

Reflective Diagnostics

A Modular Instrument for Structural Audit of Language Model Reasoning

Nicolas Holm

EnlightenedAI Research Lab

Montreal, Canada

research@enlightenedai.ai

ABSTRACT

This paper defines Reflective Diagnostics, a modular instrument for structural audit of reasoning artifacts generated by large language models. The instrument operates exclusively on produced outputs and does not access internal model states, training data, or latent representations. It measures structural properties including claim dependency organization, evidentiary anchoring, structural coverage, stability under deterministic replay, and transformation under reflective re-evaluation.

Reflective Diagnostics produces structured evidence artifacts that substantiate reported signals. Measurement is reproducible under defined audit protocols and deterministic execution conditions. The instrument does not evaluate factual correctness, alignment, semantic validity, or model intent. Its claims are limited to structural organization and transformation behavior within reasoning traces.

By formalizing structural measurement as an independent analytical layer, the instrument enables auditability distinct from benchmark performance, preference evaluation, or training-oriented process supervision. This paper establishes the definitional scope, evidentiary posture, and methodological boundaries of the instrument without disclosing implementation mechanics.

1 Why Structural Reasoning Requires Forensic Measurement

Evaluation of large language models has traditionally focused on outcome-based metrics such as task accuracy, benchmark performance, and human preference alignment [1]. These approaches assess whether a response appears correct, fluent, or useful.

Outcome correctness, however, does not guarantee structural soundness of the reasoning that produced it. A response may reach a correct conclusion while containing unsupported intermediate steps, incomplete dependency chains, or internally inconsistent justifications. Such structural properties are often invisible in aggregate outcome-based evaluations.

Prior analytical work has documented instances in which language models generate coherent and persuasive responses while exhibiting reasoning inconsistencies, unsupported inferences, or fragile dependency structures [2, 3]. These findings highlight a distinction between surface plausibility and structural validity.

Existing evaluation paradigms provide limited visibility into this distinction. Benchmarks summarize task performance but do not localize structural breakdown within a reasoning trace. Preference-based evaluations capture judgments of quality but do not expose claim-level dependency organization. Process-oriented methods evaluate intermediate reasoning steps yet remain focused on utility or correctness signals rather than structural coherence as an independent object of measurement [3, 4].

The absence of explicit structural instrumentation creates a gap between observable reasoning artifacts and formal auditability. Without a dedicated measurement layer, structural instability, unsupported claims, and coherence-first validation patterns may remain undetected.

Reflective Diagnostics addresses this gap by treating reasoning artifacts as analyzable structural objects. The instrument operates post-hoc, measuring claim organization, dependency integrity, stability under deterministic replay, and structural transformation under reflection. It does not replace existing evaluation paradigms; it introduces a forensic layer that makes structural posture observable.

The purpose of this paper is to define that measurement layer formally and delimit its scope.

2 What This Instrument Measures and Explicitly Does Not Measure

Reflective Diagnostics is a modular forensic instrument that measures structural properties of reasoning artifacts generated by large language models.

Measurement is applied after generation and operates exclusively on produced outputs. The instrument does not access internal model states, training data, system prompts, hidden chains of thought, or latent representations.

The object of analysis is the structure of reasoning as expressed in generated artifacts and its transformation under reflective re-evaluation. The instrument reports structural signals and does not infer internal cognition, intention, or mental states.

2.1 Structural Properties Measured

Across its analytical layers, the instrument measures:

- Segmentation of reasoning artifacts into analyzable claim units
- Dependency relationships between claims
- Presence or absence of explicit evidentiary anchoring
- Orphaned or unsupported claims
- Structural coverage within a reasoning trace
- Integrity and consistency of dependency chains
- Stability under deterministic replay
- Structural change between an initial reasoning artifact and its reflective counterpart
- Cross-artifact convergence or divergence under identical audit conditions

These measurements characterize structural organization and transformation behavior within reasoning artifacts. They do not evaluate semantic accuracy, task correctness, or normative validity.

2.2 Explicit Non-Claims

Reflective Diagnostics does not measure:

- Truth or falsity of external facts

- Moral correctness or value alignment
- User intent, motivation, or psychological state
- Safety policy compliance or prompt risk classification
- Model “understanding,” cognition, or consciousness
- Clinical, legal, or medical validity
- Internal decision-making processes

The instrument does not claim to determine why a structural pattern occurs, whether it generalizes beyond defined audit conditions, or how it should be mitigated. It reports structural properties under defined audit protocols.

3 System Architecture and Analytical Modules

Reflective Diagnostics is organized as a modular system rather than a linear pipeline. Each module performs a distinct analytical function and may operate independently under defined audit protocols.

The instrument consists of three primary components:

- **Probe Module**
- **Reasoning Module**
- **Reflection Module**

These modules do not form a single sequential chain. Instead, the Probe captures reasoning artifacts, which may then be analyzed within the Reasoning module, the Reflection module, or both.

3.1 Probe Module

The Probe module executes controlled prompt conditions and captures model outputs under defined parameters.

Its function is limited to:

- Recording model responses
- Preserving execution parameters
- Generating structured trace artifacts

The Probe does not perform structural evaluation. It produces artifacts that may later be analyzed by other modules.

3.2 Reasoning Module (Structural Reasoning Analysis)

The Reasoning module analyzes a single reasoning artifact or multiple artifacts generated under identical protocol conditions.

This module performs:

- Claim segmentation
- Dependency mapping
- Structural integrity measurement
- Structural coverage measurement
- Orphan identification
- Stability testing under deterministic replay

When multiple artifacts are analyzed under identical conditions, the Reasoning module also performs **cross-model synthesis** through an internal component referred to as the **Interpreter**.

The Interpreter operates within the Reasoning module and:

- Compares structural posture across models
- Aggregates integrity and coverage signals
- Applies defined aggregation rules
- Produces ranked outputs under protocol constraints

The Interpreter does not evaluate semantic correctness or task performance. It synthesizes structural measurements and may declare a structural winner under defined audit rules.

3.3 Reflection Module (Structural Transformation Analysis)

The Reflection module evaluates structural change between an initial reasoning artifact and its reflective counterpart.

This module measures:

- Structural expansion or contraction
- Changes in evidentiary anchoring
- Dependency reorganization
- Persistence of unsupported claims

Reflection analysis is comparative and intra-artifact by default, focusing on transformation rather than ranking.

Future extensions may introduce a reflective synthesis component capable of cross-model comparison within the Reflection module. However, the Reflection module is conceptually distinct from the Reasoning Interpreter and operates independently.

3.4 Modular Relationship

The system architecture is modular rather than sequential. The Probe feeds structured artifacts into analytical modules, but the Reasoning and Reflection modules are independent analytical environments.

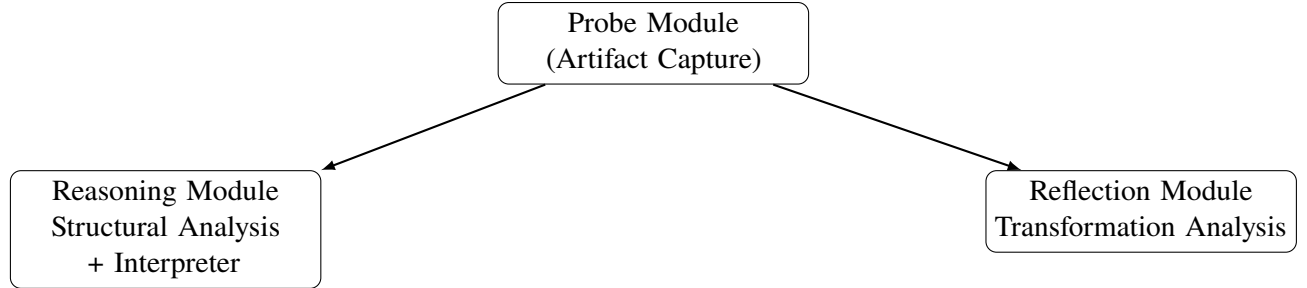


Figure 1: Modular Architecture of Reflective Diagnostics

4 Audit Protocols and Execution Modes

Each audit is executed under an explicit protocol identifier that defines analytical scope, sensitivity, and instrumentation layer.

A protocol specifies:

- Target analytical layer(s)
- Deterministic execution parameters
- Structural measurement boundaries
- Artifact aggregation conditions

An example protocol designation may include:

- RN-L2: A reasoning task evaluated under Structural Analysis Layer conditions designed to surface coherence-first validation patterns.

Protocol identifiers are required for interpretation of evidence artifacts. Structural signals are defined relative to the protocol under which they are generated.

4.1 Deterministic Mode

In deterministic mode, identical inputs and execution parameters yield identical structural measurements. Deterministic execution establishes reproducibility within defined audit conditions.

4.2 Stability Sweep Mode

In stability sweep mode, audits are repeated under controlled parameters to confirm persistence of structural posture. Stability signals distinguish invariant structural patterns from stochastic variation. Stability confirmation does not imply generalization beyond defined protocol conditions.

4.3 Protocol Isolation

Results generated under different protocols are not directly comparable without recalibration. Structural posture is defined relative to protocol scope and execution conditions.

5 Metric Families and Structural Risk Signals

Reflective Diagnostics reports structural signals grouped into defined metric families corresponding to distinct dimensions of artifact organization and transformation. Metrics are disclosed at the level of measurement intent rather than computational implementation.

The instrument reports structural posture, not semantic correctness or normative evaluation.

5.1 Structural Integrity

Structural Integrity metrics assess the internal consistency of dependency relationships within a reasoning artifact.

These signals capture:

- Coherence of claim-to-claim relationships
- Presence of unsupported inference steps
- Consistency of dependency chains across the artifact

Integrity signals indicate whether the reasoning structure maintains internal dependency logic. They do not evaluate factual accuracy or persuasiveness.

5.2 Structural Coverage

Structural Coverage metrics assess the extent to which claims participate in coherent dependency chains.

Coverage signals identify:

- Claims supported by explicit or implicit justification
- Orphaned claims lacking structural anchoring
- Distribution of structural support across the artifact

Coverage reflects structural participation, not truth or evidentiary quality.

5.3 Stability Signals

Stability metrics assess persistence of structural posture under deterministic replay.

These signals indicate:

- Whether measurements remain invariant under identical execution parameters
- Whether structural changes occur under controlled repetition

Stability signals confirm consistency within defined audit conditions. They do not imply robustness across domains or prompts.

5.4 Reflective Transformation Signals

For artifacts subjected to reflective re-evaluation, the instrument reports signals describing structural change between an initial reasoning artifact and its reflective counterpart.

These signals may capture:

- Expansion or contraction of claim structure
- Introduction or removal of evidentiary anchors
- Changes in dependency organization
- Divergence or convergence between reasoning stages

Transformation signals describe structural modification and do not evaluate self-awareness, metacognition, or intent.

5.5 Cross-Artifact Synthesis Signals

When multiple artifacts are analyzed under identical protocol conditions, the instrument may report convergence or divergence across artifacts.

These signals characterize:

- Shared structural posture
- Structural variability under consistent prompts
- Distribution of integrity and coverage signals across artifacts

Comparative reporting remains structural and does not constitute performance ranking or semantic evaluation.

6 Evidence Artifacts and Verification

Each audit conducted under Reflective Diagnostics produces a structured evidence artifact that substantiates reported signals.

An evidence artifact includes:

- A protocol identifier specifying audit conditions
- The analyzed reasoning artifact(s)
- Reported metrics at each analytical layer
- Deterministic execution parameters
- Stability confirmation under replay conditions
- Cross-artifact synthesis outputs, when applicable

The evidence artifact serves as the primary unit of verification. Structural claims are grounded in this record rather than narrative interpretation.

6.1 Deterministic Execution and Stability

Audits may be executed under deterministic conditions in which identical inputs, parameters, and protocol definitions yield identical measurements.

Where stability validation is required, deterministic replay confirms persistence of the reported structure. Stability signals indicate whether measurements remain invariant under controlled repetition.

Such confirmation establishes consistency within defined audit conditions and does not imply generalization beyond them.

6.2 Cross-Artifact and Cross-Model Aggregation

When multiple reasoning artifacts are analyzed under identical protocol conditions, the instrument may generate comparative synthesis outputs.

These outputs report convergence or divergence across artifacts. Comparative reporting does not evaluate semantic correctness, task performance, or relative model quality; it describes structural characteristics under defined measurement criteria.

6.3 Verification Boundaries

Evidence artifacts substantiate reported signals without requiring access to implementation details. Verification operates at the artifact level: identical inputs under identical protocols must yield identical outputs for reproducibility to be established.

The artifact therefore functions as the formal audit record.

7 Case Illustration: Structural Posture Under RN-L2

This section presents a single illustrative audit conducted under the RN-L2 protocol. The case is included solely to demonstrate instrument operation and artifact structure. It does not constitute comparative evaluation or general performance claims.

The audited artifact consisted of a reasoning response to a structured analytical prompt. The instrument was applied under deterministic execution conditions.

7.1 Reported Structural Signals

Table 1: Reported Structural Signals Under RN-L2

Metric Family	Reported Structural Signal
Structural Integrity	Reduced
Structural Coverage	Partial
Orphaned Claims	Present
Stability Under Replay	Stable
Reflective Transformation	Expansion without anchor increase

The instrument identified unsupported dependency steps and orphaned claims within the reasoning trace. Coverage signals indicated that a subset of claims lacked explicit structural anchoring.

Deterministic replay under identical parameters reproduced the same structural posture, confirming stability of reported signals.

7.2 Reflective Transformation

When the reasoning artifact was subjected to reflective re-evaluation under the same protocol conditions, the instrument detected structural expansion without proportional increase in evidentiary anchoring. Dependency reorganization occurred, but previously identified unsupported segments persisted.

The audit artifact therefore records:

- Stable structural posture under deterministic replay
- Presence of unsupported claims
- Reflective expansion without structural correction

This case is illustrative rather than exhaustive and is presented solely to demonstrate artifact-level reporting.

8 Instrument Grounding and Sanity Validation

The instrument was exercised against reasoning artifacts exhibiting recognized failure archetypes, including unsupported dependency chains, coherence-first validation patterns, and internally inconsistent claim structures.

This grounding was not intended to establish comparative performance, benchmark accuracy, or statistical generalization. Its purpose was to confirm that the instrument surfaces signals consistent with its defined measurement scope.

Grounding cases were executed under explicitly defined audit protocols and deterministic conditions. In each case, reported signals corresponded to observable dependency inconsistencies, orphaned claims, or instability within the reasoning artifact.

Deterministic replay confirmed persistence of the detected structure under identical inputs and fixed execution parameters, indicating that the signals were not artifacts of stochastic variation.

This grounding establishes operational plausibility and internal consistency. It does not establish completeness, statistical calibration, domain generalization, or superiority relative to other evaluation methods.

The instrument's claims remain limited to structural properties observable under defined audit conditions.

9 Relationship to Existing Evaluation Methods

Reflective Diagnostics operates at a distinct analytical layer from commonly used evaluation paradigms for large language models.

Benchmark-based evaluations measure task performance and outcome accuracy. These methods assess whether a model produces correct or preferred outputs under predefined conditions. They provide aggregate performance indicators but do not analyze the structural organization of reasoning artifacts.

Preference-based evaluations capture human judgments regarding quality, helpfulness, or alignment. While informative, such assessments remain interpretive and do not expose claim-level dependency structure within the artifact itself.

Process-oriented methods, including step-level reward modeling and chain-of-thought supervision, evaluate intermediate reasoning steps as part of optimization or training procedures. These approaches assess utility or progress toward a correct answer rather than structural coherence as an independent object of measurement.

Reflective Diagnostics is orthogonal to these paradigms. It does not measure correctness, alignment, or utility. It measures structural organization, dependency integrity, stability under deterministic replay, and structural transformation under reflection.

The instrument is not designed for training or optimization. It functions as a forensic layer intended to make structural posture observable and auditable at the artifact level.

10 Methodology Disclosure Policy

This paper defines Reflective Diagnostics at the level of measurement scope, audit protocol structure, analytical layers, and evidentiary output. It does not disclose computational implementation details.

Excluded from this document are:

- Extraction logic used for claim segmentation and dependency mapping
- Scoring functions, aggregation procedures, and normalization schemes
- Calibration parameters and threshold criteria
- Prompt construction templates or operational audit scripts
- System orchestration mechanics or execution pipelines
- Internal ranking or weighting formulas

These exclusions are intentional. The purpose of this paper is to establish the instrument’s definitional boundaries and evidentiary posture, not to enable reconstruction of operational mechanics.

Evidence artifacts produced by the instrument are designed to substantiate reported structural signals without exposing implementation details or enabling replication of the system’s internal processes.

Disclosure is therefore limited to what is necessary to evaluate measurement scope, analytical consistency, and artifact-level reproducibility.

Future publications may expand on methodological components under controlled contexts. This document serves as the boundary-setting and definitional reference for the instrument.

11 Conclusion

This paper defines Reflective Diagnostics as a modular instrument for measuring structural properties of reasoning generated by large language models.

The instrument operates after generation and reports structural signals across multiple analytical layers, including claim dependency organization, coverage, stability under deterministic replay, and transformation under reflection. It produces structured evidence artifacts sufficient to substantiate reported signals without exposing implementation details.

Reflective Diagnostics does not evaluate correctness, alignment, or semantic validity, nor does it infer internal cognition or intent. Its claims are limited to structural properties under defined audit protocols.

By formalizing structural measurement as a distinct analytical layer, the instrument establishes a reproducible basis for auditing reasoning artifacts independently of outcome-based performance metrics.

This paper serves as the definitional reference for that measurement framework.

References

- [1] Bowman, S. R., et al. (2022). *Measuring Progress on Scalable Oversight for Large Language Models*. arXiv:2211.03540.
- [2] Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903.
- [3] Turpin, M., et al. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388.
- [4] Wang, X., et al. (2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv:2203.11171.