Which Reasoning Tokens Matter? A Token-Level Study of Effort, Faithfulness, and Control in Language Model Reasoning

Anonymous ACL submission

Abstract

Chain-of-thought reasoning is typically evaluated at the level of complete traces, implicitly assuming that all reasoning tokens contribute similarly to final predictions. We challenge this assumption and ask a finer-grained question: which reasoning units actually matter? We present a unified framework that analyzes token- and clause-level reasoning units along three functional axes: effort, faithfulness, and control. Across four reasoning benchmarks and three language models, we find that reasoning traces are highly heterogeneous: a limited subset of units consistently carries inferential load, whereas many others play organizational or stylistic roles. Faithfulness is the strongest predictor of final correctness, and a combined score over the three axes substantially outperforms length, confidence, and position-based baselines. Budget-matched interventions further show that removing high-scoring units causes markedly larger accuracy drops than competing selection strategies. These results suggest that reasoning quality is better characterized by the functional distribution of reasoning units than by trace length alone.

1 Introduction

Chain-of-thought reasoning has become a common interface through which language models expose intermediate computation. When a model produces a reasoning trace before its final answer, that trace is often taken as evidence of deeper, more careful, or more interpretable reasoning. In practice, however, most evaluations still treat the trace as a whole. A chain is usually judged by its outcome, its length, or its overall plausibility, which implicitly assumes that its reasoning tokens contribute in similar ways to the final prediction. This assumption is convenient, but it leaves a basic question unanswered: within a reasoning trace, which parts actually matter?

This question is not only about granularity. Two traces of similar length can differ sharply in quality because they allocate their content differently. One may contain a small number of units that genuinely support the answer, while another may spend much of its content on connective phrases, stylistic fillers, or local restatements that make the output look more deliberate without changing the result. As a result, trace length is often a weak proxy for reasoning quality, and evaluating a trace as a whole can blur the difference between useful reasoning and rhetorically elaborate text. A similar divide appears in nearby work: effort asks whether longer traces reflect deeper computation, faithfulness asks whether the stated reasoning truly supports the answer, and controllability asks whether reasoning can be steered in a behaviorally meaningful way. Yet these questions are usually studied separately and rarely tied to a shared unit inside the trace.

We argue that reasoning traces should be analyzed as heterogeneous structures rather than monolithic text sequences. Our starting point is simple: reasoning units are not equally important. Some units directly support the model's decision, some reflect costly intermediate computation that does not reliably lead to correct outcomes, some are easy to control without carrying much inferential value, and many serve mainly organizational or stylistic roles. To make this distinction operational, we study reasoning at the level of tokens and clause spans and score each unit along three axes: effort, faithfulness, and control. Effort captures whether a unit is associated with deeper or less trivial computation, faithfulness captures whether perturbing the unit changes the final prediction in a meaningful way, and control captures whether the unit can be externally constrained in a stable and consequential manner. This shifts the focus from whether a chain of thought is good on average to which parts of the chain actually do the work.

We study this question across four reasoning

benchmarks and three language models spanning a general instruction model, a model with explicit thinking mode, and a reasoning distilled model. The evaluation suite combines datasets with different forms of intermediate supervision, including supporting facts, decompositions with evidence, entailment trees, and formal proofs. We pair this suite with a unified analysis pipeline that segments generated traces into minimal reasoning units, scores each unit along the three axes, and studies their distribution through correlation analysis, targeted interventions, taxonomy induction, and human calibration.

Our results show that reasoning traces are strongly heterogeneous. Faithfulness is the strongest predictor of final correctness, while a combined score over effort, faithfulness, and control consistently outperforms length, confidence based measures, and position based baselines. Budget matched interventions provide direct causal evidence: removing high scoring reasoning units causes substantially larger drops in accuracy than deleting units selected by competing strategies, and clause spans emerge as more stable intervention targets than isolated tokens or full reasoning steps. The three axes are related but clearly non redundant, and successful traces differ in composition from unsuccessful ones. In particular, correct traces contain more load bearing units, whereas incorrect traces contain more fragile and weakly faithful content. These findings suggest that the key variable is not how much reasoning text a model produces, but how that text distributes functional importance.

Our work reframes visible reasoning analysis around functionally meaningful units and connects three fragmented dimensions of reasoning evaluation within a single empirical framework. We introduce a token and clause span scoring pipeline for effort, faithfulness, and control; construct a multi benchmark evaluation suite with diverse forms of intermediate supervision; show through correlation and intervention analyses that these functional scores explain correctness better than standard trace proxies; derive an interpretable taxonomy that separates load bearing units from more fragile or decorative content; and validate the framework with human annotation and alignment to external structure. Taken together, these contributions support a simple claim: reasoning traces should be evaluated not only by whether they exist or how long they are, but by which parts of them actually matter.

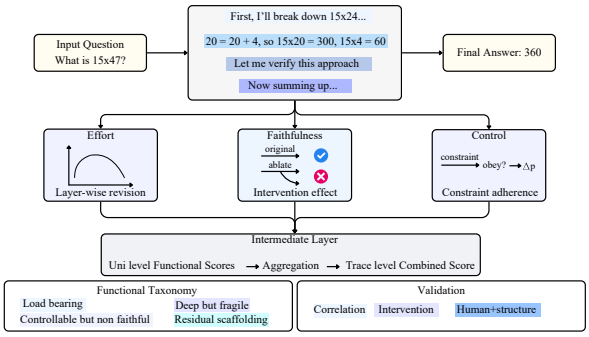

Figure 1: Overview of the proposed framework. Given an input question, a language model produces a visible reasoning trace and a final answer. We segment the reasoning trace into token- and span-level reasoning units, score each unit along three functional axes—effort, faithfulness, and control—and aggregate these unit-level scores into trace-level summaries and a combined score. The resulting structure supports two downstream views: an interpretable functional taxonomy over reasoning units and a validation pipeline based on correlation analysis, targeted interventions, and human/external-structure evaluation.

2 Related Work

Explicit reasoning in language models is commonly studied through chain-of-thought generation and test-time inference strategies. Early work showed that prompting models to produce intermediate steps can improve performance on complex tasks, while later methods explored sampling, search, and broader test-time scaling strategies (Wei et al., 2023; Wang et al., 2023; Yao et al., 2023; Zhang et al., 2025). More recent work argues that longer reasoning traces do not necessarily imply better reasoning, and that more refined measures of inference-time computation are needed (Chen et al., 2026; Tajwar et al., 2026; Ritter et al., 2026). Our work is complementary to this line: rather than proposing a new reasoning method, we ask how produced reasoning traces should be analyzed once they exist, and whether their internal functional composition explains correctness better than whole-trace proxies such as length or confidence.

A second line of work studies whether natural-language explanations or reasoning traces are faithful to the computation that leads to the final answer. Earlier work in NLP showed that attention weights and plausible rationales are not necessarily faithful explanations (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; DeYoung et al., 2020; Madsen et al., 2022). Recent

reasoning-specific work sharpens this concern for chain-of-thought models, showing that models do not always rely on the reasoning they verbalize and that plausible chains can still be weakly connected to final predictions (Paul et al., 2024; Arcuschin et al., 2025; Cornish and Rogers, 2025; Tutek et al., 2025). Other work attempts to improve explanation faithfulness directly (Wang and Atanasova, 2025; Kamp et al., 2025; Swaroop et al., 2025). Our paper shares the faithfulness perspective, but treats faithfulness as only one axis of reasoning quality, alongside effort and control.

Recent work has also moved beyond treating reasoning as an undifferentiated text sequence. Several studies emphasize that meaningful analysis often requires reasoning units above single tokens but below full traces, such as steps, clauses, or reusable substructures (Yu et al., 2026; Aggarwal et al., 2026; Pan et al., 2026; Lightman et al., 2023). At the same time, work on monitorability and controllability shows that visible reasoning occupies an ambiguous space between interpretability and behavioral influence (Korbak et al., 2025; Yang et al., 2026; Meek et al., 2025; Yueh-Han et al., 2026). This distinction is central to our work: we explicitly measure whether reasoning units are behaviorally important and whether they are controllable, and show that the two are related but non-equivalent.

A broader neighboring literature studies reasoning through hidden states, latent computation, and mechanistic representations rather than explicit traces (Amos et al., 2026; Liu et al., 2026a; Lepori et al., 2026; Cunningham et al., 2023; Luo et al., 2026; Xu et al., 2026; Liu et al., 2026b; Duvvuri et al., 2026). Our paper is narrower in scope: we do not aim to recover hidden reasoning in full, but instead analyze visible reasoning traces as behaviorally meaningful interfaces. The gap we address is the lack of a unified framework that studies fine-grained reasoning units inside those traces along the three dimensions most directly tied to reasoning quality: effort, faithfulness, and control.

3 Problem Setup

We study visible reasoning traces produced by a language model. Let x denote an input instance and y its gold answer. Under a fixed prompting and decoding protocol, a model M generates a reasoning trace r and a final answer $\hat{y}$:

$$(r, \hat{y}) = M(x).$$

Our goal is not to improve generation, but to analyze the functional structure of r: which parts of a visible reasoning trace contribute to the final prediction, and which parts are mainly stylistic or organizational.

Each benchmark instance may also include external intermediate supervision g, such as supporting facts, decomposition steps, entailment trees, or proof annotations. We do not treat g as token-level ground truth for generated reasoning; instead, it is used only for evaluation and calibration. The evaluation set is

$$\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^{N}.$$

For each generated trace r_i, we segment the reasoning span into analyzable units

$$U(r_i) = \{u_{i1}, u_{i2}, \ldots, u_{in_i}\},$$

where a unit may be a token or a clause span. This gives two complementary views of the same trace: token-level analysis for maximal resolution and span-level analysis for more coherent local structure.

Each reasoning unit u_{ij} receives three functional scores,

$$E(u_{ij}), \qquad F(u_{ij}), \qquad C(u_{ij}),$$

corresponding to effort, faithfulness, and control. At a high level, effort measures whether a unit reflects non-trivial computation, faithfulness measures whether it is behaviorally important for the final answer, and control measures whether it can be externally constrained in a stable and consequential way. From these unit-level scores, we derive trace-level summaries by aggregating over all units in a trace; the precise definitions are given in Section 4.

Final performance is represented by a binary correctness variable

$$z_i = \mathbb{I}[\hat{y}_i = y_i].$$

This setup supports two questions: do functional summaries explain correctness better than trace-level proxies such as length, confidence, or position, and do top-ranked reasoning units have greater causal importance under intervention?

To study the second question, we define an intervention operator $\mathcal{I}$ that modifies a selected subset of reasoning units while leaving the final-answer span untouched:

$$r_i' = \mathcal{I}(r_i, S_i),$$

where $S_i \subseteq U(r_i)$. After intervention, the model produces a new answer $\hat{y}_i'$ with correctness indicator z_i'. If a scoring function identifies functionally important units, then intervening on its top-ranked units should produce larger drops in answer accuracy or confidence than budget-matched baselines.

Finally, we use the three axes to induce a coarse taxonomy over reasoning units. The goal is not to claim that reasoning falls into perfectly discrete categories, but to provide an interpretable summary of how visible traces distribute functional importance. The central problem of the paper is therefore to define and evaluate unit-level measures that identify which parts of a visible reasoning trace matter for the final prediction, and whether this functional distribution explains reasoning quality better than standard trace-level proxies.

4 Method

Our framework treats a visible reasoning trace as a sequence of reasoning units and assigns each unit three functional scores: effort, faithfulness, and control. The pipeline has four stages: segment the trace into analyzable units, compute unit-level scores, aggregate them to the trace level, and derive a four-way taxonomy.

4.1 Reasoning units and functional scores

Each model output is parsed into a reasoning span and a final-answer span using a fixed delimiter-based format. Invalid parses or malformed final answers are excluded. For each valid trace, we construct two aligned views: a token-level view and a clause-span view,

$$U(r_i) = \{u_{i1}, u_{i2}, \ldots, u_{in_i}\}.$$

Clause spans are the primary analysis unit because they preserve local semantic structure better than isolated tokens while remaining more localized than full reasoning steps. Unless otherwise noted, trace-level summaries are computed over clause spans.

The effort score measures whether a reasoning unit reflects non-trivial internal computation. For token position t, let $p_\ell^{(t)}$ denote the next-token distribution induced by layer ℓ. We compute three signals: adjacent-layer revision,

$$R^{(t)} = \frac{1}{L-1} \sum_{\ell=1}^{L-1} \mathrm{JSD}\left(p_\ell^{(t)}, p_{\ell+1}^{(t)}\right),$$

late convergence,

$$S^{(t)} = \frac{K^{(t)} - 1}{L-1}$$

$$K^{(t)} = \min\left\{\ell : \hat{v}_\ell^{(t)} = \hat{v}_{\ell+1}^{(t)} = \cdots = \hat{v}_L^{(t)}\right\},$$

and hidden-state update magnitude,

$$H^{(t)} = \frac{1}{L-1} \sum_{\ell=1}^{L-1} \left(1 - \cos\left(h_\ell^{(t)}, h_{\ell+1}^{(t)}\right)\right).$$

We combine them into

$$E_{\mathrm{tok}}(t) = \alpha_R \, \widetilde{R}^{(t)} + \alpha_S \, \widetilde{S}^{(t)} + \alpha_H \, \widetilde{H}^{(t)},$$

where tildes denote within-dataset, within-model normalization and $\alpha_R + \alpha_S + \alpha_H = 1$. We use uniform weights. The effort score of a clause span is the mean over its tokens:

$$E(u) = \frac{1}{|u|} \sum_{t \in u} E_{\mathrm{tok}}(t).$$

The faithfulness score measures whether a reasoning unit is behaviorally important for the final prediction. For each unit u, we delete it from the visible reasoning trace, regenerate the final-answer span, and measure the downstream effect. Let $\hat{y}_i$ be the original answer and $\hat{y}_i^{(-u)}$ the regenerated answer after deleting u. Let $p_i(y_i \mid r_i)$ and $p_i(y_i \mid r_i^{(-u)})$ denote the normalized probabilities assigned to the gold answer before and after intervention. We define

$$D(u) = \lambda_1 \left[p_i(y_i \mid r_i) - p_i(y_i \mid r_i^{(-u)})\right]$$

$$+\lambda_2 \, \mathbb{I}\left[\hat{y}_i \neq \hat{y}_i^{(-u)}\right] + \lambda_3 \, \mathbb{I}\left[z_i > z_i^{(-u)}\right],$$

and normalize by unit size:

$$F(u) = \frac{D(u)}{\max(|u|, 1)}.$$

The main-paper faithfulness score uses deletion as the primary intervention family. Implementation details for intermediate-layer distributions, gold-answer scoring, and control prompts are given in Appendix A.4.

The control score measures whether a reasoning unit can be externally constrained in a stable and consequential way. We use lexical, structural, and semantic control prompts. For each unit u, we compute the average adherence score $a(u)$ over valid prompts and the average downstream effect $\Delta(u)$ as the absolute change in normalized gold-answer probability. We define

$$C(u) = a(u) \cdot \Delta(u).$$

4.2 Trace-level aggregation and interventions

Given unit-level scores, we compute trace-level summaries by averaging over all scored units:

$$\bar{E}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} E(u_{ij}), \qquad \bar{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} F(u_{ij}),$$

$$\bar{C}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} C(u_{ij}),$$

and, to compare scores across datasets and models, we standardize each axis within each dataset-model block and define

$$S_i = 0.30\, z_E(i) + 0.50\, z_F(i) + 0.20\, z_C(i),$$

with weights fixed on a disjoint pilot pool.

Our main causal comparison uses a budget-matched protocol. For each trace, every method removes 15% of reasoning-span tokens, rounded to the nearest integer, with a minimum of 4 and a maximum of 24. The final-answer span is never modified. Span-level methods greedily accumulate clause spans until the budget is met or minimally exceeded; token-level methods remove the highest-ranked tokens directly.

4.3 Taxonomy and evaluation

To summarize functional composition, we derive a four-way taxonomy from normalized effort, faithfulness, and control scores within each dataset-model block. Let high denote the top quartile and low denote the bottom half. Each unit is assigned, in priority order, to one of four classes: load-bearing, deep-but-fragile, controllable-but-non-faithful, and residual scaffolding.

We evaluate the framework in two ways. First, we compare derived unit labels against human annotations on a triple-annotated calibration subset. Second, for benchmarks with external intermediate supervision, we project supporting facts, decomposition steps, entailment steps, or proof steps into the same span-level comparison space and measure alignment with generated reasoning units. These signals are used only for evaluation and interpretation.

5 Experiments Setup

We test whether unit-level functional scores explain reasoning quality better than standard trace-level proxies, whether targeted interventions identify behaviorally important reasoning units, and whether the induced taxonomy is supported by human judgment and dataset-provided intermediate structure.

5.1 Benchmarks and models

We evaluate on four benchmarks with intermediate supervision: HotpotQA, StrategyQA, EntailmentBank, and ProofWriter. The final filtered suite contains 2,900 items: 800, 700, 600, and 800 examples, respectively.

We use three language models with different reasoning profiles: Llama-3.1-8B-Instruct, Qwen3-8B with thinking mode enabled, and DeepSeek-R1-Distill-Qwen-32B. Unless otherwise noted, all experiments use prompt family P1, temperature 0.2, top-p 0.95, and one trace per item, with maximum reasoning length 256 for 8B models and 384 for the 32B model. Outputs follow a fixed reasoning/answer format, and invalid traces are counted as incorrect in pooled accuracy.

5.2 Evaluation protocol

Each valid trace is analyzed at two granularities: tokens and clause spans, with clause spans as the primary unit. Trace-level summaries are obtained by averaging unit scores over clause spans.

For trace-level analyses, we compare against token count, average log-probability, self-certainty, and a position baseline, and report Spearman correlations with bootstrap confidence intervals. For EntailmentBank and ProofWriter, bootstrap is performed at the source-tree level.

For intervention experiments, we compare against position-only deletion, random token deletion, attention-saliency deletion, gradient-saliency deletion, and step-level deletion. Our main causal comparison uses a budget-matched protocol: each method removes 15% of reasoning-span tokens, rounded to the nearest integer, with a minimum of 4 and a maximum of 24, while leaving the final-answer span untouched. We report accuracy drop, answer-flip rate, change in normalized gold-answer probability, and invalid-output rate.

To study complementarity among effort, faithfulness, and control, we compute pairwise correlations and cross-axis disagreement rates. We then derive a four-way taxonomy over reasoning units and compare class proportions between correct and incorrect traces.

5.3 Human calibration and implementation

We evaluate the taxonomy on a human-annotated subset of 240 traces and 2,462 annotated spans. Each trace is independently labeled by three annotators, and agreement is measured with Krippen-

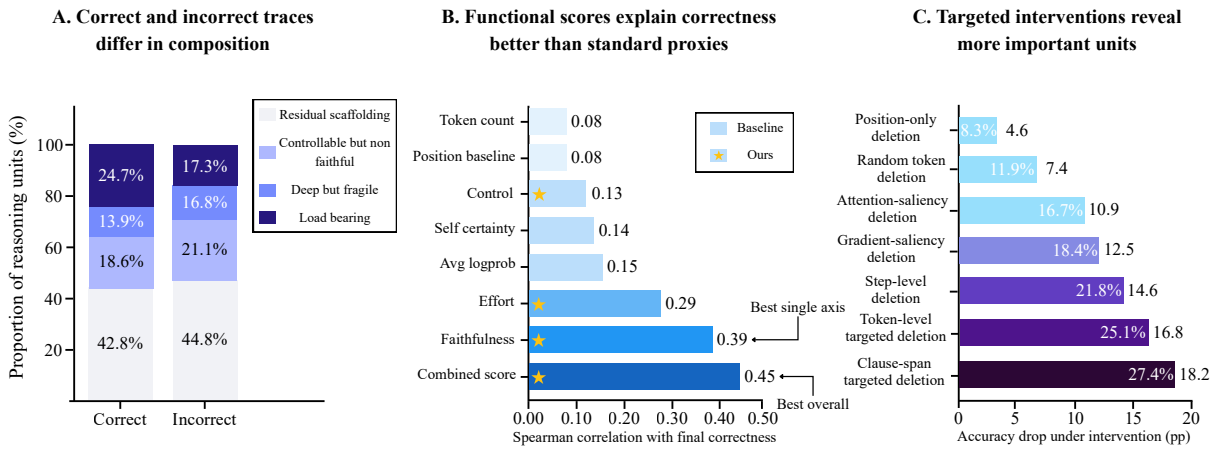

Figure 2: Core empirical findings. (A) Correct and incorrect traces differ in functional composition: correct traces contain more load-bearing units, whereas incorrect traces contain more deep-but-fragile and controllable-but-non-faithful content. (B) Functional scores explain final correctness better than standard trace-level proxies, with faithfulness as the strongest single axis and the combined score performing best overall. (C) Under budget-matched interventions, removing units selected by our method causes larger accuracy drops than competing baselines, especially at the clause-span level.

dorff's α. Automatic predictions are evaluated with macro-F1 against adjudicated human labels. For datasets with external intermediate supervision, we also project supporting facts, decomposition steps, entailment steps, or proof steps into a common span-level comparison space and measure alignment with generated reasoning units.

The pipeline is implemented in Python using PyTorch, Transformers, vLLM, Accelerate, NumPy, pandas, SciPy, scikit-learn, and a deterministic English segmentation stack. The total compute footprint is 409.2 A100-equivalent GPU hours. Full implementation details and robustness analyses are reported in the appendix.

6 Results

Figure 2 summarizes the main empirical picture. Across four reasoning benchmarks and three language models, reasoning traces are functionally heterogeneous: correct traces contain more load-bearing content, functional scores explain correctness better than standard trace-level proxies, and targeted interventions on highly ranked units cause larger performance drops than competing baselines.

Table 1 shows that standard trace-level proxies are weak predictors of final correctness. Token count and the position baseline both reach only 0.08 mean correlation, and token count is uninformative on StrategyQA (-0.03). Confidence-based metrics are slightly stronger, with mean correlations of 0.15 for average log-probability and 0.14 for self-certainty. All three functional scores are more informative: effort reaches 0.29, faithfulness is the strongest axis at 0.39, and control reaches 0.13. The combined score performs best overall, with a mean correlation of 0.45 and the strongest value on every benchmark. This suggests that reasoning quality is better explained by the functional distribution of reasoning units than by length, confidence, or position alone.

Table 2 provides causal support for this result. Under a matched deletion budget, simple baselines produce limited damage: position-only deletion reduces accuracy by 4.6 points, random deletion by 7.4, attention saliency deletion by 10.9, gradient saliency deletion by 12.5, and step-level deletion by 14.6. Both targeted methods are stronger: token-level targeted deletion causes a 16.8 point drop in accuracy and a 25.1% answer-flip rate, while clause-span targeted deletion reaches 18.2 points and 27.4%. Invalid-output rates remain low, suggesting that the framework identifies behaviorally important reasoning content rather than merely selecting late or locally salient text, and that clause spans are a more stable intervention unit than isolated tokens.

Table 3 shows that effort, faithfulness, and control are related but not redundant. Their mean pairwise correlations are moderate: 0.31 for effort and faithfulness, 0.21 for effort and control, and 0.29 for faithfulness and control. At the same time, disagreement regions are substantial: on average, 9.6% of units are high in effort but low in faith-

Table 1: Association between reasoning-related metrics and final correctness.

Metric	HotpotQA	StrategyQA	EntailmentBank	ProofWriter	Mean
Token count	0.11 [0.05, 0.17]	-0.03 [-0.09, 0.04]	0.17 [0.09, 0.24]	0.08 [0.02, 0.14]	0.08 [0.04, 0.11]
Avg logprob	0.19 [0.13, 0.25]	0.14 [0.07, 0.20]	0.16 [0.09, 0.23]	0.12 [0.05, 0.18]	0.15 [0.10, 0.19]
Self-certainty ($-$entropy)	0.17 [0.11, 0.23]	0.09 [0.02, 0.15]	0.18 [0.11, 0.25]	0.11 [0.04, 0.17]	0.14 [0.10, 0.18]
Position baseline	0.09 [0.03, 0.15]	0.02 [-0.04, 0.08]	0.13 [0.06, 0.20]	0.06 [-0.01, 0.12]	0.08 [0.04, 0.12]
Effort score (ours)	0.31 [0.25, 0.36]	0.27 [0.20, 0.34]	0.34 [0.27, 0.41]	0.22 [0.15, 0.29]	0.29 [0.25, 0.33]
Faithfulness score (ours)	0.38 [0.32, 0.44]	0.29 [0.22, 0.36]	0.41 [0.35, 0.48]	0.47 [0.41, 0.54]	0.39 [0.35, 0.43]
Control score (ours)	0.12 [0.05, 0.18]	0.06 [-0.01, 0.12]	0.19 [0.12, 0.26]	0.15 [0.08, 0.21]	0.13 [0.09, 0.17]
Combined score (ours)	0.44 [0.39, 0.50]	0.36 [0.29, 0.43]	0.49 [0.43, 0.55]	0.51 [0.45, 0.57]	0.45 [0.41, 0.48]

Table 2: Budget-matched answer sensitivity under targeted interventions.

Intervention / baseline	ΔAcc (pp)	Answer flip (%)	ΔP(correct)	Invalid output rate (%)
Position-only deletion	-4.6 [-5.8, -3.5]	8.3 [6.8, 9.9]	-0.058 [-0.071, -0.046]	0.5
Random token deletion	-7.4 [-8.9, -5.9]	11.9 [10.2, 13.6]	-0.088 [-0.103, -0.075]	1.8
Attention-saliency deletion	-10.9 [-12.5, -9.3]	16.7 [14.9, 18.6]	-0.134 [-0.150, -0.119]	1.2
Gradient-saliency deletion	-12.5 [-14.2, -10.8]	18.4 [16.5, 20.4]	-0.149 [-0.166, -0.133]	1.5
Step-level deletion	-14.6 [-16.2, -12.8]	21.8 [19.9, 23.9]	-0.171 [-0.188, -0.154]	0.8
Token-level targeted deletion (ours)	-16.8 [-18.6, -14.9]	25.1 [23.0, 27.3]	-0.193 [-0.211, -0.176]	1.3
Clause-span targeted deletion (ours)	-18.2 [-20.1, -16.4]	27.4 [25.2, 29.7]	-0.209 [-0.228, -0.191]	0.9

fulness, 8.5% are high in faithfulness but low in control, and 5.9% are high in control but low in faithfulness. Visible reasoning traces therefore cannot be reduced to a single scalar notion of reasoning quality.

Table 4 shows how this non-redundancy translates into trace composition. Correct traces contain substantially more load-bearing units than incorrect traces: 24.7% versus 17.3%, a difference of $+7.4$ points. Incorrect traces instead contain more deep-but-fragile units (16.8% versus 13.9%) and more controllable-but-non-faithful units (21.1% versus 18.6%). The residual scaffolding class is large in both groups, but the difference is not significant. In other words, successful traces are not simply those with less filler; they allocate more of their content to units that are genuinely load-bearing.

Human and external validation results, reported in the appendix, support the same picture. On the annotated subset, our method reaches $\alpha = 0.74$ agreement, macro-F1 $= 0.71$, and external-structure F1 $= 0.63$, substantially above the position baseline (0.43). Together with the intervention results, this provides converging evidence that the proposed scores capture meaningful variation within visible reasoning traces.

Overall, Figure 2 and Tables 1–4 support a consistent conclusion: reasoning quality is better characterized by the functional distribution of reasoning units than by trace length alone.

7 Discussion

Our results support a view of visible reasoning traces. Chain-of-thought is often treated as a single object whose observable properties are its presence, length, or plausibility. Under that view, longer traces are taken as evidence of deeper reasoning. Our findings suggest a selective picture. Visible traces are functionally heterogeneous: some units are tied to the final decision, some reflect effort without stable usefulness, some are controllable without carrying much inferential weight, and

Table 3: Joint analysis of effort, faithfulness, and control.

Dataset	Corr(E,F)	Corr(E,C)	Corr(F,C)	High-E low-F (%)	High-F low-C (%)	High-C low-F (%)
HotpotQA	0.34 [0.28, 0.39]	0.19 [0.12, 0.26]	0.28 [0.21, 0.34]	10.3	9.8	6.7
StrategyQA	0.27 [0.20, 0.34]	0.11 [0.03, 0.18]	0.17 [0.09, 0.24]	11.8	10.7	7.1
EntailmentBank	0.41 [0.34, 0.47]	0.24 [0.17, 0.31]	0.33 [0.26, 0.40]	8.9	7.5	5.8
ProofWriter	0.22 [0.15, 0.29]	0.29 [0.22, 0.36]	0.38 [0.31, 0.44]	7.4	5.9	4.0
Mean	0.31 [0.27, 0.35]	0.21 [0.16, 0.25]	0.29 [0.24, 0.33]	9.6	8.5	5.9

Table 4: Reasoning-unit taxonomy and final correctness.

Reasoning-unit class	Correct traces (%)	Incorrect traces (%)	Δ	Sig.
Load-bearing	24.7 [23.1, 26.2]	17.3 [15.8, 18.9]	+7.4 [+5.3, +9.4]	***
Deep-but-fragile	13.9 [12.7, 15.2]	16.8 [15.4, 18.3]	-2.9 [-4.4, -1.3]	**
Controllable-but-non-faithful	18.6 [17.2, 20.0]	21.1 [19.5, 22.7]	-2.5 [-4.4, -0.7]	*
Residual scaffolding	42.8 [40.9, 44.7]	44.8 [42.9, 46.8]	-2.0 [-4.5, 0.3]	n.s.

many serve organizational or stylistic roles. This helps explain why trace length is often a weak proxy for reasoning quality. What matters is not how reasoning text a model produces, but how functional importance is distributed within that text.

This perspective also clarifies the relation among effort, faithfulness, and control. The three are positively related, but they do not collapse into a single dimension. A reasoning unit can be effortful but fragile, faithful but hard to control, or controllable while remaining weakly connected to the answer. Treating these properties separately therefore gives a more precise account of visible reasoning than any single trace-level statistic.

The intervention results make this concrete. Under a matched budget, targeted deletion is substantially stronger than length- or position-based perturbation, and clause spans are more effective intervention targets than isolated tokens. This suggests that important reasoning content is often localized, but not necessarily at the level of single tokens. The taxonomy adds a complementary compositional view: correct traces contain more load-bearing content, whereas incorrect traces contain more fragile and weakly faithful units. Many reasoning failures therefore look less like failures of fluency and more like failures of allocation.

We do not claim that visible reasoning fully reveals hidden computation. Instead, we show that it remains a useful interface for analysis when studied at the right granularity. The same broad distinctions recur across interventions, human judgments, and benchmark-provided intermediate structure, supporting the empirical usefulness of the framework. More broadly, this suggests a different direction for reasoning analysis: instead of asking whether a model reasons more, it may be more useful to ask whether its visible reasoning concentrates on units that are faithful, consequential, and controllable.

8 Conclusion

We presented a framework for analyzing visible reasoning traces at the level of tokens and clause spans. The core claim is that reasoning units are not equally important. We introduced three functional axes—effort, faithfulness, and control—and used them to score reasoning units, compare them with standard trace-level proxies, evaluate them under targeted interventions, and derive an interpretable taxonomy. Across four reasoning benchmarks and three language models, the results are consistent: functional scores explain final correctness better than trace length, confidence, or position; intervening on highly ranked units causes larger behavioral disruption than competing baselines; and successful traces differ from unsuccessful ones in functional composition, especially in the proportion of load-bearing content. These patterns are further supported by human calibration and alignment with dataset-provided intermediate structure across datasets. Overall, visible reasoning quality is better characterized by the distribution of behaviorally meaningful units than by trace length alone.

Limitations

Our analysis focuses on visible reasoning traces rather than the full internal computation of a language model. The proposed framework should therefore be interpreted as a structured analysis of behaviorally meaningful reasoning interfaces, not as a complete account of hidden reasoning processes. A unit that appears unimportant in the visible trace may still correlate with internal computation that is not directly expressed in text.

A second limitation is that our three axes depend on operational choices. Effort is estimated from layer-wise revision dynamics, faithfulness from local interventions on visible reasoning units, and control from externally imposed lexical, structural, and semantic constraints. These choices are principled and empirically useful, but they are not unique. Different intervention families, hidden-state statistics, or control protocols could change the absolute values of the resulting scores.

Our segmentation strategy is also an approximation. Clause spans are more stable and informative than isolated tokens in our experiments, but they are still automatically derived units rather than gold reasoning primitives. In some cases, a single span may mix inferential, organizational, and stylistic content, which limits the precision of any unit-level assignment.

The empirical scope of the paper is also limited. We study four English reasoning benchmarks and three open-weight language models. The reported patterns are therefore best interpreted as evidence for cross-task regularities within this setting, rather than as a universal characterization of visible reasoning across all domains, languages, or model families.

Finally, our human calibration and external-structure evaluations should not be treated as complete ground truth for reasoning importance. Human labels, benchmark-provided intermediate structure, and intervention effects each capture different aspects of reasoning, and none is sufficient on its own. Our conclusions rely on the consistency of these signals rather than on any single source of supervision.

References

Shashank Aggarwal, Ram Vikas Mishra, and Amit Awekar. 2026. Evaluating chain-of-thought reasoning through reusability and verifiability. *Preprint*, arXiv:2602.17544.

Ido Amos, Avi Caciularu, Mor Geva, Amir Globerson, Jonathan Herzig, Lior Shani, and Idan Szpektor. 2026. Latent reasoning with supervised thinking states. *Preprint*, arXiv:2602.08332.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *Preprint*, arXiv:2503.08679.

Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. 2026. Think deep, not just long: Measuring llm reasoning effort via deep-thinking tokens. *Preprint*, arXiv:2602.13517.

Chrisanna Cornish and Anna Rogers. 2025. Examining the faithfulness of deepseek r1's chain-of-thought reasoning. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025)*, pages 11–19, Mumbai, India. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. *Preprint*, arXiv:1911.03429.

Sai Surya Duvvuri, Chanakya Ekbote, Rachit Bansal, Rishabh Tiwari, Devvrit Khatri, David Brandfonbrener, Paul Liang, Inderjit Dhillon, and Manzil Zaheer. 2026. Interleaved head attention. *Preprint*, arXiv:2602.21371.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Preprint*, arXiv:1902.10186.

Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2025. Learning from sufficient rationales: Analysing the relationship between explanation faithfulness and token-level regularisation strategies. *Preprint*, arXiv:2511.16353.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *Preprint*, arXiv:2507.11473.

Michael A. Lepori, Tal Linzen, Ann Yuan, and Katja Filippova. 2026. Language models struggle to use representations learned in-context. *Preprint*, arXiv:2602.04212.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Junchen Liu, Sven Elflein, Or Litany, Zan Gojcic, and Ruilong Li. 2026a. Test-time training with kv binding is secretly linear attention. *Preprint*, arXiv:2602.21204.

Yizhou Liu, Sara Kangaslahti, Ziming Liu, and Jeff Gore. 2026b. Inverse depth scaling from most layers being similar. *Preprint*, arXiv:2602.05970.

Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. 2026. Learning a generative meta-model of llm activations. *Preprint*, arXiv:2602.06964.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Austin Meek, Eitan Sprejer, Iván Arcuschin, Austin J. Brockmeier, and Steven Basart. 2025. Measuring chain-of-thought monitorability through faithfulness and verbosity. *Preprint*, arXiv:2510.27378.

Wenbo Pan, Zhichao Liu, Xianlong Wang, Haining Yu, and Xiaohua Jia. 2026. Towards long-horizon interpretability: Efficient and faithful multi-token attribution for reasoning llms. *Preprint*, arXiv:2602.01914.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Preprint*, arXiv:2402.13950.

Daniel Ritter, Owen Oertell, Bradley Guo, Jonathan Chang, Kianté Brantley, and Wen Sun. 2026. Llms can learn to reason via off-policy rl. *Preprint*, arXiv:2602.19362.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? *Preprint*, arXiv:1906.03731.

Anand Swaroop, Akshat Nallani, Saksham Uboweja, Adiliia Uzdenova, Michael Nguyen, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, Vasu Sharma, and Maheep Chaudhary. 2025. Frit: Using causal importance to improve chain-of-thought faithfulness. *Preprint*, arXiv:2509.13334.

Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. 2026. Maximum likelihood reinforcement learning. *Preprint*, arXiv:2602.02710.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. *Preprint*, arXiv:2502.14829.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yingming Wang and Pepa Atanasova. 2025. Self-critique and refinement for faithful natural language explanations. *Preprint*, arXiv:2505.22823.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Preprint*, arXiv:1908.04626.

Guowei Xu, Mert Yuksekgonul, and James Zou. 2026. Sparse reward subsystem in large language models. *Preprint*, arXiv:2602.00986.

Shu Yang, Junchao Wu, Xilin Gong, Xuansheng Wu, Derek Wong, Ninghao Liu, and Di Wang. 2026. Investigating cot monitorability in large reasoning models. *Preprint*, arXiv:2511.08525.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. 2026. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. *Preprint*, arXiv:2509.00190.

Chen Yueh-Han, Robert McCarthy, Bruce W. Lee, He He, Ian Kivlichan, Bowen Baker, Micah Carroll, and Tomek Korbak. 2026. Reasoning models struggle to control their chains of thought. *Preprint*, arXiv:2603.05706.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025. A survey on test-time scaling in large language models: What, how, where, and how well? *Preprint*, arXiv:2503.24235.

A Experimental Setup and Dataset Construction

This section consolidates the dataset, model, generation, and implementation details that support the main experiments. We first summarize the benchmark suite and model configuration, then document dataset provenance and filtering, and finally report generation validity, per-model performance, and compute cost. Together, these tables make the benchmark construction and experimental setup fully reproducible.

A.1 Benchmark and model summary

Table 1 summarizes the evaluation suite used throughout the paper. The four benchmarks were selected because they expose different forms of intermediate reasoning structure while covering distinct reasoning styles: multi-hop question answering in HotpotQA, decomposition-based binary reasoning in StrategyQA, multistep scientific entailment in EntailmentBank, and proof-based deductive reasoning in ProofWriter. The final filtered suite contains 2,900 evaluation items in total. The table also shows that the benchmarks differ substantially in both input length and generated reasoning length. HotpotQA is the most context-heavy benchmark, while EntailmentBank produces the longest reasoning traces on average. This diversity is important because it reduces the chance that the main findings are driven by a single task format or a narrow trace-length regime.

Table 2 reports the model set and the default generation protocol used in the main paper. We evaluate three open-weight models with different reasoning profiles: Llama-3.1-8B-Instruct, Qwen3-8B with thinking mode enabled, and DeepSeek-R1-Distill-Qwen-32B. Unless otherwise stated, all main experiments use prompt family P1, temperature 0.2, top-p 0.95, and one trace per item. Invalid traces are counted as incorrect in pooled accuracy. The valid-trace rates are high across all models, indicating that the required reasoning/final-answer format is stable rather than excessively fragile. The pooled suite accuracy also shows a clear ranking across models, which is useful for testing whether the proposed analysis captures meaningful structure across systems with substantially different reasoning quality.

A.2 Dataset provenance and filtering

Table 3 documents the public sources, reasoning supervision, and evaluation splits used for each benchmark. We distinguish the original public resource from the concrete evaluation source used in this paper because the latter is not identical across datasets. In particular, the StrategyQA evaluation split is a fixed held-out split derived from public training data and is used only for within-paper comparison and ablation consistency; it is not intended as a substitute for official hidden-test evaluation. We also include ThoughtSource as an auxiliary resource because it was used for formatting and schema sanity checks, even though it is not part of the evaluation suite itself. This table is intended to make the exact provenance of all public resources explicit and reproducible.

Table 4 reports the filtering pipeline applied before generation. Candidate pools are sampled once and then held fixed for all subsequent experiments, so every model and intervention is evaluated on the same benchmark items. Filtering removes items with excessive truncation risk, unsupported intermediate structure, or label normalization failures. These decisions are made at the dataset-item level rather than the model-output level. The table also records dataset-specific sampling seeds and candidate-pool sources. The resulting final kept counts define the exact 2,900-item benchmark used throughout the paper.

A.3 Generation validity, performance, and compute

Table 5 reports post-generation validity statistics on the fixed benchmark suite. Invalid traces are grouped into mutually exclusive primary buckets for accounting purposes, including delimiter or parse failures, empty or underlength reasoning, answer-format mismatches, and other residual invalid outputs. These statistics complement the valid-trace rates reported in Table 2 and show that the required reasoning/final-answer format is stable across all three models. In particular, the invalid rate decreases substantially from Llama-3.1-8B-Instruct to DeepSeek-R1-Distill-Qwen-32B, which is consistent with the general improvement in generation quality across models.

Table 6 gives the per-model, per-dataset performance breakdown behind the pooled results in the main paper. Each dataset column reports the benchmark-specific binary correctness criterion

Table 1: Benchmark suite summary.

Dataset	Evaluation source	Evaluation unit	External supervision	#Eval examples	Avg input tokens	Avg gen. reasoning	Avg gen. answer
HotpotQA	Official dev (distractor context)	QA item	Supporting facts	800	846	72.8	2.7
StrategyQA	Public-train-derived held-out split (seed 42)	Labeled yes/no item	Decomposition + evidence	700	164	57.3	1.0
EntailmentBank	Task 2 instances derived from official test trees	Task instance	Entailment trees	600	301	88.5	3.9
ProofWriter	Official test subset (D2/D3/D5 mix)	Proof item	Proofs / intermediate conclusions	800	153	61.9	1.4
Weighted avg. / total				2900	377.4	69.3	2.2

Table 2: Models and generation protocol.

Model	Params (B)	Reasoning status	Hidden states?	Prompt fam.	Temp.	Top-p	Max reason.	Valid trace %	Acc. (%)
Llama-3.1-8B-I.	8	Instruct, no explicit thinking	Yes	P1	0.2	0.95	256	95.9	63.3
Qwen3-8B	8	Hybrid; thinking enabled	Yes	P1	0.2	0.95	256	97.8	69.4
DS-R1-Distill-32B	32	Reasoning-distilled	Yes	P1	0.2	0.95	384	98.7	78.4

used in evaluation, and the pooled suite accuracy is the example-count-weighted average over the four benchmarks. This table is useful for checking that the overall trends are not driven by a single dataset. The model ordering is stable across the suite: DeepSeek-R1-Distill-Qwen-32B performs best overall, followed by Qwen3-8B and then Llama-3.1-8B-Instruct.

Finally, Table 7 summarizes the implementation stack, major hyperparameters, and compute footprint of the full pipeline. Panel A reports A100-equivalent GPU hours for the main experimental blocks, including generation, hidden-state extraction, intervention experiments, robustness reruns, and human-calibration preprocessing. Panel B lists software versions and non-default implementation choices relevant to reproducibility. Reporting these details is especially important here because the method combines generation, hidden-state analysis, intervention, annotation, and statistical evaluation in a single pipeline.

A.4 Additional scoring implementation details

This subsection clarifies three implementation details that are central to reproducibility: how intermediate-layer next-token distributions are computed for the effort score, how the normalized gold-answer probability is defined for faithfulness, and how control prompts and adherence scores are constructed.

Intermediate-layer next-token distributions. For the effort score, all hidden states are extracted from a teacher-forced forward pass over the generated reasoning trace under the same tokenizer and model checkpoint used for generation. Let $h_\ell^{(t)}$ denote the hidden state at token position t and layer ℓ. To obtain the intermediate-layer next-token distribution $p_\ell^{(t)}$, we apply the same frozen output projection used at the final layer to each intermediate hidden state. Concretely, if the model uses output projection matrix W_{LM} and bias b_{LM}, we compute

$$z_\ell^{(t)} = W_{\mathrm{LM}}\,\phi\Big(h_\ell^{(t)}\Big) + b_{\mathrm{LM}}$$
$$p_\ell^{(t)} = \mathrm{softmax}\Big(z_\ell^{(t)}\Big)$$

where $\phi(\cdot)$ denotes the final pre-head normalization used by the model architecture (e.g., the same final RMSNorm/LayerNorm applied before the LM

Table 3: Dataset provenance, reasoning supervision, and evaluation source.

Dataset	Official source used	Public reasoning supervision	Public size / split info	Evaluation source used
HotpotQA	Official homepage and paper (2026-03-14)	Sentence-level supporting facts	Official train/dev/test; large-scale multi-hop QA	Official dev, distractor context
StrategyQA	Official paper and release (2026-03-14)	Decomposition and evidence	Official reports 2,780 examples	Train-derived held-out (seed 42)
EntailmentBank	Official paper and AI2 (2026-03-14)	Multistep entailment trees	1,840 trees; official train/dev/test	Task 2 from official test trees
ProofWriter	Official paper and release (2026-03-14)	Deductive proofs/conclusions	Official D0/D1/D2/D3/D5 sets	Official test (D2/D3/D5 mix)
ThoughtSource	Public meta-dataset (2026-03-14)	Standardized CoT organization	Public meta-dataset	Formatting/schema sanity checks

Table 4: Filtering log for pre-sampled candidate pools.

Dataset	Candidate-pool source	Seed	Initial items	Removed: length/trunc.	Removed: struct./ann.	Removed: label/norm.	Final kept
HotpotQA	Official dev, distractor context	11	1000	91	0	109	800
StrategyQA	Public train release	42	850	18	61	71	700
EntailmentBank	Official test trees, Task 2	17	720	22	48	50	600
ProofWriter	Official D2/D3/D5 test mix	23	920	27	13	80	800
Total	–	–	3490	158	122	310	2900

head). We do not train separate per-layer decoding heads. The same frozen output head is shared across all layers, so the intermediate distributions should be interpreted as diagnostic readouts rather than as calibrated task-level probabilities. This shared-head construction is sufficient for the relative signals used in the paper, since $R^{(t)}$, $S^{(t)}$, and $H^{(t)}$ are all subsequently normalized within each dataset-model block before aggregation.

The late-convergence quantity uses the top prediction induced by these shared-head distributions. Specifically,

$$\hat{v}_\ell^{(t)} = \arg\max_v p_\ell^{(t)}(v),$$

and $K^{(t)}$ is the earliest layer whose argmax remains unchanged through the final layer. In practice, this makes $S^{(t)}$ a simple measure of how late the local prediction stabilizes under the same output projection.

Normalized gold-answer probability. For the faithfulness score, the quantity $p_i(y_i \mid r_i)$ is implemented as a length-normalized teacher-forced answer score for a canonical gold-answer string. For each example i, we first map the gold answer to a canonicalized surface form $y_i^\star$ using the same normalization rules as the benchmark-specific exact-match evaluation (case folding, whitespace normalization, punctuation stripping where applicable, and yes/no canonicalization for binary tasks). Given input x_i and visible reasoning trace r_i, we then score the canonical answer under forced decoding:

$$s_i(a \mid r_i) = \exp\left(\frac{1}{|a|}\sum_{k=1}^{|a|} \log P(a_k \mid x_i, r_i, a_{<k})\right),$$

where $a = (a_1, \ldots, a_{|a|})$ is a tokenized canonical answer string. Length normalization avoids systematically favoring shorter gold answers.

We then convert this score into a normalized gold-answer probability by comparing the canonical gold answer against the canonicalized model answer for the same item:

$$p_i(y_i \mid r_i) = \frac{s_i(y_i^\star \mid r_i)}{s_i(y_i^\star \mid r_i) + s_i(\hat{y}_i^\star \mid r_i)}.$$

After intervention on unit u, we analogously com-

Table 5: Generation validity log.

Model	Gen. traces	Valid traces	Inv. rate (%)	Primary: delim./parse	Primary: empty/short	Primary: format	Other primary
Llama-3.1-8B-I.	2900	2781	4.1	34	21	29	35
Qwen3-8B	2900	2836	2.2	13	9	18	24
DS-R1-Distill-32B	2900	2862	1.3	7	4	10	17

Table 6: Per-model, per-dataset outcome summary. All metrics are Accuracy (%).

Model	HotpotQA	StrategyQA	EntailmentBank	ProofWriter	Pooled Suite
Llama-3.1-8B-I.	59.1	60.4	64.8	68.9	63.3
Qwen3-8B	66.8	68.2	71.5	71.6	69.4
DS-R1-Distill-32B	75.9	77.8	80.4	79.8	78.4

pute

$$p_i(y_i \mid r_i^{(-u)}) = \frac{s_i(y_i^\star \mid r_i^{(-u)})}{s_i(y_i^\star \mid r_i^{(-u)}) + s_i(\hat{y}_i^{(-u)\star} \mid r_i^{(-u)})}.$$

For binary-answer benchmarks, this reduces to comparing the canonical gold label against the competing canonical model label. For multi-token open-answer benchmarks, the same teacher-forced scoring procedure is used on the normalized answer string. In all cases, this quantity should be read as a normalized relative preference for the gold answer under the current visible trace, not as a calibrated posterior over the full answer space.

Control prompt construction and adherence scoring. The control score is built from three local prompt families: lexical, structural, and semantic. All control prompts keep the original input, task instruction, and final-answer format fixed, and modify only a local requirement tied to one reasoning unit u.

For *lexical control*, we select a salient content word or short content phrase from the target unit after removing punctuation, stopwords, and purely formatting tokens. A lexical prompt either requires inclusion of that item or explicitly forbids it in the regenerated local reasoning. A prompt is considered valid only if the selected content item is unambiguous under tokenization and appears in the target unit in a way that can be checked exactly after normalization. Lexical adherence is scored as a binary indicator: 1 if the required token or phrase is included (or successfully excluded), and 0 otherwise.

For *structural control*, we require the local reasoning to follow a lightweight structural template while preserving the original task context. Typical templates include numbered-step realization, explicit connective insertion, or a constrained local ordering of two extracted content chunks from the same unit. Structural prompts are considered valid only when the unit is long enough to support the template and when the relevant chunks can be extracted deterministically. Structural adherence is scored as 1 if the generated local span satisfies the requested template after normalization, and 0 otherwise.

For *semantic control*, we first derive a normalized proposition-like target from the unit by stripping discourse markers, normalizing inflection, and retaining the main content predicate together with its local arguments or conclusion phrase. The prompt then asks the model to make this proposition explicit in the regenerated reasoning. Semantic prompts are considered valid only when the unit yields a non-empty normalized proposition under this procedure. Semantic adherence is scored by overlap with the normalized proposition target: a prompt receives adherence 1 if the regenerated local span exceeds the fixed semantic-match threshold used in our implementation (ROUGE-L F1 ≥ 0.60 after normalization), and 0 otherwise.

For each unit u, let $\mathcal{P}(u)$ denote the set of valid control prompts generated from these families. We define the average adherence score as

$$a(u) = \frac{1}{|\mathcal{P}(u)|} \sum_{p \in \mathcal{P}(u)} \mathrm{adh}(p),$$

where $\mathrm{adh}(p) \in \{0, 1\}$ is the prompt-specific adherence indicator. We then measure the downstream behavioral consequence only on successful controls:

$$\Delta(u) = \frac{1}{|\mathcal{P}^+(u)|} \sum_{p \in \mathcal{P}^+(u)} \left| p_i(y_i \mid r_i) - p_i(y_i \mid r_i^{(p)}) \right|,$$

Table 7: Hyperparameters, compute budget, and implementation summary.

Panel A: Compute budget

Experiment block	Infrastructure	Precision	Main settings	Seeds	GPU hours
Main trace gen.	1×A100 80G	BF16	Temp. 0.2, top-p 0.95, length 256/384, one trace/item	1	82.6
Hidden-state ext.	1×A100 80G	BF16	Full-layer export for scored reasoning traces	3	97.8
Faithfulness int.	1×A100 80G	BF16	Targeted deletion, substitution, and masking	3	121.4
Control exp.	1×A100 80G	BF16	Lexical, structural, and semantic control prompts	3	36.9
Robustness reruns	1×A100 80G	BF16	Greedy, temp. sweep, prompt-family sweep	3	58.7
Human-calib. pre.	Local CPU	Mix	Segmentation, export, and adjudication package	1	11.8
Total	–	–	–	–	409.2

Panel B: Implementation summary

Component	Version	Role	Non-default note
Python	3.11.x	Runtime environment	UTF-8 enforced throughout
PyTorch	2.5.x	Tensor runtime	Deterministic kernels disabled
Transformers	4.52.x	Model loading / gen.	Custom `[Final Answer]` criteria
vLLM	0.8.x	Fast batched generation	Used for generation only
Accelerate	1.4.x	Device orchestration	No distributed training
NumPy	2.1.x	Numerics	Pinned for reproducible bootstrap
pandas	2.2.x	Tabulation	Schema validation on read
SciPy	1.14.x	Statistics	Bootstrap CI and Spearman
scikit-learn	1.6.x	Agreement / F1	Macro-F1 for agreement
spaCy	3.8.x	Segmentation	English pipeline only
NLTK	3.9.x	Backup segmentation	Fallback only
Tokenizer libs	Model-spec.	Tokenization	Tokenizer hash saved with artifact

where $\mathcal{P}^+(u) = \{p \in \mathcal{P}(u) : \mathrm{adh}(p) = 1\}$ and $r_i^{(p)}$ is the regenerated reasoning trace under control prompt p. If $\mathcal{P}^+(u)$ is empty, we set $\Delta(u) = 0$. The final control score is then

$$C(u) = a(u) \cdot \Delta(u).$$

This construction intentionally separates the ease of steering a unit from the behavioral consequence of successful steering. A unit receives a high control score only when the model both obeys the local constraint reliably and exhibits a nontrivial change in gold-answer preference under that controlled regeneration.

Why these choices are sufficient for the main analysis. These implementation choices are designed to support the relative comparisons used in the paper rather than to estimate fully calibrated internal probabilities. The effort score uses a shared frozen output head only as a common probe across layers; the faithfulness score uses normalized teacher-forced answer scores only as a stable relative measure before and after intervention; and the control score separates adherence from downstream effect by construction. The resulting quantities are therefore best interpreted as reproducible functional diagnostics over visible reasoning units, which is the level of analysis targeted in this work.

B Human Calibration and External-Structure Validation

This section provides the full evaluation details behind the human-calibration and external-structure results summarized in the main paper. Our goal is to test whether the proposed unit-level framework aligns with two independent sources of evidence: human judgments over minimal reasoning spans and benchmark-provided intermediate reasoning structure. We first report the compact calibration results on the human-annotated subset, then present the corresponding external-structure alignment results on the full evaluation set, and finally document the annotation protocol, raw la-

bel inventory, deterministic taxonomy mapping, and dataset-level agreement breakdown. Together, these materials clarify what the framework is being compared against, how the comparison is carried out, and why the resulting agreement numbers are meaningful.

B.1 Summary calibration results

Table 8 reports the main human-calibration results used in the paper. The table is based on a fully human-annotated subset of 240 traces, with 60 traces sampled from each benchmark. Each trace is segmented into minimal semantically coherent spans for annotation, yielding 2,462 annotated spans in total. For each dataset, we report the number of annotated traces and spans, inter-annotator agreement measured by Krippendorff's α, our macro-F1 against adjudicated human labels, and alignment with dataset-provided intermediate structure in the same projected span-level comparison space. The final column gives a position baseline under the same evaluation protocol.

Several patterns are worth noting. First, agreement among annotators is consistently substantial, ranging from 0.67 on StrategyQA to 0.79 on ProofWriter, with an overall value of 0.74. This indicates that the distinction among functionally different reasoning spans is not arbitrary or idiosyncratic. Second, our macro-F1 is stable across all four datasets and clearly above the baseline, suggesting that the proposed functional scoring framework recovers human-interpretable distinctions at a useful level of granularity. Third, the external-structure F1 values are strongest on the more explicitly structured benchmarks, EntailmentBank and ProofWriter, which is consistent with the fact that their intermediate supervision is closer to formal reasoning steps than the weaker structural supervision provided by supporting facts or decomposition hints.

Table 9 extends the external-structure analysis from the calibration subset to the full 2,900-example evaluation suite. Here the objective is not to reproduce human labels, but to test whether spans identified as important by our framework are more likely to align with benchmark-provided intermediate structure than simple baselines. All datasets are projected into a common span-level comparison space. Generated traces are segmented into minimal semantically coherent spans, and external supervision is mapped into matched span or step targets before overlap scoring. Alignment is computed with greedy one-to-one matching using ROUGE-L F1, with a match threshold of 0.30.

The full-evaluation results mirror the pattern seen on the annotated subset. Our method consistently outperforms both the position baseline and the random baseline on every benchmark. The strongest gains again appear on EntailmentBank and ProofWriter, where the external structure is most explicit. The fact that the same pattern appears both on the small annotated subset and on the full evaluation suite strengthens the interpretation that the framework is capturing nontrivial structure rather than overfitting to a limited human-labeled sample.

B.2 Annotation protocol and taxonomy mapping

Table 10 documents the annotation protocol and the relationship between raw human labels and the derived four-way taxonomy used in the main paper. Panel A describes the annotation setup, including the annotation unit, span length distribution, annotator count, training round, adjudication procedure, labeling time, compensation, consent, and annotator background. These details are important because they show that the human-calibration subset is not a casual or weakly supervised check, but a deliberately constructed evaluation resource with full triple annotation and adjudication.

Panel B reports the raw human label inventory. These labels are intentionally not identical to the automatic taxonomy. Human annotators assign spans to categories such as load-bearing, bridge or organizational, decorative or stylistic, answer-revealing but non-explanatory, and malformed or unclear. This separation matters because it makes the evaluation stricter: the automatic taxonomy is not simply reproducing a label set chosen to match its own outputs.

Panel C then gives the deterministic mapping algorithm that converts normalized effort, faithfulness, and control scores into the derived functional classes used throughout the paper. The mapping is defined by thresholds and a fixed priority order, producing a partition over included reasoning units. Raw human labels are used only for evaluation and interpretation, not for assigning classes on the full evaluation set. This distinction is important for understanding the calibration results: the human study validates the taxonomy after the fact rather than training it directly.

Table 8: Human calibration agreement and alignment with external reasoning structure.

Dataset	Annot. traces	Annot. spans	Krip. α	Our macro-F1	Our ext. struct. F1	Pos. base. F1
HotpotQA	60	611	0.71	0.68	0.59	0.41
StrategyQA	60	538	0.67	0.63	0.54	0.37
EntailmentBank	60	724	0.76	0.73	0.66	0.44
ProofWriter	60	589	0.79	0.77	0.71	0.48
All	240	2462	0.74	0.71	0.63	0.43

Table 9: Agreement with external reasoning structure on the full evaluation set.

Dataset	Our alignment F1	Position base. F1	Random base. F1
HotpotQA	0.58	0.41	0.24
StrategyQA	0.52	0.37	0.26
EntailmentBank	0.66	0.44	0.23
ProofWriter	0.71	0.48	0.28
Mean	0.62	0.43	0.25

B.3 Dataset-level agreement breakdown

Table 11 provides the dataset-level breakdown behind the aggregate calibration result in Table 8. While the main paper reports the compact overall calibration summary, this table makes clear that the agreement pattern is not driven by a single benchmark. All datasets show moderate-to-strong annotator agreement and nontrivial alignment between the automatic labels and adjudicated human judgments. The same pattern holds for external-structure F1.

This dataset-level view is helpful for interpretation. ProofWriter and EntailmentBank achieve the strongest agreement and alignment, which is expected given their more explicit intermediate supervision and more structured reasoning traces. HotpotQA and StrategyQA are weaker but still clearly above trivial baselines, reflecting the fact that their supervision is looser and less step-like. The consistency of this trend across datasets supports the broader claim that the proposed framework captures genuine differences in local reasoning function even when the supervision signal is indirect.

C Additional Intervention and Robustness Analyses

This section reports additional analyses that extend the main intervention results in three directions. First, we examine how the choice of intervention unit affects both behavioral impact and agreement with human judgments. Second, we compare our deletion-based intervention against alternative perturbation families in their native, non-budget-matched forms. Third, we test whether the proposed framework is robust to decoding and prompt variation. Together, these tables clarify which parts of the main results depend on the intervention unit, which are specific to the matched-budget comparison used in the paper, and how stable the resulting scores and trace composition remain under reasonable generation changes.

C.1 Intervention granularity and perturbation families

Table 12 studies the effect of intervention granularity under the same budget-matched protocol used in the main paper. The table compares token, bigram, trigram, clause-span, sentence-span, and step-span interventions, all under a fixed deletion budget of 15% of reasoning-span tokens per trace, rounded to the nearest integer with a minimum of 4 and a maximum of 24. In addition to behavioral effects, the table also reports human-agreement macro-F1 on the same 240-trace calibration subset used for the human-evaluation results in the main paper, as well as across-seed variation in accuracy drop.

Several patterns are clear. Clause spans produce the strongest overall intervention effect, with the largest accuracy drop (-18.2 pp) and answer-flip rate (27.4%), while also achieving the highest human-agreement macro-F1 (0.77). Sentence spans remain strong but are slightly weaker than clause spans, suggesting that larger units preserve semantics but are less precise. Step spans are the weakest among span-based methods, both behaviorally and in human agreement, which is consistent with the claim that full reasoning steps are often too coarse to isolate where functional importance is concentrated. At the other extreme, token-level intervention is more precise but slightly less effective and less aligned with human judgments than clause spans. Overall, this table supports the use of clause spans as the default intervention unit

Table 10: Human annotation protocol, label inventory, and taxonomy mapping.

Panel A: Annotation protocol

Field	Value
Annotation unit	Minimal semantically coherent span
Span length range	1–12 tokens (Median: 4)
Scale	240 traces independently labeled by 3 annotators
Training & Adjudication	20 pilot traces; Majority vote + author arbitration
Time metrics	2.9 min/trace; Total: 36.8 person-hours
Annotators	1 author + 2 NLP-trained graduates (USD 22/hr)
Ethics & Consent	Exempt; Written opt-in consent obtained

Panel B: Raw human label inventory

Label	Description	Count	Share (%)
Load-bearing	Supports successful inference or answer selection	617	25.1
Bridge / Org.	Connective or structural span organizing inference	514	20.9
Decorative	Stylistic scaffolding without substantive role	933	37.9
Ans.-revealing	Reveals answer without functioning as explanation	231	9.4
Malformed	Ill-formed, ambiguous, or uninterpretable	167	6.8
Total	–	2462	100.0

Panel C: Deterministic partition algorithm for derived functional classes

Step	Rule	Output / Action	Incl. in partition
1	Segment traces into scored units; normalize effort, faithfulness, and control within dataset/model	Define thresholds	Yes
2	If faithfulness $\in$ top quartile AND control $\notin$ bottom half	Assign *Load-bearing*	Yes
3	Else, if effort $\in$ top quartile AND control $\in$ bottom half	Assign *Deep-but-fragile*	Yes
4	Else, if control $\in$ top quartile AND faithfulness $\in$ bottom half	Assign *Contr.-non-faithful*	Yes
5	All remaining scored units	Assign *Residual scaffolding*	Yes
6	Malformed/unclear human spans are excluded before analysis	Excluded from totals	No
7	Tie-break: Load-bearing $\rightarrow$ Deep $\rightarrow$ Controllable $\rightarrow$ Residual	Unique assignment	Yes
8	Renormalize proportions over all included units	Final four-class partition	Yes

in the main paper: they balance semantic coherence, causal strength, and stability better than either smaller token-level units or larger step-level units.

Table 13 broadens the intervention analysis by comparing alternative perturbation families in their native forms, without budget matching. The purpose of this table is diagnostic rather than strictly causal: unlike the main-paper comparison, these intervention types do not remove or alter the same amount of content, so their magnitudes should not be interpreted as directly comparable equal-budget effects. Instead, the table is useful for understanding what kinds of perturbations are especially disruptive and how our targeted interventions behave relative to other natural perturbation families.

The main pattern is that semantically destructive interventions tend to cause the largest drops. Nonsense-token substitution, antonym or contradiction substitution, and both targeted deletion variants all lead to large losses in accuracy and high answer-flip rates. By contrast, synonym substitution is less damaging, which is expected if some local semantic content is preserved. Local reordering and soft masking fall in between. Importantly, the two targeted methods remain among the strongest interventions overall even in this non-budget-matched setting, reinforcing the interpretation that they are selecting especially behaviorally important content rather than merely exploiting a favorable accounting rule in the matched-budget protocol. At the same time, the table also shows

Table 11: Annotation agreement breakdown by dataset.

Dataset	Krippendorff's α	Our macro-F1 vs. human	Our external-structure F1
HotpotQA	0.71	0.68	0.59
StrategyQA	0.67	0.63	0.54
EntailmentBank	0.76	0.73	0.66
ProofWriter	0.79	0.77	0.71
All	0.74	0.71	0.63

Table 12: Token-versus-span ablation.

Unit of intervention	Human-agr. macro-F1	ΔAcc (pp)	Answer flip (%)	Inv. rate (%)	Std. (pp)
Token	0.71	-16.8	25.1	1.3	1.6
Bigram	0.73	-17.3	25.9	1.1	1.4
Trigram	0.74	-17.7	26.5	1.0	1.3
Clause span	0.77	-18.2	27.4	0.9	1.2
Sentence span	0.75	-17.6	26.8	0.7	1.1
Step span	0.68	-14.6	21.8	0.8	1.0

that stronger numerical damage in this setting can partly reflect more destructive perturbation semantics, which is why the main paper uses the budget-matched comparison as the primary causal evidence.

C.2 Decoding and prompt sensitivity

Table 14 evaluates robustness to decoding stochasticity for Qwen3-8B with thinking mode enabled under the same filtered evaluation subset and scoring pipeline as the main analysis. The table reports valid-trace rate, pooled accuracy, average reasoning length, the three functional scores, and decorative ratio under greedy decoding and three temperature settings. Decorative ratio here denotes the trace-level average proportion of units assigned to the Residual scaffolding class.

The trend is consistent across metrics. As decoding becomes more stochastic, reasoning traces become longer, but this increase in length is not accompanied by stronger functional quality. Instead, pooled accuracy decreases, effort declines, faithfulness declines, control declines, and decorative ratio increases. Greedy decoding yields slightly shorter traces than the main setting, but its accuracy remains close and its decorative ratio is lower. The main decoding configuration (temperature 0.2, top-p 0.95) appears to strike a good balance between stable formatting, competitive accuracy, and high functional scores. More importantly, the table reinforces one of the central claims of the paper: longer visible traces do not necessarily imply higher-quality reasoning. As temperature increases, the model produces more text but a larger fraction of that text is classified as residual scaffolding rather than behaviorally meaningful content.

Table 15 reports sensitivity to prompt family under the main decoding configuration, again for Qwen3-8B with thinking mode enabled. The four prompt families differ in how much they encourage explicit or structured reasoning: P1 uses free-form chain-of-thought, P2 asks for concise reasoning, P3 encourages numbered steps, and P4 uses minimal rationale. The table reports the same summary statistics as the decoding table.

The main finding is that prompt format changes the surface form of traces more than it changes the overall qualitative pattern. P1 and P3 achieve the best pooled accuracy, while P4 produces the shortest traces and the lowest decorative ratio. P2 and P4 improve formatting stability slightly, but this comes with lower faithfulness and lower overall accuracy. P3 increases control relative to P1, which is plausible given the more explicit structural organization imposed by numbered steps, but it does not surpass P1 on the combined overall pattern of accuracy and faithfulness. The broader point is that the framework remains meaningful under prompt variation: different prompt families shift length, style, and control, but the core distinction between load-bearing and non-load-bearing reasoning content remains stable enough to support the main conclusions.

Overall, the analyses in this section support three conclusions. First, the choice of intervention unit matters, and clause spans provide the best trade-

Table 13: Native intervention ablation.

Intervention type	Meaning preserv.	Grammar preserv.	ΔAcc (pp)	Answer flip (%)	Inv. rate (%)
Random token deletion	Low	Low	-7.9	12.6	1.9
Synonym substitution	Med–High	High	-6.3	9.7	0.4
Antonym / contradiction subst.	Low	High	-17.2	25.6	0.6
Nonsense-token substitution	None	Medium	-19.8	29.4	1.2
Local reordering	Medium	Medium	-14.1	20.2	0.9
Soft masking	Medium	High	-16.4	23.6	0.3
Targeted token deletion (ours)	Low	Medium	-18.9	27.8	1.4
Targeted clause-span deletion (ours)	Low	Medium	-23.9	34.0	0.9

Table 14: Sensitivity to decoding configuration.

Decoding config	Valid trace %	Pooled acc. (%)	Avg reas. tokens	Effort score	Faith. score	Control score	Decor. ratio (%)
Greedy	98.3	68.9	61.8	0.27	0.36	0.16	42.1
T=0.2, p=.95 (main)	97.8	69.4	69.3	0.29	0.39	0.13	43.6
T=0.5, p=.95	96.9	68.1	78.6	0.26	0.34	0.11	46.4
T=0.7, p=.95	95.7	66.3	91.1	0.22	0.30	0.08	49.6

off between semantic coherence, causal strength, and human alignment. Second, our targeted methods remain highly disruptive even when compared with alternative perturbation families outside the matched-budget setting. Third, the proposed scores and taxonomy are robust to moderate variation in decoding and prompting, even though these changes affect trace length and stylistic surface form. These findings help justify the main modeling choices in the paper and clarify the scope under which the reported effects are stable.

C.3 Weight and threshold sensitivity

To test whether the combined score depends strongly on a single hand-chosen weight vector, we recompute the trace-level combined score under several alternative weight settings. For each setting, we report the trace-level Spearman correlation with final correctness on each benchmark and the macro-average over the four datasets. This analysis isolates the effect of the aggregation weights while keeping all underlying unit-level scores fixed.

The results in Table 16 show that the combined score is robust to moderate changes in the aggregation weights. The default, equal-weight, faithfulness-heavy, and no-control settings remain close, whereas settings that suppress faithfulness degrade much more clearly. In particular, removing faithfulness produces the largest drop across all four benchmarks. This pattern suggests that faithfulness is the most important axis, but that the overall benefit of aggregation is not tied to a single fragile choice of weights.

We also test the sensitivity of the four-way taxonomy to the threshold choices used to define high and low values of the normalized functional scores. For each threshold setting, we recompute the automatic class assignment, compare class proportions between correct and incorrect traces, and evaluate macro-F1 on the human-annotated subset.

Table 17 shows that the main compositional conclusions are stable across a reasonable range of threshold choices. Correct traces consistently contain more load-bearing units, whereas incorrect traces consistently contain more deep-but-fragile and controllable-but-non-faithful content. Human macro-F1 also changes only slightly across settings. This suggests that the taxonomy results are not an artifact of one specific quantile choice.

D Failure Analysis and Reporting Checklist

This section complements the main results with two forms of supporting analysis. We first summarize the main failure modes observed across generation, control, and reasoning evaluation. These results help clarify where the framework and the underlying model outputs break down in practice, and which settings are most likely to produce unstable or misleading traces. We then provide a checklist-oriented disclosure summary that maps the empirical, annotation, and reproducibility contents of the paper to a reviewer-facing reporting structure. The goal of this section is therefore twofold: to

Table 15: Sensitivity to prompt family.

Prompt family	Valid trace %	Pooled acc. (%)	Avg reas. tokens	Effort score	Faith. score	Control score	Decor. ratio (%)
P1 Free-form CoT	97.8	69.4	69.3	0.29	0.39	0.13	43.6
P2 Concise reas.	98.4	68.6	47.6	0.25	0.35	0.16	36.9
P3 Numbered steps	98.1	69.1	63.4	0.27	0.37	0.19	40.8
P4 Minimal rat.	98.7	67.4	39.2	0.22	0.31	0.18	31.7

Table 16: Combined-score weight ablation. Each row reports the trace-level Spearman correlation between the combined score and final correctness under a different weight setting for effort, faithfulness, and control. The default setting used in the main paper is shown in the first row.

Weights (w_E, w_F, w_C)	HotpotQA	StrategyQA	EntailmentBank	ProofWriter	Mean
Default $(0.30, 0.50, 0.20)$	0.44	0.36	0.49	0.51	0.45
Equal $(0.33, 0.33, 0.33)$	0.42	0.34	0.46	0.48	0.43
Faithfulness-heavy $(0.20, 0.60, 0.20)$	0.43	0.35	0.48	0.50	0.44
Effort-heavy $(0.60, 0.20, 0.20)$	0.38	0.31	0.41	0.38	0.37
Control-heavy $(0.20, 0.20, 0.60)$	0.30	0.22	0.34	0.30	0.29
No control $(0.40, 0.60, 0.00)$	0.42	0.34	0.47	0.49	0.43
No effort $(0.00, 0.70, 0.30)$	0.40	0.31	0.44	0.48	0.41
No faithfulness $(0.50, 0.00, 0.50)$	0.27	0.18	0.30	0.25	0.25

document failure patterns that are relevant for interpreting the main findings, and to make the reporting status of the work explicit and easy to audit.

D.1 Failure modes

Table 18 summarizes the main failure modes observed across the generation and control pipeline. The table includes both formatting failures and reasoning-specific failures, and the categories are intentionally non-exclusive: a single trace may exhibit more than one problem, so the rates do not sum to the overall invalid-trace rate. Two different denominators are used. The first is the full set of 8,700 traces generated by three models over the 2,900-item benchmark suite. The second is a stratified control subset of 1,740 valid traces used for control-specific analysis.

The most basic failure modes are formatting-related. Premature answer leakage, empty or under-length reasoning, and delimiter or parse failures are all relatively rare, each occurring in only a small fraction of traces. These statistics help explain why the main-paper results are not dominated by formatting instability. Even so, they are not negligible: free-form prompting still produces a modest number of malformed outputs, especially for weaker models or longer contexts.

A second group of failures reflects low-quality reasoning traces even when formatting succeeds. Degenerate repetition or filler loops appear in 3.2% of all traces, and self-contradictory reasoning in 2.8%. The largest reasoning-specific category is unsupported reasoning leap, which appears in 6.1% of all traces after a two-stage procedure that combines heuristic flagging with full manual confirmation of all flagged cases. This is consistent with the broader interpretation of the paper: visible reasoning often contains spans that look structurally plausible but are weakly grounded in either preceding premises or benchmark-provided intermediate structure.

The control-specific rows are also informative. Control instruction ignored occurs in 46.8% of the control subset, and forbidden-token leakage in 31.5%. These rates are much higher than the formatting failure rates, indicating that controllability is substantially harder than simple format compliance. This supports one of the main conclusions of the paper: local controllability is a meaningful axis of variation that should not be assumed to coincide with either trace well-formedness or reasoning usefulness. In other words, even when a model produces a valid and fluent chain of thought, it may still resist local control or violate explicit constraints.

From an interpretive standpoint, these failures help bound the claims of the paper. First, they show that the proposed framework is not simply rediscovering formatting artifacts, since the dominant failure modes in the main experiments are not malformed outputs but weaknesses in local grounding, controllability, and reasoning consistency. Sec-

Table 17: Taxonomy-threshold ablation. Each row reports the effect of changing the thresholds that define high and low normalized axis values for the derived four-way taxonomy. We report the correct-minus-incorrect difference in class proportion for the three most diagnostic classes, together with macro-F1 on the human-annotated calibration subset.

Threshold setting	Load-bearing Δ	Deep-but-fragile Δ	Controllable-but-non-faithful Δ	Human macro-F1
Strict: high = top 20%, low = bottom 40%	+7.8	-2.4	-2.1	0.70
Default: high = top 25%, low = bottom 50%	+7.4	-2.9	-2.5	0.71
Moderate: high = top 33%, low = bottom 50%	+6.9	-3.0	-2.6	0.70
Loose-low: high = top 25%, low = bottom 60%	+7.1	-3.4	-2.9	0.69

Table 18: Failure mode inventory.

Failure mode	Denom.	Count	Rate (%)	Detected by	Most affected setting
Premature answer leakage	8700	181	2.1	Rule-based parser	Llama on PW / Hotpot
Empty/short reasoning	8700	92	1.1	Rule-based parser	Llama on StrategyQA
Delimiter / parse failure	8700	109	1.3	Rule-based parser	Free-form variants
Degenerate repetition	8700	278	3.2	Heuristic detector	High-temp decoding
Self-contradictory	8700	241	2.8	Manual coding	Hotpot / StrategyQA
Unsupported leap	8700	531	6.1	Heuristic + manual	StrategyQA
Control ignored	1740	815	46.8	Constraint check	Control prompts
Forbidden leakage	1740	548	31.5	Constraint check	High-difficulty settings

ond, they indicate that some apparent limitations of visible reasoning analysis may stem not from the framework itself but from the fact that model-generated traces often mix valid local reasoning with unsupported or weakly constrained content. This is one reason the paper emphasizes functional composition rather than treating an entire trace as uniformly informative.

D.2 Checklist-oriented disclosure

Table 19 provides a checklist-oriented disclosure summary for the full paper. Its purpose is not to introduce new empirical claims, but to help reviewers and readers locate the parts of the manuscript and supplementary material that address common reproducibility and reporting concerns. The table maps each checklist-style item to a status flag and a supporting section or artifact.

The first group of rows covers data, experimental setup, and compute. These rows point readers to the benchmark and filtering tables, the generation protocol, and the compute summary. The second group covers annotation and human evaluation, including instructions, compensation, consent, and annotator background. The final group addresses artifact availability and AI-assistance disclosure. In this paper, the checklist table is especially useful because the work spans several methodological components at once: benchmark construction, generation, hidden-state analysis, intervention, human annotation, and external-structure validation. A compact disclosure table helps make clear that each of these pieces has been reported in an auditable form.

One important feature of this table is that it distinguishes between material already included in the paper and material that is planned for release with the artifact package. This is relevant for the reproducibility status of the work. The core experimental settings, annotation details, and empirical statistics are already documented in the paper and appendix, while code, prompts, and analysis scripts are marked as planned release materials. The table therefore functions as both a reviewer aid and a concise reproducibility map for the full submission.

The two tables in this section serve a practical function for interpreting the rest of the paper. The failure-mode inventory shows where the generation and control pipeline is most brittle, while the disclosure summary shows where the relevant supporting information has been documented. In combination, they help situate the main empirical claims within a fuller picture of model behavior, annotation procedure, and reporting completeness.

Table 19: Checklist-oriented disclosure summary.

Checklist item	Status	Supporting section / artifact
Data statistics and split information	Yes	Appendix Tables 3 and 4
Model parameters, infrastructure, and compute	Yes	Appendix Table 7
Experimental settings and hyperparameters	Yes	Appendix Tables 2 and 7
Generation validity accounting	Yes	Appendix Table 5
Descriptive statistics and uncertainty intervals	Yes	Main Tables 1–4
Existing packages and software settings	Yes	Appendix Table 7
Annotator instructions, protocol, and mapping	Yes	Appendix Table 10
Recruitment and compensation	Yes	Appendix Table 10
Consent statement	Yes	Appendix Table 10
Ethics statement	Yes	Limitations / ethics section
Annotator demographics and background	Yes	Appendix Table 10
Artifact availability	Planned	Code, prompts, and scripts in supplementary repository
AI assistance disclosure	Yes	Acknowledgments and reproducibility statement