Which Reasoning Units Matter? A Label-Free Diagnostic Framework for Visible Reasoning Traces

Anonymous ACL submission

Abstract

Chain-of-thought reasoning is usually evaluated at the level of complete traces, which obscures how different parts of a trace support a model's answer. We ask which reasoning units matter in visible reasoning traces. We present a label-free diagnostic framework that segments generated reasoning into tokens and clause spans, then scores each unit along three axes: effort, answer support, and control. Effort measures internal revision dynamics, answer support measures how much a unit affects the model's original answer under local intervention, and control measures whether a unit can be constrained with downstream effect. Across four reasoning benchmarks and three language models, we find that traces are heterogeneous: a limited subset of units carries most functional support, while many others organize or decorate the reasoning. The proposed scores explain final correctness better than length, confidence, and position baselines, with correctness used only for validation. Interventions under a matched deletion budget further show that removing high-ranked units causes larger accuracy drops than competing selection strategies. These results suggest that visible 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 before producing a final answer (Wei et al., 2023; Wang et al., 2023; Yao et al., 2023). These traces are often treated as evidence of deeper, more careful, or more interpretable reasoning. In practice, however, most evaluations still treat a trace as a whole. A chain is usually judged by its outcome, length, or overall plausibility, which implicitly assumes that its parts contribute in similar ways to the model's answer. This leaves a basic question unanswered: within a visible reasoning trace, which parts actually matter?

The 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 strongly support the model's answer, while another may spend much of its text on connective phrases, stylistic fillers, or local restatements that make the output look more deliberate without changing the answer. This helps explain why trace length is often a weak proxy for reasoning quality. Related work has studied reasoning effort, explanation faithfulness, and controllability, but these questions are often examined separately and at different levels of analysis (Zhang et al., 2025; Chen et al., 2026; Arcuschin et al., 2025; Tutek et al., 2025; Yueh-Han et al., 2026). A shared unit-level view is still missing.

We argue that visible reasoning traces should be analyzed as heterogeneous structures rather than monolithic text sequences. Our goal is not to replace final-answer accuracy when gold labels are available. Instead, we aim to diagnose how a model's own answer is supported within the trace. To make this possible without using ground-truth labels, we introduce a label-free framework that segments generated reasoning into tokens and clause spans and scores each unit along three axes: effort, answer support, and control. Effort measures internal revision dynamics across layers, answer support measures how much a unit affects the model's original answer under local intervention, and control measures whether a unit can be externally constrained with downstream effect. Gold answers are used only to validate whether these diagnostics explain success or failure.

This framing changes what a reasoning evaluation metric is expected to do. If labels are known, final-answer correctness already tells us whether the model succeeded. A unit-level diagnostic in-

stead asks where the visible trace concentrates the support for that answer, which parts are fragile, and which parts can be steered. This makes the analysis useful for auditing long traces, prioritizing spans for human review, studying failures that look fluent, and comparing whether different models allocate their reasoning text in similar ways.

We study this question across four reasoning benchmarks and three language models: HotpotQA, StrategyQA, EntailmentBank, and ProofWriter, using Llama-3.1-8B-Instruct, Qwen3-8B with thinking mode enabled, and DeepSeek-R1-Distill-Qwen-32B. The benchmarks provide different forms of intermediate structure, including supporting facts, decompositions with evidence, entailment trees, and formal proofs. Our pipeline segments each generated trace into reasoning units, computes the three label-free scores, aggregates them to trace-level summaries, and derives a taxonomy of reasoning units. We evaluate the resulting diagnostics with final correctness, matched-budget interventions, human calibration, and alignment to external reasoning structure.

Our results show that visible reasoning traces are strongly heterogeneous. The proposed scores explain final correctness better than length, confidence, and position baselines, even though they do not use gold answers in their definition. Interventions under a matched deletion budget show that removing high-ranked units causes larger drops in accuracy than competing selection strategies, with clause spans emerging as stable intervention targets. The three axes are related but not redundant, and successful traces differ from unsuccessful ones in functional composition: correct traces contain more load-bearing units, while incorrect traces contain more fragile or weakly supporting content. These findings support the central claim of the paper: visible reasoning quality is better characterized by the distribution of functionally meaningful units than by trace length alone.

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 (Wei et al., 2023; Wang et al., 2023; Yao et al., 2023; Zhang et al., 2025).

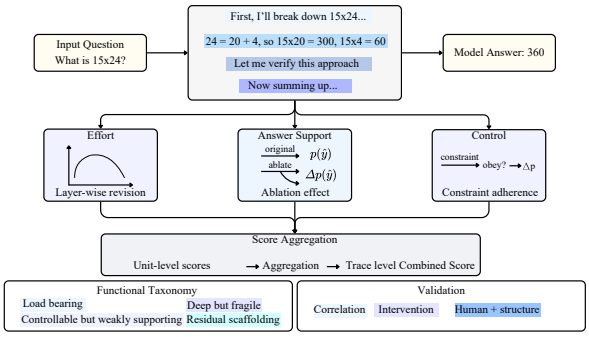

Figure 1: Overview of the proposed framework. Given an input question, a language model produces a visible reasoning trace and a model answer. We segment the trace into token and clause-span units, then score each unit along three label-free axes: effort, answer support, and control. Unit-level scores are aggregated into trace-level summaries and used to derive an interpretable taxonomy of reasoning units. The framework is evaluated through final-answer correctness, matched-budget interventions, human calibration, and alignment with external reasoning structure.

Recent work further argues that longer traces do not necessarily imply better reasoning, and that inference-time computation needs more refined measures than length alone (Chen et al., 2026; Tajwar et al., 2026; Ritter et al., 2026). Our work is complementary to these methods. We do not propose a new reasoning strategy; instead, we study how already generated visible traces can be diagnosed at the level of tokens and clause spans.

A related line of work studies explanation faithfulness, rationales, and feature attribution. Earlier NLP work 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). Similar concerns have been raised for chain-of-thought reasoning, where plausible traces can be weakly connected to the final prediction (Paul et al., 2024; Arcuschin et al., 2025; Cornish and Rogers, 2025; Tutek et al., 2025). Attribution methods such as LIME, Integrated Gradients, DeepLIFT, gradient saliency, and deletion or insertion tests also assign importance to inputs or rationales (Ribeiro et al., 2016; Sundararajan et al., 2017; Shrikumar et al., 2017; DeYoung et al., 2020). Our work draws on this literature but differs in object and objective: we analyze generated reasoning units, and our main behavioral axis is label-free answer support, measured by how interventions change support for the model's own answer rather than for a gold label.

Recent work has also moved beyond treating reasoning as an undifferentiated text sequence. Several studies emphasize that meaningful analysis often requires 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). Work on monitorability and controllability further shows that visible reasoning may help oversight, but is not always easy to monitor, constrain, or causally interpret (Korbak et al., 2025; Yang et al., 2026; Meek et al., 2025; Yueh-Han et al., 2026). Our control axis is motivated by this distinction. It asks whether a local reasoning unit can be constrained and whether successful control has downstream effect, rather than trying to enforce constraints during generation.

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). We do not aim to recover hidden reasoning in full. Instead, we treat visible reasoning as a behavioral interface: partial, imperfect, but still useful when analyzed at the right granularity. The gap we address is the lack of a label-free unit-level framework that separates internal effort, answer support, and control inside visible reasoning traces.

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 model 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 the trace support the model's answer, 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 the generated trace. 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. The token view gives maximal resolution, while the span view gives more coherent local structure.

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

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

corresponding to effort, answer support, and control. Effort measures whether a unit reflects nontrivial computation. Answer support measures whether the unit affects support for the model's own answer $\hat{y}_i$ under local intervention. Control measures whether the unit can be constrained in a stable way with downstream effect. These scores are label-free: they are computed from the input, generated trace, model answer, and model probabilities, not from the gold answer y_i. Gold labels are used only for validation.

For validation, final performance is represented by a binary correctness variable

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

This setup supports two questions. First, do label-free trace summaries explain correctness better than trace-level proxies such as length, confidence, or position? Second, do top-ranked reasoning units have greater behavioral importance under intervention?

To study the second question, we define an intervention operator $\mathcal{I}$ that modifies a selected subset of reasoning units:

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

where $S_i \subseteq U(r_i)$. The model then regenerates an answer $\hat{y}_i'$ from the same input and the modified trace. If a scoring rule identifies important units, intervening on its top-ranked units should produce larger changes in answer support and, when labels are available, larger drops in correctness than baselines with the same deletion budget.

Finally, we use the three axes to derive a coarse taxonomy over reasoning units. The taxonomy is not meant to make reasoning discrete. It provides a compact summary of how visible traces distribute functional importance. The central problem is therefore to define label-free unit-level measures

that identify which parts of a visible reasoning trace matter for the model's answer, and to test 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 label-free scores: effort, answer support, and control. These scores use the input, generated trace, model answer, and model probabilities, but not the gold answer. The pipeline segments the trace into units, computes unit scores, aggregates them to the trace level, and derives 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 format. Invalid parses or malformed final answers are excluded. For each valid trace, we construct two aligned views: a token 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.

The effort score measures whether a unit reflects substantive internal computation. For token position t, let $h_\ell^{(t)}$ be the hidden state at layer ℓ. We obtain $p_\ell^{(t)}$, the next token distribution induced by layer ℓ, by applying the frozen final LM head to $h_\ell^{(t)}$, with no trained per-layer decoding head. Let $\mathcal{V}$ be the vocabulary and

$$\hat{v}_\ell^{(t)} = \arg\max_{v \in \mathcal{V}} p_\ell^{(t)}(v).$$

We compute adjacent layer revision, late convergence, and hidden state update magnitude:

$$\begin{aligned}
R^{(t)} &= \frac{1}{L-1} \sum_{\ell=1}^{L-1} \mathrm{JSD}\left(p_\ell^{(t)}, p_{\ell+1}^{(t)}\right), \\
K^{(t)} &= \min\left\{\ell : \hat{v}_\ell^{(t)} = \hat{v}_{\ell+1}^{(t)} = \cdots = \hat{v}_L^{(t)}\right\}, \\
S^{(t)} &= \frac{K^{(t)} - 1}{L-1}, \\
H^{(t)} &= \frac{1}{L-1} \sum_{\ell=1}^{L-1} \left(1 - \cos\left(h_\ell^{(t)}, h_{\ell+1}^{(t)}\right)\right).
\end{aligned}$$

The token effort score is

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

where tildes denote normalization within each dataset and model block. 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 answer support score measures whether a unit supports the model's original answer. For each unit u, we delete u, regenerate the final answer span, and compare support for the original model answer before and after deletion. Let $\hat{y}_i$ be the original answer and $\hat{y}_i^{(-u)}$ the answer after deleting u. We score any candidate answer $a = (a_1, \ldots, a_{|a|})$ under a trace r by length-normalized forced decoding:

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

The deletion effect is

$$\begin{aligned}
D(u) = &\lambda_1 \left[\log s_i(\hat{y}_i \mid r_i) - \log s_i(\hat{y}_i \mid r_i^{(-u)})\right]_+ \\
&+ \lambda_2 \mathbb{I}\left[\hat{y}_i \neq \hat{y}_i^{(-u)}\right],
\end{aligned}$$

where $[\cdot]_+$ keeps only positive drops in support. We normalize by unit size:

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

Gold labels are not used in this score; they enter only when validating whether the scores explain final correctness.

The control score measures whether a unit can be constrained in a stable way with downstream effect. We use lexical, structural, and semantic control prompts. For unit u, let $a(u)$ be the average adherence score over valid prompts, and let $\Delta(u)$ be the average change in support for the original model answer under successful control. We define

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

Further implementation details for intermediate layer distributions, answer scoring, and control prompts are given in Appendix A.4.

4.2 Trace aggregation and interventions

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

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

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

Each axis is standardized within each dataset and model block. The main combined score uses fixed pilot weights:

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

The weights are fixed on a disjoint pilot pool and are not tuned on the reported evaluation sets. Weight sensitivity analyses are reported in Appendix C.

Our main intervention analysis uses a matched deletion budget. For each trace, every selection rule 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. Token methods remove the highest ranked tokens directly. Span methods greedily accumulate clause spans until the budget is met or minimally exceeded. Unless otherwise stated, units are ranked by answer support $A(u)$.

4.3 Taxonomy and validation

To summarize functional composition, we derive a four-way taxonomy from normalized effort, answer support, and control scores within each dataset and 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 weakly supporting, and residual scaffolding. Load-bearing units have high answer support and are not low in control; deep but fragile units have high effort and low answer support; controllable but weakly supporting units have high control and low answer support; all remaining units are residual scaffolding.

We validate the framework in three ways. First, we test whether label-free trace summaries explain final correctness better than standard trace-level proxies. Second, we use matched-budget interventions to test whether top-ranked units have greater behavioral effect. Third, we compare derived labels with human annotations and align generated units with benchmark-provided intermediate structure. Human labels and external structure are used only for evaluation and interpretation.

5 Experimental Setup

We test whether label-free unit-level scores explain reasoning quality better than trace-level proxies, whether targeted interventions identify behaviorally important units, and whether the induced taxonomy aligns with human judgment and dataset-provided structure.

5.1 Benchmarks and models

We evaluate on four benchmarks with intermediate supervision: HotpotQA, StrategyQA, EntailmentBank, and ProofWriter. The filtered suite contains 2,900 items: 800, 700, 600, and 800 examples, respectively. Gold answers are used only to evaluate final correctness, while intermediate supervision is used for calibration and external-structure alignment.

We use three 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 token and clause-span granularities, 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 final correctness using 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. 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 support for the original model answer, and invalid-output rate. Unless otherwise stated, units are ranked by label-free answer support.

To study complementarity among effort, answer support, and control, we compute pairwise correlations and cross-axis disagreement rates. We then derive a four-way taxonomy 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 spans. Each trace is labeled by three annotators, with agreement measured by Krippendorff's α. Automatic predictions are evaluated with macro-F1 against adjudicated labels. For datasets with external intermediate supervision, we project supporting facts, decomposition steps, entailment steps, or proof steps into a common span-level comparison space and measure alignment with generated 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, visible reasoning traces are functionally heterogeneous: correct traces contain more load-bearing content, label-free 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. The proposed functional scores are more informative: effort reaches 0.29, answer support is the strongest single 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. Since these scores do not use gold answers, this result shows that label-free trace diagnostics carry information about success and failure beyond length, confidence, or position.

Table 2 gives behavioral evidence for the unit rankings. Under the same 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%. The support for the original model answer also drops most under the targeted methods, and invalid-output rates remain low. This suggests that answer support identifies units that are behaviorally important for the model's prediction, rather than merely selecting late or locally salient text.

Table 3 shows that effort, answer support, and control are related but not redundant. Their mean pairwise correlations are moderate: 0.31 for effort and answer support, 0.21 for effort and control, and 0.29 for answer support and control. The disagreement regions are also substantial: on average, 9.6% of units are high in effort but low in answer support, 8.5% are high in answer support but low in control, and 5.9% are high in control but low in answer support. Visible reasoning traces therefore cannot be reduced to a single scalar notion of reasoning quality.

Table 4 shows how non-redundancy among the axes appears in 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 contain more deep but fragile units (16.8% versus 13.9%) and more controllable but weakly supporting units (21.1% versus 18.6%). The residual scaffolding class is large in both groups, but the difference is not significant. Successful traces are not simply those with less filler; they allocate more content to units that support the model's answer.

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-

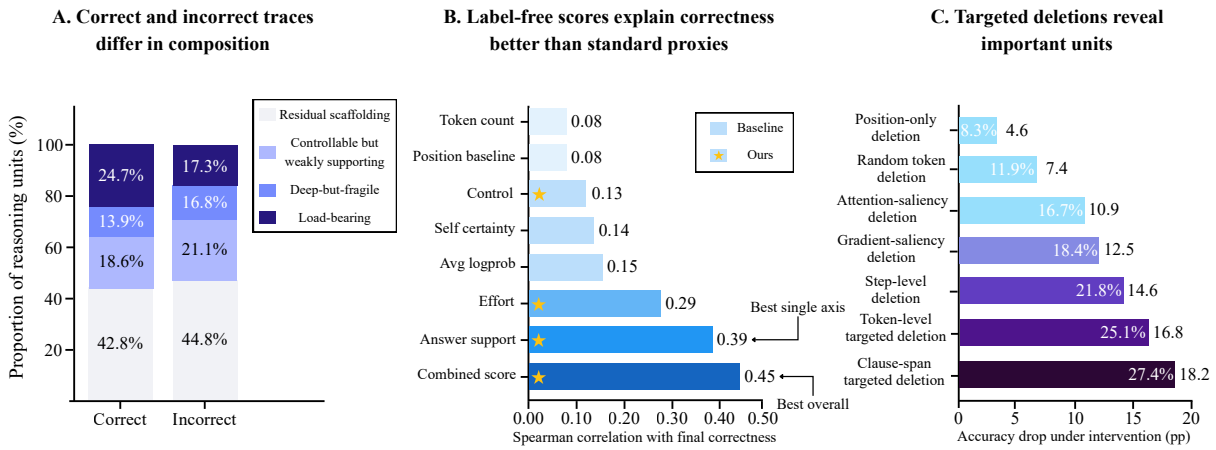

Figure 2: Core empirical findings. (A) Correct and incorrect traces differ in functional composition: correct traces contain more load-bearing units, whereas incorrect traces contain more deep-but-fragile and controllable but weakly supporting content. (B) Label-free functional scores explain final correctness better than standard trace-level proxies, with answer support as the strongest single axis and the combined score performing best overall. (C) Under a matched deletion budget, removing units ranked by label-free answer support causes larger accuracy drops than competing baselines, especially at the clause span level.

Table 1: Association between reasoning-related metrics and final correctness. Proposed scores are label-free; correctness is used only for validation. The combined score uses fixed pilot weights over effort, answer support, and control. A per-model breakdown is in Appendix C.1.

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]
Answer support 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]

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: visible reasoning quality is better characterized by the functional distribution of reasoning units than by trace length alone.

7 Discussion and Conclusion

Our results support a more selective view of visible reasoning traces. Chain-of-thought is often treated as a single object, judged by its presence, length, or plausibility. Our findings suggest that this view is incomplete. Visible traces are functionally heterogeneous: some units support the model's answer, some reflect effort without stable usefulness, some are controllable without much answer support, and 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 much reasoning text a model produces, but how functional importance is distributed within that text.

This view also clarifies the relation among effort, answer support, and control. The three axes are positively related, but they do not collapse into one dimension. A unit can be effortful but fragile, supportive but hard to control, or controllable

Table 2: Answer sensitivity under a matched deletion budget. Units selected by our method are ranked by label-free answer support. Additional attribution baselines show the same matched-budget pattern in Appendix C.2.

Intervention / baseline	ΔAcc (pp)	Answer flip (%)	Δ support for $\hat{y}$	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

Table 3: Joint analysis of effort, answer support, and control.

Dataset	Corr(E,A)	Corr(E,C)	Corr(A,C)	High-E low-A (%)	High-A low-C (%)	High-C low-A (%)
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. Significance markers indicate the correct-vs.-incorrect difference: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. not significant.

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 weakly supporting	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.

while remaining weakly tied to the model's answer. The intervention results make this concrete: under a matched deletion budget, targeted deletion is stronger than length or position based perturbation, and clause spans are more effective targets than isolated tokens. The taxonomy adds a compositional view, showing that correct traces contain more load-bearing content, whereas incorrect traces contain more fragile and weakly supporting units.

We do not claim that visible reasoning fully reveals hidden computation. Instead, we show that it remains useful when analyzed at the right granularity. The same broad distinctions appear across interventions, human judgments, and benchmark-provided intermediate structure, which supports the empirical value of the framework without treating visible reasoning as a complete account of internal reasoning.

We presented a label-free 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. Across four reasoning benchmarks and three language models, the proposed 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. Visible reasoning quality is therefore better characterized by the distribution of behaviorally meaningful units than by trace length alone.

Limitations

Our analysis is limited to visible reasoning traces. The framework should not be read as a full account of hidden model computation: a unit that appears weak in the generated trace may still correspond to internal computation not expressed in text.

The three scores depend on operational choices. Effort uses layer-wise revision dynamics, answer support uses local interventions on the model's original answer, and control uses lexical, structural, and semantic constraints. Other probes, intervention families, or control protocols may change the absolute scores.

The empirical scope is also limited. We study four English reasoning benchmarks and three open-weight models, and our reasoning units are automatically segmented tokens or clause spans rather than gold primitives. Human labels, benchmark structure, and intervention effects are useful validation signals, but none is a complete ground truth for reasoning importance.

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.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

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.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

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.

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.

Appendix: Supplementary Material

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 reproducible.

A.1 Benchmark and model summary

Table A1 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.

Table A2 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 across the evaluated models.

A.2 Dataset provenance and filtering

Table A3 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.

Table A4 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 resulting final kept counts define the exact 2,900-item benchmark used throughout the paper.

A.3 Generation validity, performance, and compute

Table A5 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 A2 and show that the required reasoning/final-answer format is stable across all three models.

Table A6 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 used in evaluation, and the pooled suite accuracy is the example-count-weighted average over the four benchmarks. 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 A7 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.

Table A1: 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 A2: 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

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 model-answer support is computed, 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. We do not train separate per-layer decoding heads. The same frozen output head is shared across all layers, so the intermediate distributions are diagnostic readouts rather than calibrated task-level probabilities.

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

$$\hat{v}_\ell^{(t)} = \arg\max_{v\in\mathcal{V}} p_\ell^{(t)}(v),$$

and $K^{(t)}$ is the earliest layer whose argmax remains unchanged through the final layer.

Model-answer support. For the answer support score, we score the model's original answer rather than the gold answer. Let $\hat{y}_i^\star$ denote the canonicalized surface form of the model answer produced from the original trace. Canonicalization uses the same normalization rules as benchmark evaluation, including case folding, whitespace normalization, punctuation stripping where applicable, and yes/no canonicalization for binary tasks.

Given input x_i and reasoning trace r_i, we score any candidate answer $a = (a_1, \ldots, a_{|a|})$ by length-

Table A3: 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 A4: 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

normalized forced decoding:

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

The answer support score uses the change in support for the original model answer before and after intervention:

$$\Delta_{\hat{y}}(u) = \left[\log s_i(\hat{y}_i^\star \mid r_i) - \log s_i(\hat{y}_i^\star \mid r_i^{(-u)})\right]_+.$$

This quantity does not use the gold answer. Gold labels are used only when evaluating final correctness.

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. Lexical adherence is scored as 1 if the required token or phrase is included, or successfully excluded, after normalization, 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 adherence is scored as 1 if the generated local span satisfies the requested template after normalization, and 0 otherwise.

For *semantic control*, we 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. Semantic adherence is scored by overlap with the normalized proposition target: a prompt receives adherence 1 if the regenerated local span has ROUGE-L F1

Table A5: 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 A6: 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

≥ 0.60 after normalization, and 0 otherwise.

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

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

where $\mathrm{adh}(p) \in \{0, 1\}$. We then measure downstream effect only on successful controls:

$$\Delta(u) = \frac{1}{|\mathcal{P}^+(u)|} \sum_{p \in \mathcal{P}^+(u)} \Big| \log s_i(\hat{y}_i^\star \mid r_i) - \log s_i(\hat{y}_i^\star \mid r_i^{(p)}) \Big|.$$

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

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

A unit receives a high control score only when the model obeys the local constraint and the controlled regeneration changes support for the model's original answer.

Why these choices are sufficient for the main analysis. These implementation choices support the relative comparisons used in the paper rather than estimating fully calibrated internal probabilities. The effort score uses a shared frozen output head as a common probe across layers; answer support uses length-normalized teacher-forced scores of the model's original answer before and after intervention; and the control score separates adherence from downstream effect. The resulting quantities are reproducible functional diagnostics over visible reasoning units.

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 label inventory, deterministic taxonomy mapping, and dataset-level agreement breakdown.

B.1 Summary calibration results

Table B1 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.

Agreement among annotators is consistently substantial, ranging from 0.67 on StrategyQA to 0.79 on ProofWriter, with an overall value of 0.74. Our macro-F1 is stable across all four datasets and clearly above the baseline. The external-

Table A7: 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
Answer support 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

structure F1 values are strongest on the more explicitly structured benchmarks, EntailmentBank and ProofWriter.

Table B2 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.

B.2 Annotation protocol and taxonomy mapping

Table B3 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. Panel B reports the raw human label inventory. 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. Panel C gives the deterministic mapping algorithm that converts normalized effort, answer support, and control scores into the derived func-

Table B1: 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 B2: 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

tional classes used throughout the paper. Raw human labels are used only for evaluation and interpretation, not for assigning classes on the full evaluation set.

B.3 Dataset-level agreement breakdown

Table B4 provides the dataset-level breakdown behind the aggregate calibration result in Table B1. 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.

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.

C Additional Intervention and Robustness Analyses

This section reports additional analyses that extend the main intervention results. We first examine how the choice of intervention unit affects behavioral impact and agreement with human judgments. We then compare our deletion-based intervention against alternative perturbation families in their native, non-budget-matched forms. Finally, we test whether the proposed framework is robust to decoding and prompt variation.

C.1 Per-model breakdown of main effects

Table C1 reports the main effects separately for each model. This table complements the pooled results in the main paper and checks whether the observed patterns are driven by a single model. For each model, we report the mean correlation of answer support with final correctness, the mean correlation of the combined score with final correctness, the accuracy drop under clause-span targeted deletion, and the correct-minus-incorrect difference in load-bearing unit proportion.

C.2 Attribution baseline comparison

To compare our ranking rule with stronger attribution baselines, we evaluate several label-free rankers under the same matched deletion budget used in the main paper. All rankers target the model's original answer rather than the gold answer. For gradient-based baselines, we define the scalar target as the length-normalized support for the original model answer,

$$f_i(r_i) = \log s_i(\hat{y}_i \mid r_i),$$

and compute token importance with respect to reasoning-token embeddings. Token scores are aggregated to clause spans by averaging over the tokens in each span. Each method removes the same amount of reasoning content: 15% of reasoning-span tokens, rounded to the nearest integer, with a minimum of 4 and a maximum of 24 removed tokens.

The baselines include attention saliency, gradient saliency, Gradient $\times$ Input, Integrated Gradients, and a pure leave-one-out support variant. The pure leave-one-out baseline ranks a unit only by the drop in support for the original model answer after deleting that unit, without using effort or control signals. This separates the effect of simple local support removal from the full unit-level diagnostic used in our framework.

Table B3: 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, answer support, and control within dataset/model	Define thresholds	Yes
2	If answer support $\in$ top quartile AND control $\notin$ bottom half	Assign *Load-bearing*	Yes
3	Else, if effort $\in$ top quartile AND answer support $\in$ bottom half	Assign *Deep-but-fragile*	Yes
4	Else, if control $\in$ top quartile AND answer support $\in$ bottom half	Assign *Controllable but weakly supporting*	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

Table C2 shows the expected pattern for this comparison. Gradient $\times$ Input and Integrated Gradients are stronger than raw gradient saliency, but they remain below the pure leave-one-out support variant and the proposed rankers. Pure leave-one-out support is a strong baseline because it directly measures the local change in support for the model's original answer, but it lacks the additional structure provided by the full unit-level framework. The combined unit score and the answer-support ranker remain stronger under the matched deletion budget, with clause spans producing the largest accuracy drop and answer-flip rate. This pattern would support the interpretation that the proposed diagnostic does not merely reproduce standard gradient attribution, while keeping the comparison label-free and budget matched.

C.3 Intervention granularity and perturbation families

Table C3 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 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. Clause spans produce the strongest

Table B4: 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 C1: Per-model breakdown of main effects. Correlations are macro-averaged over the four benchmarks. Clause deletion reports the accuracy drop under matched-budget clause-span targeted deletion. Load-bearing Δ is the correct-minus-incorrect difference in load-bearing unit proportion.

Model	Answer support corr.	Combined corr.	Clause deletion ΔAcc (pp)	Load-bearing Δ
Llama-3.1-8B-Instruct	0.34	0.40	-15.9	+5.9
Qwen3-8B	0.38	0.44	-18.7	+7.6
DeepSeek-R1-Distill-Qwen-32B	0.45	0.50	-20.0	+8.7
Mean	0.39	0.45	-18.2	+7.4

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). Step spans are weaker, consistent with the view that full reasoning steps are often too coarse to isolate where functional importance is concentrated.

Table C4 broadens the intervention analysis by comparing alternative perturbation families in their native forms, without budget matching. The 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. Semantically destructive interventions tend to cause the largest drops. Nonsense-token substitution, antonym or contradiction substitution, and both targeted deletion variants lead to large losses in accuracy and high answer-flip rates. The two targeted methods remain among the strongest interventions overall even in this non-budget-matched setting.

C.4 Decoding and prompt sensitivity

Table C5 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. As decoding becomes more stochastic, traces become longer, but this increase in length is not accompanied by stronger functional quality. Pooled accuracy decreases, effort, answer support, and control decline, and decorative ratio increases. This reinforces the main claim that longer visible traces do not necessarily imply higher-quality reasoning.

Table C6 reports sensitivity to prompt family under the main decoding configuration, again for Qwen3-8B with thinking mode enabled. 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 answer support and lower overall accuracy. P3 increases control relative to P1, plausibly because numbered steps impose more explicit structure. The broader pattern remains stable: different prompt families shift length, style, and control, but the distinction between load-bearing and non-load-bearing content remains meaningful.

Overall, the analyses in this section support three conclusions. Clause spans provide the best trade-off between semantic coherence, causal strength, and human alignment. Targeted methods remain highly disruptive even outside the matched-budget setting. The proposed scores and taxonomy are robust to moderate variation in decoding and prompting, even though such changes affect trace length and style.

C.5 Weight and threshold sensitivity

To assess sensitivity to the aggregation weights, we recompute the trace-level combined score under several alternative weight settings while keeping all unit-level scores fixed. For each setting, we report the Spearman correlation with final correctness on each benchmark and the macro-average over the four datasets.

Table C2: Attribution baseline comparison under a matched deletion budget. All rankers are label-free and target support for the model's original answer.

Ranker / selection rule	ΔAcc (pp)	Answer flip (%)	Δ support for $\hat{y}$	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	-10.9 [-12.5, -9.3]	16.7 [14.9, 18.6]	-0.134 [-0.150, -0.119]	1.2
Gradient saliency	-12.5 [-14.2, -10.8]	18.4 [16.5, 20.4]	-0.149 [-0.166, -0.133]	1.5
Gradient × Input	-13.1 [-14.7, -11.2]	19.3 [17.2, 21.5]	-0.157 [-0.174, -0.139]	1.4
Integrated Gradients	-13.8 [-15.6, -12.1]	20.6 [18.6, 22.7]	-0.164 [-0.181, -0.147]	1.3
Step-level deletion	-14.6 [-16.2, -12.8]	21.8 [19.9, 23.9]	-0.171 [-0.188, -0.154]	0.8
Pure LOO support	-15.4 [-17.2, -13.7]	23.0 [20.9, 25.1]	-0.181 [-0.199, -0.163]	1.0
Combined unit score (ours)	-17.1 [-18.9, -15.2]	25.5 [23.4, 27.6]	-0.198 [-0.218, -0.180]	1.1
Answer support, token level (ours)	-16.8 [-18.6, -14.9]	25.1 [23.0, 27.3]	-0.193 [-0.211, -0.176]	1.3
Answer support, clause span (ours)	-18.2 [-20.1, -16.4]	27.4 [25.2, 29.7]	-0.209 [-0.228, -0.191]	0.9

Table C3: 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

The results in Table C7 show that the main conclusion is not tied to a narrow choice of weights. The pilot-fixed setting used in the main paper performs best, but the equal-weight, answer-support-heavy, and no-control variants remain close. In contrast, settings that suppress answer support degrade more clearly, with the no-answer-support variant producing the largest drop across all four benchmarks. This pattern suggests that answer support is the most important axis, while the benefit of multi-axis aggregation remains stable across reasonable weighting choices.

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 C8 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 weakly supporting content. Human macro-F1 also changes only slightly across settings, suggesting 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. 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.

Table C4: 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 C5: Sensitivity to decoding configuration.

Decoding config	Valid trace %	Pooled acc. (%)	Avg reas. tokens	Effort score	Ans. support	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

D.1 Failure modes

Table D1 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. Two denominators are used: the full set of 8,700 traces generated by three models over the 2,900-item benchmark suite, and a stratified control subset of 1,740 valid traces used for control-specific analysis.

Formatting failures are relatively rare. Premature answer leakage, empty or underlength reasoning, and delimiter or parse failures each occur in only a small fraction of traces. A second group of failures reflects low-quality reasoning even when formatting succeeds. Degenerate repetition appears in 3.2% of all traces, self-contradictory reasoning in 2.8%, and unsupported reasoning leaps in 6.1%. The control-specific rows are larger: control instruction ignored occurs in 46.8% of the control subset, and forbidden-token leakage in 31.5%. This supports the claim that local controllability is a distinct and difficult axis of variation.

D.2 Checklist-oriented disclosure

Table D2 provides a checklist-oriented disclosure summary for the full paper. Its purpose is not to introduce new empirical claims, but to help 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 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 supporting information has been documented.

Table C6: Sensitivity to prompt family.

Prompt family	Valid trace %	Pooled acc. (%)	Avg reas. tokens	Effort score	Ans. support	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 C7: 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, answer support, and control. The pilot-fixed setting used in the main paper is shown in the first row.

Weights (w_E, w_A, 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
Answer-support-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 answer support $(0.50, 0.00, 0.50)$	0.27	0.18	0.30	0.25	0.25

Table C8: 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 weakly supporting 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 D1: 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

Table D2: Checklist-oriented disclosure summary.

Checklist item	Status	Supporting section / artifact
Data statistics and split information	Yes	Appendix Tables A3 and A4
Model parameters, infrastructure, and compute	Yes	Appendix Table A7
Experimental settings and hyperparameters	Yes	Appendix Tables A2 and A7
Generation validity accounting	Yes	Appendix Table A5
Descriptive statistics and uncertainty intervals	Yes	Main Tables 1–4
Existing packages and software settings	Yes	Appendix Table A7
Annotator instructions, protocol, and mapping	Yes	Appendix Table B3
Recruitment and compensation	Yes	Appendix Table B3
Consent statement	Yes	Appendix Table B3
Ethics and limitations statement	Yes	Limitations section and annotation protocol
Annotator demographics and background	Yes	Appendix Table B3
Artifact availability	Yes	Anonymous supplementary artifact with prompts, split IDs, derived scores, intervention outputs, annotations, and reproduction scripts
AI assistance disclosure	Yes	Acknowledgments and reproducibility statement