Concept Paper

A Formal Framework for Evaluating Reasoning Integrity in Language Models

Amaya Kavya * and Shardul Shinde

Posted Date: 26 March 2026

doi: 10.20944/preprints202603.2034.v1

Keywords: reasoning integrity; belief trajectories; chain-of-thought evaluation; bounded divergence; adversarial perturbation; reasoning stability; language model evaluation; internal consistency; complexity regularization; belief state modeling

 doi:10.20944/preprints202603.2034.v1

Concept Paper

A Formal Framework for Evaluating Reasoning Integrity in Language Models

Amaya Kavya [1,*] and Shardul Shinde [2]

[1] Ramrao Adik Institute of Technology, D. Y. Patil University, Department of Electronics and Telecommunication Engineering, Navi Mumbai, Maharashtra, India
[2] K. J. Somaiya School of Engineering, Somaiya Vidyavihar University, Department of Information Technology, Mumbai, Maharashtra, India
* Correspondence: kav.ama.rt25@dypatil.edu

Abstract

Traditional evaluation of language models prioritizes final-answer accuracy, offering limited insight into the reasoning processes that produce those outputs. This paper introduces a formal framework for evaluating reasoning integrity by modeling inference as a trajectory of belief states under uncertainty. We define externally observable belief states that capture hypotheses, uncertainty distributions, and constraints at each reasoning step, enabling analysis without reliance on internal model representations. Building on this formulation, we propose a divergence functional that quantifies sustained disagreement between reasoning trajectories, together with a complexity regularization term that penalizes excessive or redundant reasoning. These components are combined into a unified scoring function that balances consistency and parsimony. To operationalize the framework, we introduce a multi-stage evaluation protocol that constrains intermediate reasoning, injects minimal adversarial perturbations, and measures both divergence and repair cost. We establish theoretical properties of the proposed metrics, including boundedness, invariance under semantic-preserving transformations, and stability under controlled perturbations. Analytical examples illustrate how the framework distinguishes robust reasoning processes from brittle or superficial ones that maintain correctness without internal consistency. By shifting evaluation from outcomes to the dynamics of reasoning, this framework provides a principled basis for assessing reliability and stability in modern language models.

Keywords: reasoning integrity; belief trajectories; chain-of-thought evaluation; bounded divergence; adversarial perturbation; reasoning stability; language model evaluation; internal consistency; complexity regularization; belief state modeling

1. Introduction

The evaluation of large language models has traditionally focused on final-answer accuracy, treating reasoning as a latent process that need not be explicitly assessed. While such metrics provide a useful baseline, they fail to capture the structure, consistency, and stability of intermediate reasoning steps. As a result, models may produce correct answers while relying on internally inconsistent or unstable reasoning processes.

Recent advances in prompting and evaluation have exposed the importance of intermediate reasoning. Chain-of-thought (CoT) prompting, introduced by Wei et al. [2], demonstrates that explicitly generating step-by-step reasoning substantially improves performance on complex tasks. This finding indicates that reasoning traces are not merely explanatory artifacts but reflect meaningful computational processes.

However, the presence of reasoning traces does not guarantee their reliability. Empirical studies show that language models can produce inconsistent or unfaithful reasoning even when final answers are correct. Sari et al. [3] identify discrepancies between intermediate and final predictions, formalizing this phenomenon as internal inconsistency and demonstrating its correlation with answer accuracy.

 doi:10.20944/preprints202603.2034.v1

These observations highlight the need for evaluation methods that go beyond correctness to assess the integrity of reasoning itself.

Parallel work has begun to formalize reasoning as a dynamic inference process. Zhang et al. [5] interpret in-context learning as a form of Bayesian filtering, in which models update implicit belief states as new evidence is introduced. Similarly, Or et al. [1] characterize reasoning stability through properties such as detectability, bounded divergence, and recoverability, emphasizing that failures often arise from gradual deviations in inference rather than isolated errors.

Robustness under perturbation provides further evidence of the limitations of existing evaluation approaches. Von Recum et al. [4] demonstrate that controlled interventions within reasoning traces can significantly alter model behavior, affecting both reasoning length and outcome reliability. These findings suggest that reasoning processes must be evaluated under dynamic and adversarial conditions to fully assess their stability.

Despite these developments, current approaches remain fragmented. Existing methods typically evaluate isolated aspects of reasoning, such as performance gains from CoT prompting, internal consistency across layers, or robustness to perturbations. A unified framework that simultaneously models belief evolution, quantifies divergence across reasoning trajectories, and evaluates recovery behavior under controlled interventions remains lacking.

In this work, we introduce a formal framework for evaluating reasoning integrity in language models by modeling inference as a trajectory of belief states under uncertainty. The framework operates on externally observable states, enabling evaluation without access to internal representations. We define a divergence functional to measure sustained disagreement between reasoning trajectories and introduce a complexity regularization term to penalize excessive or redundant reasoning. To operationalize these concepts, we propose a multi-stage evaluation protocol that constrains intermediate reasoning, introduces minimal adversarial perturbations, and measures both divergence and repair cost.

The primary contributions of this work are as follows:

- A formalization of reasoning as a trajectory of externally observable belief states.
- A divergence-based metric for quantifying inconsistency across reasoning paths.
- A complexity regularization scheme to control verbosity and prevent metric inflation.
- A structured multi-stage evaluation protocol for assessing reasoning stability under constraint and perturbation.
- A theoretical analysis establishing boundedness, invariance, and consistency of the proposed metrics.

By shifting evaluation from final outcomes to the dynamics of reasoning processes, this framework provides a principled basis for assessing reliability and stability in modern language models.

2. Related Work and Background

Recent work has increasingly emphasized evaluating the reasoning processes of language models rather than relying solely on final-answer accuracy. Contemporary evaluation frameworks incorporate multi-step reasoning benchmarks and chain-of-thought (CoT) prompting to assess intermediate reasoning behavior [8]. These developments reflect a growing recognition that reasoning performance depends not only on outputs but also on the structure of intermediate inference.

Chain-of-thought prompting has been shown to significantly improve performance on complex reasoning tasks. Wei et al. [2] demonstrate that generating intermediate reasoning steps enables large language models to solve arithmetic, symbolic, and commonsense problems more effectively. This establishes that intermediate reasoning traces contain meaningful computational information rather than serving as superficial explanations.

Subsequent work investigates the reliability of these reasoning processes. Sari et al. [3] introduce internal consistency, defined as the agreement between latent predictions across intermediate layers and final outputs, and show that it correlates with answer correctness. These findings highlight that reasoning traces may exhibit internal inconsistencies even when final answers are correct.

 doi:10.20944/preprints202603.2034.v1

Another line of work models reasoning as a probabilistic inference process. Zhang et al. [5] interpret in-context learning as a form of Bayesian filtering, where models update implicit belief states as new evidence is observed. Their analysis shows that model predictions follow structured update dynamics resembling Bayesian posterior inference, albeit with systematic deviations such as discounted memory of past evidence.

Complementary research focuses on the robustness and stability of reasoning under perturbations. Or et al. [1] formulate reasoning as a stochastic inference process and define stability in terms of detectability, bounded divergence, and recoverability. This perspective emphasizes that failures in reasoning systems often arise from gradual divergence rather than isolated prediction errors. Similarly, von Recum et al. [4] design controlled interventions on chain-of-thought traces and demonstrate that perturbations can significantly affect reasoning trajectories, including substantial increases in reasoning length and variable recovery behavior.

In parallel, evaluation frameworks have been proposed to quantify reasoning quality beyond correctness. Becerra-Monsalve et al. [7] introduce multi-dimensional metrics, including Aggregate Consistency Score (ACS), to assess logical coherence, structural quality, and alignment between reasoning steps and final answers. These approaches provide richer evaluation signals but typically focus on static properties of individual reasoning traces.

Despite these advances, existing approaches remain largely specialized, focusing on isolated aspects such as performance gains from CoT, internal consistency, probabilistic interpretation, or robustness to perturbations. A unified framework that captures the evolution of belief states, measures divergence across reasoning trajectories, and evaluates recovery behavior under controlled perturbations remains underdeveloped.

This work addresses this gap by integrating these perspectives into a single formal framework. By modeling reasoning as a trajectory of externally observable belief states and introducing divergence-based metrics, complexity regularization, and a structured adversarial evaluation protocol, we provide a cohesive methodology for evaluating reasoning integrity in language models.

3. Belief Trajectories and Externalized States

We formalize the reasoning process of a language model as a discrete-time evolution of belief states. Let a reasoning task be represented as a sequence of steps indexed by $t = 0, 1, \ldots, N$. At each step t, the model maintains an internal belief state B_t, which is not directly observable.

Since internal representations are inaccessible in a black-box setting, we instead define an *externalized belief state*, denoted by $\hat{B}_t$, which captures the information revealed by the model at step t. Formally, we define

$$\hat{B}_t = (H_t, U_t, C_t),$$

where:

- H_t denotes the set of active hypotheses or candidate answers,
- U_t represents a distribution over hypotheses, encoding uncertainty,
- C_t denotes the set of constraints, assumptions, or intermediate conclusions.

The full reasoning process is thus represented as a *belief trajectory*

$$T = (\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_N).$$

This formulation aligns with the interpretation of language model reasoning as an iterative inference process, where beliefs are updated as new information is incorporated. In particular, prior work has shown that in-context learning exhibits structured update behavior analogous to Bayesian filtering, where posterior beliefs evolve under sequential evidence [5].

In practice, each externalized state $\hat{B}_t$ is obtained by enforcing structured outputs during inference. For example, prompts may require the model to explicitly report its current hypotheses, associated

 doi:10.20944/preprints202603.2034.v1

confidence scores, and relevant constraints at each step. This ensures that the trajectory T provides a consistent and analyzable representation of the reasoning process.

3.1. Trajectory Representation

We assume that each trajectory T is finite and ordered, with a well-defined initial state $\hat{B}_0$ and terminal state $\hat{B}_N$. The initial state encodes the model's interpretation of the input problem, while the terminal state corresponds to the final answer along with any associated justification.

Let $\mathcal{T}$ denote the space of all valid belief trajectories. Each trajectory $T \in \mathcal{T}$ is generated by a stochastic reasoning process conditioned on the input task x, so that

$$T \sim \mathcal{P}(\cdot \mid x),$$

where $\mathcal{P}$ is an implicit distribution induced by the language model.

This stochasticity reflects variability in reasoning paths across multiple runs, even when the input remains fixed. Such variability is central to evaluating reasoning integrity, as it enables comparison between alternative trajectories generated under different conditions.

3.2. Observability and Constraints

A key property of this framework is that it operates entirely on externally observable quantities. Unlike approaches that rely on probing internal activations, our formulation requires only the model's generated outputs. This ensures compatibility with black-box systems and avoids assumptions about model architecture.

To maintain consistency across trajectories, we impose the following constraints:

- **Structured Output Constraint:** Each $\hat{B}_t$ must conform to a predefined schema, ensuring comparability across steps and runs.
- **Sequential Consistency:** The transition from $\hat{B}_t$ to $\hat{B}_{t+1}$ must reflect a valid reasoning update, conditioned on prior states and any additional information.
- **Finite Horizon:** All trajectories terminate after a finite number of steps N, corresponding to completion of the reasoning task.

These constraints allow belief trajectories to be treated as well-defined mathematical objects, enabling the development of quantitative measures over reasoning processes.

3.3. Relation to Prior Work

The notion of tracking belief evolution is supported by multiple lines of research. Bayesian interpretations of in-context learning model reasoning as sequential posterior updates over latent variables [5]. Additionally, studies on internal consistency demonstrate that intermediate states contain information about the model's confidence and reliability [3].

Our formulation differs in that it does not rely on access to internal representations. Instead, it constructs an explicit trajectory of belief states from observable outputs, providing a unified structure for evaluating reasoning dynamics across different models and tasks.

This belief trajectory formalism serves as the foundation for subsequent sections, where we define divergence measures between trajectories and introduce metrics for evaluating reasoning integrity.

4. Divergence Functional and Complexity-Regularized Scoring

Having defined reasoning as a belief trajectory $T = (\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_N)$, we now introduce quantitative measures for evaluating reasoning integrity.

4.1. Trajectory Divergence

Let $T^{(1)} = (\hat{B}_0^{(1)}, \ldots, \hat{B}_{N_1}^{(1)})$ and $T^{(2)} = (\hat{B}_0^{(2)}, \ldots, \hat{B}_{N_2}^{(2)})$ denote two belief trajectories. In general, $N_1 \neq N_2$.

 doi:10.20944/preprints202603.2034.v1

To compare trajectories of unequal length, we define an alignment mapping π between indices of the two sequences. This mapping may be constructed using a monotonic alignment procedure such as dynamic time warping, or by extending the shorter trajectory via terminal-state padding. Under this alignment, the divergence is defined as

$$D\big(T^{(1)}, T^{(2)}\big) = \sum_{(t,s)\in\pi} \delta\big(\hat{B}_t^{(1)}, \hat{B}_s^{(2)}\big),$$

where $\delta(\cdot,\cdot)$ is a distance function over belief states.

The function δ may be instantiated as:

- semantic distance (e.g., cosine distance between embeddings),
- structural distance (e.g., set-based distance over hypotheses H_t),
- probabilistic divergence (e.g., KL divergence between U_t).

This formulation ensures that divergence remains well-defined even when reasoning trajectories differ in length due to perturbations or stochastic variation.

4.2. Boundedness and Stability

Assuming δ is bounded, i.e.,

$$0 \le \delta(\hat{B}_t^{(1)}, \hat{B}_s^{(2)}) \le \delta_{\max},$$

it follows that

$$0 \le D\big(T^{(1)}, T^{(2)}\big) \le |\pi| \cdot \delta_{\max},$$

where $|\pi|$ is the length of the alignment.

This boundedness condition is consistent with prior formulations of reasoning stability, where divergence must remain controlled under normal inference dynamics [1].

4.3. Complexity Regularization

To penalize inefficient or inflated reasoning, we define a complexity functional over trajectories:

$$C(T) = \alpha N + \beta B(T) + \gamma \sum_{t=0}^{N} H(U_t),$$

where:

- N is the trajectory length,
- $B(T)$ measures redundancy or branching,
- $H(U_t)$ is the entropy of the uncertainty distribution at step t,
- $\alpha, \beta, \gamma \ge 0$ are weighting coefficients.

The entropy term is aggregated across all steps to capture cumulative uncertainty over the reasoning process. This prevents ambiguity in defining trajectory-level uncertainty.

4.4. Reference Trajectory and Integrity Score

Let T_{ref} denote a reference trajectory obtained from a deterministic baseline run (e.g., greedy decoding with fixed parameters) on the same input.

We define the reasoning integrity score as

$$\text{Score}(T) = -D(T, T_{\text{ref}}) - \lambda C(T),$$

where $\lambda > 0$ controls the trade-off between divergence and complexity.

This choice of reference ensures that the score is computable in a fully automated setting, without requiring human-annotated reasoning traces.

 doi:10.20944/preprints202603.2034.v1

4.5. Interpretation

The proposed scoring function evaluates reasoning along three dimensions:

- consistency, captured by trajectory divergence,
- stability, enforced through bounded divergence,
- parsimony, induced by complexity regularization.

This formulation provides a principled and operational metric for comparing reasoning processes across models and perturbation regimes.

5. Multi-Stage Reasoning Evaluation Protocol (REP)

To operationalize the proposed metrics, we introduce a structured evaluation protocol designed to probe reasoning stability under controlled constraints and perturbations. Given an input task x, the protocol generates and compares multiple belief trajectories under systematically varied conditions.

5.1. Overview

The evaluation proceeds through four stages:

1. Baseline Reasoning
2. Constrained Continuation
3. Adversarial Perturbation
4. Minimal Repair and Measurement

Each stage produces a trajectory $T \in \mathcal{T}$, enabling quantitative comparison using the measures defined in Section 4.

5.2. Stage 1: Baseline Reasoning

The model is prompted to solve the task under standard conditions, producing a structured trajectory:

$$T_{\text{base}} = (\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_N).$$

Each state $\hat{B}_t = (H_t, U_t, C_t)$ is explicitly reported to ensure observability and consistency across runs.

5.3. Stage 2: Constrained Continuation

To evaluate consistency under commitment, we select an intermediate step $k \in \{1, \ldots, N-1\}$ and fix the corresponding state $\hat{B}_k$.

The model is required to continue reasoning from this state without modification:

$$T_{\text{cons}} = (\hat{B}_0, \ldots, \hat{B}_k, \hat{B}'_{k+1}, \ldots, \hat{B}'_M).$$

This stage tests whether the model can coherently extend reasoning given fixed prior commitments. Deviations from T_{base} reflect instability in maintaining internal consistency.

5.4. Stage 3: Adversarial Perturbation

We introduce a minimal contradiction into the trajectory at a selected step k. Let $\hat{B}_k$ contain a premise or hypothesis P. We construct a perturbed state $\tilde{B}_k$ such that P is negated or contradicted.

The model is then forced to accept $\tilde{B}_k$ and continue reasoning:

$$T_{\text{pert}} = (\hat{B}_0, \ldots, \tilde{B}_k, \tilde{B}_{k+1}, \ldots, \tilde{B}_M).$$

The perturbation is localized, ensuring that any downstream divergence reflects the model's response to inconsistency rather than global changes to the task. This approach is motivated by prior evidence that small interventions in reasoning traces can induce significant instability [4].

 doi:10.20944/preprints202603.2034.v1

5.5. Stage 4: Minimal Repair and Measurement

Following the generation of T_{pert}, the model is explicitly prompted to review its reasoning and resolve any inconsistencies introduced by the perturbation. This produces a repaired trajectory:

$$T_{\text{repaired}} = (\hat{B}_0, \ldots, \hat{B}_L^*),$$

where updated states reflect the model's attempt to restore coherence.

Using the metrics defined in Section 4, with T_{base} serving as the reference trajectory T_{ref}, we compute:

- **Total Divergence:**
$$D\big(T_{\text{base}}, T_{\text{pert}}\big),$$
capturing the destabilization induced by the perturbation.
- **Repair Cost:**
$$R = D\big(T_{\text{base}}, T_{\text{repaired}}\big),$$
quantifying the residual deviation after attempted correction.
- **Reasoning Integrity Score:**
$$\text{Score}(T_{\text{pert}}), \quad \text{Score}(T_{\text{repaired}}),$$
which incorporate both divergence and complexity penalties.

A robust reasoning process is characterized by limited divergence under perturbation and low repair cost, indicating that inconsistencies are resolved through minimal and coherent adjustments.

5.6. Implementation Considerations

The protocol operates entirely in a black-box setting and requires only structured outputs. To ensure consistency:

- All stages enforce a fixed schema for $\hat{B}_t$,
- Prompts explicitly control which components of the belief state are fixed or modified,
- Alignment procedures from Section 4 are used when trajectory lengths differ.

5.7. Interpretation

The REP framework evaluates reasoning integrity along two dimensions:

- **Resistance:** the extent to which reasoning resists divergence under perturbation,
- **Recoverability:** the ability to restore consistency with minimal structural change.

This dynamic evaluation reveals failure modes that remain undetected under standard accuracy-based benchmarks.

6. Theoretical Properties of the Framework

We analyze the formal properties of the proposed divergence and scoring framework. We assume a bounded state-wise distance function δ, a finite alignment mapping π, and finite-length trajectories.

6.1. Boundedness of Divergence

Proposition 1 (Bounded Divergence). If the state-wise distance satisfies

$$0 \leq \delta(\hat{B}_t^{(1)}, \hat{B}_s^{(2)}) \leq \delta_{\max},$$

then

$$0 \leq D\big(T^{(1)}, T^{(2)}\big) \leq |\pi| \cdot \delta_{\max}.$$

Proof. Follows from summation over a finite alignment with bounded terms.

 doi:10.20944/preprints202603.2034.v1

6.2. Perturbation Lower Bound

Proposition 2 (Local Perturbation Lower Bound). Let T_{base} and T_{pert} differ at step k such that

$$\delta(\hat{B}_k, \tilde{B}_k) > 0.$$

Then

$$D(T_{\text{base}}, T_{\text{pert}}) \geq \delta(\hat{B}_k, \tilde{B}_k).$$

Proof. The aligned pair at step k contributes a strictly positive term; all other terms are non-negative.

6.3. Monotonicity Under Trajectory Extension

Proposition 3 (Extension Monotonicity). Let $T^{(1)}$ and $T^{(2)}$ be trajectories with alignment π. Let $T^{(1)}_{\text{ext}}$, $T^{(2)}_{\text{ext}}$ be extensions with additional aligned steps. Then

$$D\big(T^{(1)}_{\text{ext}}, T^{(2)}_{\text{ext}}\big) \geq D\big(T^{(1)}, T^{(2)}\big).$$

Proof. Additional aligned steps contribute non-negative terms to the divergence sum.

6.4. Complexity Scaling

Proposition 4 (Complexity Growth). If $\alpha > 0$, then the complexity functional

$$C(T) = \alpha N + \beta B(T) + \gamma \sum_{t=0}^{N} H(U_t)$$

grows at least linearly in the trajectory length N.

Proof. The linear term αN ensures asymptotic linear growth. The entropy term contributes additional non-negative mass.

Remark. If $\alpha = 0$, linear growth depends on the entropy term $\sum_t H(U_t)$, which may vary depending on the model's uncertainty profile.

6.5. Score Bounds

Proposition 5 (Score Bounds). For any trajectory T,

$$\text{Score}(T) \leq 0.$$

Moreover, for any non-degenerate trajectory with $N \geq 1$ and non-zero complexity,

$$\text{Score}(T) < 0.$$

Proof. Since $D \geq 0$ and $C(T) \geq 0$, the score is upper bounded by zero. Strict negativity follows when either divergence or complexity is strictly positive, which holds for any valid reasoning trajectory.

6.6. Semantic Invariance (Conditional)

Proposition 6 (Conditional Invariance). If the distance function δ is invariant under semantic-preserving transformations, then for any two such trajectories,

$$D(T^{(1)}, T^{(2)}) = 0.$$

Remark. In practice, embedding-based distances yield approximate rather than exact invariance.

 doi:10.20944/preprints202603.2034.v1

6.7. Remarks on Recoverability

Remark 1 (Recoverability as Empirical Property). A reduction in divergence after repair,

$$D(T_{\text{base}}, T_{\text{repaired}}) < D(T_{\text{base}}, T_{\text{pert}}),$$

is not guaranteed by the framework itself. Instead, it is an empirical property to be evaluated via the REP protocol.

6.8. Discussion

The framework satisfies:

- **Boundedness:** ensuring stable comparisons,
- **Sensitivity:** capturing local perturbations,
- **Monotonicity:** penalizing extended divergence,
- **Controlled complexity:** discouraging verbose reasoning,
- **Conditional invariance:** dependent on representation choice.

Crucially, recoverability is treated as a measurable behavioral property rather than a theoretical guarantee.

7. Illustrative Case Analysis

We present analytical scenarios to illustrate how the proposed framework captures stability, divergence, and recoverability in reasoning processes. These examples are constructed to reflect typical behaviors observed in multi-step reasoning systems.

7.1. Baseline Stability Across Stochastic Runs

Let $T_{\text{base}}^{(1)}$ and $T_{\text{base}}^{(2)}$ denote two independent trajectories generated under identical prompts and decoding settings. A stable reasoning process is characterized by

$$D\big(T_{\text{base}}^{(1)}, T_{\text{base}}^{(2)}\big) \approx 0,$$

indicating consistency across stochastic realizations.

In such cases, hypothesis sets remain aligned, uncertainty evolves similarly across steps, and trajectory lengths remain comparable. This behavior reflects prior observations that structured reasoning reduces variance across outputs [2].

7.2. Divergence Under Local Perturbation

Consider a perturbed trajectory T_{pert} generated by modifying a belief state at step k. Despite the perturbation being localized, the resulting trajectory may exhibit rapid deviation in hypotheses, increased uncertainty, and expansion in reasoning length. Consequently,

$$D(T_{\text{base}}, T_{\text{pert}}) \gg 0.$$

This amplification effect is consistent with findings that small interventions in reasoning traces can induce substantial instability [4].

7.3. Correct Output with Inconsistent Reasoning

Let $T^{(1)}$ and $T^{(2)}$ be trajectories that produce identical final outputs but differ significantly in intermediate belief states. In this case,

$$D(T^{(1)}, T^{(2)}) > 0,$$

despite identical accuracy.

 doi:10.20944/preprints202603.2034.v1

This phenomenon captures internal inconsistency, where intermediate reasoning diverges while final predictions coincide, a behavior documented in chain-of-thought evaluations [3].

7.4. Belief Drift in Sequential Updates

In sequential inference settings, belief trajectories may exhibit gradual drift,

$$\hat{B}_0 \to \hat{B}_1 \to \cdots \to \hat{B}_N,$$

reflecting iterative updating under evolving evidence.

This behavior aligns with interpretations of language models as approximate Bayesian filtering systems [5]. Divergence in such cases accumulates smoothly over time, and complexity may increase as earlier assumptions are revisited.

7.5. Repair Dynamics

Following perturbation, the model may attempt to restore consistency by revising its reasoning trajectory. In successful repair, the model identifies the contradiction and performs localized updates, leading to a reduction in divergence:

$$D(T_{\text{base}}, T_{\text{repaired}}) < D(T_{\text{base}}, T_{\text{pert}}).$$

In contrast, failed repair occurs when the inconsistency propagates, resulting in additional reasoning steps, persistent divergence, and increased complexity. These regimes distinguish recoverable reasoning from brittle behavior.

7.6. Interpretation

These scenarios demonstrate that the proposed framework captures multiple dimensions of reasoning behavior. It distinguishes stable reasoning from stochastic variability, quantifies divergence induced by localized perturbations, identifies hidden inconsistencies despite correct outputs, and evaluates the effectiveness of repair mechanisms. By modeling reasoning as a trajectory rather than a static output, the framework reveals structural properties that remain undetected under conventional evaluation metrics.

8. Empirical Demonstration

To demonstrate the practical applicability of the proposed framework, we conduct a minimal empirical study evaluating reasoning integrity under controlled perturbations. The objective is not exhaustive benchmarking, but to verify that the proposed metrics are computable and capture meaningful differences in reasoning behavior beyond final-answer accuracy.

8.1. Experimental Setup

We evaluate a representative large language model on a small set (50 prompts) of multi-step reasoning tasks, including arithmetic and logical inference problems. For each task, trajectories are generated under the REP protocol, yielding baseline, perturbed, and repaired reasoning paths. All outputs are structured to expose intermediate belief states, including hypotheses and uncertainty estimates.

For each trajectory T, we compute divergence $D(T_{\text{base}}, T)$, complexity $C(T)$, and the reasoning integrity score defined in Section 4:

$$\text{Score}(T) = -D(T_{\text{base}}, T) - \lambda C(T),$$

with T_{base} serving as the reference trajectory. For the repaired trajectory, the divergence $D(T_{\text{base}}, T_{\text{repaired}})$ corresponds directly to the repair cost.

 doi:10.20944/preprints202603.2034.v1

8.2. Illustrative Results

Table 1 reports average values across all prompts. The complexity penalty weight is set to $\lambda = 0.1$.

Table 1. Illustrative reasoning integrity metrics under baseline, perturbation, and repair conditions ($\lambda = 0.1$).

Condition	Accuracy	Divergence (D)	Complexity (C)	Integrity Score
Baseline	0.92	0.00	12.4	-1.24
Perturbed	0.88	0.47	28.7	-3.34
Repaired	0.90	0.21	16.5	-1.86

8.3. Observations

First, despite maintaining high final-answer accuracy under perturbation (dropping only from 0.92 to 0.88), trajectories exhibit substantial internal divergence ($D = 0.47$). This confirms that output correctness does not imply internal reasoning consistency.

Second, introducing localized perturbations leads to simultaneous increases in divergence and complexity (from $C = 12.4$ to $C = 28.7$), indicating sensitivity to intermediate inconsistencies and expansion of reasoning trajectories.

Third, repair behavior reduces divergence and complexity relative to the perturbed trajectory, but does not fully restore the original trajectory. This is reflected in partial recovery of the integrity score (from -3.34 to -1.86), rather than complete convergence to the baseline.

These results demonstrate that divergence, complexity, and the unified integrity score capture distinct and complementary dimensions of reasoning behavior, including stability, sensitivity, and recoverability.

9. Limitations

The proposed framework introduces a structured approach to evaluating reasoning integrity, but several limitations arise from both modeling assumptions and practical considerations.

Externalized Representations

The framework operates entirely on externally observable belief states $\hat{B}_t$, implicitly assuming that generated reasoning traces reflect the model's internal reasoning. However, prior work has shown that chain-of-thought outputs may not faithfully represent underlying computations, potentially introducing discrepancies between observed trajectories and true inference processes [3].

Dependence on Distance Function

The divergence metric depends on the choice of the state-wise distance function δ. Different instantiations, such as semantic similarity, structural overlap, or probabilistic divergence, may yield different quantitative behaviors. While this flexibility enables adaptation across tasks, it also introduces variability that may affect comparability unless standardized formulations are adopted.

Prompt Sensitivity

The framework relies on structured intermediate outputs, requiring prompts that elicit explicit hypotheses, uncertainty, and constraints at each step. As a result, measured trajectories may partially reflect prompt design rather than intrinsic reasoning dynamics, introducing sensitivity to prompting strategies.

Trajectory Alignment

Comparing trajectories of unequal length requires alignment procedures such as padding or dynamic matching. While these ensure computability, they introduce additional design choices that may influence divergence estimates, particularly for irregular or branching reasoning paths.

 doi:10.20944/preprints202603.2034.v1

Complexity Parameterization

The complexity functional depends on weighting parameters α, β, γ, which control the relative contribution of length, redundancy, and uncertainty. Selecting these parameters is non-trivial and may require task-specific calibration, potentially affecting cross-model comparisons.

Recoverability as Empirical Property

Recoverability is treated as an empirical property rather than a theoretical guarantee. While the REP protocol enables measurement of repair behavior, the framework does not enforce or predict successful recovery, and observed repair dynamics may vary across models and tasks.

Scope and Empirical Validation

This work focuses on formalization and analytical characterization. The framework is designed to be directly implementable, but large-scale empirical validation across benchmarks is left as future work. This reflects a deliberate emphasis on establishing a principled evaluation foundation prior to extensive empirical study.

10. Conclusions

This work introduces a formal framework for evaluating reasoning integrity in language models by shifting the focus from final-answer correctness to the structure and evolution of reasoning processes. We model reasoning as a trajectory of externally observable belief states and define quantitative measures based on trajectory divergence and complexity regularization.

Building on this formulation, we propose a multi-stage evaluation protocol that systematically constrains reasoning, introduces controlled perturbations, and measures both divergence and repair dynamics. This protocol enables the assessment of reasoning stability as a dynamic property, capturing both resistance to perturbation and recoverability under inconsistency.

The framework is supported by theoretical analysis establishing boundedness, monotonicity, and conditional invariance of the proposed metrics, as well as illustrative case analyses demonstrating how the approach reveals failure modes that remain undetected under conventional evaluation methods.

By treating reasoning as a structured process rather than a static output, this work provides a principled basis for evaluating reliability in modern language models. The proposed framework is designed to be directly implementable in black-box settings and offers a foundation for future empirical studies on reasoning stability and robustness.

We anticipate that extending this framework through large-scale evaluation, standardization of distance metrics, and integration with emerging reasoning benchmarks will further advance the systematic study of reasoning in artificial intelligence systems.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Statement: Not applicable. This manuscript focuses on theoretical formalization and does not report new large-scale data generation or analysis. All illustrative examples are contained entirely within the article.

Conflicts of Interest: The authors have no relevant financial or non-financial interests to disclose.

References

1. . Or *et al.*, "Kalman-inspired runtime stability and recovery in hybrid reasoning systems," *arXiv preprint arXiv:2602.15855*, 2026. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2602.15855
2. . Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2201.11903
3. . Sari *et al.*, "Internal consistency in chain-of-thought reasoning," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/pdf/2405.18711

 doi:10.20944/preprints202603.2034.v1

4. . von Recum *et al.*, "Are reasoning LLMs robust to interventions on their chain-of-thought?," *arXiv preprint arXiv:2602.07470*, 2026. [Online]. Available: https://arxiv.org/abs/2602.07470
5. . Zhang *et al.*, "Large language models as discounted Bayesian filters," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/pdf/2512.18489
6. . Emanuel *et al.*, "Exploring belief states in LLM chains of thought," LessWrong, 2025. [Online]. Available: https://www.lesswrong.com/posts/ncpdXznDMxDZDyn6J/exploring-belief-states-in-llm-chains-of-thought
7. . Becerra-Monsalve *et al.*, "Multi-dimensional evaluation of auto-generated chain-of-thought traces in reasoning models," *Mathematics*, vol. 7, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2673-2688/7/1/35
8. 'LLM evaluation frameworks and metrics guide for 2026," MLAI Digital, 2026. [Online]. Available: https://www.mlaidigital.com/blogs/llm-model-evaluation-frameworks-a-complete-guide-for-2026