Governance-First Experimental Framework for Replayable Recursive Propagation Measurements

Ivan Silva
Carlonoscopen, LLC
ORCID: 0009-0005-2284-8891

Reserved DOI: 10.5281/zenodo.20564072 *(pre-publication reservation; not yet published)*

Citation status: Not yet published. The DOI has been reserved for manuscript tracking, review, and publication preparation. Final publication metadata and archival record will be established upon formal publication and Zenodo release.

Abstract

Modern large language models exhibit a familiar set of behaviors that resist reduction to local next-token statistics. These include confident hallucination, attractor collapse, recursive repetition, and contextual fragmentation. A prior theoretical framework [@silva2025recursive] proposed interpreting these behaviors as failures of recursive propagation, stabilization, and return projection within a state-conditioned latent geometry. Studying that framework empirically requires a measurement apparatus whose outputs can be trusted by an external reviewer: replay-verifiable, provenance-complete, and bounded against silent semantic escalation.

This paper documents the construction of such an apparatus. Across five sealed Macro Gates (A through E), the work assembles a cryptographically anchored evidence chain in which each gate references its predecessor by SHA-256 and is sealed ARC-1-immutable. Nine binding invariants are enforced structurally. Two of these, MYTH-FC and M7-SEAL, were adopted during Phase 5 and are intended to constrain metric vocabulary, to require closed-enum classification, and to keep the decision-stage module structurally absent from the codebase. Replay verification operates across six artifact stream types using canonical-JSON digest recomputation, with verifiers structurally unable to invoke compute functions (RP-2). A shadow-mode metric layer produces outputs that cannot influence the token selection process they observe; this property is verified by both behavioral anchor tests and AST-scan structural tests. During Phase 5 implementation and the Gate E review, five implementation-stage findings were surfaced and dispositioned, including a transcription typo caught at the moment of sealing.

The primary contribution of this work is the construction of the laboratory itself. The work does not claim truth detection, hallucination detection, semantic-correctness measurement, decoder superiority, or any operationalization of the reference paper's hypotheses. Two run-scale metrics, `R_info_v0_scalar_A` and `R_info_v0_ribbon_D`, compute end-to-end on real-GPT-2 evidence and produce replay-verifiable records classified `EXPERIMENTAL`. Whether the values mean anything in semantic terms is an open question for future controlled experiments inside the laboratory.

The work provides a stable substrate on which the reference paper's hypotheses, and others, can be studied under controlled, replayable, and auditable conditions. The original research question concerning the distinction between reality and illusion in language-model inference is not answered here. The apparatus that could honestly attempt to answer it now exists.

Keywords: Recursive Propagation Geometry; Replay Verification; Governance Architecture; Measurement Provenance; Shadow Metrics; Audit-Honest Systems; Experimental AI Infrastructure.

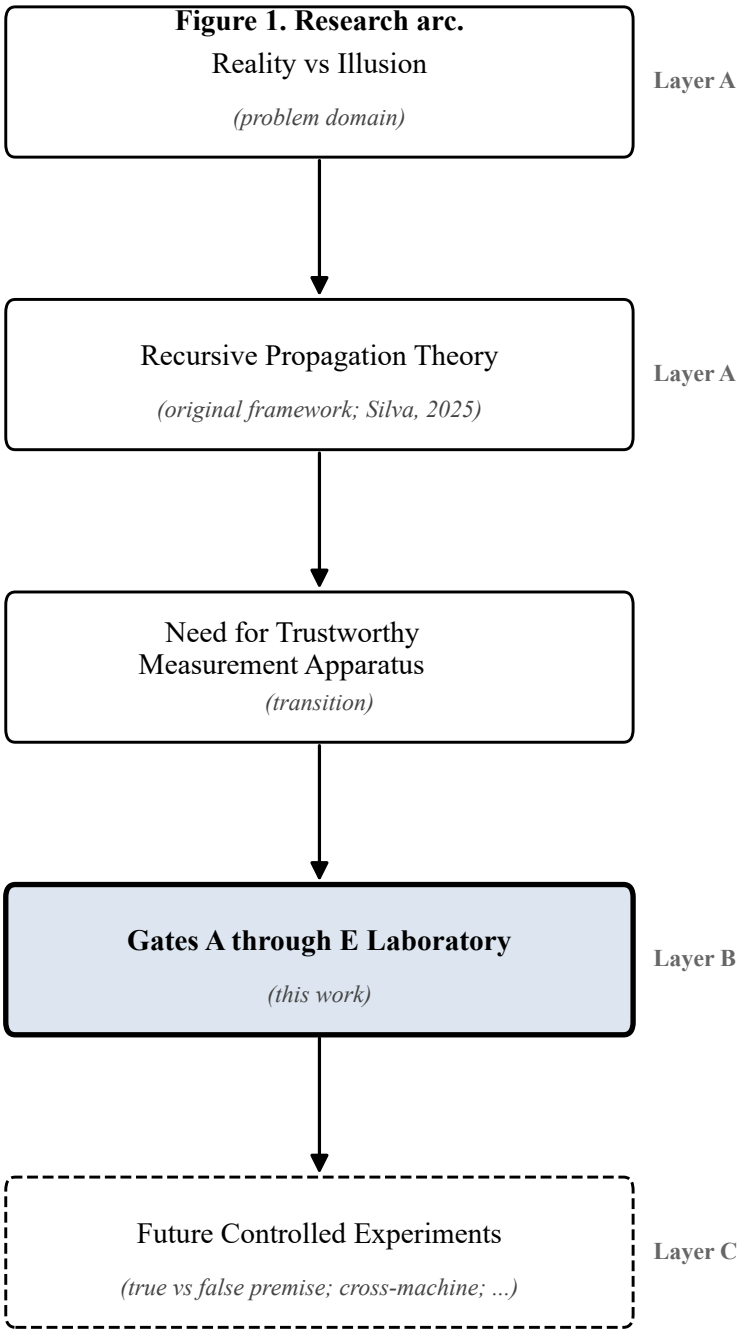
1. Introduction

The original research program from which this work emerges set out to investigate whether internal propagation dynamics in language models differ systematically between conditions that humans would describe as reality-consistent and illusion-consistent. The framing was deliberately broad. The specific operationalizations (true-versus-false premises, hallucinated versus grounded continuations, stabilization-faithful versus stabilization-collapsed trajectories) were treated as candidate windows on a more general phenomenon: how the same architecture, with the same parameters, can produce inferences that humans would classify in qualitatively different ways.

Early exploratory observations on GPT-2, reported in a prior theoretical framework [@silva2025recursive], suggested that such differences might be present in measurable internal-surface trajectories. They were not, however, observations that could carry a formal empirical program. The experimental scripts were ad-hoc, the records were not replay-verifiable, the provenance from input through computation to logged value was incomplete, and the interpretive vocabulary was free to drift between framing and finding. Any further empirical claim made under that framework would require, in addition to the experimental result itself, an apparatus whose outputs could be trusted by an external reviewer.

The present work addresses that prerequisite. It constructs the laboratory required to study the original research questions under controlled, replayable, provenance-complete, audit-honest conditions. The laboratory is the Layer B contribution of this paper. The original research questions remain Layer A. The controlled experiments that the laboratory makes possible remain Layer C and are not undertaken here.

Figure 1 shows the resulting arc: from the original *Reality versus Illusion* research question, through the *Recursive Propagation Theory* framework that proposed an interpretive model for the observed phenomena, to the recognition of a *need for trustworthy measurement* infrastructure that would allow those questions to be investigated rigorously, to the *Gates A through E laboratory* that this paper documents, and finally to the *future controlled experiments* that the laboratory enables.



The laboratory was not constructed independently of the original research program. It emerged from it. The original exploratory observations on GPT-2 exposed measurement limitations that ad-hoc experimentation could not address. Those limitations produced specific governance requirements: replay-verifiability, provenance completeness, structural separation of measurement from intervention, bounded interpretive vocabulary, and audit-honest discipline at the moment of drift. Those requirements, in turn, produced this laboratory. The future controlled experiments inside the laboratory are the same experiments that the original observations could not, by themselves, support.

The remainder of this section frames the empirical territory that motivated the original research question and the structural requirements that the laboratory was built to satisfy.

Large language models exhibit a familiar set of behavioral difficulties that resist reduction to local next-token statistics. Confident hallucination, attractor collapse, recursive repetition, semantic fragmentation, and pronounced sensitivity to recursive prompting structure persist across model scales and decoding strategies. Existing operational interpretations of inference as next-token argmax over a learned probability distribution

capture the surface mechanics but do not explain why these failure modes have the structural properties observed in practice.

A prior theoretical framework [@silva2025recursive] proposed an alternative interpretation. Inference, on that account, can be modeled as recursive propagation through state-conditioned latent geometry, with observable token streams interpreted as lower-dimensional projections of deeper recursive trajectories. The framework decomposed hallucination into three failure modes (stabilization failure, return-projection failure, and informational-flow failure) and proposed recursive lookahead decoding as a candidate architecture optimizing invariant persistence, curvature stability, and return-projection quality rather than local token probability alone. The framework reported preliminary empirical observations from GPT-2 experiments, including measurable internal-surface differences between true and false premises, altered token trajectories under future-branch simulation, and low-information attractor collapse modes when stability was optimized without informational-flow constraints. The framework was explicitly characterized as exploratory and as establishing a research program rather than a finalized theory.

Studying these hypotheses further, rather than relying on preliminary observations from ad-hoc experimental runs, requires a measurement apparatus with properties that ad-hoc experimentation cannot provide. The apparatus needs to be replayable, so that any reported measurement can be independently recomputed from sealed evidence. It needs to be provenance-complete, so that every record carries the cryptographic and configuration context necessary to determine what was measured, under what conditions, and with what code. It needs to maintain structural separation between measurement and intervention, so that the act of measuring cannot influence what is being measured. This is a non-trivial property when the measurement subject is the same generation process whose dynamics are under study. It also needs bounded vocabulary and bounded claims, so that interpretive language does not silently accrete into the record format and convert preliminary observations into apparent findings. Finally, the apparatus needs an audit-honest discipline, so that drift events are surfaced and dispositioned rather than absorbed silently into successive revisions.

This paper documents the construction of such an apparatus. Across five sealed Macro Gates, labeled A through E, the work assembles, tests, seals, and verifies a measurement laboratory for the study of language-model internal dynamics under the constraints listed above.

1.1 Contribution

The primary contribution of this work is the construction of the laboratory itself, rather than a metric or a finding about language models. The work is best understood as the construction of an apparatus that supports future controlled empirical study, rather than as a validation of the underlying theoretical hypotheses.

In particular, the work produces:

1. A cryptographically anchored chain of five sealed Macro Gates (A through E) under ARC-1 immutability.
2. Nine binding invariants enforced structurally through code-level tests, cryptographic digest material, and the structural absence of forbidden modules.
3. A replay architecture across six artifact stream types in which replay is structurally independent of compute.
4. A shadow-mode metric layer whose outputs cannot influence the token-selection process they observe.
5. Two run-scale aggregate metrics that compute end-to-end on real-GPT-2 evidence and produce replay-verifiable records, classified `EXPERIMENTAL` with the classification carried in the canonical-JSON digest material, so that any re-classification after sealing changes the record digest and invalidates the seal.

6. An audit-honest discipline that surfaced and dispositioned five implementation-stage findings during the cycle, including a transcription typo caught at the moment of sealing, without silent absorption.

1.2 What this paper does not claim

The work maintains structural separation between what was demonstrated and what was not. The reference paper's hypotheses, including true-versus-false-premise propagation differences, hallucination decomposition, and attractor collapse modes, served as motivation for building the laboratory. They were not tested by it. This paper therefore relies on explicit negative-space claims throughout.

In particular, this paper does not claim:

- truth detection;
- hallucination detection;
- semantic-correctness measurement;
- informational validity in any technical sense;
- reasoning superiority of any decoder;
- decoder superiority of any form;
- measurement of "reality" or "illusion" in any sense beyond the metaphor;
- M7 (decision/intervention) capability;
- that the run-scale metric values mean anything in semantic terms;
- that any observation generalizes beyond the specific prompts, model, and machine represented in the Gate E evidence cycle.

These are fixed non-claims, not provisional gaps. Section 5 enumerates them in their full form alongside the demonstrations.

1.3 Paper organization

Section 2 establishes the historical bridge from the original theoretical framework to the present infrastructure, with explicit separation between original hypotheses (Layer A), laboratory construction (Layer B), and future experimental program (Layer C). Section 3 documents the laboratory: the gate chain, the binding invariants, the replay architecture, the shadow-mode metric layer, and the Gate E evidence. Section 4 documents the audit methodology, including the named failure modes F-α through F-ε. Section 5 places demonstrations and explicit non-demonstrations side by side. Section 6 outlines the future experimental program that the laboratory enables. Section 7 provides methods, including the disclosure of partial reviewer independence. Sections 8 and 9 address limitations and significance. Appendices A through E provide the full claim audit, the complete Phase 5 governance disposition record, the test-suite composition, the reproducibility checklist, and the Layer-1 / Layer-A glossary.

---

2. From Recursive Propagation Geometry to Governance-First Experimental Infrastructure

The transition from the original theoretical framework to the present infrastructure was not a sharp break. It was a gradual recognition that the framework's empirical study would require a measurement apparatus that did not yet exist. This section traces that transition in three layers: the original hypotheses that motivated the research, the laboratory that was eventually built, and the future experiments that the laboratory now makes possible.

2.1 The original hypotheses (Layer A)

The reference paper [@silva2025recursive] proposed interpreting inference in large language models as recursive propagation through a state-conditioned latent geometry. Its principal theoretical constructs included:

- *Recursive propagation*, with inference modeled as trajectory propagation through a manifold $\mathcal{M}_t = \mathcal{G}(W, C_t, A_t, H_t)$ induced by model parameters, context, attention, and hidden-state history.
- *Recursive metabolization*, with reasoning modeled as impedance-matched resolution of asymmetry, where successful metabolization requires $\Delta\Gamma < 0$.
- *Tubular propagation geometry*, with local reasoning trajectories propagating within local tubular regions around dominant recursive geodesics.
- *Recursive curvature*, with instability modeled through noncommuting propagation transformations.
- *Hallucination decomposition*, with hallucination interpreted as $F_S \vee F_P \vee F_I$, where F_S is stabilization failure, F_P is return-projection failure, and F_I is informational-flow collapse.
- *Recursive lookahead decoding*, with a proposed decoder objective combining invariant survival S_I, recursive drift D, curvature instability K, return-projection quality Q, informational progression F_info, and repetition collapse penalty R_loop.

The reference paper reported preliminary experimental observations consistent with the framework, including measurable internal-surface differences between true and false premises, altered token trajectories under future-branch simulation, and punctuation resonance loops under stability-only optimization. The reference paper explicitly characterized these observations as preliminary and operationally significant, rather than as a finalized theory.

These are the prior hypotheses of Layer A. They are not the findings of the present work.

2.2 What remained unchanged

Across the transition from the reference paper to the present infrastructure, several things remained unchanged. The underlying scientific question remained the same: how internal propagation dynamics in language models give rise to hallucination, stabilization failures, and informational-flow collapse. The methodological orientation toward GPT-2 as a tractable, reproducible subject for empirical investigation was preserved. The recognition that internal-surface trajectories (entropy, attention, hidden-state evolution, and branch behavior) carry information not visible in the observable token stream alone was preserved. The principle that lookahead-based decoding reveals structure that single-step decoding does not, regardless of whether such decoding is operationalized, was preserved. The author identity and intellectual provenance were preserved. One thing that did not change across the transition is worth noting explicitly: the evidence

base of the reference paper itself remained at its original status. The preliminary observations reported in the reference paper were produced by ad-hoc experimental scripts that were not, at the time, replay-verifiable, and the present work does not retroactively replay-verify them. The reference paper's empirical posture remains as it was. What changed is the apparatus available for future work. The laboratory is therefore not a parallel construction but a structural consequence of what the original work could not establish on its own.

2.3 What evolved

Two evolutions occurred between the reference paper and the present work.

The first evolution was methodological. The reference paper's preliminary observations were obtained from ad-hoc experimental scripts whose results were not, at the time, replay-verifiable, provenance-complete, or auditable. As the framework's scientific commitments became sharper, the lack of structural guarantees around the measurement process became the binding constraint on further empirical work. Any further empirical claim made under the framework would require, in addition to the experimental result itself, several supporting properties: cryptographic anchoring of the evidence, complete provenance of how the result was computed, structural separation between measurement and the generation process being measured, and a vocabulary discipline preventing the interpretive vocabulary of the framework from silently accreting into the record formats. This requirement led to the construction of the laboratory.

The second evolution was structural. As implementation began, it became apparent that the framework's Layer-2 interpretive vocabulary (recursive propagation, metabolization, tubular geometry, curvature, invariant compression) carried a specific risk. If these terms entered the record format, even as field names, the records would silently propagate interpretive commitments into every downstream artifact. The discipline of strict Layer-1 vocabulary in the record format, accompanied by Layer-2 substring scanning tests (INV-2 and later MYTH-FC clause 1), emerged from this realization. Internal records describe what was computed; they do not describe what the computation means.

2.4 What was deferred

Substantial portions of the reference paper's program were deferred to future work. Recursive lookahead decoding as an operationalized decoder was deferred. The Gates A through E laboratory implements lookahead branches, but uses them only as an observational substrate. It does not commit branches based on lookahead scores. Implementing the reference paper's decoder objective would require modifying token selection based on metric outputs, which is forbidden under M6-SHADOW-1 and structurally precluded under M7-SEAL. Validation of the hallucination decomposition was deferred. The laboratory does not classify hallucination types, does not detect hallucination, and contains no `metric_classification = HALLUCINATION_TYPE_X` enumeration. Operationalization of curvature, metabolization, invariant compression, and tubular geometry was deferred. None of these constructs appears as a field, variable, or test in the laboratory codebase. They remain Layer A. True-versus-false-premise controlled experiments were deferred. The reference paper's preliminary observations about premise-conditional propagation differences were not tested by Gates A through E. They are the kind of controlled experiment that the laboratory now makes possible (Section 6).

2.5 Why the laboratory became necessary

The laboratory became necessary because no further empirical claim about the reference paper's framework could be made honestly without a measurement apparatus whose outputs could be trusted by an external reviewer. The word "trusted" here is operationally specific. The outputs must be recomputable from sealed evidence. Every record must carry full provenance to its source code and configuration. The measurement act

must be structurally insulated from the subject being measured. The vocabulary used to record measurements must not accrete interpretive commitments. The Gates A through E laboratory was constructed to satisfy these specific requirements.

2.6 How Gates A through E emerged

The five sealed Macro Gates emerged sequentially. Each gate is anchored to its cryptographic predecessor by SHA-256 and closed with an explicit ARC-1 seal. The progression was not arbitrary. Each gate addresses a specific class of structural guarantee:

- Gate A, architecture freeze under INV-1 (measurement independence of interpretation).
- Gate B, vocabulary freeze, condition-blindness, and operational-limit boundaries (INV-2, BR-0, OL-1 through OL-5).
- Gate C, replay-neutrality enforcement at the lookahead layer (RP-2).
- Gate D, per-commit metric layer under shadow mode (M6-SHADOW-1).
- Gate E, run-scale aggregate metric layer under MYTH-FC and M7-SEAL, with closure framed explicitly as an *infrastructure gate*, asserting only infrastructure properties.

Each gate's evidence is referenced cryptographically by its successor. No archive is opened, modified, or embedded in a successor bundle. Section 3 describes the chain in detail.

---

3. Laboratory Construction (Layer B)

The laboratory is what this paper most centrally documents. It is a small body of code, a sequence of sealed archives, and a discipline that connects the two. The sections below describe each in turn, beginning with the architectural foundation and proceeding through the gate chain, the binding invariants, the replay and shadow-mode mechanisms, and finally the Gate E evidence that the apparatus now carries.

3.1 Architectural foundation

The architectural foundation rests on two principles, frozen at Phases A and B respectively and carried forward unchanged through every subsequent gate.

INV-1 (measurement independence) states that the code that measures must not encode commitments about what the measurement means. The codebase contains the harness, the surface extractor, and the metric compute functions. It does not contain an interpretive layer, and INV-1 is the architectural commitment that no interpretive layer be introduced. The principle is structural rather than runtime-enforced, and it has shaped most of the design decisions that follow.

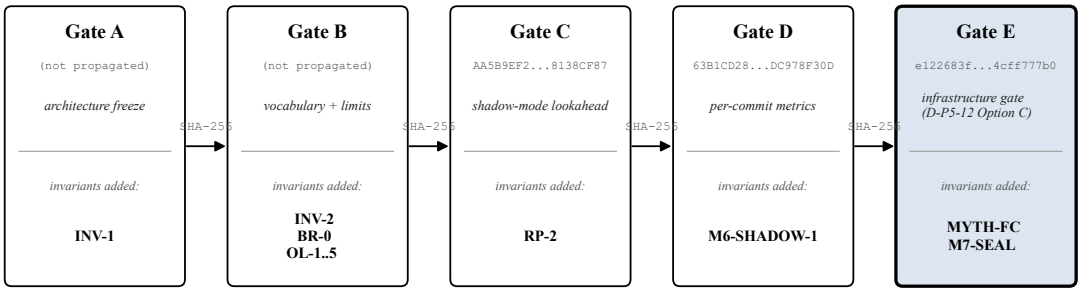
INV-2 (Layer-1 vocabulary) states that record field names use Layer-1 vocabulary only. Layer-2 interpretive terms, including the reference paper's central constructs, are excluded from appearing as field names in any artifact stream. The principle is enforced by a Layer-2 substring scan over emitted records (`test_run_scale_field_names_layer1_only` and predecessors), which finds zero violations in the Gate E sealed evidence [@phase5closure §7].

3.2 Gate lineage and ARC-1 immutability

Each Macro Gate produces a sealed ARC-1 archive. The chain is anchored cryptographically. Each successor bundle's manifest records the predecessor's archive SHA-256 in `manifests/sealed_archives.json`. The reference-not-embed doctrine, which holds that successor bundles reference predecessor archives by SHA-256 only and never open or embed them, was established at Phase 4 (D-P4-5 Option A) and remains binding [@phase5closure].

Figure 2 presents the gate chain and its invariant accretion.

Figure 2. Gate lineage. Each successor references its predecessor by SHA-256 (ARC-1).



The canonical archive SHA-256 values for the Gate C, D, and E archives appear in Table 1.

The Gate A and Gate B archive SHA-256 values were not propagated into the reviewer-side archive chain available at the time of manuscript preparation. They reside in the operator's standing master ledger but were not recovered for inclusion in this manuscript's Table 1. The chain integrity claim of this manuscript is therefore bounded to the cryptographically anchored portion from Gate C through Gate E, whose SHA-256 values appear in Table 1 and are independently verifiable against the sealed archive files. Gates A and B remain program milestones with documented architectural and vocabulary-freeze roles (see §3.3 and the master ledger), but their canonical SHA-256 values are recorded here as a known gap in the reviewer-side archive chain rather than concealed.

Table 1: Gate registry.

Gate	Bundle	Archive SHA-256 (canonical, 64-char)	Closure framing
A	(pre-bundle artifacts)	*(not propagated into reviewer-side archive chain; see §3.2)*	architecture freeze; INV-1
B	(pre-bundle artifacts)	*(not propagated into reviewer-side archive chain; see §3.2)*	INV-2; BR-0; OL-1 through OL-5
C	v3.0-gateC	`AA5B9EF287DA569165EAFB4479FE50829D7BC1EA1D19005A38B1FFA38138CF87`	RP-2; shadow-mode lookahead
D	v4.0-phase4	`63B1CD28D66CB3F81987259A8E6E0FB11DE40968122E1480CE9E341DC978F30D`	M6-SHADOW-1; per-commit metric layer
E	v5.0-phase5	`e122683fa28e0539846bfece64b0c4c85cb5690101a6fdfc3ec26524cff777b0`	infrastructure gate (D-P5-12 Option C); MYTH-FC + M7-SEAL adopted

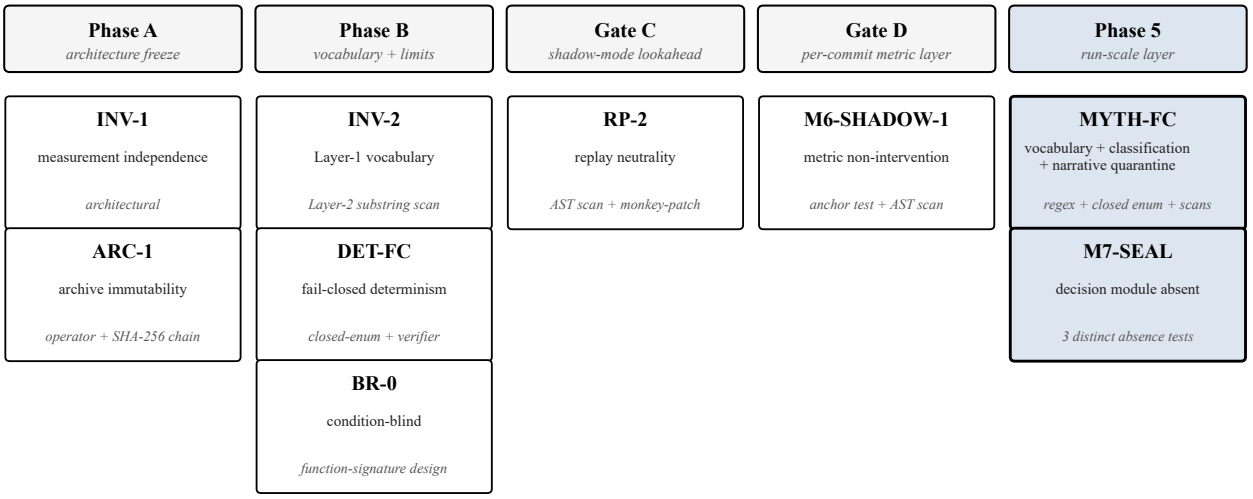
ARC-1 immutability is enforced through operator-side read-only filesystem storage. This is a procedural enforcement and should be understood as such. The cryptographic chain makes any modification of a sealed archive detectable. The read-only storage makes such modification operationally unlikely.

3.3 The nine binding invariants

At the post-Gate-E snapshot, nine binding invariants are in force. Each is anchored to its origin phase, its enforcement surface, and its current verification mechanism.

Figure 3 organizes the invariants by origin phase and enforcement mechanism.

Figure 3. The nine binding invariants in force, organized by origin phase.



MYTH-FC and M7-SEAL (shaded) were adopted in Phase 5 and tested for the first time in the Phase 5 / Gate E cycle.

Table 2: Binding invariants summary.

#	Invariant	Origin phase	Governing purpose	Enforcement surface
1	INV-1	Phase A	source code measurement-independent of interpretation	architectural
2	INV-2	Phase B	record field names use Layer-1 vocabulary only	Layer-2 substring scan
3	DET-FC	Phase B / early implementation	coded nulls; no silent default; fail-closed on malformed input	closed-enum + verifier handling
4	BR-0	Macro Gate B	compute signatures accept no truth label or correctness signal	function-signature design
5	RP-2	Macro Gate C	replay verifiers never invoke compute functions	AST scan + dynamic monkey-patch test

#	Invariant	Origin phase	Governing purpose	Enforcement surface
6	ARC-1	Macro Gate A	sealed archives are immutable; referenced by SHA-256, never modified	operator-side read-only storage + SHA cross-reference
7	M6-SHADOW-1	Macro Gate D	metric outputs are observational only; cannot influence token selection	anchor test + AST scan for forbidden symbols
8	**MYTH-FC**	D-P5-5 (Phase 5)	three clauses preventing metric interpretive mythology	identifier regex + closed-enum + banned-substring scan
9	**M7-SEAL**	D-P5-4 (Phase 5)	decision/intervention module is structurally absent	three distinct absence tests

The two invariants adopted in Phase 5, MYTH-FC and M7-SEAL, represent the largest single-phase expansion of the invariant set in the program's history to date. Both were operator-dispositioned during Phase 5 planning. Both remain binding under explicit operator-signed retraction only. Both were tested for the first time in the present cycle. This co-occurrence is worth naming directly. MYTH-FC and M7-SEAL were introduced during Phase 5, first tested during Phase 5, and the evidence they are intended to constrain was also produced during Phase 5. The cycle therefore constitutes a first demonstration of the two new invariants rather than an independent validation of them: an independent validation would require a future cycle in which the invariants are in force from the outset and the evidence is produced under stresses that did not co-occur with their adoption. The bounded statement of Section 5.1 is scoped to what this single cycle supports. The more general statement (that MYTH-FC and M7-SEAL hold under future stress classes including operationalization pressure, multi-engineer maintenance, retrospective documentation at scale, and semantic pressure) is among the explicit non-demonstrations of Section 5.3. This does not invalidate the work, and it does not retroactively weaken the seven inherited invariants that were in force before Phase 5. It is a precise statement of what the Phase 5 cycle does and does not establish about the two invariants Phase 5 introduced.

3.4 Replay architecture (RP-2)

Replay-verifiable artifact continuity is an operationally important property of the laboratory. Every record in every artifact stream carries a canonical-JSON SHA-256 digest computed over a precisely specified subset of its fields (the "digest material"). Replay verification consists of recomputing the digest from the on-disk record and asserting equality with the logged digest. A mismatch is treated as a release blocker.

The RP-2 invariant, replay-neutrality, strengthens this further. Replay verifiers must never invoke metric compute functions. The invariant is enforced at two independent layers.

Structural enforcement is performed by an AST scan. The replay module `m6_run_scale_replay.py` is statically prohibited from importing any of four forbidden module targets: `m6_run_scale` (compute), `m6_run_scale_runner` (orchestrator that transitively imports compute), `m6_metrics`, or `m6_metric_runner`. The test `test_rp2_no_run_scale_compute_import_in_replay_module` parses the replay module's AST and asserts the absence of any matching `ImportFrom` or `Import` node.

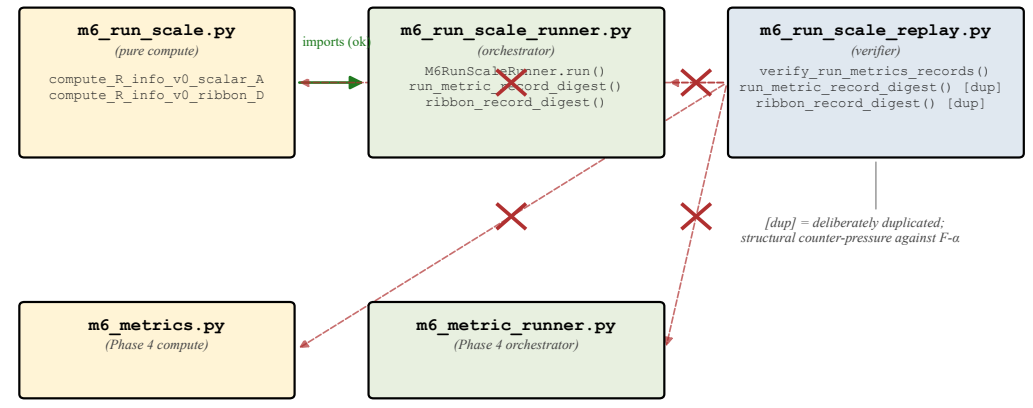
Dynamic enforcement is performed by a monkey-patch test. The test `test_rp2_no_compute_calls_during_replay` monkey-patches the compute functions to raise on invocation, then runs the replay verifier, and asserts that no exception is raised. Replay-verifiable records produced under this regime carry an explicit `compute_calls_observed: 0` field in their replay summaries.

A structural consequence of RP-2 is the deliberate duplication of digest functions between `m6_run_scale_runner.py` (which computes digests when emitting records) and `m6_run_scale_replay.py` (which recomputes digests when verifying records). Both modules carry independent copies of `run_metric_record_digest` and `ribbon_record_digest`. The duplication serves as structural counter-pressure. If the verifier imported the digest function from the runner, the verifier would transitively import compute and would violate RP-2. The duplication is tested for synchronization by `test_runner_and_replay_digest_functions_agree`. This test is part of the sealed test suite at every gate, and the sealed test count (161 at v5.0-phase5; see §7.1) is part of what re-verifying the bundle confirms.

Figure 4 illustrates the resulting module structure and the four forbidden RP-2 import edges.

Figure 4. RP-2 module separation in the Phase 5 run-scale layer.

Forbidden import edges are AST-scan-enforced; the digest function is deliberately duplicated (F-α counter-pressure).



✕ Forbidden import edges (RP-2): m6_run_scale_replay must NOT import any of: m6_run_scale, m6_run_scale_runner, m6_metrics, m6_metric_runner.

Enforced by test_rp2_no_run_scale_compute_import_in_replay_module (AST scan) + test_rp2_no_compute_calls_during_replay (dynamic).

3.5 Shadow-mode metric layer (M6-SHADOW-1)

The metric layer operates in shadow mode. Metric outputs are observational only and cannot influence token selection, commit choice, or branch ranking. M6-SHADOW-1 is enforced at two independent layers.

Behavioral enforcement is performed by an anchor test. The test `test_R_info_v0_has_zero_influence_on_committed_tokens` runs the harness twice from identical seeds. The first run executes without the run-scale metric runner; the second run executes with the run-scale metric runner operating post-hoc. The test then asserts byte-identical committed token sequences. The Gate E evidence cycle additionally confirms the property at runtime. Both Gate E summary records report `m3_influenced_selection: false` [@gateEclosure].

Structural enforcement is performed by an AST scan for forbidden symbols. The test `test_runner_has_no_callback_into_harness` parses the orchestrator module's AST and asserts the absence of any symbol named `observer`, `callback`, `on_commit`, `decide`, `select`, or `intervene` at function, method, or class level. The orchestrator has no API surface through which a harness influence could be expressed.

The two enforcement layers are independent. A violation would require both the behavioral and the structural test to be circumvented.

3.6 MYTH-FC (vocabulary, classification, narrative quarantine)

MYTH-FC, adopted in Phase 5 (D-P5-5) and binding from that point forward, consists of three clauses, each operationalized through a distinct enforcement mechanism.

Clause 1 governs identifier vocabulary. Metric identifiers conform to the regular expression `^[A-Z][A-Za-z0-9_]*_v[0-9]+(_[a-z]+_[A-Z])?$`. The version suffix `_v0` and the framing suffix (such as `_scalar_A` or `_ribbon_D`) are part of the identifier itself and are not removable without re-versioning. Layer-2 substrings (`curvature`, `metabolization`, `tubular`, `manifold`, `invariant_persistence`, `recursive_geometry`, `truth`, `correctness`, `semantically`, `information_content`) are excluded from identifier strings and are verified by `test_metric_ids_contain_no_layer2_substrings`.

Clause 2 requires mandatory closed-enum classification. Every metric record carries a `metric_classification` field drawn from a closed enumeration of `EXPERIMENTAL`, `HEURISTIC`, and `OPERATIONAL`. The classification is included in the canonical-JSON digest material. Any modification of the classification of a sealed record changes the record's digest and therefore fails RP-1 verification against the sealed archive. This closes the silent-promotion path. A metric cannot be re-classified from `EXPERIMENTAL` to `OPERATIONAL` after sealing without invalidating the seal.

Clause 3 enforces narrative quarantine. Records carry no interpretive prose. The `inputs_used` dictionary keys are restricted to a frozen allow-list of 13 entries; any other key fails `test_inputs_used_keys_from_allowlist_only`. A banned-substring scan runs over every string-valued leaf in every emitted record, prohibiting words including `means`, `indicates`, `suggests`, `implies`, `shows`, `demonstrates`, `truth`, `correctness`, `semantically`, and `information content`. Replay summary fields are constrained to a closed vocabulary (D-P5-9): `replay_kind`, `records_checked`, `mismatch_count`, `mismatch_records`, `parse_error_lines`, `all_replay_clean`, `replay_match`, `blocker`, `compute_calls_observed`, `rp2_status`, `run_scale_kind`, and `reason`.

The Gate E sealed evidence passes all three clauses. All four Phase 5 records (two scalar, two ribbon, across two prompts) classify as `EXPERIMENTAL` and contain no banned substrings.

Both the forbidden-symbol list (§3.5) and the banned-substring list above are maintained, finite enumerations. They are not complete with respect to every possible variant that might convey the same meaning. The forbidden-symbol list, for example, does not include synonyms such as `arbitrate` or `make_decision`, and the banned-substring list does not include synonyms such as `attests`, `evidences`, or `denotes`. Completeness against synonymy is a discipline question rather than a structural guarantee, and the discipline depends on the lists being maintained as Layer-2 vocabulary evolves. The Phase 5 cycle did not surface any list-evasion event under the implementations tested, but the absence of evasion in this cycle is not evidence of completeness, and the lists should be treated as living documents subject to amendment under the same operator-disposition discipline that governs the invariants themselves.

3.7 M7-SEAL (structural absence)

The decision/intervention module M7 is structurally absent from the codebase. M7-SEAL, adopted in Phase 5 (D-P5-4), enforces this absence through three independent tests:

1. `test_no_m7_module_present_in_rsp_package` asserts that the `rsp/` package directory contains no file named `m7.py` or matching `m7_*.py`.
2. `test_no_m7_imports_anywhere_in_rsp` parses every Python file in `rsp/` and asserts that no `ImportFrom` or `Import` node references any module beginning with `m7`.
3. `test_no_m7_decision_in_rsp_all` asserts that `rsp.__all__` exposes no symbol containing `m7`.

In addition, boundary tests inherited from earlier phases (renamed in v5.0-phase5 from `test_phase4_boundary_m7_absent` to `test_phase5_boundary_m7_absent` in `test_phase1.py`, `test_phase2.py`, and `test_phase3.py`) maintain the M7-absence assertion across phase boundaries.

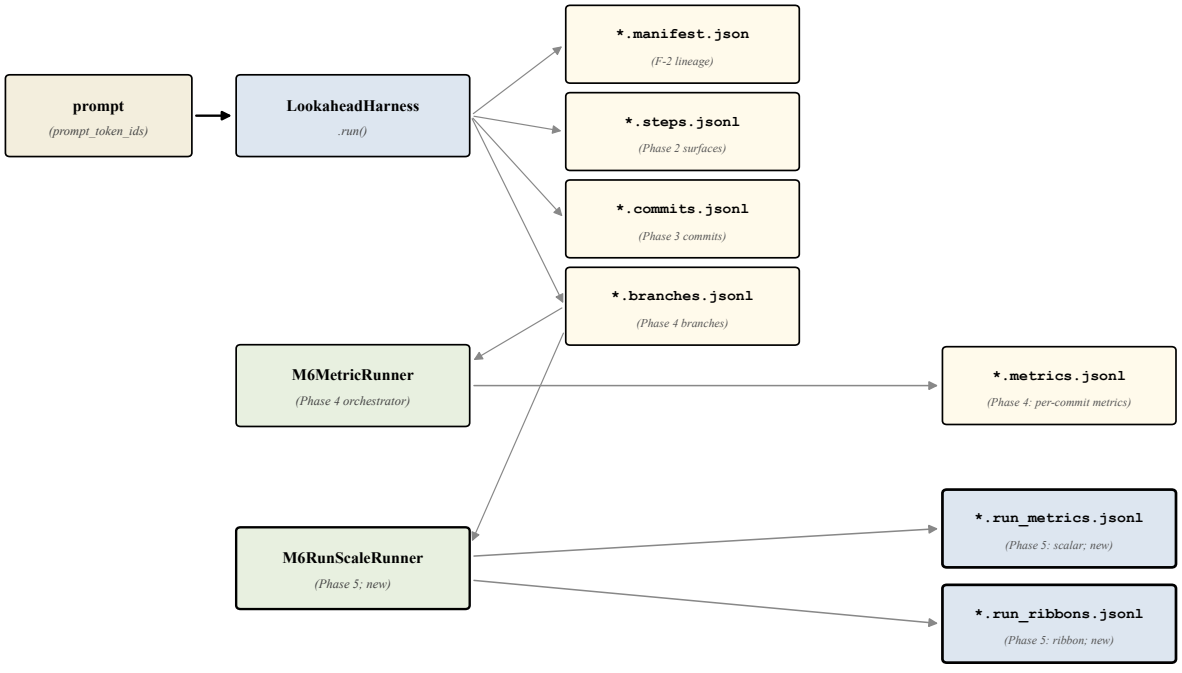
M7-SEAL is reversible only by explicit operator-signed retraction. The retraction path is documented but has not been exercised, and the question of whether the retraction discipline would hold under future pressure remains untested.

3.8 Six artifact stream types

The laboratory produces six distinct artifact stream types. Each is independently replay-verifiable by canonical-JSON digest recomputation. Figure 5 illustrates the pipeline.

Figure 5. The six artifact stream types produced by the laboratory.

Each stream is independently replay-verifiable by canonical-JSON digest recomputation. Phase 5 split scalar and ribbon streams (D-P5-7).



6 streams + manifest lineage anchor: 3 carry-forward (steps, commits, branches), 1 Phase 4 (metrics), 2 new in Phase 5 (run_metrics, run_ribbons).

Table 3: Six artifact stream types.

#	Stream	Schema version	Introduced	RP-1 verifier
1	`*.steps.jsonl`	Phase 2 schema	Phase 2	`verify_surface_steps`

#	Stream	Schema version	Introduced	RP-1 verifier
2	`*.commits.jsonl`	Phase 3 schema	Phase 3	`verify_commit_records`
3	`*.branches.jsonl`	Phase 4 schema	Phase 4	`verify_branch_records`
4	`*.metrics.jsonl`	`metrics.jsonl@1.0.0`	Phase 4	`verify_metric_records`
5	`*.run_metrics.jsonl`	`run_metrics.jsonl@1.0.0`	Phase 5 (this work)	`verify_run_metrics_records`
6	`*.run_ribbons.jsonl`	`run_ribbons.jsonl@1.0.0`	Phase 5 (this work)	`verify_run_ribbons_records`

The decision to split scalar and ribbon records into two streams (D-P5-7, State 1) was made during Phase 5 planning. Both stream types use independent digest functions. Tampering with one is detected without affecting the other.

3.9 The Phase 5 run-scale metric layer

Phase 5 introduced two run-scale aggregate metrics over the Layer-1 surface trajectories produced by lookahead rollouts:

- `R_info_v0_scalar_A` (D-P5-1 framing A, version 1.0.0) is a surface-aggregated cross-commit drift quantity, scalar per run. The metric uses the committed-branch trajectory across commits and a within-commit branch-spread baseline as denominator. Classification: `EXPERIMENTAL`.
- `R_info_v0_ribbon_D` (D-P5-1 framing D, D-P5-8 committed-branch only, version 1.0.0) is a per-surface ribbon at run scale, returning a vector of values across commits without scalar collapse. Classification: `EXPERIMENTAL`.

The plural-version decision (D-P5-1) was made deliberately during Phase 5 planning to resist scalar-collapse pressure. A single scalar metric could have invited interpretive simplification. By committing to two framings, one scalar and one ribbon, the design preserves the dimensional information that a single number would have discarded.

Both metrics ship as observational outputs only. Their classification (`EXPERIMENTAL`) is included in the canonical-JSON digest material. Promotion to `HEURISTIC` or `OPERATIONAL` requires operator-signed amendment per MYTH-FC clause 2.

3.10 Gate E evidence

The Gate E (infrastructure) evidence cycle produced two complete artifact sets from two prompts on a single device. Table 4 summarizes the results.

Table 4: Gate E artifact summary.

Run identifier	Prompt	F-2 token IDs	`R_info_v0_scalar_A`	normalized	`metric_status`	classification
`gateE_cpu_12600`	"The Titanic sank in"	4 tokens	4.577	0.821	ok	EXPERIMENTAL
`gateE_cpu_3756`	"Argentina won the 2022 World Cup"	8 tokens	1.586	0.613	ok	EXPERIMENTAL

Run identifier	Prompt	F-2 token IDs	`R_info_v0_scalar_A`	normalized	`metric_status`	classification
Interpretation	*not claimed for individual values or for the difference between values; the values are descriptive contents of the sealed artifacts. See §3.10 narrative and §5.3 items 9 and 10.*					

Both runs additionally produced `R_info_v0_ribbon_D` records, each containing 4 surfaces by 8 commits, with `metric_status: ok` and classification `EXPERIMENTAL`.

The values are recorded here as descriptive contents of the sealed artifacts. The paper does not interpret the values, does not interpret the difference between the two values, and does not claim that the values measure anything in particular. The values' semantic meaning, if any, is the subject of future controlled experiments inside the laboratory (Section 6).

The full Gate E evidence (commits, branches, run-scale metrics, run-scale ribbons, manifests, replay verification summaries, gate summaries) is sealed inside the ARC-1 archive `ARC-1_Macro_Gate_E_sealed.zip` (SHA-256 per Table 1). All four Phase 5 record streams were independently verified RP-1 clean by an independent re-verification using the bundle's own verifier code. The replay summaries report `compute_calls_observed = 0` across both new metric streams.

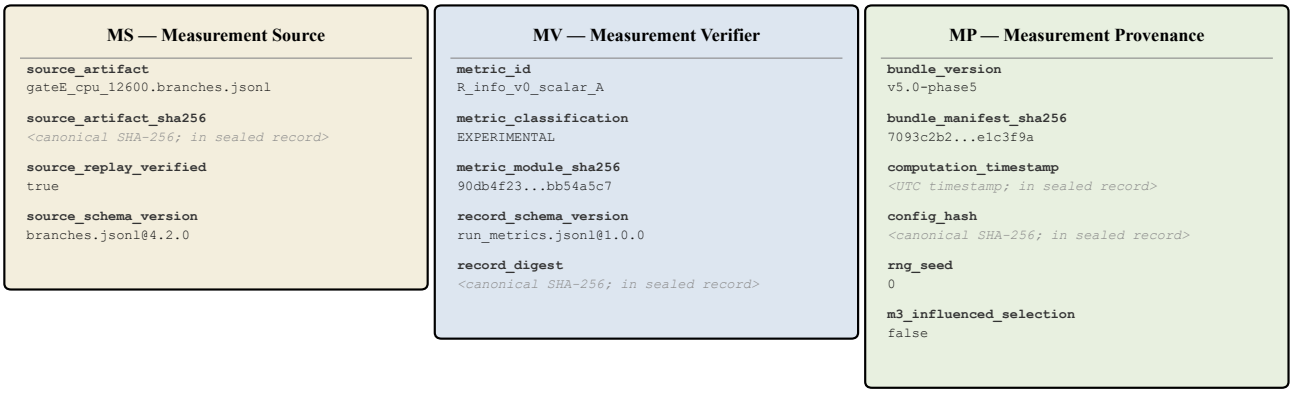
The bundle-identity cross-check confirmed that the operator ran the v5.0-phase5 bundle as delivered, without modification. The `metric_module_sha256` recorded in all four Phase 5 records (`90db4f23f66d567afc0c4d7badc7795725ff694cb49a192dcade7948bb54a5c7`) matches the byte-hash of `rsp/m6_run_scale.py` in the shipped bundle.

Every record in every Phase 5 stream additionally carries the full MV / MS / MP provenance block. Figure 6 shows the block structure for one specific record (the scalar metric on the first Gate E run), with confirmed values shown literally and schema-only fields marked as placeholders for the values that live inside the sealed archive.

Figure 6. Provenance block carried by every sealed Phase 5 record.

Confirmed values shown literally; <angle-bracket> placeholders mark fields whose exact value lives in the sealed archive and is not reproduced here.

Example record: R_info_v0_scalar_A on prompt "The Titanic sank in" (gateE_cpu_12600)



Every record in every Phase 5 stream carries this full MV / MS / MP block. The block is part of the canonical-JSON digest material and any modification to any field changes the record's digest, which fails RP-1 verification against the sealed archive.

3.11 The F-2 amendment (manifest prompt lineage)

A small but consequential amendment landed during Phase 5. The manifest schema gained two fields: `prompt_token_ids` (an authoritative integer-list lineage anchor) and `prompt_text` (an optional decoded form). Before F-2, runs were identified only by their committed-token signatures, which made cross-prompt comparison awkward and ad-hoc, and F-2 closes that gap directly in the manifest.

The amendment was deliberately structured to be additive only. Neither of the two new fields appears in any commit or branch digest material, and the test `test_f2_does_not_affect_existing_commit_or_branch_digests` verifies that two runs with identical configuration but different prompt text produce byte-identical commit and branch digests. The new fields therefore live in the manifest alone, where they serve as lineage anchors without affecting the existing digest chain. The decoded `prompt_text` field depends on the tokenizer associated with the model checkpoint specified in the manifest (in Gate E, the standard `gpt2` BPE tokenizer); the authoritative lineage anchor is the integer-list `prompt_token_ids`, which is tokenizer-independent given the same checkpoint.

The F-2 amendment was operator-dispositioned during Phase 5 planning (D-P5-6, D-P5-11) and landed in v5.0-phase5 directly.

---

4. Audit Methodology

The laboratory enforces audit honesty through a small number of procedural patterns that were operationalized during implementation. This section documents the patterns that operated during the Phase 5 and Gate E cycle.

4.1 Implementation-stage findings

Five findings were surfaced, diagnosed, and dispositioned during the cycle. Each is recorded in the sealed archive without silent absorption.

A reader might note that five findings is a low count for a complete Phase 5 implementation cycle. The count is not normalized against any baseline of expected findings, and the discipline does not commit to a target rate. The procedural pattern (defer-and-surface, no silent action, audit-honest framing; see §4.3) is intended to produce a faithful count rather than a particular count. Findings are recorded because they were not absorbed silently, regardless of how many that turns out to be. The five reported here are the findings that surfaced; the manuscript does not claim that no other findings could have surfaced under a more adversarial implementation regime or under different operator-tooling configurations.

Table 5: Implementation-stage findings (Phase 5 and Gate E review cycle).

#	Origin	Class	Resolution
1	Step A	test-side error in F-2 digest-invariance test; the test held distinct `experiment_id` values, and `branch_record_digest` includes `experiment_id`	the test was corrected to hold `experiment_id` fixed; the code was already correct
2	Step D	carry-forward count refinement; three boundary-test renames were applied, not four as anticipated in planning	the manifest records six modified-from-v4.0 files
3	Step E	`verify_bundle.py` print logic was bundle-version-specific	the print logic was generalized; `scripts/` is not in the manifest's tracked file list
4	Step C / Gate E	`R_info_v0_scalar_A` on the smoke backend produced approximately 29.17 (random weights); on real GPT-2 the runs produced 4.58 and 1.59	recorded as a structural property of the framing-A denominator; no semantic claim
5	Gate E review	partial reviewer independence (the same instance produced bundle and review)	disclosed in `GATE_E_READINESS_ASSESSMENT.md` §10 and carried forward

Findings 1 through 3 are local code, test, or script issues that surfaced and were corrected during implementation. Finding 4 is a structural observation about formula behavior, recorded without semantic interpretation. Finding 5 is a governance-shape observation about reviewer independence and is the most structurally significant of the five.

4.2 Failure-mode discipline (F-α through F-ε)

Five named failure modes describe shapes that the discipline is designed to detect or prevent. The first four were anticipated in implementation-planning documentation [@phase5implplan §05] prior to implementation. The fifth was named during the post-Gate-E consolidation cycle.

Table 6: Named failure modes.

Mode	Shape	Anticipated	Triggered in this cycle?
F-α	refactor-driven RP-2 erosion (duplicate digest helpers fused into a shared module)	impl-plan doc 05	no; duplication preserved; an AST scan would have caught fusion
F-β	closed-vocabulary leakage via `diagnostic` field	impl-plan doc 05	no; no record carried `diagnostic`; the banned-substring scan was clean
F-γ	silent classification drift (a metric ships HEURISTIC or OPERATIONAL)	impl-plan doc 05	no; both Phase 5 metrics ship EXPERIMENTAL; classification is in digest material
F-δ	M7-SEAL drift in prose	impl-plan doc 05	no; closure documents use bounded language
F-ε	**retrospective mythology**: bounded observations editorially converting to general claims	post-Gate-E consolidation (this work)	**observed as a shape; not formally adopted as a binding invariant**

The "anticipated, not triggered" status for F-α through F-δ admits two readings, and both are consistent with the evidence in hand: that the structural counter-pressures operated, or that the relevant stresses did not occur in this cycle. The manuscript does not claim the first reading over the second; the discipline is to record the observation, not to interpret its absence.

The fifth failure mode, F-ε, deserves explicit comment. The shape is documentary, not behavioral. With a complete cycle in hand, retrospective documents (closure reports, observations, paper drafts) can quietly convert specific bounded observations into general claims through small editorial steps. The final step in such a drift can be an authorization decision that rests on a foundation the earlier evidence never actually supported. F-ε is not about bad faith. It is about how the natural shape of retrospective documentation tends toward generalization, unless explicit counter-pressure is applied. The shape was named during the consolidation cycle (`IMPLEMENTATION_GOVERNANCE_OBSERVATIONS_v1.md` §3) and is recorded here without formal adoption as a binding invariant. Whether F-ε becomes binding is an operator disposition that has not been issued.

The present manuscript itself sits within the F-ε attack surface. A long-form retrospective document is exactly where the shape would emerge. The structural counter-pressure throughout this manuscript is the claim audit (Appendix A), the explicit non-claims (Section 5), and the editorial rule that every quantitative claim cites its sealed-evidence source by filename.

4.3 Procedural patterns

Three patterns operated during the cycle at the human-procedure layer.

Defer-and-surface. Ambiguity encountered during implementation was surfaced to the operator with explicit framing, rather than resolved silently. Every governance-shape decision in the Phase 5 cycle (D-P5-1 through D-P5-12) appears on the operator's record.

No silent action. Every finding that surfaced was recorded: test-side bugs, count refinements, script generalizations, the transcription typo at seal time. The five findings of Section 4.1 exist as records because no event was absorbed silently.

Audit-honest framing. Statements about the program were bounded to what the cycle's evidence supported. The canonical bounded statement (Section 5.1) is shorter and weaker than any of its plausible generalizations. That boundedness was preserved through closure and consolidation.

4.4 Bundle-identity cross-check

A specific cross-check was used at Gate E to confirm that the operator ran the shipped bundle without modification. The `metric_module_sha256` recorded in every Phase 5 record was independently recomputed against the v5.0-phase5 bundle's `rsp/m6_run_scale.py`. The hashes matched byte-for-byte across all four records. This binds the Gate E evidence to a specific code surface that any reproducer can independently verify.

4.5 The master-ledger SHA correction

At the moment of Gate E sealing, the operator-supplied `archive_sha256` in the master ledger entry contained 65 hexadecimal characters. An independent SHA-256 computation of the uploaded archive yielded the canonical 64-character value. The discrepancy was a transcription typo, with one extra digit at position 30 of the recorded string. The discrepancy was surfaced in the reviewer-side seal receipt before the master ledger entry hardened, and the operator corrected the entry to the canonical SHA-256.

This event is recorded for two reasons. First, the audit-honest discipline operated at a consequential transcription moment of the cycle. Second, it is the only drift event surfaced during the cycle. The bounded reading of the event is that the discipline surfaced one transcription-typo class of drift under one specific set of conditions. The bounded reading does not support generalizing to "the discipline surfaces drift in general." Future stress classes are not addressed by this single event.

---

5. What Was Demonstrated and What Was Not Demonstrated

A paper whose contribution is the construction of an apparatus carries an unusual obligation. It needs to say what the apparatus was demonstrated to do, but it also needs to say, equally clearly, what it was not demonstrated to do. The two enumerations below follow the claim audit (Appendix A) and are intended to carry equal weight. A reader who reads only Section 5.2 (demonstrations) without Section 5.3 (non-demonstrations) will have a partial and misleading picture of what the work supports.

5.1 The canonical bounded statement

The single substantive claim the work supports is the following.

> The infrastructure pipeline operates and remains governance-consistent under the tested implementation-stage conditions.

This is the closure language sealed in the Gate E archive (`docs/GATE_E_CLOSURE.md` §1). It is operative through Gate E. It is the only operative substantive claim the work currently makes about its own outputs.

5.2 What was demonstrated

The evidence in this cycle supports the following claims, each traceable to specific sealed evidence.

1. *Replay integrity.* Six artifact stream types are independently replay-verifiable by canonical-JSON digest recomputation. The Gate E evidence cycle produced six clean stream verifications.
2. *RP-2 replay neutrality.* Replay verifiers do not invoke compute functions; `compute_calls_observed = 0` in all sealed Gate E run-scale records.
3. *Provenance continuity.* Every Phase 5 record carries the full MV/MS/MP provenance block.
4. *ARC-1 archive immutability.* Each gate's sealed archive is referenced by SHA-256 in successor bundles; the predecessor lineage is cryptographically verifiable.
5. *Shadow-mode metric execution.* Metric computation does not influence token selection in Gate E runs; this is verified by both a behavioral anchor test and an AST-scan structural test.
6. *Invariant preservation under nine binding invariants.* All nine invariants held throughout Phase 5 implementation and Gate E evidence.
7. *Audit-honest defect surfacing.* Five implementation-stage findings were surfaced, diagnosed, and dispositioned without silent absorption.
8. *Implementation-stage governance survival under tested stress classes.* The governance model designed during Phase 5 planning held under the following stress classes encountered during the cycle: (a) operator-side execution of the delivered bundle without modification, verified by the `metric_module_sha256` cross-check (§4.4); (b) re-verification of all six artifact streams by independent verifier code paths under RP-2; (c) execution under documented operator dispositions D-P5-1 through D-P5-12, with all dispositions traceable in the master ledger; (d) surfacing and dispositioning of five implementation-stage findings without silent absorption (§4.1); (e) correction of the master-ledger SHA transcription typo at the moment of sealing (§4.5). Stress classes that were *not* encountered in this cycle (including operationalization pressure, multi-engineer maintenance, retrospective documentation at scale, semantic pressure, and cross-machine execution on real GPT-2) are enumerated as explicit non-demonstrations in §5.3.

5.3 What was not demonstrated

The work explicitly does not demonstrate any of the following. These are not provisional gaps. They are fixed non-demonstrations that define the empirical ceiling of the work.

1. The work does not demonstrate truth detection.
2. The work does not demonstrate hallucination detection.
3. The work does not demonstrate semantic-correctness measurement.
4. The work does not demonstrate informational validity in any technical sense.
5. The work does not demonstrate reasoning superiority of any decoder.
6. The work does not demonstrate decoder superiority of any form.
7. The work does not demonstrate measurement of "reality" or "illusion."
8. The work does not demonstrate any M7 (decision/intervention) capability.
9. The work does not demonstrate that the run-scale metric values mean anything in semantic terms.
10. The work does not demonstrate generalization beyond the two Gate E prompts, the `gpt2` model

checkpoint, or the operator's specific machine.

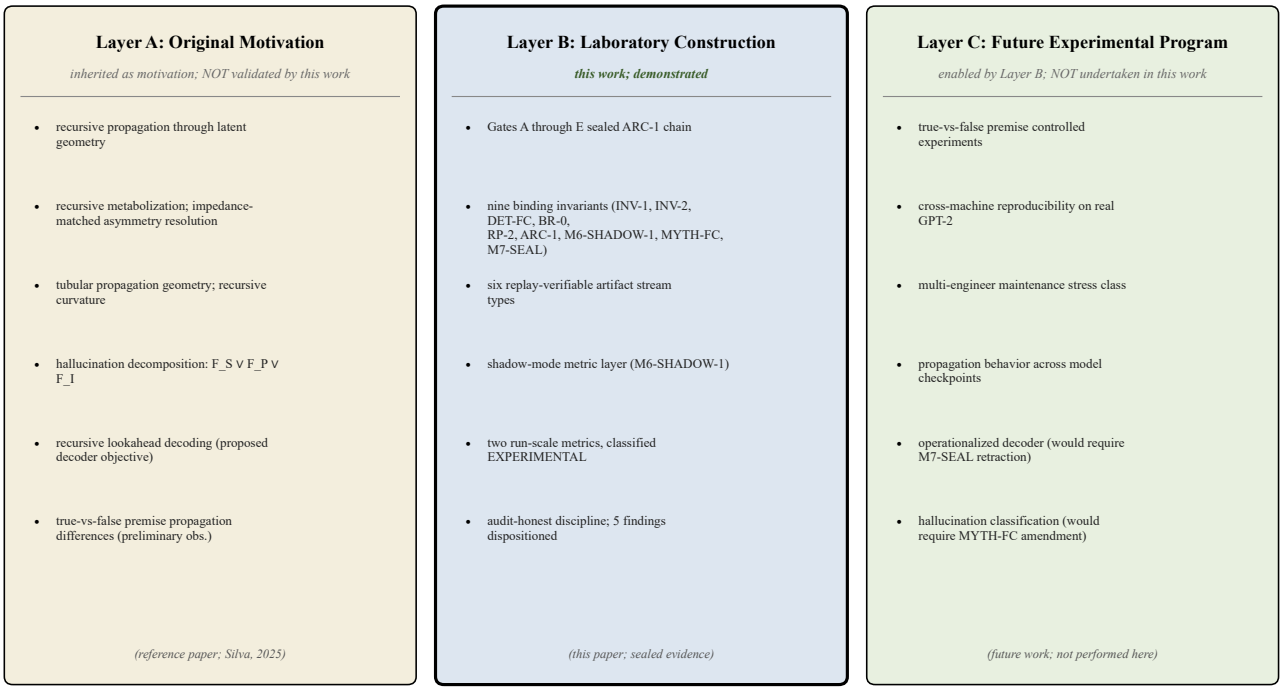
11. The work does not demonstrate cross-machine reproducibility on real GPT-2.
12. The work does not demonstrate that the structural counter-pressures (RP-2, MYTH-FC, M7-SEAL) will hold under future stress classes, including operationalization pressure, multi-engineer maintenance, retrospective documentation at scale, semantic pressure, or other classes not yet encountered.
13. The work does not demonstrate that past operator dispositions remain correct under conditions not yet tested.
14. The work does not validate the reference paper's hypotheses. It provides the apparatus that could, in future controlled experiments, test some of them.

5.4 The layer boundary

Figure 8 visualizes the boundary between Layer A (motivation), Layer B (laboratory construction, the present work), and Layer C (future experimental program).

Figure 8. The three layers maintained throughout the manuscript.

The regions are non-overlapping. Layer A constructs are not validated by Layer B; Layer C proposals are not findings of this work.



Maintaining this separation is the structural counter-pressure against F-ε (retrospective mythology). See CLAIM AUDIT (Appendix A).

The three layers are non-overlapping. Reading the paper without preserving the layer boundary undermines the work's structural purpose.

---

6. Future Experimental Program (Layer C)

With the laboratory in place, a number of experiments become possible that were not possible before. The most important of these is the controlled comparison that motivated the reference paper, but several others follow naturally from the same substrate. The descriptions below are presented as proposed work, not as findings of the present paper, and each is explicit about what would need to be true (in evidence, in invariants, or in operator disposition) before the work could be undertaken honestly.

6.1 True-versus-false-premise controlled experiments

The reference paper [@silva2025recursive] reported preliminary observations suggesting measurable internal-surface differences between true and false premises. That observation was the historical motivation for the laboratory. It has not been tested under the laboratory's governance regime.

A controlled experiment inside the laboratory would proceed as follows. A balanced set of paired prompts that differ only in the truth value of an asserted proposition would be prepared. Each prompt pair would be executed under identical lookahead-harness configuration. The full set of run-scale metrics would be computed on the sealed evidence. The records would be replay-verified under RP-1 and RP-2. The resulting `R_info_v0_*` record set would be statistically analyzed as a function of the premise condition. Pre-registered statistical methods would form part of that future experimental design and are not specified here. The classification of the metrics would remain `EXPERIMENTAL` throughout. Any move to `HEURISTIC` or `OPERATIONAL` would require operator-signed amendment under MYTH-FC.

Such an experiment is not undertaken in this paper. The conditions under which it could be attempted honestly now exist, but the experiment itself, including the design of the paired-prompt set, the statistical pre-registration, and the subsequent analysis, is work that remains to be done.

6.2 Cross-machine reproducibility

Gate E ran on a single device. Cross-machine reproducibility on real GPT-2, the property that two independently configured machines running the same bundle on the same prompts produce byte-identical records, is asserted by determinism notes in the bundle but is not exercised by this work. Establishing cross-machine reproducibility is a precondition for any conclusion about behavior of the system, rather than behavior on a single machine.

6.3 Other research directions enabled

The laboratory supports, in principle, the controlled study of several other questions:

- propagation dynamics conditioned on prompt structure;
- branch-trajectory differences under variations of lookahead width and depth;
- run-scale metric behavior across model checkpoints, with explicit re-disposition if checkpoints other than `gpt2` are introduced;
- the behavior of replay-verifiable metrics under multi-engineer maintenance (a stress class that is explicitly listed as untested in Section 5.3).

Each of these is work that the laboratory makes possible without itself undertaking. None is performed in the present paper.

6.4 Out of scope under the current invariant set

Some research directions anticipated by the reference paper remain out of scope under the current binding invariants.

Recursive lookahead decoding as an operationalized decoder would require modifying token selection based on metric outputs, which is forbidden under M6-SHADOW-1 and structurally precluded under M7-SEAL. Operationalizing such a decoder would require operator-signed retraction of both invariants, and no such retraction has been issued or is anticipated under the present scope.

Hallucination detection or classification would require a `metric_classification = HALLUCINATION_TYPE_X` entry. Introducing such an entry would be a substantive amendment to MYTH-FC clause 2's closed enum and would require operator disposition.

Operationalization of Layer-A constructs as L-1 measurables (recursive metabolization, tubular geometry, curvature, invariant compression) cannot occur as record field names under INV-2 and MYTH-FC clause 1. Introducing them would require explicit invariant amendment.

These are not failures of the laboratory. They are bounded scope decisions consistent with the work's stated contribution.

---

7. Methods

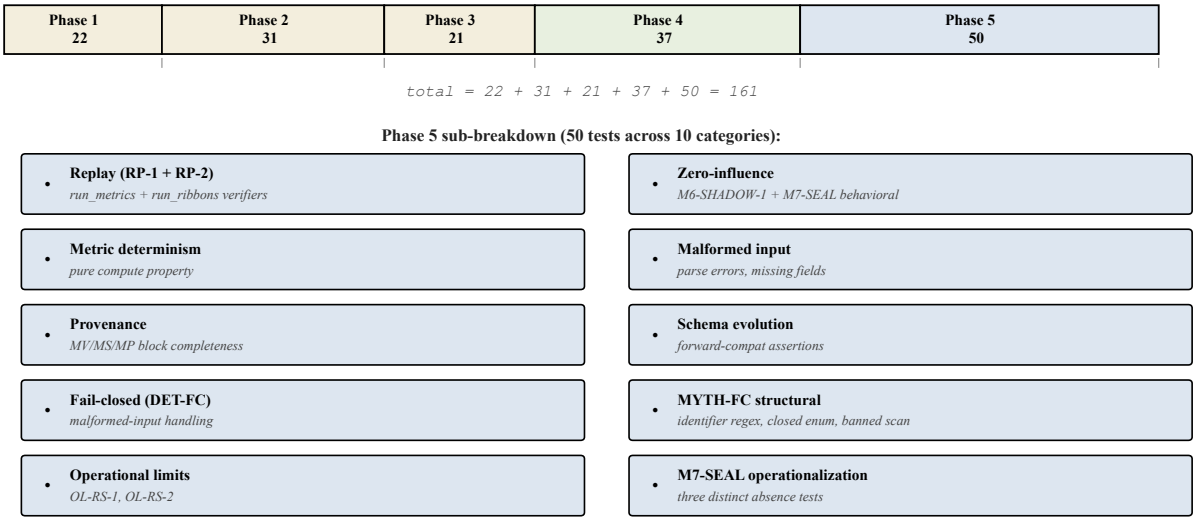
7.1 Reproducibility

Full reproducibility instructions appear in Appendix D. In summary, the v5.0-phase5 bundle is the canonical implementation artifact. Its manifest SHA-256 is `7093c2b23e09786b22d8d65b6ea20f5529095d23310adb77fed1da553e1c3f9a`. The Gate E evidence is sealed inside the ARC-1 archive `ARC-1_Macro_Gate_E_sealed.zip`, with canonical SHA-256 per Table 1. Replay verification proceeds via `scripts/verify_replay.py` against the sealed evidence directory. The expected outcome is `ALL CLEAN` with `compute_calls_observed_total = 0`.

The full 161-test suite at v5.0-phase5 (Phase 1: 22, Phase 2: 31, Phase 3: 21, Phase 4: 37, Phase 5: 50) was passing at Gate E seal, and the same test suite is the standard pre-evidence gate for any future Gate E re-execution. Figure 7 shows the composition by phase, with the Phase 5 sub-breakdown into the ten test categories introduced in this cycle.

Figure 7. Test-suite composition at the v5.0-phase5 seal (161 tests, all passing).

Counts come directly from the sealed test pass evidence; the Phase 5 sub-breakdown shows the ten new test categories introduced in this cycle.



All 161 tests pass at v5.0-phase5 seal. Cross-phase boundary tests (renamed in Phase 5) keep the M7-absence assertion in force.

Phase 5 sub-category labels summarized from PHASE5_CLOSURE_REPORT.md §3.5; refer to the closure report for verbatim names and per-category test counts.

7.2 Software and hardware

The Gate E evidence was produced on a single device with the following stack: Python 3.12.4, PyTorch 2.6.0+cu124 (CPU build), HuggingFace Transformers 4.57.6, GPT-2 small (`gpt2` checkpoint), and `CUBLAS_WORKSPACE_CONFIG=:4096:8` for determinism. The full set of pinned versions appears in `configs/requirements.txt` of the v5.0-phase5 bundle, whose bundle manifest SHA-256 (`7093c2b23e09786b22d8d65b6ea20f5529095d23310adb77fed1da553e1c3f9a`) together with the `requirements.txt` content hash provides the reproducible specification.

Both Gate E runs used `seed=0`, `n_commits=8`, `lookahead_depth=6`, and `lookahead_width=6`. Both prompts were tokenized using the standard `gpt2` BPE tokenizer. The resulting token IDs are recorded in the F-2 fields of each manifest.

7.3 Authorship and AI-assisted tooling disclosure

The author (Ivan Silva, ORCID 0009-0005-2284-8891) is the principal author, governance authority, and sole substantive author of the work. The author's authorship statement is sealed within the Gate E archive (`docs/GATE_E_CLOSURE.md`, "Project Director and Authorship Signature" section).

AI-assisted tooling was used during the production of this work under the author's direction and review. The specific system used was Claude, an AI assistant developed by Anthropic, accessed through the Claude product interface, and used as a coding assistant, document-drafting assistant, and figure-script-generation assistant. Contributions of the tooling included: (a) code implementation under operator direction, with all governance dispositions (D-P5-1 through D-P5-12 and the prior gate-level dispositions) issued by the operator and recorded in the master ledger; (b) drafting of governance, planning, and closure documents under operator direction, with substantive content, scope, and structural decisions supplied by the operator; (c) drafting of this manuscript and an internal review pass on the draft; (d) generation of the figure-production scripts and the figures themselves under the provenance discipline documented in `docs/FIGURE_MANIFEST.md`. The tooling did not act as an independent rights holder, an author, or a governance principal, and all substantive decisions remained operator-dispositioned. The seal-time authorship statement (sealed within the Gate E archive, `docs/GATE_E_CLOSURE.md`, "Project Director and Authorship Signature" section) is the binding authorship record. This posture is consistent with the seal-time authorship statement and is recorded here in the methods section per operator disposition.

7.4 Internal review independence disclosure

The Gate E reviewer-side readiness assessment (`docs/GATE_E_READINESS_ASSESSMENT.md` §10) disclosed that the same AI-assisted tooling instance produced the v5.0-phase5 bundle and reviewed the operator's Gate E artifacts. The same partial-independence finding applies to the present manuscript. The same AI-assisted tooling produced the manuscript draft and performed an initial internal review pass on that draft. This is structurally analogous to a single-engineer review of the engineer's own work.

At each governance stage, the division of labor between the AI tooling and the operator was as follows. *Planning* dispositions (which gates to seal, which invariants to bind, which D-P5-N options to accept, which findings to record): operator-issued, AI-tooling-drafted-but-only-when-asked. *Implementation* (code, tests, scripts): AI-tooling-drafted under operator review, with all substantive structural decisions issued by the operator. *Sealing* (running the harness, computing archive SHA-256s, sealing the ARC-1 archive): operator-side only. *Review* (the Stage 4 internal adversarial pass; the readiness assessment; this manuscript): AI-tooling-drafted under operator direction and review, with the structural limitation that the same instance produced both the artifact and its review.

The structural counter-pressure operating against this limitation includes several distinct mechanisms. The verifier code itself is structurally independent of the compute code by RP-2 construction, AST-scan-enforced. The operator provides substantive direction and review at every governance disposition. The claim audit (Appendix A) constrains the manuscript independently of the prose itself.

The author recommends, but does not require, that at least one independent human reviewer outside Carlonoscopen review the manuscript prior to publication. The recommendation is operationally a non-blocker, and the manuscript can proceed without it, but the review chain would be measurably tighter with an external reviewer in it than without.

7.5 Statistical methods

The Gate E evidence cycle is not a statistical study. The two prompts do not constitute a sample for hypothesis testing. The run-scale metric values are not subjected to inferential statistics in this paper. Section 6 outlines the controlled experiments that would produce a sample appropriate for statistical analysis. Such analyses are not performed in this work.

---

8. Limitations

The work has explicit limitations beyond those enumerated in Section 5.3.

Single-cycle stress exposure. The nine binding invariants were tested under one implementation-stage cycle (Phase 5 / Gate E). MYTH-FC and M7-SEAL specifically were adopted during Phase 5 and have only been tested in this one cycle. As noted in §3.3, this means the cycle constitutes a first demonstration of those two invariants rather than an independent validation. Whether all nine invariants hold under future stress classes, including operationalization pressure, multi-engineer maintenance, cross-organizational use, and semantic-escalation pressure, is not established.

Single-machine evidence. Gate E produced evidence on a single device. The asserted determinism of the pipeline is documented (`DETERMINISTIC_EXECUTION_NOTES.md`) but is not exercised across machines.

Partial reviewer independence. Section 7.4 disclosed the structural overlap between the implementer of the bundle and the reviewer of its evidence. The recommendation for an external human reviewer outside Carlonoscopen remains open.

Single-model evidence. Gate E used GPT-2 small. Whether the run-scale metrics behave structurally similarly on other model checkpoints is unknown. Introducing a different checkpoint would be a substantive experimental decision requiring re-disposition.

Two-prompt sample. The Gate E evidence consists of two prompts. This is sufficient to confirm that the pipeline operates end-to-end on real-GPT-2 evidence. It is insufficient for any statistical claim about behavior across prompts.

F-ε is named but not binding. The fifth failure mode was named during consolidation but has not been adopted as a binding invariant. The retrospective-mythology shape is acknowledged structurally in this paper through the claim audit and the explicit non-claims. Formal binding-invariant status is an operator disposition that has not been issued.

The reference paper's framework remains exploratory. The Layer A constructs underwriting the original research motivation remain at the same evidential status they had in the reference paper. The laboratory does not validate them. It does not invalidate them. It makes their controlled study possible.

---

9. Discussion

The substantive contribution of this paper is the laboratory itself, rather than a finding about language models or a metric value. The strongest claim the work supports is the bounded form of Section 5.1: the infrastructure pipeline operates and remains governance-consistent under the tested implementation-stage conditions. A weaker statement would not have been honest given the work that was done; a stronger statement is not warranted by the evidence in hand.

Two observations are worth recording in this section.

The first is that the apparatus exists, and the experiments remain to be conducted. This is by design rather than by oversight. The laboratory's contribution is precisely that future experiments, including the controlled true-versus-false-premise experiments that motivated the reference paper, can now be performed under conditions where every measurement is replay-verifiable, provenance-complete, and structurally insulated from interpretive escalation. This is a different kind of scientific contribution than the report of an empirical finding. It is the construction of the substrate on which empirical findings could honestly rest, and that substrate has its own properties (replayability, provenance, audit-honesty, bounded vocabulary) that can be evaluated independently of whatever findings eventually use it.

The second observation is that the negative space of this work carries more weight than the positive space. The explicit non-claims (Section 5.3), the structural absence of M7, the structural absence of Layer-2 vocabulary in record formats, and the structural absence of compute calls during replay are all absences that the laboratory makes operationally checkable. The discipline of testing absences (no `m7_*` module exists, no compute call occurred during replay, no banned substring appeared in any record) rather than only testing presences is, in this author's view, one of the laboratory's more consequential structural commitments. It is also one of the easier commitments to underweight for readers who expect a measurement apparatus to be characterized by what it measures, rather than by what it refuses to measure.

The reference paper's original research question, concerning how reality and illusion might be distinguished within recursive propagation dynamics, is not answered by this work. The laboratory that could honestly attempt to answer it now exists, and whether the answers (when they eventually come) support the reference paper's framework, partially support it, or undermine it, is an open question that the laboratory now makes empirically tractable. Either outcome would be scientifically valuable, and either outcome would be honest under the discipline established here. The present work commits to neither, and that non-commitment is itself a deliberate property of the apparatus.

---

10. Conclusion

This paper documents the construction of a governance-bounded, replay-verifiable, provenance-complete experimental laboratory for the study of language-model internal dynamics. Across five sealed Macro Gates (A through E) under nine binding invariants, the work assembles a cryptographically anchored evidence chain, a structurally enforced separation between measurement and intervention, a closed-vocabulary discipline against silent interpretive escalation, and an audit-honest record of five implementation-stage findings that surfaced and were dispositioned during the cycle.

The original Reality versus Illusion research question that motivated the program remains open. The Recursive Propagation Theory framework that proposed an interpretive model for the observed phenomena remains a hypothesis-generating framework, not a validated theory. The laboratory required to study the framework rigorously is now operational. It has demonstrated replayability across six artifact stream types, provenance completeness on every record, governance survival under the implementation-stage stress classes

encountered during Phase 5, and audit-honest discipline that surfaces drift rather than absorbing it silently. It has not demonstrated the underlying hypotheses of the reference paper, and this manuscript does not claim that it has.

The significance of this work is structural rather than empirical. Future controlled experiments, including the true-versus-false-premise comparison that motivated the reference paper, can now be performed under conditions of replay-verifiability, provenance completeness, and bounded interpretation that were not previously available. The laboratory does not predict whether those future experiments will support the reference paper's framework, partially support it, or undermine it. Either outcome would be scientifically valuable, and either outcome would be honest under the discipline established here.

The laboratory now exists. The original questions remain open, but they can be approached under more controlled and reproducible conditions than were previously available.

References

[silva2025recursive] Silva, I. (2025). *Recursive Propagation Geometry and Hallucination Dynamics: A State-Conditioned Framework for Recursive Inference, Stabilization, and Return Projection.* Carlonoscopen Journal of Coherence Intelligence.

[phase5closure] Silva, I. (2026). *RSP Phase C / Phase 5, Closure Report.* `docs/closure_reports/PHASE5_CLOSURE_REPORT.md`, v5.0-phase5 bundle, sealed within Macro Gate E ARC-1 archive (SHA-256: `e122683f…777b0`).

[gateEclosure] Silva, I. (2026). *Macro Gate E, Closure Document (Infrastructure Gate).* `docs/GATE_E_CLOSURE.md`, sealed within Macro Gate E ARC-1 archive.

[phase5implplan] Silva, I. (2026). *RSP Phase 5 Implementation-Planning Package.* `governance/phase_5_impl_planning/`, v5.0-phase5 bundle, sealed within Macro Gate E ARC-1 archive.

[masterledger] Silva, I. (2026). *RSP Master Program Ledger v1.* Post-Gate-E consolidation snapshot.

[obsdoc] Silva, I. (2026). *RSP Implementation Governance Observations v1.* Post-Gate-E consolidation snapshot. Introduces F-ε.

[structreview] Silva, I. (2026). *RSP Structural Review Pass v1.* Post-Gate-E consolidation snapshot.

Acknowledgments

The author thanks Claude, an AI assistant developed by Anthropic, used during implementation, document drafting, figure-script generation, and an initial internal review pass under the author's direction and review. Per the seal-time authorship statement, this tooling operates as part of the workflow and is not an independent rights holder, author, or governance principal. The granular contribution scope is documented in §7.3.

Data and Code Availability

The complete v5.0-phase5 bundle (24 tracked source and test files, plus supporting scripts, manifests, governance documents, and closure reports) is sealed within the ARC-1 archive `ARC-1_Macro_Gate_E_sealed.zip` (canonical SHA-256: `e122683fa28e0539846bfece64b0c4c85cb5690101a6fdfc3ec26524cff777b0`).

The Gate E evidence cycle's 14 artifact files (commits, branches, manifests, gate summaries, run-scale records, replay verification summaries from two runs) are sealed within the same archive under `evidence/phase5_gateE/`.

All figures in this manuscript are reproducible from Python scripts in the `scripts_figures/` directory of the manuscript source tree. Each figure ships in two formats with identical base filenames: a Scalable Vector Graphics file (`figure_NN_*.svg`, used as the canonical reference in the manuscript Markdown) and a Portable Network Graphics raster (`figure_NN_*.png`, provided as a companion for tooling that does not render SVG natively). The two formats are produced by the same generation script in a single matplotlib backend call, so they render identical content. Section 7.1 and Appendix D specify the reproducibility steps for both the figures and the underlying Gate E evidence.

Reproducibility instructions are provided in Appendix D.

Author Contributions

Ivan Silva: conceptualization, methodology, software, formal analysis, investigation, writing of original draft, writing review and editing, project administration, supervision, and funding acquisition. Sole substantive author.

AI-assisted tooling was used in code production, document drafting, and an internal review pass under the author's direction. This tooling does not constitute authorship per the seal-time authorship statement.

Competing Interests

The author declares no competing financial interests. The author holds three concurrent roles that should be disclosed together: principal author of the manuscript, principal of Carlonoscopen LLC (which operates the Carlonoscopen Journal of Coherence Intelligence as the publication venue), and sole governance authority over the program documented in the manuscript (including operator dispositions, gate-sealing authorization, and invariant binding). This is a structural concentration of authority and is acknowledged here as such, not minimized. The structural counter-pressures operating against this concentration include: (a) the claim audit (Appendix A), which constrains the manuscript independently of the prose; (b) the partial-independence disclosure in §7.4; (c) the author's standing recommendation for an external human reviewer outside Carlonoscopen prior to publication; (d) the seal-time authorship statement, which is cryptographically anchored and immutable; and (e) the public reservation of the DOI (10.5281/zenodo.20564072), which fixes the manuscript's identity prior to publication. None of these counter-pressures fully substitutes for an independent editorial process. The author considers the external-reviewer recommendation to be the central outstanding mitigation and would welcome its activation.

Appendices

The following appendices accompany this manuscript as separate Markdown files in the manuscript source tree at `appendices/`:

- Appendix A: Full Claim Audit, `appendices/appendix_A_claim_audit.md`
- Appendix B: Phase 5 Governance Disposition Record (D-P5-1 through D-P5-12), `appendices/appendix_B_phase5_dispositions.md`
- Appendix C: Test-Suite Composition, `appendices/appendix_C_test_suite.md`
- Appendix D: Reproducibility Checklist, `appendices/appendix_D_reproducibility.md`
- Appendix E: Glossary (Layer-1 vs Layer-A Vocabulary), `appendices/appendix_E_glossary.md`

Appendix A (the claim audit) and Appendix D (the reproducibility checklist) are required reading for any substantive review of this manuscript. Appendices B, C, and E are recommended for governance, testing, and terminology orientation.

End of manuscript v0.2-draft. Status: Draft for Internal Review. Reserved DOI: 10.5281/zenodo.20564072. Awaiting operator review prior to internal peer review pass and freeze.