Markovian ODE-guided scoring can assess the quality of offline reasoning traces in language models

Anonymous authors
Paper under double-blind review

Abstract

Reasoning traces produced by generative language models are increasingly used for tasks ranging from mathematical problem solving to automated fact checking. However, existing reasoning evaluation metrics are typically validated using either synthetic perturbations or benchmark-specific correlations. Moreover, the design choices underlying many metrics are often closely tied to the perturbation sets used for their evaluation, raising concerns about their generalizability. Additionally, despite employing multiple perturbation types, prior work frequently relies on binary grading schemes that provide only a coarse assessment of reasoning quality. Consequently, such evaluations offer limited evidence that a metric remains sensitive to realistic degradations in reasoning quality or that it generalizes across reasoning domains. As a result, it remains unclear whether existing metrics capture properties that align with human judgments of reasoning quality. To this end, we introduce MarODE, an offline evaluation framework that assigns quality scores to reasoning traces. Its effectiveness is assessed using *human-centric perturbations* and human judgments, which jointly evaluate the fundamental dimensions of an evaluation metric – *goodness* and *soundness*. The approach is grounded in a Markovian formulation of reasoning progression and an ordinary differential equation based characterization of trace dynamics, enabling efficient evaluation of reasoning quality. In a large-scale evaluation, MarODE outperforms existing baselines consistently under Somers' D correlation. Our results emphasize the value of theory-driven evaluation frameworks as reasoning traces become central to language model-based systems.

1 Introduction

With the rise of large language models (LLMs), there has been growing interest in leveraging their generative capabilities to produce human-like, multi-step reasoning traces. As such traces become increasingly prevalent, evaluating their quality has emerged as a critical challenge. Recent work shows that this challenge arises because there are a few reliable automatic methods to evaluate reasoning chains, and no common framework to systematically compare different reasoning strategies across tasks (Hao et al., 2024). Reasoning traces, often expressed as chain-of-thought or intermediate rationales, have been shown to substantially improve performance on complex tasks involving arithmetic, symbolic manipulation, multi-hop question answering, and logical inference (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022b). These traces enable models to decompose complex problems into simpler substeps, making them effective across a broad range of domains, including mathematics, factual verification, programming, and decision support. Empirical analysis demonstrates that explicitly generating intermediate reasoning steps is a key factor behind these improvements, leading to substantial gains in multi-step reasoning performance even under noisy or imperfect reasoning traces (Wang et al., 2023). Prior work shows that generated rationales are not always coherent with the model's underlying decision process, raising concerns about reliability and interpretability (Jacovi & Goldberg, 2020; Yao et al., 2023). As reasoning traces increasingly mediate interaction between humans and intelligent systems, systematically defining and evaluating their quality remains essential for reliable and effective deployment.

Existing approaches to reasoning evaluation can broadly be divided into two paradigms (Lee & Hockenmaier, 2025). *Generation-time* or *online* evaluation assesses reasoning quality during decoding, typically by leveraging intrinsic model signals such as token probabilities, uncertainty estimates, or intermediate verifier feedback to guide or score the generation process (Qiu et al., 2025; Chen et al., 2023; Ankner et al., 2024). In contrast, *post-hoc* or *offline* evaluation operates on completed reasoning traces, treating them as textual artefacts and assessing their quality using external criteria such as step-level validity, coherence, factual grounding, and consistency towards an answer (Golovneva et al., 2023; Prasad et al., 2023; Kotonya & Toni, 2020). In this work, we focus on post-hoc evaluation of reasoning traces, aiming to characterize and measure the quality of completed reasoning processes independently of the generation mechanism.

Prior studies often struggle to capture a key aspect of reasoning evaluation – natural language reasoning is inherently human-centric and cannot be reliably assessed using purely mechanical perturbations or metrics designed to detect such perturbations (Chrysostomou & Aletras, 2021). As language and reasoning styles evolve, evaluation metrics must remain generalizable, rather than being tightly coupled to specific surface-level patterns (Jacovi & Goldberg, 2020). To address this gap, we introduce human-centric perturbations that better reflect realistic variations in reasoning. These perturbations are designed around three key principles: (i) stochastic degradation, which captures the inherent variability of reasoning errors; (ii) perturbation design that is largely independent of the evaluation metric itself, reducing the risk of metric-specific bias; and (iii) a graded, rather than binary, assessment framework that enables modeling of varying levels of reasoning deterioration. The resulting perturbation scores provide a ground-truth signal for evaluating whether metric scores correlate with the severity of reasoning degradation. We show that the performance of established benchmarks, such as ROSCOE (Golovneva et al., 2023), degrades sharply under these conditions.

Motivated by these observations, we propose MarODE, a novel framework for post-hoc evaluation of reasoning traces that aims to capture coherence, quality, and consistency. The framework is grounded in principled formulations based on Markov chains and an ordinary differential equation (ODE)-based model of directional consistency. In developing MarODE, we place reasoning quality at the center of the evaluation process and seek to characterize its underlying indicators through a mathematical lens. This allows us to translate qualitative aspects of reasoning quality into a quantitative and principled framework.

Complementing the metric design, our human-centric perturbation framework is constructed to capture local, global, and surface-level forms of reasoning degradation. The perturbations are motivated by failure modes identified in prior studies of reasoning generation and are designed from a human evaluation perspective rather than being tailored to the proposed metric. By intentionally decoupling perturbation design from metric design, we provide an independent evaluation setting that reduces benchmark-specific bias. Together, these components enable a more reliable and generalizable assessment of reasoning quality.

Across data distributions, MarODE consistently outperforms existing baselines in sensitivity to perturbations and shows stronger alignment with human judgments on expert-evaluated reasoning traces, supporting the claims of goodness and soundness. Owing to the limited availability of high-quality long-form reasoning traces, we construct traces for perturbation analysis using the claim–verdict–evidence structure of factuality datasets such as LIAR (Wang, 2017) and PolitiFact (Põldvere et al., 2023). We further assess alignment with human judgments on diverse datasets spanning multiple reasoning domains.

Our contributions are:

- We propose **MarODE**, a post-hoc evaluation framework for reasoning traces that jointly models *Markovian coherence*, *ODE-guided directional consistency*, and *evidence alignment*, providing a unified measure of reasoning quality. The PyPI package and usage instructions are provided in Appendix A.
- We introduce a **human-centric perturbation framework** that progressively degrades reasoning traces through stochastic degradation, enabling fine-grained assessment of evaluation metrics beyond binary perturbation settings.
- We formulate **directional consistency as a continuous-time belief evolution process**, deriving an ODE-based update rule from entailment–contradiction dynamics and demonstrating its stability relative to multiplicative update schemes across diverse reasoning scenarios.

- We conduct a **large-scale empirical evaluation** on both synthetically perturbed and human-evaluated reasoning traces spanning factual, logical, and mathematical reasoning tasks, showcasing the **goodness** and **soundness** of our metric.

2 Related Works

Quality of reasoning traces. Evaluating the quality of reasoning traces requires moving beyond final answer correctness or surface-level fluency. Prior work has consistently shown that high-quality reasoning is characterized by coherent step-to-step progression, global directional consistency, and explicit grounding in evidence or premises (Dalvi et al., 2021; Jiang et al., 2025). These principles underpin a wide range of reasoning benchmarks and evaluation frameworks, including ROSCOE (Golovneva et al., 2023), EntailmentBank (Dalvi et al., 2021), ProofWriter (Tafjord et al., 2021), and verifier-based approaches for mathematical and factual reasoning (Cobbe et al., 2021; Geva et al., 2021).

Evaluation of reasoning traces. Reasoning traces can be evaluated using rule-based matching, where ground truth is represented as graph-based entities (Nguyen et al., 2024; Li et al., 2023). In some cases, reasoning traces are assessed using intrinsic metrics during trace generation. Common examples include uncertainty estimation via token probability entropy (Qiu et al., 2025; Farquhar et al., 2024). V-information is also used to measure how much of a trace contributes to the final answer (Chen et al., 2023). Reinforcement learning approaches can further be employed to learn value functions for evaluating traces (Lai et al., 2024). Another class of evaluation methods operates post hoc, after the full trace has been generated, focusing on aspects such as semantic correctness, coherence, and completeness (Golovneva et al., 2023; Kotonya & Toni, 2020; Prasad et al., 2023).

Failure modes in reasoning traces. The perturbations in our framework are motivated by reasoning failures repeatedly identified in prior work. Faithfulness studies have shown that chain-of-thought explanations may diverge from the true basis of a model's prediction, producing unsupported conclusions, weak verdict justifications, and evidence-grounding failures despite appearing plausible to human readers (Turpin et al., 2023). Formal reasoning benchmarks further reveal systematic weaknesses in logical inference, including errors involving quantifiers and logical operators (Li et al., 2024). Temporal reasoning has also emerged as a persistent challenge: Holtermann et al. (2025) demonstrated inconsistent temporal generalization across language models, while Bazaga et al. (2025) distinguished between temporal knowledge representation and the ability to reason correctly over temporal information. Beyond local reasoning errors, studies of causal reasoning show that language models often rely on superficial patterns rather than underlying causal structure. The CLadder benchmark (Jin et al., 2023) highlights failures in formal causal inference, while Yamin et al. (2025) showed that performance degrades as causal graph complexity increases and event ordering becomes less aligned with narrative structure. Prior work has also identified vulnerabilities to irrelevant information and redundant reasoning steps, demonstrating that language models can be distracted by irrelevant content and exhibit instability in the presence of unnecessary intermediate reasoning (Cheng et al., 2025; Chowdhury & Manocha, 2026). Finally, failures involving logical modality, reasoning closure, evidence grounding, and conceptual consistency have been documented in studies of reasoning faithfulness and causal representation learning (Wang et al., 2025; Zhang et al., 2023). Collectively, these findings motivate our perturbation suite, which targets local, global, and surface-level degradations in reasoning quality while remaining grounded in empirically observed failure modes.

3 Datasets

We evaluate reasoning trace quality across both generated and human evaluated datasets, covering fact verification, logical reasoning, and mathematical reasoning tasks. The dataset construction follows a unified pipeline for reasoning trace generation, extraction, and validation, ensuring comparability across models and prompting conditions.

3.1 Generated reasoning traces

We construct a large-scale reasoning trace dataset using two widely adopted fact-checking benchmarks: LIAR (Wang, 2017) and PolitiFact (Põldvere et al., 2023). Together, these datasets comprise a total of 8,600 factual claims spanning political statements with ground truth veracity labels. Each claim is paired with structured prompts designed to elicit step-by-step reasoning traces. Reasoning traces are generated using five reasoning optimised large language models: DeepSeek-Qwen-14B, DeepSeek-LLaMA-8B, DeepSeek-Qwen-7B (Guo et al., 2025), Qwen-3B-CoT (Bai et al., 2023) and GPT-OSS-20B (Kumar et al., 2025). The details are provided in Appendix B. For each model, we employ three few-shot prompting configurations of 1-shot, 2-shot, and 4-shot, resulting in three distinct reasoning traces per claim. This yields a total of 25,800 generated reasoning traces across all prompting settings for every model.

Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ denote the set of factual claims, and let $\mathcal{P}_k(c)$ denote the prompt constructed for claim c under a k-shot configuration, where $k \in \{1, 2, 4\}$. For a given language model $m \in \mathcal{M}$, let $p_m(y \mid \mathcal{P}_k(c))$ denote the conditional probability distribution over output sequences induced by m when conditioned on $\mathcal{P}_k(c)$. A generated response is then obtained by sampling

$$y_m^{(t)}(c, k) \sim p_m(\cdot \mid \mathcal{P}_k(c)), \tag{1}$$

where $t \in \{1, \ldots, T\}$ indexes repeated decoding attempts, with a maximum of $T = 5$ retries.

$$\mathcal{E} : y \mapsto r, \tag{2}$$

Each generated response is post-processed using a deterministic extraction operator, which isolates the reasoning span r delimited by explicit reasoning markers. A reasoning trace r is considered valid if and only if it satisfies the structural validity predicate

$$\mathcal{V}(r) = \mathbb{I}\big[r = \{R_0, R_1, \ldots, R_K, V_f\},\ K \geq 1\big], \tag{3}$$

where R_0 denotes the initial reasoning step, $\{R_1, \ldots, R_K\}$ denote intermediate reasoning steps, and V_f denotes the terminal `Final Verdict`.

If $\mathcal{V}(r) = 0$, generation is repeated by incrementing $t \leftarrow t + 1$ until either a valid reasoning trace is obtained or $t = T$. For each tuple (c, k, m), the retained reasoning trace is defined as

$$r_m^*(c, k) = \arg\min_t \left\{ t \mid \mathcal{V}\Big(\mathcal{E}\Big(y_m^{(t)}(c, k)\Big)\Big) = 1 \right\}. \tag{4}$$

The prompts used for the generation of these traces are described in Appendix C.

3.2 Human-evaluated reasoning traces

In addition to model-generated traces, we include a curated set of reasoning traces to serve as high-quality reference data. This subset is drawn from four established reasoning benchmarks: EntailmentBank (Dalvi et al., 2021), ProofWriter (Tafjord et al., 2021), GSM8K (Cobbe et al., 2021) and StrategyQA (Geva et al., 2021). From each dataset, we have 150 instances, resulting in a total of 600 human-evaluated reasoning traces. These datasets span deductive, logical, numerical, commonsense, and scientific reasoning domains, enabling evaluation under heterogeneous reasoning styles. They cover diverse reasoning phenomena, including multi-hop logical entailment, formal proof construction, arithmetic and symbolic reasoning and commonsense question answering. All selected traces follow explicit step-by-step reasoning formats with clearly stated conclusions, making them suitable for fine-grained evaluation of reasoning trace quality under our proposed methodology. We evaluate reasoning traces from these datasets using three expert human evaluators. Evaluations are conducted along multiple dimensions, including logical and consistent progression, clarity of expression, absence of redundancy, grounding in the given context, coherence of reasoning flow supporting the final judgment, and overall reasoning quality.

4 Methodology

In this section, we describe how we perturb and evaluate reasoning traces, and how these elements come together in MarODE. We begin by outlining the perturbation methodology, followed by an explanation of how reasoning quality is modeled through coherence, directional consistency, and evidence alignment.

4.1 Human-centric perturbation

To evaluate the sensitivity of reasoning traces under controlled degradation, we apply a set of human-centric perturbations to the generated reasoning traces. Perturbations are designed to randomly and selectively disrupt logical structure, semantic coherence, or evidential grounding while preserving surface fluency, thereby enabling fine-grained evaluation of reasoning quality.

Let $r = \{R_0, R_1, \ldots, R_K, V_f\}$ denote a reasoning trace consisting of ordered reasoning steps $\{R_i\}_{i=0}^{K}$ and a terminal verdict V_f. Perturbations operate on the step sequence $\{R_i\}_{i=0}^{K}$, with V_f preserved unless explicitly perturbed.

We define a perturbation function set (details are given in Appendix, Table 4) as

$$\mathcal{F} = \{f_1, \ldots, f_{14}\},$$

where each $f_j \in \mathcal{F}$ is a partial operator

$$f_j : \{R_0, \ldots, R_K\} \to \{\tilde{R}_0, \ldots, \tilde{R}_K\},$$

returning a success indicator $\sigma_j \in \{0, 1\}$.

To balance perturbation intensity, traces are randomly permuted and assigned a target perturbation count

$$p \in \{0, \ldots, 6\}.$$

For each trace, up to p distinct perturbations are sampled uniformly without replacement from $\mathcal{F}$ and applied sequentially until

$$\sum_j \sigma_j = p \quad \text{or} \quad \mathcal{F} \text{ is exhausted.}$$

The perturbed trace $\tilde{r}$ is reconstructed with preserved step indices. A perturbation score is defined as

$$s(p) = 1.0 - 0.1p, \qquad p \in \{0, \ldots, 6\},$$

yielding $s(p) \in [0.4, 1.0]$. Each data instance stores $(\tilde{r}, s(p), \{f_j\})$ for downstream evaluation.

Continuous modeling of reasoning degradation. In contrast to prior perturbation-based evaluation frameworks such as ROSCOE, which primarily employ binary labeling of perturbed reasoning traces, our methodology models reasoning degradation as a gradual and structured process. Through randomly composed and graded human-centric perturbations targeting local step inconsistencies, global structural disruptions, and surface-level noise, we model gradual degradation in reasoning that reflects partial logical breakdowns, semantic drift, and evidential weakening observed in human reasoning (Dasgupta et al., 2022). The stochastic application of multiple perturbation operators prevents biases from fixed transformation patterns, enabling fine-grained assessment of reasoning robustness while preserving surface fluency. This design supports nuanced evaluation of reasoning quality beyond discrete failure detection.

4.2 MarODE: Markovian–ODE reasoning evaluation

We introduce MarODE, a modular metric for evaluating the quality of reasoning traces. Given a reasoning trace $r = \{R_0, R_1, \ldots, R_K, V_f\}$, MarODE computes quality through three components: coherence, quality, and evidence alignment, which are aggregated into a final score. Figure 1 presents an overview of MarODE.

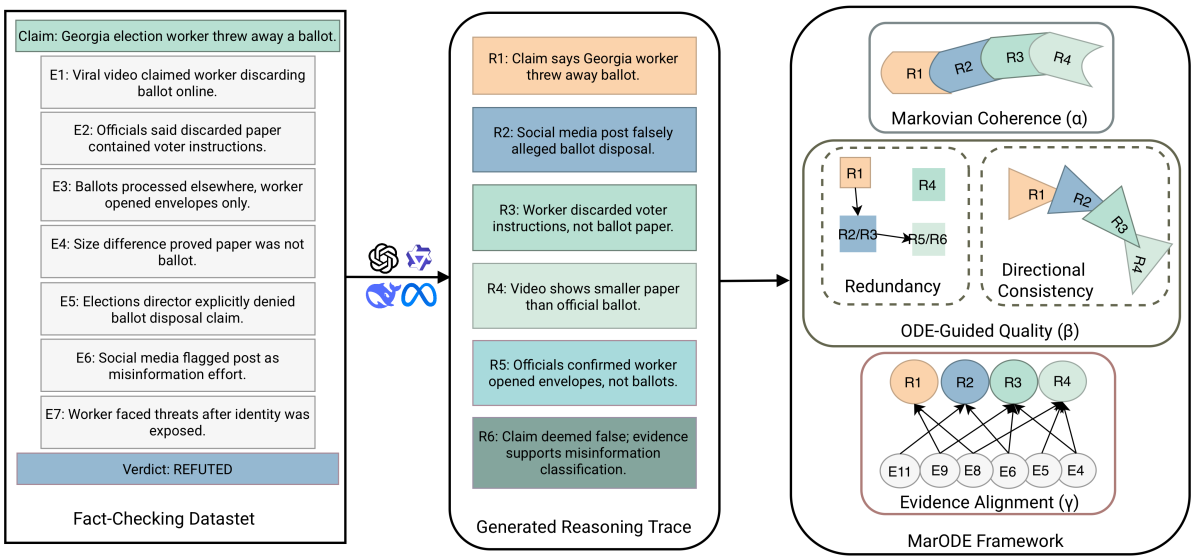

Figure 1: **An overview of the MarODE framework**, which integrates three complementary components – *Markovian Coherence*, *Directional Consistency*, and *Evidence Alignment*, to capture a generalist notion of completeness and soundness in reasoning traces. (a) *Markovian Coherence* models the local flow of reasoning, where each step in a high-quality reasoning trace is meaningfully connected to its predecessor while also facilitating progression toward a successor, often with controlled informational overlap. (b) *Directional Consistency* enforces a crisp, one-directional progression of reasoning, ensuring that each step advances toward a conclusion rather than oscillating or regressing. (c) *Redundancy* caters to steps that fail to contribute significant new information, not only increasing the length of the reasoning chain but also compromising its overall quality and its ability to converge on a certain conclusion. (d) *Evidence Alignment* grounds reasoning steps in supporting evidence, mitigating hallucinations that can otherwise lead to confident yet incorrect conclusions.

4.2.1 Markovian coherence

A good reasoning trace ensures coherent connections and commonality between adjacent trace elements. The flow of information should be gradual and continuous, avoiding abrupt or non-justifiable jumps. Each reasoning step R_i is embedded into a normalized semantic space, yielding embeddings $\{\mathbf{e}_i\}_{i=0}^{K}$. A transition matrix $\mathbf{P} \in \mathbb{R}^{K \times K}$ is constructed using cosine similarity,

$$P_{ij} = \frac{\exp\left(\cos(\mathbf{e}_i, \mathbf{e}_j)/\tau\right)}{\sum_{j'} \exp\left(\cos(\mathbf{e}_i, \mathbf{e}_{j'})/\tau\right)}, \tag{5}$$

where $\tau > 0$ denotes a temperature parameter and self-transitions are suppressed.

This formulation is inspired by prior works that leverage random walks over semantic representations to capture coherence and relatedness, particularly the random-walk–based semantic similarity framework of (Ramage et al., 2009) and markov process–based models of coherence over discourse entities proposed by (Petersen et al., 2015; Aghajohari et al., 2026; Bi et al., 2026). Reasoning coherence is estimated by simulating 256 random walks over the induced Markov chain. Let, $\pi = (i_0, i_1, \ldots, i_L)$ denote a sampled walk. Each walk is scored by rewarding monotonic transitions, where $i_{t+1} = i_t \pm 1$. To obtain a higher quality set of walks, we ensure that the start state lies within the first third of the reasoning steps. The coherence score is defined as

$$C(r) = \mathbb{E}_{\pi \sim \mathbf{P}} \left[\frac{1}{|\pi| - 1} \sum_t \mathbb{I}\big(|i_{t+1} - i_t| = 1\big) \right], \tag{6}$$

approximated via Monte Carlo averaging. This value is recorded as coherence score.

4.2.2 ODE-guided quality modeling

Global reasoning quality is modeled by combining redundancy analysis with directional consistency through a continuous time formulation.

Redundancy. Each reasoning step R_i is tokenized into a set of alphanumeric tokens $\mathcal{T}(R_i)$. For all unordered step pairs (i, j), a redundancy penalty is computed using symmetric differences,

$$d_{ij} = |\mathcal{T}(R_i) \triangle \mathcal{T}(R_j)| \,.$$

Penalties are assigned as

$$\rho_{ij} = \begin{cases} 1, & d_{ij} = 0, \\ \exp(-d_{ij}/4), & \text{otherwise.} \end{cases}$$

The redundancy score is then defined as

$$R(r) = 1 - \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \rho_{ij}, \tag{7}$$

where $\mathcal{P}$ denotes the set of upper-triangular step pairs. This formulation penalizes reasoning traces that repeat the same or very similar words across steps, encouraging each step to add new information. As a result, higher redundancy scores indicate more lexical variety and less repetition throughout the reasoning trace.

Directional consistency via ODE integration. Directional Consistency is modeled by assuming that reasoning trace elements follow a uniform direction toward a specific conclusion. This direction is represented as the evolution of a latent belief variable $p(t) \in (0, 1)$, initialized at $p(0) = 0.5$. For each reasoning step R_i, a Natural Language Inference (NLI) model produces entailment and contradiction probabilities with respect to R_{i-1}, yielding a normalized signal:

$$S_i = \text{clip}\left(\frac{(e_i - c_i) + 1}{2},\, 0,\, 1\right). \tag{8}$$

Here, e_i and c_i denote the entailment and contradiction probabilities, respectively. We employ an ordinary differential equation (ODE) formulation to compute the latent belief variable, rather than using multiplicative probability chains from classical probability theory. Inspired by prior work that formulates latent state trajectories through continuous dynamics, we adapt the continuous time framework of neural ODEs (Chen et al., 2018) and recent applications of neural differential equations for latent sequence dynamics (Li et al., 2025) to define directional consistency via ODE integration over reasoning step signals. The rationale and derivation for the following design choice are discussed in Section 4.2.3.

$$\frac{dp}{dt} = \rho(S_i - 0.5)p(1 - p),$$

Here, $p \in [0, 1]$ denotes the evolving belief strength. We model belief evolution through the continuous dynamics, which is numerically integrated across reasoning steps using a stable Runge – Kutta (RK4) scheme (see Appendix D).

The integration yields a sequence $\{p_i\}_{i=0}^{K}$, with $p_{i+1} = \Phi(p_i, S_i)$, where Φ denotes the numerical RK4 update operator. The final directional consistency score is defined as

$$DC(r) = p_K,$$

where K is the length of the reasoning trace.

Quality aggregation. The overall quality score is computed as

$$Q(r) = \tfrac{1}{2}R(r) + \tfrac{1}{2}DC(r), \tag{9}$$

4.2.3 Derivation of directional consistency dynamics

Let $\mathcal{R}_i$ denote the i-th reasoning step in a sequence. We define the local entailment and contradiction probabilities between consecutive reasoning steps as

$$P_i^e = \mathbb{P}(\mathcal{R}_i \models \mathcal{R}_{i-1}), \qquad P_i^c = \mathbb{P}(\mathcal{R}_i \perp \mathcal{R}_{i-1}), \tag{10}$$

where $\models$ and $\perp$ denote entailment and contradiction, respectively.

Cumulative entailment and contradiction. The cumulative entailment probability up to step i is defined as

$$E_i = \mathbb{P}(\mathcal{R}_i \models \mathcal{R}_{i-1} \models \ldots \models \mathcal{R}_0) = \prod_{k=1}^{i} P_k^e, \tag{11}$$

Similarly, the cumulative contradiction probability is

$$C_i = \mathbb{P}(\mathcal{R}_i \perp \mathcal{R}_{i-1} \perp \cdots \perp \mathcal{R}_0) = \prod_{k=1}^{i} P_k^c. \tag{12}$$

Odds formulation. We define the local odds for step $\mathcal{R}_i$ as

$$\Lambda(\mathcal{R}_i) = \frac{P_i^e}{P_i^c}. \tag{13}$$

The cumulative odds up to step i are then given by

$$O_i = \frac{E_i}{C_i} = \frac{P_i^e P_{i-1}^e \ldots P_1^e}{P_i^c P_{i-1}^c \ldots P_1^c} = \Lambda(\mathcal{R}_i)\, O_{i-1}. \tag{14}$$

Taking logarithms yields the additive form

$$\log O_i = \log O_{i-1} + \log \Lambda(\mathcal{R}_i). \tag{15}$$

Letting $L_i = \log O_i$, we obtain

$$L_i - L_{i-1} = \log \Lambda(\mathcal{R}_i) = \log P_i^e - \log P_i^c. \tag{16}$$

NLI-based approximation. We approximate the log-odds increment using natural language inference (NLI) scores. Let

$$S_i^{\text{raw}} = \text{Entail}_{\text{NLI}}(\mathcal{R}_i, \mathcal{R}_{i-1}) - \text{Contradict}_{\text{NLI}}(\mathcal{R}_i, \mathcal{R}_{i-1}), \tag{17}$$

and normalize it to the unit interval:

$$S_i = \max\left(0, \min\left(1, \frac{S_i^{\text{raw}} + 1}{2}\right)\right). \tag{18}$$

We then approximate

$$\log \Lambda(\mathcal{R}_i) \approx c\left(S_i - \frac{1}{2}\right), \tag{19}$$

where c is a scaling constant. This approximation preserves the qualitative behavior of $\log P_i^e - \log P_i^c$ and captures the relative dominance of entailment over contradiction using NLI signals.

Continuous formulation. Interpreting the reasoning index i as a continuous variable, we rewrite the update equation as

$$\frac{\partial L}{\partial i} = r\left(S_i - \frac{1}{2}\right), \tag{20}$$

where r is the rate-of-change constant.

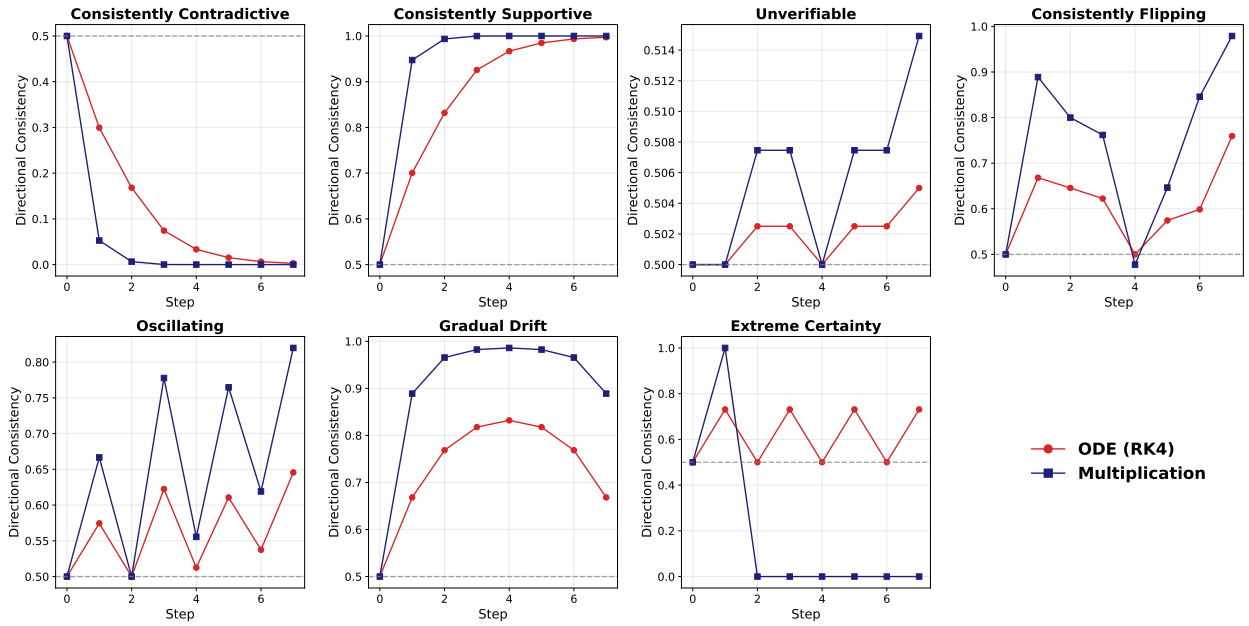

Figure 2: Directional consistencies across seven reasoning scenarios. Each subplot shows the evolution of belief under two update methods: a logistic ODE solved via Runge–Kutta (blue circles) and a multiplicative Bayesian-style update (orange squares). Scenarios include consistently supportive, consistently contradictory, unverifiable, flipping, oscillating, gradual drift, and extreme certainty. The figure highlights differences in stability and volatility across reasoning dynamics.

Directional consistency probability. We define the directional consistency (entailment consistency) probability as

$$\pi_i = \sigma(L_i) = \frac{1}{1 + e^{-L_i}} = \frac{E_i}{E_i + C_i}, \tag{21}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Differentiating π_i with respect to i gives

$$\frac{\partial \pi_i}{\partial i} = \frac{d\sigma(L_i)}{dL_i} \cdot \frac{\partial L_i}{\partial i} = \pi_i(1 - \pi_i) \cdot \frac{\partial L_i}{\partial i}. \tag{22}$$

Substituting from Eq. (20), we obtain the governing equation used in our method:

$$\frac{\partial \pi_i}{\partial i} = \pi_i(1 - \pi_i)\, r \left(S_i - \frac{1}{2} \right) \tag{23}$$

This equation models how directional consistency evolves across a reasoning sequence, where π_i represents the probability that the sequence up to step i maintains a consistent entailment direction.

Directional consistency analysis. To examine how different reasoning dynamics influence directional consistency updating, we simulated seven scenarios of NLI signals, ranging from consistently supportive to consistently contradictory, unverifiable, oscillating, flipping, gradual drift, and extreme certainty as shown in Figure 2. For each sequence of entailment, contradiction, and neutral probabilities, we computed directional consistency trajectories using two methods: (i) a logistic ordinary differential equation solved via a fourth–order Runge–Kutta scheme, and (ii) a multiplicative Bayesian-style update that accumulates entailment and contradiction likelihoods. The resulting plots show how directional consistency evolves step by step under each scenario, highlighting the contrast between *smooth ODE dynamics* and *sharper multiplicative*

updates. The ODE-based formulation produces smoother and steadier trajectories, particularly for longer reasoning chains, while multiplicative updates tend to magnify small local changes and can quickly become extreme. Extensions like neural controlled differential equations further generalize this continuous modeling perspective for dynamic signals (Kidger et al., 2020). Continuous-time modeling has been highlighted in recent machine learning research as a way to capture smoother and more stable temporal dynamics than traditional discrete update methods (Hasani et al., 2022). This comparison highlights why a continuous ODE-based approach is better suited for tracking reasoning progression.

4.2.4 Evidence alignment

To assess evidential grounding, each reasoning step R_i is aligned with a set of evidence sentences $\mathcal{E}$. For each step, the most similar evidence sentence is selected based on cosine similarity in embedding space. This evidence alignment formulation is inspired by prior work on evidential grounding that aligns reasoning steps with supporting sentences and refines alignment using natural language inference signals, as explored in ERASER (DeYoung et al., 2020) and FEVER style verification models (Nie et al., 2019). If the similarity exceeds a threshold θ, a base alignment score is assigned and refined using NLI probabilities,

$$E_i = 0.5 + \lambda_e \Pr_i(\text{entail}) - \lambda_c \Pr_i(\text{contradict}), \tag{24}$$

clipped to the interval $[0, 1]$. The evidence score is defined as

$$E(r) = \frac{1}{K+1} \sum_{i=0}^{K} E_i, \tag{25}$$

4.2.5 Final MarODE aggregation

The final MarODE score is computed as a convex combination of the three components,

$$\text{MarODE}(r) = w_c\, C(r) + w_q\, Q(r) + w_e\, E(r) \tag{26}$$

with weights distributed equally, for overall metric, we have, $w_c = w_q = w_e = \frac{1}{3}$. MarODE provides a unified measure of reasoning coherence, quality, and evidential grounding.

5 Experimental setup

We evaluate the quality of reasoning traces using a Markovian-ODE based metric and compare it against established baselines. The evaluation relies on synthetic perturbation-based labeling and human judgments as gold standards to assess the goodness and soundness of the proposed metric. Our analysis spans multiple datasets, prompting regimes model families and evaluation metrics.

5.1 Evaluation for soundness and goodness

We perform experiments under two complementary settings:

Perturbation-based labeling (generated diagnostics dataset). We construct labeled datasets by applying controlled human-centric perturbations (Section 4.1) to generated reasoning traces and assigning graded quality labels according to the degree of perturbation. As the number and severity of perturbations increase, the expected quality of a reasoning trace decreases monotonically (Ribeiro et al., 2020). This setup enables scalable and controlled evaluation of a metric's sensitivity to systematic degradation in reasoning quality (Gardner et al., 2020). We then score the perturbed traces using the proposed metric and competing baselines. We assess alignment between metric scores and perturbation-based quality labels using Somers' D to evaluate sensitivity to perturbation induced reasoning degradation.

Human-annotated labeling (existing benchmarks). We evaluate alignment with human judgment using expert-evaluated reasoning traces drawn from established benchmarks and quantify this alignment using Somers' D correlation. We prepare a set of questions to obtain a holistic human judgment of the quality of each reasoning trace (Figure 4 in Appendix). These expert annotations serve as an external gold standard for assessing both the goodness of the proposed metric and its validity in real-world reasoning evaluation settings. This evaluation complements synthetic perturbation-based analyses by grounding metric performance in expert human assessments of reasoning quality (Chen et al., 2023).

5.2 In-Context supervision and reasoning quality

To examine the impact of contextual supervision on reasoning quality, we conduct a systematic shot-wise analysis across all datasets, models, and evaluation metrics. Reasoning traces are generated under 1-shot, 2-shot, and 4-shot prompting configurations, and metric scores are computed on the corresponding original traces. Shot-wise trends are quantified via pairwise comparisons between (4-shot, 1-shot), (2-shot, 1-shot), and (4-shot, 2-shot) configurations (Table 3). We employ the Wilcoxon signed rank test to assess whether observed differences between shot settings are statistically significant (Dror et al., 2018). This analysis enables us to evaluate whether increasing the number of in-context examples systematically influences the quality of generated reasoning traces (Brown et al., 2020), and whether such effects are consistent across different evaluation metrics. We further analyze score distributions by plotting violin plots (Figure 3) for reasoning traces generated under different shot settings. The shape of each distribution is then used to examine how shot counts influence the quality of the resulting reasoning traces across all baselines.

5.3 Baselines

We compare MarODE against several established baselines that evaluate reasoning traces in an offline setting. Below, we briefly describe the baselines used in our experimental study. Details are provided in the Appendix E. **ROSCOE** (Golovneva et al., 2023) is a collection of interpretable, reference-free metrics designed to assess the quality of step-by-step reasoning, focusing on semantic alignment (ROSCOE-SA), semantic similarity (ROSCOE-SS), logical inference (ROSCOE-LI), and language coherence (ROSCOE-LC), using theory-driven error categories and diagnostic datasets. **ReCEval** (Prasad et al., 2023) evaluates reasoning chains in terms of both correctness and informativeness, leveraging entailment and information-based measures to quantify how individual reasoning steps contribute to the final answer. Prior work (Kotonya & Toni, 2020) evaluates explanation quality through the notions of **Local and Global Coherence**, measuring whether individual reasoning steps are well justified and whether the explanation forms a coherent whole. We also include an **LLM-as-a-Judge** baseline, in which a large language model is prompted to directly assign a scalar quality score to each reasoning trace using implicitly learned evaluation heuristics.

For entailment-based metrics and related NLI components, we use the `deberta-xlarge-mnli` model across all baselines and MarODE, due to its strong NLI performance. For sentence embeddings, we use the `all-MiniLM-L6-v2` (Yin & Zhang, 2024) model across all baselines and MarODE. The model was chosen for its strong performance across semantic similarity, retrieval, and clustering while maintaining high embedding quality with low-latency, efficient inference. For LLM-as-a-Judge, we use `prometheus-7b-v2.0`, chosen for its comparable evaluation quality to larger open models such as GPT-4 (Kim et al., 2024). For ROSCOE's linguistic consistency (ROSCOE-LC) components, we use models including `gpt2-large` and `tinybertcola` to capture lexical coherence within reasoning chains. All baselines are evaluated on both perturbation-based and human-annotated datasets using Somers' D correlation.

5.4 Meta-evaluation using Somers' D

To quantify alignment between metric scores and ground truth labels, derived either from controlled perturbations or expert human evaluations, we employ Somers' D correlation, following prior work such as ROSCOE. Somers' D is a rank-based measure of ordinal association that is well suited for meta-evaluation settings involving an *ordinal* ground truth variable and a *continuous* metric score. Somers' D has been widely used in statistics and applied evaluation settings as an asymmetric, rank-based association measure designed for ordinal outcomes (Somers, 1962; Newson, 2002).

Formally, Somers' D can be expressed as a normalized form of Kendall's rank correlation coefficient:

$$D(Y \mid X) = \frac{\tau(X, Y)}{\tau(X, X)},$$

where $\tau(X, Y)$ denotes Kendall's rank correlation between the ordinal ground truth variable X and the metric score Y. As $\tau(X, X)$ quantifies the number of pairs with unequal X values, Somers' D corresponds to the difference between the number of concordant and discordant pairs, normalized by the number of pairs for which the ground truth labels differ.

6 Results

Evaluation overview. We evaluate MarODE against established reasoning quality baselines across datasets, models, and different shot settings. Performance is examined along two dimensions: *goodness*, reflecting sensitivity to controlled degradation in reasoning through human-centric perturbations, and *soundness*, capturing agreement with expert human judgments. Synthetic reasoning traces (Section 3.1) are generated using five reasoning optimized LLMs – DeepSeek-Qwen-14B, DeepSeek-LLaMA-8B, DeepSeek-Qwen-7B (Guo et al., 2025), Qwen-3B-CoT (Bai et al., 2023) and GPT-OSS-20B (Kumar et al., 2025), on 8,600 factual claims drawn from LIAR and PolitiFact, under 1, 2, and 4-shot prompting, yielding 25,800 structured reasoning traces for every model. The prompts used are detailed in Appendix C. Evaluation is conducted on both synthetically perturbed reasoning traces and human-annotated benchmarks spanning factual, logical, and mathematical reasoning (Section 3.2). Correlation between metric scores and perturbation levels is quantified using Somers' D. MarODE is compared against ROSCOE, ReCEval, Local and Global Coherence, and LLM-as-a-Judge, enabling systematic assessment of sensitivity, significance, and alignment across evaluation settings.

Correlation analysis under human-centric perturbations. Table 1 reports Somers' D correlations across datasets, prompting settings, and backbone models, measuring alignment with human-centric perturbations. Across LIAR and PolitiFact, different shot counts, and model families, MarODE consistently achieves the strongest correlations, outperforming all baselines. We find that ROSCOE (Golovneva et al., 2023) variants exhibit limited alignment across datasets generated by different model variants. Semantic alignment and similarity (ROSCOE-SA, ROSCOE-SS) yield moderate correlations ($\sim$0.10–0.14), while logical inference (ROSCOE-LI) remains weak and logical consistency (ROSCOE-LC) is near zero or negative, with several non-significant results. Aggregation improves robustness and statistical significance (ROSCOE_MEAN $\sim$0.06–0.10) but remains markedly below MarODE. Other baselines, including LLM-as-a-Judge (Kim et al., 2024), Local and Global Coherence (Kotonya & Toni, 2020), and ReCEval (Prasad et al., 2023), consistently show low correlations (typically < 0.05) and limited sensitivity to increased shot count. In contrast, MarODE components demonstrate consistently high and statistically significant correlations across datasets, prompting regimes, and models ($\sim$0.24–0.29). Relative to ROSCOE_MEAN, MarODE yields a 235%–279% increase in correlation strength (Table 1), with gains remaining stable across shot counts and architectural variations.

Block-wise ablations of MarODE. We conduct block-wise ablations of MarODE to assess the contribution and interaction of components – Markovian Coherence (α), ODE-guided Quality (β) and Evidence Alignment (γ) across datasets and models (Table 1). *Markovian Coherence* (α) is the dominant factor, achieving high standalone correlations ($\sim$0.24–0.29). Removing other components often improves performance: pairwise combinations excluding evidence grounding, especially ($\alpha\beta$), consistently outperform ($\alpha\gamma$) and attain the strongest pairwise correlations overall ($\sim$0.21–0.33). This suggests that evidence signals can reduce alignment. *Evidence Alignment* (γ) shows moderate standalone performance ($\sim$0.20–0.24) but offers limited gains when combined with other blocks. In several cases, adding (γ) yields marginal or negative changes relative to(α) based combinations, indicating a secondary or context-dependent role. *ODE-guided Quality* (β) has weak standalone alignment (mostly < 0.06) but acts as a stabilizer when paired with (α), improving consistency and correlation strength. However, removing (β) from the full aggregation can increase correlations, indicating that its benefits are not universal. Overall, while MarODE remains strong ($\sim$0.24–0.29), ablations show that peak performance often arises from coherence-centered combinations.

Evaluation Metric	Models					Reasoning Trace Datasets					
	Qwen-3B-CoT	Deep-Seek-Qwen-7B	Deep-Seek-Qwen-14B	Deep-Seek-LLaMA-8B	GPT-OSS-20B	LIAR (1-shot)	LIAR (2-shot)	LIAR (4-shot)	PolitiFact (1-shot)	PolitiFact (2-shot)	PolitiFact (4-shot)
ROSCOE-SA	0.119	0.105	0.111	0.112	0.105	0.100	0.098	0.108	0.115	0.115	**0.128**
ROSCOE-SS	0.129	0.124	0.130	0.126	0.128	0.117	0.115	**0.139**	0.126	0.135	0.133
ROSCOE-LI	0.032	0.016	0.050	0.017	0.015	0.025	0.028	0.029	0.028	0.026	0.020
ROSCOE-LC	-0.019	-0.001	-0.022	0.023	-0.017	-0.007	-0.008	-0.018	0.000	-0.005	-0.007
ROSCOE_MEAN	0.082	0.069	0.096	0.069	0.062	0.069	0.073	0.078	0.079	0.079	0.074
LLM-as-a-Judge	0.044	0.027	0.042	0.038	0.033	0.037	0.028	0.014	0.052	0.046	0.041
Local and Global Coherence	0.046	0.019	0.052	0.032	0.032	0.042	0.040	0.035	0.030	0.038	0.033
ReCEval	0.046	0.005	0.031	-0.002	0.007	0.008	0.008	0.021	0.025	0.024	0.018
MarODE	**0.294**	**0.237**	**0.288**	**0.292**	**0.263**	**0.262**	**0.264**	**0.260**	**0.290**	**0.284**	**0.279**
MarODE COHERENCE(α)	**0.286**	0.175	**0.280**	0.283	0.201	0.240	0.236	0.243	0.217	0.228	0.230
MarODE QUALITY(β)	0.029	0.064	0.033	0.008	0.047	0.018	0.034	0.036	0.037	0.041	0.041
MarODE EVIDENCE(γ)	0.235	0.196	0.225	0.237	0.214	0.208	0.212	0.208	0.240	0.234	0.228
MarODE($\alpha\beta$)	**0.327**	**0.209**	**0.328**	**0.312**	**0.231**	**0.276**	**0.274**	**0.277**	**0.258**	**0.265**	**0.264**
MarODE($\beta\gamma$)	0.240	0.199	0.231	0.239	0.217	0.210	0.216	0.213	0.244	0.238	0.232
MarODE($\alpha\gamma$)	0.286	**0.230**	0.279	**0.286**	**0.257**	**0.255**	**0.256**	**0.253**	**0.280**	**0.276**	**0.271**

Table 1: **Somers' D correlations measuring the association between human-centric perturbation scores and evaluation metrics across different backbone models and prompting settings.** Columns correspond to reasoning traces generated by a range of LLMs as well as fact-checking benchmarks (LIAR and PolitiFact) under varying shot configurations. Higher values indicate stronger sensitivity to perturbation-induced changes in reasoning quality. For each column, the three highest correlations are highlighted in **bold**, while underlined values denote correlations that are not statistically significant ($p \geq 0.05$). Results include ROSCOE variants, LLM-based and coherence-oriented baselines, ReCEval, and MarODE with its component-wise and pairwise ablations, enabling a detailed analysis of goodness across models and prompting regimes.

Evaluation against human judgments. Table 2 reports Somers' D correlations between metric scores and human judgments across four human-evaluated benchmarks: EntailmentBank (Dalvi et al., 2021), ProofWriter (Tafjord et al., 2021), GSM8K (Cobbe et al., 2021) and StrategyQA (Geva et al., 2021). Overall, prior baselines exhibit weak, inconsistent, and often non-significant alignment with human assessments, with several correlations near zero or negative across tasks. ROSCOE components display highly variable behavior. Although ROSCOE-SA and ROSCOE-SS achieve moderate positive correlations on ProofWriter, their performance degrades on other datasets, including negative correlations on EntailmentBank and GSM, with many results lacking statistical significance. Aggregation (ROSCOE_MEAN) improves alignment on ProofWriter but remains near zero or negative elsewhere. Other baselines (LLM-as-a-Judge, Local and Global Coherence) and ReCEval similarly show limited alignment, with ReCEval exhibiting strong correlation only on StrategyQA. In contrast, MarODE and its components consistently align with human judgments across benchmarks.

Shot-wise analysis of metric scores. Figure 3 analyzes the effect of varying the number of in-context examples (1, 2, and 4-shot prompting) on metric scores for non-perturbed reasoning traces. Across all evaluated metrics, changing the shot count does not lead to significant shifts in score distributions. The violins exhibit similar shapes, medians, and inter-quartile ranges, indicating that the overall quality of generated reasoning traces remains stable regardless of the number of shots provided. Wilcoxon signed rank test also failed to conclude any remarkable significant shift in quality of reasoning traces generated using different shots (Table 3). Some baselines, such as LLM-as-a-Judge and Local and Global Coherence, show significant dispersion in central tendencies across different shot settings. Other baselines, such as ROSCOE and ReCEval, while exhibiting lower dispersion, suffer from asymmetric score distributions. In contrast, MarODE displays a notably tighter and more symmetric distribution across all shot configurations, reflecting greater stability.

Evaluation Metric	Reasoning Datasets			
	ENTAILMENT BANK	PROOFWRITER	GSM	STRATQA
ROSCOE-SA	-0.10337	0.17031	-0.04427	0.06290
ROSCOE-SS	-0.04991	0.15439	-0.03248	0.01250
ROSCOE-LI	0.03013	0.05386	-0.02203	0.06036
ROSCOE-LC	0.01167	0.09714	-0.05359	0.01563
ROSCOE_MEAN	-0.00730	0.21145	-0.03248	0.08869
LLM-as-a-Judge	0.07436	0.00000	0.00277	0.00449
Local and Global Coherence	0.03185	-0.00670	0.01650	0.04894
ReCEval	-0.02211	-0.02183	**0.05052**	**0.25611**
MarODE	**0.16981**	**0.33896**	**0.18578**	**0.22563**
MarODE_COHERENCE(α)	0.03114	0.21407	**0.18086**	0.12542
MarODE_QUALITY(β)	0.02810	**0.23998**	0.02121	0.22133
MarODE_EVIDENCE(γ)	**0.16870**	0.13692	-	0.07443
MarODE($\alpha\beta$)	0.02293	**0.33857**	-	**0.27349**
MarODE($\beta\gamma$)	0.14841	0.23823	-	0.19516
MarODE($\alpha\gamma$)	**0.17732**	0.21572	-	0.11604

Table 2: **Somers' D correlations measuring alignment between human (manual) judgments and evaluation metrics across four reasoning benchmarks: EntailmentBank, ProofWriter, GSM, and StrategyQA**. Higher values indicate stronger agreement with human assessments. For each dataset, the three highest correlations are highlighted in **bold**. Underlined values denote correlations that are not statistically significant ($p \geq 0.05$). Results include ROSCOE variants, aggregated and LLM-based baselines, ReCEval, and MarODE, along with their component-wise and pairwise ablations.

7 Discussion

Goodness of MarODE. The goodness of an evaluation metric refers to how well it captures meaningful changes in the underlying phenomenon it is designed to measure, especially under realistic variations (Yeh, 2000; Amidei et al., 2018). For reasoning evaluation, a metric should meaningfully reflect changes in reasoning quality under realistic variations – such as responding to differences in logical coherence, evidence use, and inferential consistency, while remaining insensitive to superficial or stylistic changes (Lan et al., 2025; Lee & Hockenmaier, 2025). Across all experimental settings, MarODE consistently demonstrates goodness as an evaluation metric, evidenced by its high and stable correlations with human-centric perturbations (Table 1 and Figure 7 in Appendix). Unlike prior baselines whose correlations remain weak or degrade across datasets, shot counts, and backbone models, MarODE maintains stable performance across diverse conditions. The magnitude of improvement over existing methods is substantial. Relative to aggregated ROSCOE variants, MarODE achieves a 235%–279% increase in correlation strength, indicating a stronger sensitivity to perturbations that affect reasoning quality rather than surface-level textual properties. The relative ordering of metrics remains preserved across architectures and prompting regimes.

Soundness of MarODE. Beyond raw performance, a sound reasoning metric should align with human judgments across diverse reasoning domains (Chen et al., 2023; DeYoung et al., 2020). Results on human-evaluated benchmarks (Table 2) show that MarODE achieves the strongest and most consistent alignment with expert annotations among all evaluated metrics. Existing baselines, including ROSCOE, LLM-as-a-Judge, and Local and Global Coherence, exhibit inconsistent and often non-significant alignment with human judgments. In contrast, MarODE achieves significant correlations across symbolic, mathematical, multi-hop, and strategic reasoning benchmarks, indicating that it captures reasoning properties that generalize across tasks and align with human assessments. Additional evidence for soundness comes from the distributional properties of metric scores. As shown in Figure 3, MarODE exhibits smooth, unimodal, and approximately symmetric distributions, whereas several baselines display skewed, multi-modal, or heavy-tailed patterns, often associated with instability, heuristic saturation, or sensitivity to spurious cues (Fomicheva & Specia,

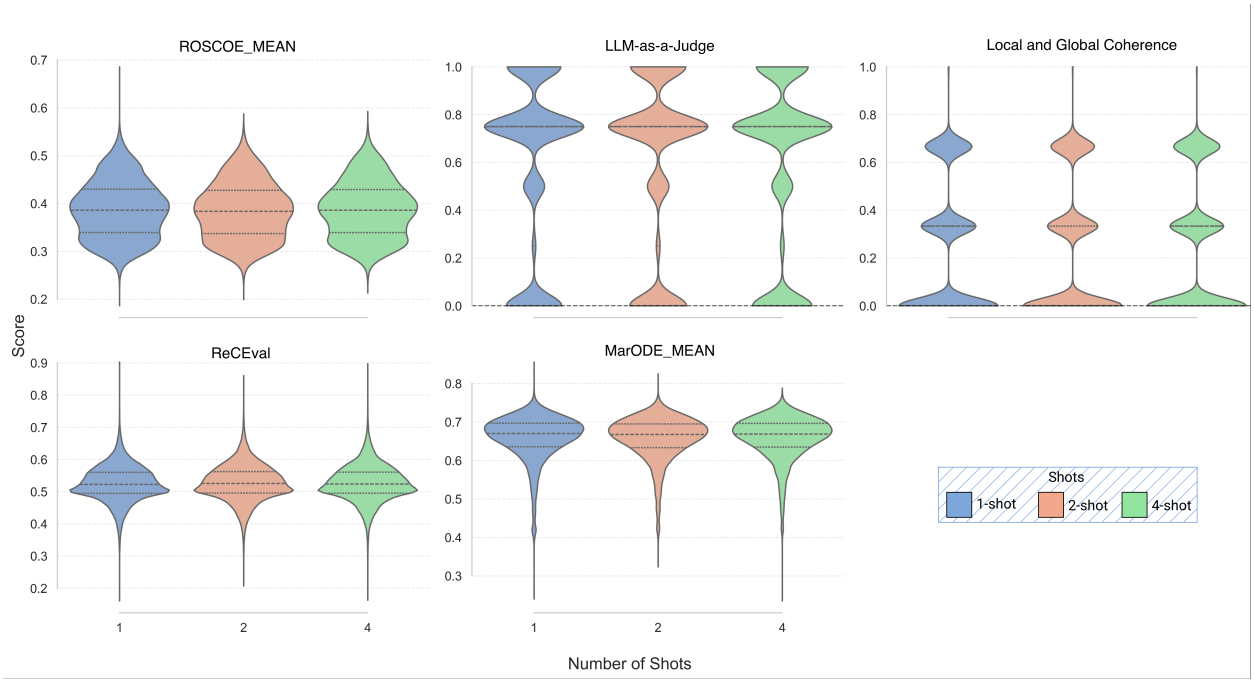

Figure 3: **Distribution of reasoning trace quality scores across different prompting regimes.** Violin plots illustrate the score distributions for five evaluation metrics – ROSCOE, LLM-as-a-Judge, Local and Global Coherence, ReCEval, and MarODE, under three prompting settings (1, 2, and 4-shot) used to generate reasoning traces. For a given baseline, the overall shape of the distribution remains largely consistent across different shot counts, suggesting relative stability of the metric with respect to prompting variation. In contrast, distinct baselines exhibit markedly different distributional shapes, reflecting differences in scale, sensitivity, and scoring behavior across different evaluation approaches.

2019). Such behavior is characteristic of metrics that aggregate multiple weakly dependent reasoning factors, producing more stable and regular score distributions (Longjohn et al., 2025; Soh & Zhao, 2024). Prior work further suggests that well-structured score distributions are more closely aligned with human judgments and yield more consistent and interpretable evaluations (Pillutla et al., 2021; Rei et al., 2023). In contrast, the asymmetric distributions observed for LLM-as-a-Judge, Local and Global Coherence, and ROSCOE indicate greater sensitivity to surface cues or specific reasoning styles. The tighter and more symmetric behavior of MarODE therefore supports its interpretation as a stable measure of reasoning quality.

Prompt sensitivity and stability of reasoning quality. The shot-wise analysis (Section 6) indicates that varying the number of in-context examples has a limited effect on the quality of generated reasoning traces. As shown by the paired Wilcoxon signed-rank tests in Table 3, no consistent or statistically significant differences are observed across shot settings. Across datasets and evaluation metrics, score distributions remain largely similar for 1, 2, and 4-shot prompting, suggesting that increasing demonstrations does not substantially alter reasoning behavior. This pattern suggests that reasoning quality is driven more by model and task characteristics than by prompt length. While several baseline metrics exhibit variability or skewness across shot settings, MarODE remains comparatively stable, with consistent and symmetric score distributions. This stability indicates that MarODE captures gradual changes in reasoning quality in a controlled manner rather than amplifying prompt-induced fluctuations. Overall, these findings challenge the assumption that increasing in-context examples reliably improves reasoning quality (Agarwal et al., 2024).

Further evidence of this stability appears in the belief trajectory analysis in Figure 2. Across diverse reasoning dynamics, including consistently supportive, contradictory, oscillating, and gradual drift scenarios, the ODE-based update underlying MarODE produces smooth and continuous transitions in directional consistency. This behavior aligns with prior work on neural ODE, which shows that continuous time formulations

Evaluation Metric	Prompting Configuration Comparison					
	(4-shot − 1-shot)		(2-shot − 1-shot)		(4-shot − 2-shot)	
	Median Diff	Significant	Median Diff	Significant	Median Diff	Significant
ROSCOE-SA	-0.000769	TRUE	0.000777	TRUE	-0.001501	TRUE
ROSCOE-SS	0.001300	TRUE	-0.001400	TRUE	0.002700	TRUE
ROSCOE-LI	-0.000336	FALSE	-0.002918	TRUE	0.004092	TRUE
ROSCOE-LC	-0.001100	TRUE	-0.001100	TRUE	0.000200	TRUE
ROSCOE_MEAN	-0.000374	FALSE	-0.001979	TRUE	0.001659	TRUE
LLM-as-a-Judge	0.000000	FALSE	0.000000	TRUE	0.000000	TRUE
Local and Global Coherence	0.000000	TRUE	0.000000	FALSE	0.000000	TRUE
ReCEval	0.001079	TRUE	0.002021	TRUE	-0.001183	TRUE
MarODE	-0.000479	TRUE	-0.001809	TRUE	0.001725	TRUE
MarODE_COHERENCE(α)	-0.002344	TRUE	-0.000930	TRUE	-0.001465	TRUE
MarODE_QUALITY(β)	0.000188	TRUE	-0.000102	TRUE	0.000257	TRUE
MarODE_EVIDENCE(γ)	-0.000658	FALSE	-0.004585	TRUE	0.003825	TRUE
MarODE($\alpha\beta$)	-0.000588	TRUE	-0.000589	TRUE	-0.000115	FALSE
MarODE($\beta\gamma$)	-0.000095	FALSE	-0.002524	TRUE	0.002290	TRUE
MarODE($\alpha\gamma$)	-0.001165	TRUE	-0.002749	TRUE	0.002195	TRUE

Table 3: **Median score differences and statistical significance ($p < 0.05$) from paired Wilcoxon signed-rank tests across three comparison settings: (4–1), (2–1), and (4–2).** Each setting contrasts reasoning traces generated under different shot conditions, with positive or negative median differences indicating higher or lower metric scores respectively for the former condition in each pair. The *Significant* column denotes whether the observed median difference is statistically reliable.

yield smooth trajectories and reduced sensitivity to discretization effects (Chen et al., 2018). In contrast, multiplicative updates often produce abrupt jumps, collapse to extreme values, or exhibit high volatility, particularly in flipping and extreme certainty regimes. Such behavior is consistent with findings from belief and opinion dynamics literature, where multiplicative rules are sensitive to noise, ordering effects, and local inconsistencies, frequently leading to instability or early certainty (Su & Wu, 2015). The stability of the ODE-based formulation is further supported by work on stable neural flows, which demonstrates that continuous dynamical systems with appropriate constraints exhibit bounded, predictable behavior and improved sensitivity to perturbations (Massaroli et al., 2020). MarODE's distributional stability appears to emerge from its continuous treatment of belief evolution across reasoning steps. Further discussion provided in Appendix F

8 Conclusion

We propose **MarODE**, a metric for evaluating reasoning quality, and systematically assess its performance using both synthetic perturbations and human-annotated ground truth. Our evaluation demonstrates its effectiveness, robustness, and overall stability as a reliable measure of reasoning quality. Although the absolute correlation values remain moderate, our evaluation exposes important limitations of existing metrics and demonstrates that MarODE achieves substantially stronger performance under a more generalized and human-centric evaluation setting. We find that, MarODE attains the highest alignment among all evaluated methods, improving correlation by 235%–279% over prior metrics.We find that coherence acts as the primary driver of reasoning quality, with ODE-guided quality providing additional synergy, while evidence alignment plays a limited role. MarODE also produces smooth and symmetric score distributions compared to baselines, indicating stable evaluation behavior. Overall, these findings show that reasoning quality is strongly determined by the coherent and consistent progression of steps, establishing MarODE as a reliable and general framework for evaluating the quality of reasoning traces.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76930–76966. Curran Associates, Inc., 2024. doi: 10.52202/079017-2447. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8cb564df771e9eacbfe9d72bd46a24a9-Paper-Conference.pdf.

Milad Aghajohari, Kamran Chitsaz, Amirhossein Kazemnejad, Sarath Chandar, Alessandro Sordoni, Aaron Courville, and Siva Reddy. The markovian thinker: Architecture-agnostic linear scaling of reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3As6AQ9ELI.

Jacopo Amidei, Paul Piwek, and Alistair Willis. Evaluation methodologies in automatic question generation 2013-2018. In Emiel Krahmer, Albert Gatt, and Martijn Goudbeek (eds.), *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 307–317, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6537. URL https://aclanthology.org/W18-6537/.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models, 2024. URL https://arxiv.org/abs/2408.11791.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Adrián Bazaga, Rexhina Blloshmi, Bill Byrne, and Adrià de Gispert. Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28014–28033, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1358. URL https://aclanthology.org/2025.acl-long.1358/.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Reasoning self-evaluation via trajectory dynamics modeling, 2026. URL https://openreview.net/forum?id=Qtq5YjnIlB.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. REV: Information-theoretic evaluation of free-text rationales. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2007–2030, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.112. URL https://aclanthology.org/2023.acl-long.112/.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 10400–10408. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1155. URL https://doi.org/10.24963/ijcai.2025/1155. Survey Track.

Sanjoy Chowdhury and Dinesh Manocha. Uncovering the representation geometry of minimal cores in overcomplete reasoning traces, 2026. URL https://arxiv.org/abs/2605.14358.

George Chrysostomou and Nikolaos Aletras. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 477–488, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*, 2022.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8472–8494, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.496. URL https://aclanthology.org/2024.findings-emnlp.496/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Marina Fomicheva and Lucia Specia. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558, September 2019. doi: 10.1162/coli_a_00356. URL https://aclanthology.org/J19-3004/.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh,

Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xYlJRpzZtsY.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. LLM reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=b0y6fbSUG0.

Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4 (11):992–1003, Nov 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00556-7. URL https://doi.org/10.1038/s42256-022-00556-7.

Carolin Holtermann, Paul Röttger, and Anne Lauscher. Around the world in 24 hours: Probing LLM knowledge of time and place. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 22875–22897, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1115. URL https://aclanthology.org/2025.acl-long.1115/.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6490–6514, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.329. URL https://aclanthology.org/2025.emnlp-main.329/.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31038–31065. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/631bb9434d718ea309af82566347d607-Paper-Conference.pdf.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623/.

Deepak Kumar, Divakar Yadav, and Yash Patel. Gpt-oss-20b: A comprehensive deployment-centric analysis of openai's open-weight mixture of experts model. *arXiv preprint arXiv:2508.16700*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms, 2024. URL https://arxiv.org/abs/2406.18629.

Tian Lan, Yang-Hao Zhou, Zi-Ao Ma, Fanshu Sun, Rui-Qing Sun, Junyu Luo, Rong-Cheng Tu, Heyan Huang, Chen Xu, Zhijing Wu, et al. A survey of automatic evaluation methods on text, visual and speech generations. *arXiv preprint arXiv:2506.10019*, 2025.

Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1789–1814, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.94. URL https://aclanthology.org/2025.findings-emnlp.94/.

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, and Fakhri Karray. Hd-ndes: Neural differential equations for hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6173–6186, 2025.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

Yiyuan Li, Shichao Sun, and Pengfei Liu. FRoG: Evaluating fuzzy reasoning of generalized quantifiers in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7239–7256, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.411. URL https://aclanthology.org/2024.emnlp-main.411/.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint arXiv:2501.04234*, 2025.

Stefano Massaroli, Michael Poli, Michelangelo Bin, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Stable neural flows. *arXiv preprint arXiv:2003.08063*, 2020.

Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1558–1570, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.117. URL https://aclanthology.org/2021.emnlp-main.117/.

Roger B. Newson. Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. *The Stata Journal*, 2:45 – 64, 2002. URL https://api.semanticscholar.org/CorpusID:53974402.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2862–2883, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.168. URL https://aclanthology.org/2024.findings-acl.168/.

Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6859–6866, 2019.

Casper Petersen, Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Entropy and graph based modelling of document coherence using discourse entities: An application to ir. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 191–200, 2015.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman

Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4816–4828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=NtHfJrjkiv.

Nele Põldvere, Zia Uddin, and Aleena Thomas. The politifact-oslo corpus: A new dataset for fake news analysis and detection. *Information*, 14(12), 2023. ISSN 2078-2489. doi: 10.3390/info14120627. URL https://www.mdpi.com/2078-2489/14/12/627.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. Entropy-based decoding for retrieval-augmented large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4616–4627, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.236. URL https://aclanthology.org/2025.naacl-long.236/.

Daniel Ramage, Anna N. Rafferty, and Christopher D. Manning. Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-4, pp. 23–31, USA, 2009. Association for Computational Linguistics. ISBN 9781932432541.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. The inside story: Towards better understanding of machine translation neural evaluation metrics. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1089–1105, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.94. URL https://aclanthology.org/2023.acl-short.94/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Yun Joon Soh and Jishen Zhao. A step towards mixture of grader: Statistical analysis of existing automatic evaluation metrics. *arXiv preprint arXiv:2410.10030*, 2024.

Robert H. Somers. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799, 1962. URL https://api.semanticscholar.org/CorpusID:144232738.

Qinliang Su and Yik-Chung Wu. On convergence conditions of gaussian belief propagation. *Trans. Sig. Proc.*, 63(5):1144–1155, March 2015. ISSN 1053-587X. doi: 10.1109/TSP.2015.2389755. URL https://doi.org/10.1109/TSP.2015.2389755.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.153. URL https://aclanthology.org/2023.acl-long.153/.

Meng Wang, Jinshuo Liu, Victor Gutierrez-Basulto, Lina Wang, and {Jeff Z.} Pan. Aqe-rf: An adaptive quantifier extension and rule-filtering graph network for logical reasoning of text. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18517–18529, July 2025. ISSN 2162-237X. doi: 10.1109/TNNLS.2025.3588525.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067/.

Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in NLP models: A survey. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4569–4586, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.339. URL https://aclanthology.org/2022.naacl-main.339/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Khurram Yamin, Shantanu Gupta, Gaurav R. Ghosal, Zachary C. Lipton, and Bryan Wilder. Failure modes of llms for causal reasoning on narratives, 2025. URL https://arxiv.org/abs/2410.23884.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, 2000. URL https://aclanthology.org/C00-2137/.

Chen Yin and Zixuan Zhang. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pp. 677–684. Atlantis Press, 2024. ISBN 978-94-6463-540-9. doi: 10.2991/978-94-6463-540-9_69. URL https://doi.org/10.2991/978-94-6463-540-9_69.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. Causal reasoning of entities and events in procedural texts. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 415–431, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.31. URL https://aclanthology.org/2023.findings-eacl.31/.

A Installation of MarODE

1. Install the package:

```
pip install marode
```

2. Import and initialize the evaluator:

```
from marode.evaluator import MarODEEvaluator, EvaluatorConfig, get_device

config = EvaluatorConfig()
device = get_device(-1)  # -1 for CPU, 0 for GPU
evaluator = MarODEEvaluator(config, device)
```

3. Prepare an entry with reasoning trace and evidence:

```
entry = {
    "id": "example_1",
    "reasoning_trace": "R0: ...\nR1: ...\nR2: ...",
    "evidence_text": ["evidence 1", "evidence 2"]
}
```

4. Score the entry and print only the scores:

```
scored_entry = evaluator.score_entry(entry)
scores = scored_entry["ourmetric"]
for k, v in scores.items():
    print(f"{k}: {v:.4f}")
```

B Models Used

Five widely used reasoning and instruction-following language models were selected as prompting backbones, including:

- **DeepSeek-R1-Distill-Qwen-14B** (Guo et al., 2025): a 14B distilled reasoning model based on Qwen2.5-14B. It was fine-tuned using reasoning data generated by DeepSeek-R1, allowing it to inherit advanced chain-of-thought reasoning capabilities discovered through large-scale reinforcement learning.
- **DeepSeek-R1-Distill-LLaMA-8B** (Guo et al., 2025): an 8B knowledge distilled reasoning model built upon Llama-3.1-8B. Using reasoning trajectories generated by DeepSeek-R1, it acquires advanced reasoning capabilities.
- **DeepSeek-R1-Distill-Qwen-7B** (Guo et al., 2025): a 7B distilled reasoning model based on Qwen2.5-Math-7B. The model is trained on reasoning samples generated by DeepSeek-R1.
- **Qwen2.5-3B** (Bai et al., 2023): a 3B model from the Qwen2.5 family of large language models. Qwen2.5 provides instruction-following capabilities, including strong mathematics and coding skills. It also supports context lengths of up to 128K tokens and multilingual processing.
- **GPT-OSS-20B** (Kumar et al., 2025): a 21B open-weight mixture-of-experts language model released by OpenAI, with 3.6B active parameters per token. The model is designed for reasoning-intensive and agentic tasks, supports configurable reasoning effort, and provides native capabilities for structured output generation, function calling, web browsing, and code execution.

Reasoning Trace Human Evaluation Questionnaire

Instruction: Please read the reasoning trace carefully and answer the following questions. For each question, select one value from **1 (Strongly Disagree)** to **5 (Strongly Agree)** by filling in one circle.

Q1: Do the reasoning steps follow a logical and consistent progression without contradictions?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q2: Are the reasoning steps clearly expressed and easy to understand?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q3: Do the reasoning steps each add new, useful information rather than repeating earlier steps?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q4: Are the reasoning steps well-grounded in the provided evidence or context, without introducing unsupported claims?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q5: Does the final verdict logically follow from the reasoning steps?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q6: Considering all aspects (flow, clarity, novelty, evidence, and verdict), how would you rate the overall quality of this reasoning trace?	○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 4: Human evaluation questionnaire used to assess the quality of reasoning traces. Annotators rate each question on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Perturbation Name	Scope	Description	Failure Mode Targeted
Penultimate Ambiguity	Local	Introduces ambiguity in the second-to-last reasoning step	Verdict justification
Temporal Confusion	Local	Alters or misorders temporal references	Temporal coherence
Unsupported Conclusion	Global	Introduces conclusions not supported by prior steps	Evidence grounding
Irrelevant Elaboration	Local	Adds tangential but fluent content	Logical focus
Duplication	Local	Repeats reasoning steps verbatim	Redundancy
Underspecification	Local	Removes critical explanatory details	Step completeness
Antonym Insertion	Local	Injects semantic opposites into key claims	Semantic consistency
Definitional Redundancy	Local	Repeats definitions without adding information	Conciseness
Modal Logic Confusion	Local	Confuses necessity, possibility, or certainty	Logical modality
Ordering	Global	Reorders reasoning steps	Causal structure
Quantifier Abuse	Local	Distorts universal or existential claims	Quantitative validity
Random Hyphenation	Surface	Inserts spurious hyphenation artifacts	Surface noise
Key Concept Swap	Local	Replaces central entities or concepts	Conceptual integrity
Final Verdict Insertion	Global	Injects premature verdicts into reasoning	Reasoning closure

Table 4: Summary of reasoning trace perturbation functions used in this study. Each perturbation targets a specific failure mode in multi-step reasoning, spanning local step level inconsistencies, global structural disruptions, and surface level noise. Perturbations are applied sequentially without replacement to each reasoning trace, with a maximum of six successfully applied perturbations per trace, ensuring controlled yet diverse degradations of reasoning quality.

C Prompt Design for Reasoning Trace Generation

To generate structured, evidence-grounded reasoning traces, we employ a standardized prompt template that explicitly constrains both the logical flow and output format of model responses (Figure 5). The prompt frames the model as an analyst tasked with reconstructing a step-by-step justification linking a factual claim and its associated evidence texts to a provided ground-truth verdict label, without relying on external knowledge or assumptions.

Each prompt instance consists of three components: (i) a factual claim, (ii) a set of evidence snippets retrieved from the fact-checking benchmark, and (iii) the corresponding veracity label. The model is instructed to produce a reasoning trace beginning with an initial claim restatement (R_0), followed by a variable-length sequence of evidence-grounded inference steps (R_1 through R_K), and concluding with a terminal `Final Verdict` explicitly restating the given label.

To ensure deterministic extraction and structural validity, reasoning spans are delimited using explicit markers `<Rstart>` and `<Rend>`. Few-shot exemplars embedded within the prompt illustrate progressive evidence attribution, contradiction resolution, and logical progression toward the verdict. The number of intermediate steps is left unconstrained, allowing trace length to adapt to claim complexity.

D Numerical Integration of Directional Consistency Dynamics

To stably integrate the directional consistency ODE

$$\frac{dp}{dt} = \rho(S_i - 0.5)p(1-p),$$

We employ a classical fourth-order Runge–Kutta (RK4) method, which provides a balance between numerical accuracy and computational efficiency for nonlinear dynamics.

Given the current belief state p_i and step signal S_i, the intermediate slopes are computed as:

$$\begin{aligned}
k_1 &= f(p_i, S_i), \\
k_2 &= f\left(p_i + \tfrac{\Delta t}{2} k_1, S_i\right), \\
k_3 &= f\left(p_i + \tfrac{\Delta t}{2} k_2, S_i\right), \\
k_4 &= f\left(p_i + \Delta t\, k_3, S_i\right),
\end{aligned}$$

where

$$f(p, S_i) = \rho(S_i - 0.5)p(1-p).$$

The belief state is then updated as

$$p_{i+1} = p_i + \frac{\Delta t}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right),$$

followed by clipping to the unit interval. This procedure is repeated across all reasoning steps, producing the final directional consistency score $DC(r) = p_K$.

RK4 is widely used for integrating nonlinear dynamical systems due to its fourth-order accuracy and stability properties, making it suitable for modeling smooth belief trajectories across discrete reasoning steps.

E Baseline Methodologies

ROSCOE (Reasoning and Stepwise Coherence Evaluation). ROSCOE (Golovneva et al., 2023) is a modular framework for diagnosing multiple failure modes in chain-of-thought reasoning without requiring gold references. It decomposes reasoning quality into interpretable components capturing semantic alignment, logical validity, and linguistic coherence. ROSCOE-SA and ROSCOE-SS compute embedding-based semantic similarity between adjacent steps and between steps and conclusions to assess topical consistency and semantic progression. ROSCOE-LI leverages natural language inference models to quantify entailment and contradiction relations across reasoning steps, enabling detection of logical breaks or unsupported transitions. ROSCOE-LC evaluates surface-level fluency and syntactic coherence using pretrained language acceptability models. These component scores can be aggregated to yield a holistic quality measure, while maintaining diagnostic transparency across distinct reasoning error categories.

RecEval (Reasoning Chain Evaluation). RecEval (Prasad et al., 2023) evaluates reasoning traces by jointly modeling correctness and informativeness of intermediate steps. Logical correctness is estimated using entailment-based verification between reasoning steps and final conclusions, ensuring that intermediate claims are supported. Informativeness quantifies the degree to which each step contributes novel, non-redundant information toward solving the task. By integrating these dimensions, RecEval penalizes reasoning chains that are either logically flawed or unnecessarily repetitive, framing reasoning quality as a contribution-based process rather than purely semantic similarity.

Local and Global Coherence Metrics. Local and Global Coherence (Kotonya & Toni, 2020) approaches assess explanation quality by separating step-wise justification from overall narrative consistency. Local coherence evaluates whether individual reasoning steps are supported by preceding context, typically through semantic similarity or entailment relations between adjacent steps. Global coherence measures whether the complete reasoning trace forms a logically consistent and goal-directed explanation supporting the final conclusion. This decomposition enables detection of both localized reasoning gaps and holistic inconsistencies, though it generally relies on discrete pairwise comparisons without modeling cumulative reasoning dynamics.

LLM-as-a-Judge. The LLM-as-a-Judge paradigm formulates reasoning trace evaluation as a direct qualitative scoring task performed by a large language model. Given a generated reasoning trace, the evaluator model is prompted to assign a scalar quality score. Rather than relying on explicit symbolic metrics, this approach leverages the implicit reasoning and evaluative capabilities of modern LLMs to approximate human judgment at scale.

As illustrated in Figure 6, the evaluation prompt constrains outputs to a single continuous numeric score, promoting consistency and ease of aggregation across large trace collections. While effective in practice, this paradigm remains sensitive to prompt phrasing, evaluator model biases, and domain shifts, and offers limited transparency due to reliance on latent evaluation heuristics rather than explicit reasoning criteria.

F Additional Discussion

F.1 Perturbation sensitivity and metric generalization.

ROSCOE (Golovneva et al., 2023) exhibited perfect correlation across many datasets under its curated perturbations; however, this performance does not generalize when the perturbation set changes. In our experiments, altering the perturbation set leads to severe and inconsistent drops in correlation, indicating that ROSCOE and similar metrics are strongly coupled to the specific mechanical perturbation patterns they are designed to detect (Moradi & Samwald, 2021). As a result, high correlation under a given perturbation set may reflect perturbation-specific sensitivity rather than generalized reasoning evaluation. Effective reasoning metrics should remain largely independent of particular perturbation operators and instead capture general structural properties of reasoning traces (Wang et al., 2022a). MarODE is designed to capture this motivation, focusing on coherence, directionality, redundancy, and evidential grounding rather than perturbation-specific cues. Although this general formulation yields lower raw correlation scores, MarODE demonstrates higher effectiveness in detecting diverse human-centric perturbations and maintains stable behavior under different distributional settings compared to other baselines, making it better suited for generalized reasoning evaluation.

F.2 Significance of evidence in reasoning trace evaluation.

An important insight from our analysis concerns the role of explicit evidence alignment in reasoning evaluation. While evidence grounding is often assumed to be a central component of high-quality reasoning, our results suggest a more nuanced picture. Evidence alignment (γ) exhibits moderate standalone performance across perturbation-based evaluations and human judgments. However, its contribution is inconsistent when combined with other components. In several settings, adding evidence alignment to coherence-centred formulations yields marginal or even negative effects on correlation strength (Table 1). This pattern indicates that explicit evidence signals can sometimes introduce noise, particularly when evidence relevance or attribution is ambiguous. Empirically, the coherence-quality pairing ($\alpha\beta$) achieves higher correlations than the

combination of all three components. These findings suggest that structural coherence, captured through Markovian transitions, plays a more foundational role in perceived reasoning quality than evidence matching. Evidence acts as a secondary, context-dependent signal, strengthening evaluation only when it directly supports intermediate steps. This highlights important implications for metric design, cautioning against over-reliance on evidence signals. An alternative interpretation relates to the perturbations themselves. Reasoning traces are often evaluated in isolation, without access to the full evidence used during generation (Prasad et al., 2023; Fang et al., 2024). As a result, perturbations applied only to reasoning steps may not adequately capture groundedness. Measuring evidence sensitivity may therefore require perturbations that directly target factual consistency and evidence attribution, beyond local, global, or surface-level trace manipulations.

Prompt: Reasoning Trace Generation

You are an analyst. Your task is to generate a clear, logically structured reasoning trace that explains and justifies a given verdict for a factual claim. You will be provided with a factual claim, a set of evidence texts, and a verdict label (Supported, Refuted, Partially Supported, Misleading, or Unverifiable). Your task is to reconstruct the reasoning that connects the provided evidence to the final verdict, without relying on any external knowledge or assumptions. The number of reasoning steps (R1…Rk) can vary based on complexity.

Output Format (use exactly):
<Rstart>
R0: [Interpret or restate the claim in clearer terms]
R1: [First reasoning step grounded in evidence, building toward the given verdict]
R2: [Second reasoning step grounded in evidence, continuing the logical path]
...
Rk: [Final reasoning step that solidifies the justification for the verdict]
Final Verdict: [Restate the provided verdict label here]
<Rend>

Here are some examples for the task:
Example 1
Input
Claim: " Georgia election worker threw away a ballot"
Label: "REFUTED"
Evidence Text: [
"Viral video claimed worker discarding ballot online.",
"Officials said discarded paper contained voter instructions.",
"Ballots processed elsewhere, worker opened envelopes only.",
"Size difference proved paper was not ballot.",
"Elections director explicitly denied ballot disposal claim.",
"Social media flagged post as misinformation effort.",
"Worker faced threats after identity was exposed.",
"Officials rated claim false after investigation review."
]
Output
<Rstart>
R0: Claim says Georgia worker threw away ballot.
R1: Social media post falsely alleged ballot disposal.
R2: Worker discarded voter instructions, not ballot paper.
R3: Video shows smaller paper than official ballot.
R4: Officials confirmed worker opened envelopes, not ballots.
R5: Claim deemed false; evidence supports misinformation classification.
<Rend>

Task: {{ test claim, label and evidence texts }}

Please provide the reasoning traces. Ensure that the reasoning output is between <Rstart> and <Rend>.

Figure 5: Standardized prompt template used for reasoning trace generation. The prompt provides the factual claim, associated evidence texts, and ground truth verdict label, and constrains model outputs to a structured multi step reasoning format delimited by explicit markers. Few-shot exemplars guide evidence grounding and logical progression toward the final verdict, enabling deterministic extraction and structural validation of generated reasoning traces.

Prompt: LLM as a Judge

Evaluate the reasoning trace using the scoring rubric below.
Assume the evidence to be true.
Scoring Rubric
1.00 (Outstanding)
0.75 (Good)
0.50 (Adequate)
0.25 (Poor)
0.00 (Critically Flawed)

CLAIM: {claim}
LABEL: {label}
EVIDENCE: {evidence_str}
REASONING TRACE: {reasoning_trace}

Provide ONLY a numerical score between 0.00 and 1.00 with exactly two decimal places. Do not include any text, explanation, or formatting.

Score:

Figure 6: Prompt template used in the LLM-as-a-Judge evaluation paradigm. The evaluator model is instructed to assess the quality of a given reasoning trace and output a single scalar score on a continuous scale, without producing explanatory text. This formulation enables scalable qualitative assessment but relies on the model's implicit evaluation heuristics.

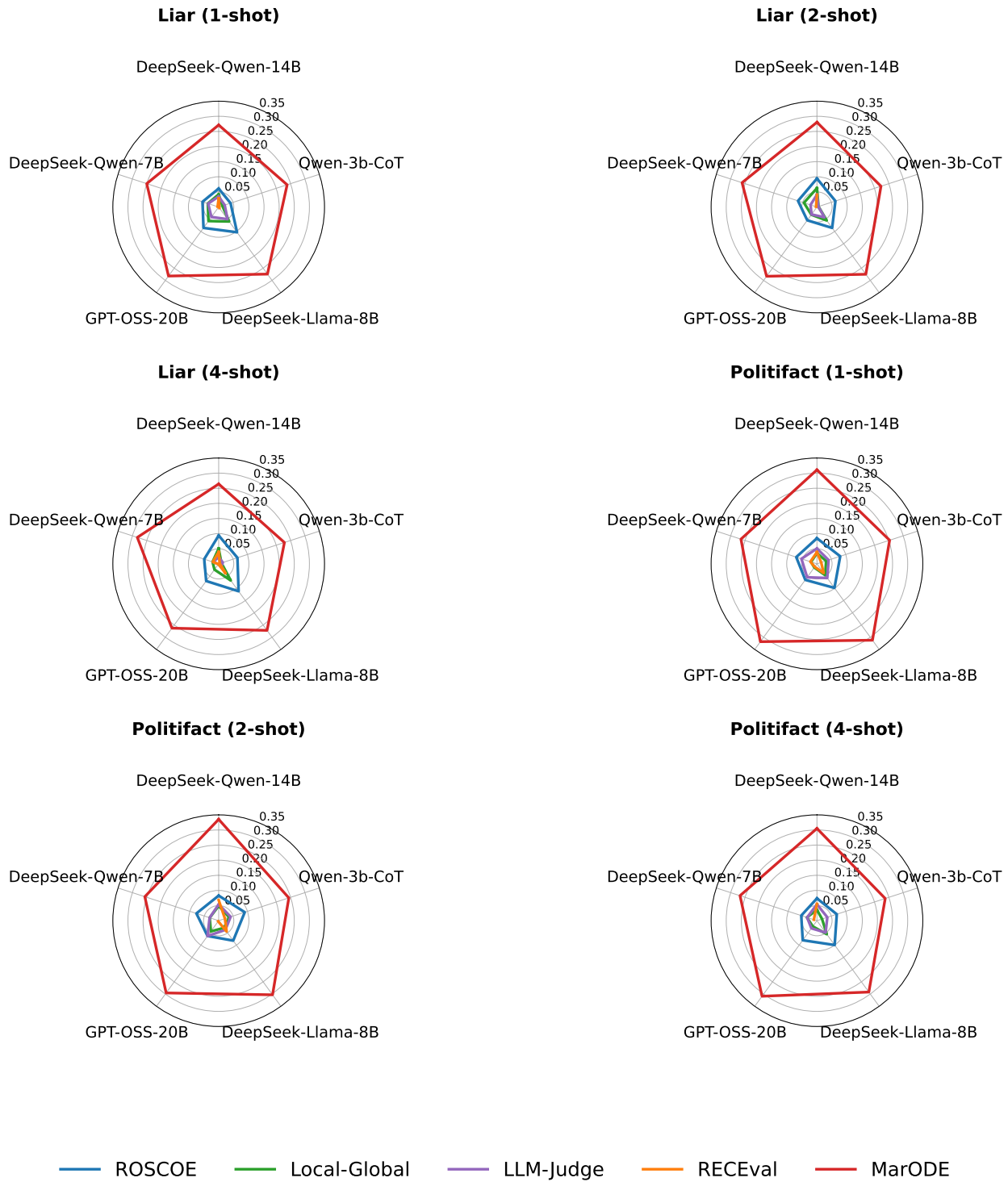

Figure 7: **Comparative performance of post-hoc reasoning evaluation metrics across datasets and models, visualized as radar charts.** Each subplot corresponds to a specific dataset (LIAR and PolitiFact under 1, 2, and 4-shot prompting), with axes representing the five backbone models. Colored lines denote metric-wise Somers' D correlations with human-centric perturbations for ROSCOE, Local-Global, LLM-as-a-Judge, ReCEval, and our proposed MarODE. MarODE consistently exhibits the highest correlations across datasets, shot counts, and model families, indicating stronger sensitivity to perturbations that affect reasoning quality.