid-training, and post-training (RL), associated with distinct data regimes	Intervention chain and artefact governance surface	Phase progression does not imply increasing $\Phi_i$or R_g. It remains compatible with distributional optimisation	Use as a governance ledger for reproducibility, traceability, and phase-resolved measurement additions
Compute budget T, RL allocation ratio $\beta, T_{mid} = (1-\beta)T, T_{RL} = \beta T$	Proxy budget split between mid-training and RL to compare allocations	Resource-governance proxy axis	Compute proxy budgeting is not joule-level energy measurement, therefore it cannot substitute for CPJ denominators	Use for within-proxy allocation comparisons, require joule-level accounting before any efficiency claims
Token-equivalent RL cost approximation, FLOPs model	RL cost expressed via computational equivalence or FLOPs-based approximations	Efficiency-proxy accounting	FLOPs and token equivalents are not joules, they do not capture hardware dependence, memory movement, and system overhead	Use as a transparency mechanism to identify where energy measurement must be inserted, not as an energy claim
Reward components, R_{out}, R_{pv}, mixing coefficient α, strict gating	Composite reward shaping that mixes outcome and process verification, including schemes where outcome reward is gated on perfect process	Reliability instrument, incentive-topology governance	Process-aware rewards enforce trace compliance, they do not establish comprehension, and they do not instrument $\Phi_i$or R_g	Use to suppress opportunistic policies and reduce verifier-relative epistemic forgery within the task universe
"Reward hacking"	Failure mode where optimisation exploits scoring loopholes, mitigated by process-aware signals	Governance hazard, epistemic-forgery analogue within constrained domains	Mitigating reward hacking improves reliability relative to the evaluator, it does not establish integration	Use as rationale for process constraints and adversarial verifier testing, while prohibiting cognition escalation

B.2 Non-equivalence ledger, high-risk conflations to be explicitly prohibited

The crosswalk above implies a short non-equivalence ledger that should be treated as a standing interpretive constraint when the paper is used in institutional reporting or procurement narratives:

1. The paper's rendering function $\Phi(G, \tau)$is not CIITR Φ_i. The former is a generator mapping from latent structure to text, the latter is an internal integration construct that must be instrumented as a system property.
2. Strict process-verified pass@k is not comprehension C_s. It is a competence statistic under a verifier-defined correctness criterion, and remains compatible with $R_g \approx 0$.
3. Operation depth $op(G)$is not rhythmic continuity R_g. It parameterises task complexity, not temporal recurrence or re-entry capacity.
4. Compute normalisation and FLOPs-based equivalence are not CPJ. They are accounting devices for proxy budgets, they do not measure physical energy and do not define comprehension yield.
5. Process-aware rewards and verifier compliance mitigate reward hacking, they do not establish integrated comprehension. They enforce externally defined trace fidelity relative to a known generator.

B.3 Practical implication of the crosswalk

The crosswalk is intended to make the paper more usable, not less. It clarifies that the paper offers unusually strong evidence about curriculum governance, reward topology governance, and verifier-mediated reliability within a generator-defined reasoning universe. It also clarifies that, under CIITR, escalation from that evidence base to comprehension adjudication requires additional instrumentation, specifically cross-context binding and perturbation-resilience protocols for Φ_i-proxies, re-entry and delayed recurrence protocols for R_g-proxies, and joule-level energy measurement linked to a CIITR-aligned yield definition for CPJ.

Appendix C. Proposed re-entry test suite for R_g

This appendix specifies a re-entry and recurrence test suite designed to instrument R_g as rhythmic continuity and re-entry capacity over time, rather than as trace length, chain-of-thought volume, or within-episode token coherence. The suite is constructed to be implementation-ready. It defines protocol classes, perturbation types, recurrence windows, success and failure criteria, and reporting requirements that permit audit-grade attribution. The core design requirement is non-reducibility to static pattern matching: the test must force state reactivation under controlled interruption such that success cannot be achieved by merely re-reading the full prior trace, re-solving from scratch using complete information available at the re-entry stage, or exploiting superficial lexical similarity.

C.1 Purpose, scope, and admissibility posture

The suite has three purposes.

First, to distinguish **episode-bound competence** from **cross-episode continuity**. A system can satisfy strict process verification within one uninterrupted episode while remaining non re-entrant when the episode is fractured.

Second, to provide an **attribution boundary** between model-intrinsic continuity and continuity externalised through orchestration. Re-entry may be achieved by injecting stored state, retrieval augmentation, or a tool chain, and this must be reported as composite-system continuity rather than intrinsic R_g.

Third, to yield a **graded R_g-proxy** as a function of recurrence window length and perturbation severity, enabling governance comparisons across training regimes, model families, and deployment configurations.

Admissibility under CIITR is strictly conditional. Results from this suite are admissible as R_g-proxy evidence only when the test is executed under a declared profile that controls state persistence, and when leakage channels are constrained.

C.2 Definitions and controlled artefacts

The suite requires explicit definitions of what constitutes the “state” that must persist across interruption, and what constitutes a permissible cue at re-entry.

C.2.1 Latent relational object

A task instance must have a latent relational object $S^{\setminus *}$ that is fully known to the evaluator, and only partially observed by the system at each stage. In the paper’s ontology this can be a DAG with values, but the suite is not limited to DAGs. The defining property is that correctness can be assessed by checking whether the system reconstructs and applies $S^{\setminus *}$.

C.2.2 System state representation

The suite defines an evaluated internalised state proxy S_t as the structured object that the model would need to carry forward to avoid recomputation. For DAG-style tasks, S_t can be operationalised as:

- A dependency structure hypothesis (graph skeleton and directed edges)
- A set of resolved intermediate values
- A set of unresolved subgoals with dependency constraints

C.2.3 Re-entry cue

A re-entry cue Q_{re} is the minimal prompt provided at Stage 2. It may reference the task identity and a partial boundary condition, but it must not contain enough information to solve the task from scratch. The cue must be constructed so that successful completion requires reactivating S_t rather than recomputing S^{*} from complete information.

C.2.4 Interruption

An interruption is an enforced discontinuity between Stage 1 and Stage 2 consisting of:

- Removal of the full prior chain (unless explicitly permitted in a control condition)
- An intervening distractor workload that consumes context capacity
- A delay window defined either in tokens, in wall-clock time, or both

C.3 Test suite structure and governance profile requirements

The suite is organised into three protocol classes, each targeting a distinct continuity competency.

1. **Reconstruction re-entry**: state must be reconstructed from partial cues after interruption.
2. **Recurrence reactivation**: state must be reactivated and reused in a later task where it is required but not restated.
3. **Rhythmic continuity under interference**: state must remain stable under misleading near-neighbour stimuli and controlled corruption.

Each protocol class must be executed under two governance profiles:

- **Profile M0, model-only**: no external memory, no retrieval, no tool-based state persistence, no hidden system prompt state carryover beyond the explicit context window.
- **Profile M1, composite-system**: external memory and orchestration allowed, but must be declared, logged, and measured separately.

Only Profile M0 can support claims about intrinsic R_g. Profile M1 can support claims about system-level continuity under governance controls.

C.4 Protocol class 1, reconstruction re-entry

C.4.1 Stage design

- Stage 1: system solves Part A of the instance, producing an explicit structured state output S_{out} in a standard schema.

- Interruption: the full Stage 1 transcript is withheld from Stage 2 in M0. A distractor task is inserted.
- Stage 2: system receives Q_{re} plus a bounded fragment of its own earlier S_{out}, with controlled deletion or corruption depending on the perturbation type. The system must complete Part B.

C.4.2 Required outputs

- Reconstructed state S'_{out}
- Completion trace for Part B
- A minimal consistency justification in a fixed format, limited-length, to support auditing of failure modes without inviting narrative inflation

C.4.3 Core success criterion

Success requires both:

- Structural validity: S'_{out} matches S^{*} to a declared tolerance
- Task validity: final answers and intermediate constraints for Part B are satisfied

C.5 Protocol class 2, delayed recurrence reactivation

C.5.1 Paired task structure

- Task A establishes a latent object S^{*} and requires the system to solve a partial objective.
- Task B occurs after a recurrence window and context shift, and requires reuse of the same S^{*} to solve a different objective that cannot be solved from Task B alone.

C.5.2 Non-reducibility requirement

Task B must be constructed so that the missing relational object is not reconstructable from its own text without reusing what was established in Task A. If Task B is solvable independently, the protocol is invalid for R_g adjudication.

C.6 Protocol class 3, rhythmic continuity under interference

This protocol targets the distinction between continuity and mere cue-following. It is designed to detect cases where the system "re-enters" by attaching to the most recent surface cues rather than preserving the latent relational object.

C.6.1 Interference injection

Between Stage 1 and Stage 2, insert one of the following:

- A near-neighbour task with high lexical overlap but a different latent object
- A contradictory narrative that tempts state overwrite
- A partial-state corruption that forces error correction rather than blind continuation

C.6.2 Continuity requirement

Success requires that the system's reconstructed S'_{out}aligns with the original $S^{\setminus *}$, not with the interfering near-neighbour structure.

C.7 Perturbation catalogue

Perturbations are standardised "test operators" applied at the interruption boundary. Each perturbation must be applied at multiple severity levels.

Perturbation ID	Perturbation type	Operational description	Intended failure mode detected	Severity levels
P1	Cue deletion	Remove key fragments from Q_{re}or from the bounded state fragment	Overreliance on shallow cues, inability to reconstruct state	Low, medium, high
P2	Cue paraphrase	Paraphrase Q_{re}while preserving referential identity	Template dependence, lexical anchoring	Low, high
P3	Distractor load	Insert intervening content consuming context budget	Fragility under context displacement	256, 2048, 8192 tokens
P4	Near-neighbour interference	Insert a similar task with different latent structure	State overwrite, recency capture	Low similarity, high similarity
P5	State corruption	Corrupt parts of the bounded state fragment provided at Stage 2	Error correction capacity versus confabulation	Single-field, multi-field
P6	Schema shift	Change the required output schema at Stage 2	Rigidity, inability to rebind state into new form	Minor, major
P7	Partial observability	Hide some variables and require inference from preserved constraints	Dependence on full restatement	Low, medium
P8	Temporal delay	Enforce wall-clock delay between stages, with controlled resumption	Time sensitivity, session boundary fragility	Seconds, minutes, hours

C.8 Recurrence windows and scheduling

Recurrence windows must be specified in two orthogonal units.

1. **Context displacement window** W_c, measured in intervening tokens inserted between Stage 1 and Stage 2.
2. **Wall-clock window** W_t, measured as elapsed time between the two stages.

A minimal implementation may use only W_c. A governance-grade implementation should include both, since operational deployments often introduce time gaps independent of token displacement.

Recommended default schedule:

Window ID	$W_c$intervening tokens	$W_t$elapsed time	Primary interpretation
W0	0	0	Baseline, no interruption
W1	256	0 to 10 seconds	Short displacement
W2	2048	1 to 5 minutes	Medium displacement

Window ID	W_c intervening tokens	W_t elapsed time	Primary interpretation
W3	8192	30 to 120 minutes	Long displacement
W4	16384 or new session	24 hours	Cross-session recurrence

C.9 Failure criteria and classification taxonomy

To avoid ambiguity, failures are classified into distinct categories with explicit triggers.

F1, recomputation masquerading as continuity
Triggered when Stage 2 performance is achieved despite the absence of any dependency on Stage 1 state, demonstrated by a control where Stage 1 is omitted and performance remains unchanged.

F2, state overwrite under interference
Triggered when the reconstructed state aligns with the near-neighbour injected structure rather than the original latent object.

F3, confabulated state completion
Triggered when the system emits a structurally coherent state that fails verification against $S^{\setminus *}$, coupled with high confidence or unjustified invariance.

F4, brittle schema coupling
Triggered when a schema shift prevents re-entry despite unchanged latent structure, indicating continuity is tied to output formatting rather than preserved relational object.

F5, externalised continuity misattribution
Triggered when M1 succeeds but M0 fails, indicating continuity is supplied by orchestration rather than intrinsic R_g.

F6, leakage channel dependence
Triggered when success depends on prohibited leakage such as restating the full Stage 1 trace, hidden system prompts, or persistent memory not declared in the governance profile.

C.10 Metrics and R_g-proxy construction

The suite reports both point metrics and curves.

C.10.1 Point metrics

- Re-entry success rate p_{re} per window and perturbation
- State reconstruction accuracy A_S, measured as structural match to $S^{\setminus *}$
- Interference robustness A_I, success conditional on P4 being present
- Correction robustness A_C, success conditional on P5 being present

C.10.2 Continuity curve

For each perturbation family P_i, estimate $p_{re}(W_c)$across windows. The curve itself is treated as the primary continuity signature.

C.10.3 R_g-proxy index

A minimal R_g-proxy can be derived as a normalised area under the re-entry curve across a declared window range:

$$\hat{R}_g(P_i) = \frac{1}{W_{c,max}} \int_0^{W_{c,max}} p_{re}\,(w)\,dw$$

A governance-grade report should provide $\hat{R}_g$by perturbation type and an aggregate $\hat{R}_g$weighted by operational relevance.

C.11 Controls, audit logs, and reporting template

To prevent interpretive drift, each run must generate a minimal audit log including:

- Governance profile declaration M0 or M1
- Full specification of perturbations and windows
- Full specification of what Stage 2 was permitted to observe
- Verifier versions, parsing rules, and schema versions
- Sampling policy, including any use of pass@k at Stage 2

Reporting should be standardised to an appendix-level form:

- Protocol class identifier, window schedule, perturbation schedule
- Re-entry success matrix $p_{re}(W, P)$
- Failure taxonomy distribution across F1 to F6
- Attribution statement distinguishing intrinsic continuity from orchestrated continuity
- A limitation clause stating what the suite does not establish, including that behavioural continuity proxies do not, by themselves, establish integrated $\Phi_i$unless paired with binding and perturbation-resilience instrumentation

C.12 Administrative conclusion

This re-entry suite is designed to convert rhetorical claims about “deliberation,” “reasoning persistence,” or “memory” into adjudicable evidence about re-entry and rhythmic continuity. It accomplishes this by forcing the system to succeed under interruption, displacement, interference, and controlled corruption, with explicit controls that detect recomputation, leakage, and orchestration dependence. Under CIITR, successful performance in this suite under the model-only profile is the minimum evidentiary prerequisite for escalating from competence narratives to any claim that $R_g$is nontrivial.

Appendix D. CPJ measurement protocol

This appendix specifies an implementation-grade measurement pipeline for CPJ, Comprehension per Joule, under CIITR. The protocol is designed to close the principal measurement gap exposed in the paper's otherwise disciplined compute budgeting regime, namely that compute normalisation and token-equivalent accounting are not physical energy measurement and therefore cannot support CPJ claims. The protocol is deliberately procedural. It defines measurement boundaries, energy instrumentation, phase segmentation, data schema, uncertainty handling, and a standard reporting format suitable for audit and reproducibility.

D.1 Definitions and admissibility requirements

D.1.1 CPJ definition

Under CIITR, comprehension is defined as:

$$C_s = \Phi_i \times R_g$$

CPJ is defined as comprehension yield per unit energy:

$$\mathrm{CPJ} = \frac{C_s}{E}$$

where E is physical energy measured in joules at a declared measurement boundary.

D.1.2 Immediate consequence

A CPJ claim is non-admissible if either of the following holds:

- Energy is not measured in joules using a declared boundary and a calibrated pipeline, or
- Comprehension yield is substituted by task accuracy, pass@k, or any other competence metric without an explicit Φ_i and R_g operationalisation.

The paper's compute budgeting and token-equivalent accounting can support allocation comparisons under compute proxies. They cannot substitute for the joule denominator and cannot define the CIITR numerator.

D.2 Measurement boundary specification

D.2.1 Boundary options

The protocol supports three boundary classes. The selection must be declared in advance and held constant across compared runs.

- Boundary B0, accelerator-only, measures energy drawn by the primary compute device, for example GPU, TPU, NPU. This boundary is acceptable only if memory movement, host CPU, and interconnect overhead are stable and explicitly treated as excluded.

- Boundary B1, node-level, measures total energy of the compute node, including accelerator, host CPU, RAM, storage, and cooling within the node envelope. This is the recommended boundary for most laboratory studies.
- Boundary B2, system-level, measures total facility energy attributable to the run, including networking, rack infrastructure, and external cooling. This boundary is recommended for procurement and sustainability-oriented governance decisions.

D.2.2 Boundary declaration requirements

Every CPJ report must include:

- Boundary identifier, B0, B1, or B2.
- Physical measurement point, for example per-device telemetry, on-board power sensors, external inline meter.
- Inclusion and exclusion statement, explicitly listing what the boundary includes and what it excludes.
- Baseline treatment, whether gross energy is reported, or net energy with idle baseline subtraction, or both.

D.3 Instrumentation stack and calibration

D.3.1 Instrument hierarchy

The protocol defines a hierarchy of acceptable instruments. Use the highest feasible tier, and document downgrades.

- Tier 1, external power analyser, an inline AC or DC meter with timestamped sampling, suitable for node-level or system-level boundaries. This is the preferred method when high assurance is required.
- Tier 2, platform power sensors, for example BMC, IPMI, rack PDU telemetry, accelerator board sensors, vendor power APIs, suitable when external meters are unavailable.
- Tier 3, software-estimated power, derived from utilisation and model-based estimates. This tier is not admissible for CPJ claims, it is permitted only for exploratory engineering comparisons and must be labelled as non-CPJ.

D.3.2 Sampling and time synchronisation

- Sampling rate should be at least 1 Hz for training phases and at least 10 Hz for inference micro-benchmarks, unless hardware limits impose a lower rate. Lower rates must be justified, and uncertainty must be reported.
- All measurement channels, power, phase markers, and evaluation events, must share a time base. If perfect synchronisation is not possible, the protocol requires explicit clock drift estimation and correction.

D.3.3 Calibration and error model

- External meters must be calibrated according to vendor specification, with calibration date recorded.
- Telemetry sources must be validated against an external meter for a representative workload segment, at minimum one training segment and one inference segment.
- The report must provide an uncertainty estimate for energy, including the dominant error sources, for example sampling rate aliasing, sensor quantisation, telemetry lag, baseline instability.

D.4 Phase accounting model

D.4.1 Phase ledger

The protocol requires phase-resolved accounting. Each run must be segmented into a phase ledger with explicit start and stop markers.

The minimum phase set is:

- Phase P1, pre-training
- Phase P2, mid-training
- Phase P3a, RL rollouts, sampling and environment interaction if applicable
- Phase P3b, RL updates, gradient computation and parameter updates
- Phase P4, evaluation inference, including pass@k sampling where used
- Phase P5, verification and parsing overhead, if process verification is externalised or materially nontrivial

This phase decomposition aligns with the paper's intervention framing, while making explicit that RL often has heterogeneous cost components that compute proxies may compress into a single estimate.

D.4.2 Phase marker implementation

Each phase transition must emit a machine-readable marker with:

- phase_id
- run_id
- timestamp_start, timestamp_end
- workload descriptor, for example batch size, sequence length, sampling k, verifier mode
- hardware descriptor, for example device type, precision mode, number of devices

Markers must be logged to an immutable run ledger.

D.4.3 Overhead control

During measurement runs:

- Non-essential processes should be disabled.
- Thermal and frequency management settings should be recorded.
- Cooling configuration should be stable.
- Background network traffic should be minimised, or measured and modelled if it cannot be controlled.

The protocol requires reporting of thermal steady-state attainment, because thermal throttling can materially alter joule-per-token characteristics.

D.5 Energy computation method

D.5.1 Energy integration

Energy per phase is computed by integrating power over time:

$$E_{phase} = \int_{t_0}^{t_1} P(t)\, dt$$

In discrete sampled form:

$$E_{phase} \approx \sum_{i=1}^{n} P_i\, \Delta t_i$$

D.5.2 Gross and net energy

The protocol distinguishes:

- Gross energy, the measured energy during the phase interval.
- Net energy, gross energy minus idle baseline power multiplied by phase duration.

Both may be reported. If net energy is reported, baseline measurement must be performed under the same boundary and instrumentation stack, over a duration sufficient to estimate baseline variance.

D.5.3 Multi-node aggregation

For distributed runs, energy is computed per node and aggregated. The report must include:

- Energy per node per phase
- Aggregate energy per phase
- Any imbalance, for example straggler nodes, that changes effective energy per unit work

D.6 Comprehension yield measurement, numerator requirements

D.6.1 Numerator policy

The CPJ numerator must be a CIITR-aligned comprehension yield. The protocol permits two admissible numerator forms, depending on the maturity of Φ_i and R_g instrumentation.

- Numerator N1, proxy comprehension score, $\hat{C}_s = \widehat{\Phi}_i \times \hat{R}_g$, derived from a fixed evaluation suite that includes cross-context binding and perturbation-resilience protocols for $\widehat{\Phi}_i$, and re-entry and recurrence protocols for $\hat{R}_g$. This is admissible as proxy-CPJ and must be labelled accordingly.
- Numerator N2, comprehension score with stronger instrumentation, permitted only if the project provides a defensible operationalisation of Φ_i and R_g that exceeds proxy status within the declared ontology. This is uncommon and should be treated as an advanced track.

Absent $\widehat{\Phi}_i$ and $\hat{R}_g$ instrumentation, pass@k cannot be used as a numerator for CPJ. It may be reported as a separate competence indicator, but not substituted for comprehension yield.

D.6.2 Evaluation set freezing

To prevent metric drift, the evaluation suite used to compute $\widehat{\Phi}_i$ and $\hat{R}_g$ must be frozen, versioned, and held constant across phase comparisons. The report must include suite version identifiers, generator parameters, and verifier versions.

D.7 CPJ reporting forms

The protocol defines three CPJ reporting forms, each with a distinct governance meaning.

D.7.1 Inference CPJ, per-instance

Measures comprehension yield achieved during inference per joule consumed by inference plus verification overhead:

$$\mathrm{CPJ}_{inf} = \frac{\hat{C}_{s,inf}}{E_{inf}}$$

This form is relevant for operational deployment efficiency and service-level cost models.

D.7.2 Training marginal CPJ, phase increment

Measures comprehension yield improvement attributable to a training phase per joule consumed by that phase:

$$\mathrm{CPJ}_{train,phase} = \frac{\hat{C}_{s,post} - \hat{C}_{s,pre}}{E_{phase}}$$

This form is relevant for deciding how to allocate budget among pre-training, mid-training, and RL, which is directly aligned with the paper's phase-allocation question, but expressed in thermodynamic terms rather than compute proxies.

D.7.3 Verified-correct-per-joule, auxiliary operational KPI

Because the paper uses strict process verification and pass@k, the protocol recommends an auxiliary KPI:

$$\mathrm{VCJ} = \frac{\Pr\,(\text{verified-correct})}{E}$$

This is not CPJ, it is an operational competence efficiency measure, useful for linking sampling policies, verification overhead, and energy exposure.

D.8 Standard reporting format

Every CPJ report should include the following sections, in this order.

D.8.1 Executive measurement statement

- Objective of the measurement, inference CPJ, training marginal CPJ, or both
- Boundary class B0, B1, or B2
- Instrument tier and instrumentation stack
- Numerator form N1 or N2, proxy status if N1

D.8.2 Phase energy table

A phase-resolved table with:

- Phase identifier
- Duration
- Gross energy, joules
- Net energy, joules, if reported
- Average power
- Work units, for example tokens processed, rollouts generated, verification operations
- Uncertainty interval

D.8.3 Comprehension yield panel

A panel reporting:

- $\widehat{\Phi}_i$ indices, including binding stability and perturbation resilience scores
- $\widehat{R}_g$ indices, including re-entry success curves by window and perturbation class
- $\hat{C}_s$ computed as the product of the above, with explicit caveats

D.8.4 CPJ computations

- CPJ_{inf} with numerator and denominator values
- $\mathrm{CPJ}_{train,phase}$ per phase

- VCJ as auxiliary, if reported
- Confidence intervals and sensitivity to baseline subtraction

D.8.5 Claims discipline statement

A mandatory paragraph that states:

- Whether the result is proxy-CPJ or CPJ
- Which substitution fallacies are prohibited, specifically that compute proxies and pass@k are not substitutes for joules and C_s
- Scope boundary of the ontology, including that the generator-defined verifier supports certain reliability conclusions but does not by itself establish integration across open contexts

D.9 Acceptance criteria for CPJ admissibility

A result may be labelled proxy-CPJ only if all conditions below are satisfied:

- Joule-level energy measurement is performed at a declared boundary, with calibrated instrumentation and uncertainty reported.
- Phase accounting includes at minimum the phases that materially change energy exposure, including RL rollouts if RL is used, and verification overhead if process verification is used.
- Comprehension yield is computed as $\hat{C}_s = \hat{\Phi}_i \times \hat{R}_g$ on a frozen evaluation suite, with $\hat{R}_g$ derived from a re-entry protocol that is not reducible to static pattern matching.

If any condition is not satisfied, the report must not present CPJ, and must restrict itself to compute-proxy budgeting and competence metrics, which is the regime the paper itself remains within.

D.10 Administrative conclusion

The protocol converts the paper's budget framing into a thermodynamic governance instrument by replacing compute normalisation with joule-level accounting and by requiring a CIITR-aligned comprehension yield definition. This preserves the utility of phase-decomposed intervention analysis, while preventing a category error in which improved task scores under process verification are interpreted as improved epistemic efficiency. Under CIITR, capability can rise while CPJ declines, particularly when gains rely on sampling, verification overhead, or rollout-intensive reinforcement regimes. The measurement pipeline specified here is therefore a necessary condition for any institutional narrative that intends to speak about efficiency, sustainability, or substrate-level optimisation in comprehension terms.

Appendix E. Reproducibility and configuration ledger

This appendix specifies a reproducibility and configuration ledger suitable for controlled replication of the paper's core results and for the CIITR extensions proposed in Parts IV and V. The ledger is not presented as a general "best practice" checklist. It is defined as a

regulated artefact register intended to make behavioural outcomes reconstructible, comparable across runs, and auditable under claims discipline. The ledger is therefore structured around the paper's central governance surfaces, namely the synthetic generator, the phase-decomposed training pipeline, the reward topology, and the verifier-supported evaluation harness.

The organising principle is that any variable which can plausibly alter (i) pass@k under strict process verification, (ii) sensitivity to depth and breadth regimes, (iii) RL efficacy at the competence boundary, or (iv) the operational resource profile of the programme, must be versioned, logged, and reported as part of the experiment's identity. The aim is to prevent a common failure mode in replication, where the "same model" is assumed to be the same system even though curriculum placement, reward definitions, or verifier scripts differ materially.

E.1 Ledger scope and audit posture

The ledger is defined as a set of immutable records, each record corresponding to one experimental run and one published artefact family. A run is considered reproducible only if:

- The generator can regenerate the identical dataset splits and instance identities.
- The training pipeline can be re-executed with identical budgets and data orderings.
- The reward pipeline can be re-instantiated with identical components and mixing coefficients.
- The evaluation harness can reproduce pass@k under strict process verification using the same parser, verifier, and sampling policy.
- Hardware and precision modes are reported sufficiently to interpret compute-proxy and energy measurements, especially when the CIITR CPJ protocol is applied.

E.2 Ledger format and record structure

Each run is represented as a structured record with the following sections. The structure below is intentionally administrative and is suitable for copy-pasting into a version-controlled repository as a human-readable and machine-parseable manifest.

E.2.1 Run identity and provenance

- run_name: a descriptive label, including the intervention class (baseline, mid, RL, mid+RL, reward variant)
- run_owner: responsible person or lab group
- run_timestamp: start and end time
- claims_scope_tag: performance-only, performance+mechanism, CIITR-extension, proxy-CPJ enabled
- code_provenance: repository commit hash, submodule hashes, dependency lockfile digest

- environment_fingerprint: container image hash or environment export, including OS, CUDA or relevant backend, compiler versions

E.2.2 Model identity ledger

- base_model_name: architecture identifier
- base_model_checkpoint: exact checkpoint ID or hash
- tokenizer_version: tokenizer hash and vocabulary version
- precision_mode: fp16, bf16, fp8, int8, etc.
- context_window: maximum context tokens and truncation policy
- decoding_policy_default: temperature, top-p, top-k, repetition penalties, stop conditions

Where training produces multiple phase snapshots, the ledger must include:

- checkpoint_pre: hash and storage location
- checkpoint_mid: hash and storage location
- checkpoint_post: hash and storage location
- checkpoint_delta_notes: any non-standard operations, such as parameter freezing, LoRA modules, or adapter stacks

E.2.3 Synthetic data generator ledger

This section is mandatory because the paper's ontology is generator-defined.

- generator_version: commit hash or package version
- seed_global: global RNG seed
- seed_split: dataset splitting seed
- seed_instance: instance generation seed schema, including whether per-instance seeds are derived deterministically from a master seed
- graph_family: DAG generation algorithm identifier
- graph_parameters: node counts, edge distribution constraints, any acyclicity enforcement rules
- difficulty_parameter: definition of $op(G)$used in the run, including whether it is $|E|$or a different operation proxy, and any constraints on edge counts
- value_generation: distributions for numeric values and any constraints
- template_catalog_version: template set identifier and hash
- template_assignment_policy: mapping rules from instances to templates

- rendering_rules: explicit versus implicit dependency settings, natural-language phrasing toggles
- train_mid_rl_split_spec: explicit specification of which structural and template regimes go into each phase dataset
- dataset_manifest: file hashes for generated datasets, plus the deterministic recipe to regenerate them

E.2.4 Training pipeline ledger, phase budgets and scheduling

This section aligns with the paper's phase decomposition.

- training_framework: library and version
- optimiser: type, hyperparameters, schedule
- batch_size: global and per-device
- gradient_accumulation: steps
- sequence_length: training sequence length and padding/truncation policy
- learning_rate_schedule: warmup, decay, final LR
- regularisation: dropout, weight decay, gradient clipping

Phase segmentation:

- phase_pre:
    - dataset_id
    - budget_units: tokens, steps, epochs
    - budget_value
    - data_order_seed
    - stopping_criteria
- phase_mid:
    - dataset_id
    - budget_units
    - budget_value
    - mid_training_objective_notes: any special loss terms or curriculum staging inside mid-training
- phase_rl:
    - dataset_id
    - budget_units: rollouts, steps, token-equivalent estimate if used

    - budget_value
    - rollout_policy: sampling parameters, maximum trace length, termination rules
    - update_policy: PPO or other, number of epochs per batch, KL constraints, clipping
    - separation_of_costs: explicit split between rollout cost and update cost, if instrumented

If the paper's compute-normalised budget split parameter β is used, record:

- total_budget_T: value and units
- beta: value
- derived_T_mid: value and units
- derived_T_RL: value and units
- equivalence_model: derivation method for RL token-equivalent cost, including any FLOPs assumptions if used as a proxy

E.2.5 Reward topology ledger

This section is mandatory when RL is used and when process-aware rewards are used.

- reward_scheme_id: outcome-only, mixed, strict-gated, or other
- R_out_definition: exact computation of outcome reward, including scaling
- R_pv_definition: exact computation of process-verification reward, including step weights and partial credit policy
- alpha: mixing coefficient, if mixed
- gating_rule: formal rule for strict gating, if used
- verifier_in_reward: whether verifier is invoked online during RL, and its computational placement
- reward_normalisation: any normalisation or clipping
- reward_logging_policy: what is logged per rollout, including per-step rewards and verifier outcomes

E.2.6 Evaluation harness ledger

This section defines what "performance" means for reproducibility purposes.

- evaluation_suite_version: identifier for the test set, including OOD splits by depth and breadth
- evaluation_seed: RNG seed for sampling

- decoding_policy_eval: sampling parameters for evaluation, must be fixed for pass@k comparability
- k_values: list of k used, e.g., 1, 8, 32, 128
- parsing_script_version: commit hash and parser configuration
- verifier_version: commit hash and verifier configuration
- strictness_mode: definition of strict correctness, including whether both step-level and final answer must match
- error_taxonomy_version: if error classification is reported, include taxonomy script version
- reporting_outputs: exact metrics produced, pass@k, process accuracy, graph match metrics, etc.

If CIITR extensions are executed, include additional evaluation artefacts:

- Phi_i_proxy_suite_version: binding and perturbation protocols version
- Rg_reentry_suite_version: re-entry suite version, including windows and perturbations
- energy_measurement_enabled: boolean, and if true, link to CPJ ledger fields in E.2.7

E.2.7 Energy and resource measurement ledger (for CPJ-enabled runs)

If CPJ or proxy-CPJ reporting is enabled, the run must include:

- energy_boundary: B0, B1, or B2
- instrument_tier: external meter, platform sensors, or exploratory estimate
- sampling_rate_hz: for power telemetry
- baseline_policy: gross, net, or both
- phase_markers_enabled: yes/no, and marker schema version
- phase_energy_table: pointer to per-phase joule log files
- uncertainty_model: error bounds and dominant sources

This section is mandatory for any claim that uses “efficiency” language beyond compute proxies.

E.3 Controlled replication procedure, administrative workflow

To support controlled replication, the ledger is paired with a procedural replication workflow that specifies what must be reproduced and what may vary.

E.3.1 Mandatory invariants for replication

A replication is considered controlled only if the following are held invariant:

- Generator version and all seeds sufficient to regenerate identical datasets
- Training budgets per phase and data ordering seeds
- Reward scheme definitions and coefficients
- Evaluation decoding policy, k values, parsing and verification scripts
- Claims scope tag and admissibility posture

E.3.2 Permitted variation classes

The following may vary, but must be declared:

- Hardware platform and precision mode, provided that compute proxy reporting is interpreted accordingly
- Training framework versions, provided that numerical equivalence is validated on a small reference run
- Batch size adjustments, provided that effective token budget and gradient update counts remain equivalent

E.3.3 Minimum replication outputs

A controlled replication must publish:

- The full run record manifest
- A dataset manifest with regeneration recipe
- A metrics report that reproduces the paper's headline results within declared tolerance bands
- A divergence analysis if tolerance bands are not met, including which ledger fields differed and why

E.4 Ledger templates for publication

For publication-grade transparency, it may be advantageous to attach two artefacts to any release:

1. A human-readable "run card" that summarises the ledger at high level for reviewers and institutional stakeholders.
2. A machine-readable manifest, for example JSON or YAML, that contains the full record as specified above.

The essential constraint is that both artefacts must be derivable from the same immutable ledger entries, ensuring that summary documents cannot drift from the underlying configuration reality.

E.5 Administrative conclusion

The paper's core contribution is inseparable from its controlled ontology and phase-decomposed intervention logic. Without a regulated configuration ledger, that contribution becomes difficult to replicate and easy to over-interpret, because small changes in generator parameters, mid-training distributions, reward mixing, or verifier strictness can materially reshape outcomes. This appendix therefore defines a reproducibility artefact that treats the training and evaluation pipeline as a governed system, enabling controlled replication, credible cross-study comparison, and defensible claims governance in both research and institutional contexts.