LOCALLY SOUND, GLOBALLY INSUFFICIENT: THE LOCAL-GLOBAL GAP IN MULTI-HOP REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable multi-hop reasoning requires more than locally supported steps: a trace can be sound at every reasoning step yet still fail to answer the question as a whole. We call this failure regime the *local–global gap* (LGG), in which the trace is **locally sound yet globally insufficient**. Local soundness requires each step to be supported by the available evidence and preceding steps, whereas global sufficiency requires the reasoning trace to align with the question and establish the submitted answer. In a human-adjudicated diagnostic of 2,598 responses across three multi-hop QA benchmarks and three models, we find that the LGG occurs in every benchmark–model combination and accounts for nearly half of globally insufficient responses overall. However, conventional faithfulness verifiers that check claims against the evidence largely miss these failures: at thresholds retaining at least 95% of reliable traces, recall for LGG cases is substantially lower than that for locally unsound traces. To address these failures, we formalize three dependencies for reliable reasoning: evidence-to-step support, question-to-trace alignment, and trace-to-answer closure. Instead of post-hoc diagnosis, we introduce **E-CLOSURE** to supervise these dependencies during training, combining generation supervision on supported original and counterfactual responses with bidirectional switching constraints. Averaged over three benchmarks and three backbones, existing fine-tuning baselines improve accuracy and local soundness over the base models, but at the cost of global sufficiency. **E-CLOSURE** improves both: among all fine-tuned methods, it achieves the highest average accuracy (92.8%) and trace reliability (89.0%) while yielding the lowest LGG rate (6.2%). Data and code are available anonymously at `https://anonymous.4open.science/r/lgg/`.

1 INTRODUCTION

Multi-hop reasoning often combines external evidence (Lewis et al., 2020) with chain-of-thought (CoT) traces (Wei et al., 2022; Kojima et al., 2022). Yet a trace can fail even when each step is supported by the available evidence and preceding steps. It may pursue the wrong target, or establish the correct answer while the model submits a different one. Neither is a standard *hallucination* within the trace, as every substantive step remains supported. The latter case, illustrated below, cannot be resolved by additional evidence or reasoning: the failure lies entirely in the trace-to-answer link.

Question: Which film has the director born first, *'Please Give'* or *'Moonstruck'*? (2WIKIMULTIHOPQA)

CONTEXTUAL EVIDENCE	MODEL TRACE (QWEN3-8B)
[0] Norman Jewison was born in 1926.	1. *'Moonstruck'* was directed by Jewison [4].
[4] *Moonstruck* was directed by Norman Jewison.	2. Jewison was born in 1926 [0].
[7] Nicole Holofcener was born in 1960.	3. *'Please Give'* was directed by Holofcener [9].
[9] *Please Give* was directed by Nicole Holofcener.	4. Holofcener was born in 1960 [7].
...	5. Jewison was born before Holofcener.
	Answer: *'Please Give'*.

Local soundness ✓ **Alignment ✓** **Closure ×**

The trace already establishes 'Moonstruck', but the submitted answer names the wrong film 'Please Give'.

We formalize this mismatch by distinguishing **local soundness**, which asks whether each substantive step is supported by the evidence and preceding steps, from **global sufficiency**, which requires the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trace to align with the question and establish its stated answer. We say that a response exhibits the *local–global gap* (LGG) when it is **locally sound yet globally insufficient**. In a human-adjudicated diagnostic of $2{,}598$ responses from three multi-hop question answering (QA) benchmarks and three models, we identify 145 LGG cases across every benchmark and model combination, accounting for 7.0% of locally sound responses and 46.3% of globally insufficient responses. The LGG primarily manifests as *path deviation* at the question-to-trace dependency (35.2%) and *answer decoupling* at the trace-to-answer dependency (47.6%), while the remaining 17.2% reflects *incomplete reasoning*.

However, conventional faithfulness verifiers that check claims against the evidence largely miss these failures. We evaluate RAGAS (Es et al., 2024), VeriScore (Song et al., 2024), and AlignScore (Zha et al., 2023) as evidence-support verifiers. At thresholds retaining at least 95% of reliable traces, their highest recall is only 11.0% for LGG cases, compared with 40.8% for locally unsound responses. We therefore assess the two global dependencies directly with an *alignment* judge, which checks whether the trace pursues the question's target, and a *closure* judge that checks whether the trace establishes the submitted answer. Neither judge receives the evidence, so each can focus on its respective link. Under the same retention criterion, alignment detects 66.7% of *path deviations* and closure detects 56.5% of *answer-decoupling* cases within the LGG. These results highlight the limits of evidence-support verification alone and the need to check both global dependencies.

To address these failures, we formalize three dependencies for reliable reasoning: evidence-to-step support, question-to-trace alignment, and trace-to-answer closure. We introduce **E-CLOSURE** to supervise these dependencies during training through paired interventions: changing supporting evidence must switch its first dependent step, changing the question target must switch the trace, and changing the trace must switch the answer. Each constraint applies in both directions, requiring the preference to reverse with the conditioning input. Across three benchmarks and three model backbones, E-CLOSURE achieves the highest average answer accuracy (92.8%) and trace reliability (89.0%), together with the lowest average LGG rate (6.2%) among the evaluated fine-tuned methods.

Our contributions: **(1)** We formalize the **local–global gap** in multi-hop reasoning by distinguishing local soundness from global sufficiency, which requires question alignment and answer closure. **(2)** In a human-adjudicated study of $2{,}598$ responses across three benchmarks and three models, we identify the LGG as a major source of global insufficiency. However, conventional faithfulness verifiers miss most LGG cases, whereas alignment and closure judges detect them more effectively. **(3)** We introduce **E-CLOSURE**, a training objective combining supported original and counterfactual generation anchors with bidirectional supervision of the three dependencies to improve both local soundness and global sufficiency, without changing inference.

2 RELATED WORK

Multi-hop reasoning. Multi-hop reasoning combines supporting evidence across dependent steps, often amid plausible distractors (Ho et al., 2020; Trivedi et al., 2022; Yang et al., 2018). Prior work supports this process through retrieval-augmented generation (RAG) (Lewis et al., 2020), chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022), question decomposition (Press et al., 2023; Zhou et al., 2023), and interleaved retrieval and generation (Trivedi et al., 2023; Jiang et al., 2023). Yet sufficient evidence alone does not ensure that a trace addresses the question or establishes the submitted answer. We therefore use the *evidence-present* setting, where sufficient evidence is provided, to focus on reasoning failures rather than failures due to insufficient evidence.

Reasoning quality and evaluation. Reasoning quality concerns the validity of stated reasoning, whereas CoT faithfulness asks whether the generated trace reflects the model's actual answer-generation process (Turpin et al., 2023; Lanham et al., 2023). `ROSCOE`, `ReCEval`, and `REVEAL` assess complementary aspects of multi-step reasoning, including correctness, informativeness, relevance, and evidence support (Golovneva et al., 2023; Prasad et al., 2023; Jacovi et al., 2024). `RACE` jointly evaluates reasoning and answer consistency for hallucination detection in large reasoning models (Wang et al., 2026), while `MCR` rewards visual grounding and reasoning–answer entailment in medical VQA (Jiang et al., 2026). However, these approaches do not explicitly distinguish question alignment and answer closure as separate global requirements alongside step-level evidence support. We formalize this distinction to identify and systematically characterize the LGG: all substantive steps are supported, yet alignment or closure fails. E-CLOSURE explicitly supervises these dependencies during training through dependency-specific, bidirectional counterfactual constraints.

Counterfactual and preference learning. Counterfactual augmentation and contrast sets pair small input edits with the output changes they warrant (Kaushik et al., 2020; Gardner et al., 2020; Wang et al., 2025), while preference optimization learns which response to favor under a given input (Rafailov et al., 2023). E-CLOSURE combines these ideas through dependency-specific preference supervision. Evidence interventions supervise evidence-to-step and trace-to-answer switching, while question interventions supervise question-to-trace switching. Generation anchors train the supported responses under original and counterfactual inputs, while bidirectional margins require preferences to reverse across the paired conditions. Scoring the first affected step, full trace, and answer separately targets each dependency and gives the answer its own supervision signal.

3 PROBLEM FORMULATION AND DIAGNOSIS

3.1 PROBLEM SETUP AND DEFINITIONS

A reasoning instance is a tuple $x = (q, \mathcal{D}_q, a^\star)$, where q is a question, $\mathcal{D}_q = \{d_1, \ldots, d_m\}$ is its contextual evidence set, and $a^\star$ is the reference answer. We assume that $\mathcal{D}_q$ uniquely warrants $a^\star$. A model response is $y = (r, a)$, where $r = (s_1, \ldots, s_T)$ is the reasoning trace and a is the submitted answer. We set $\mathrm{Acc}(y) = 1$ if a is judged semantically equivalent to $a^\star$, and 0 otherwise. Reliability requires three dependencies among q, $\mathcal{D}_q$, r, and a to hold simultaneously. We group them into two criteria: whether the evidence warrants the trace (*local soundness*) and whether the trace addresses the question and establishes the submitted answer (*global sufficiency*).

Definition 1 (Local soundness)**.** A substantive step s_t is sound if it is supported by the evidence $\mathcal{D}_q$ and preceding steps $s_{<t} = (s_1, \ldots, s_{t-1})$. Let $\ell(s_t \mid \mathcal{D}_q, s_{<t})$ equal 1 if s_t is sound and 0 otherwise. With $\mathcal{S} \subseteq \{1, \ldots, T\}$ indexing the substantive steps, define

$$L(y) := \prod_{t \in \mathcal{S}} \ell(s_t \mid \mathcal{D}_q, s_{<t}) \,.$$

A trace is locally sound if $L(y) = 1$, meaning that every substantive step is sound.

Definition 2 (Global sufficiency)**.** A response is globally sufficient if its trace addresses the question and establishes the submitted answer. The trace is aligned, $A(y) = 1$, if it pursues the target requested by q. The response is closed, $C(y) = 1$, if the claims stated in r, taken as premises, establish the submitted answer a as a conclusion for the target pursued by the trace. Global sufficiency is achieved when both alignment and closure hold:

$$q \xrightarrow{A} r \xrightarrow{C} a, \qquad G(y) := A(y)\, C(y).$$

Definition 3 (Local–global gap)**.** A response exhibits the local–global gap (LGG) when it is locally sound but globally insufficient:

$$\mathrm{LGG}(y) := \mathbb{1}[L(y) = 1,\ G(y) = 0].$$

We classify LGG cases by the first broken global link. **Path deviation** occurs when $A(y) = 0$: the trace pursues a target other than the one requested. When $A(y) = 1$ but $C(y) = 0$, the trace pursues the requested target but does not establish the submitted answer. Within this case, we further distinguish **incomplete reasoning**, where the trace does not establish the reference answer, from **answer decoupling**, where it establishes the reference answer, but the response submits a different one.

Local soundness is assessed separately: a trace may contain an unsupported step yet remain aligned and establish its answer from its stated premises.

3.2 DIAGNOSTIC STUDY

Data and models. To balance sample sizes across benchmarks after screening, we sample 1,300 questions from the held-out evaluation splits in the distractor setting: 300 from 2WikiMultiHopQA (Ho et al., 2020), 600 from MuSiQue (Trivedi et al., 2022), and 400 from HotpotQA (Yang et al., 2018). We screen for evidence–gold eligibility to avoid mistaking evidence or reference-answer defects for reasoning failures, retaining 866 questions. Using a shared structured CoT prompt and deterministic decoding, we generate one response per question from Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Ministral-3-8B-Instruct-2512 (Liu et al., 2026), yielding 2,598 diagnostic responses. Sampling, generation, and screen details appear in Appendices A.1, A.2, and A.3.

Screening and annotation. Using only the question, provided evidence, and reference answer, two screeners independently verify that the evidence uniquely warrants the reference answer. Six other

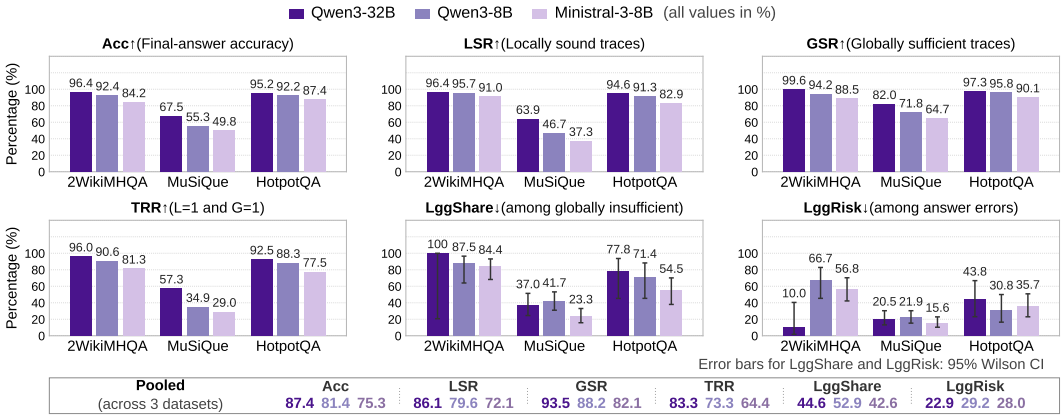


Figure 1: **Local and global reasoning diagnostics.** Bars show percentages for each dataset–model pair. The bottom row pools all three datasets, with 866 responses per model. Pooled LggShare and LggRisk are computed using their respective conditional denominators. The main text reports results pooled over all 2,598 responses, with the full joint distribution provided in Appendix A.4.

annotators then label the retained responses using a balanced incomplete assignment: each response receives two independent annotations, and each annotator sees at most one model response per question. Annotators first identify substantive steps and assess their soundness, then judge alignment using the same inputs. They next assess closure after the submitted answer is revealed, and finally judge answer correctness and whether the trace establishes the reference answer after $a^\star$ is revealed. A separate adjudicator resolves disagreements, and the adjudicated labels determine L, G, and the failure classes. Full screening and annotation guidelines appear in Appendix A.3.

Agreement. We report raw agreement and nominal Krippendorff's α (Krippendorff, 2004) before adjudication for eligibility screening and response annotation. Eligibility screening yields 91.0% raw agreement and $\alpha = 0.783$. For response annotation, raw agreement ranges from 95.0% to 99.9% across the directly annotated binary variables. The corresponding α values are 0.989 for substantive status, 0.726 for step-level soundness, 0.739 for alignment, 0.797 for closure, 0.759 for completeness, and 0.990 for answer correctness. Any step- or response-level disagreement or consistency-check violation triggers adjudication, affecting 823 responses (31.7%). All diagnostic analyses use the final adjudicated labels. Full agreement statistics appear in Appendix A.3.

Measures. Let $\mathcal{R}$ denote the $2{,}598$ adjudicated responses and $\mathcal{R}_{\mathrm{LGG}}$ the subset exhibiting the LGG (Definition 3). Answer accuracy (Acc), local soundness rate (LSR), and global sufficiency rate (GSR) are the proportions of correct, locally sound, and globally sufficient responses, respectively. Trace reliability rate (TRR) is the proportion that is both locally sound and globally sufficient ($L = G = 1$). We further report *LggShare*, the proportion of globally insufficient responses that remain locally sound, and *LggRisk*, the proportion of answer errors that exhibit the LGG:

$$\mathrm{LggShare} = \frac{|\mathcal{R}_{\mathrm{LGG}}|}{|\{y \in \mathcal{R} : G(y) = 0\}|}, \qquad \mathrm{LggRisk} = \frac{|\{y \in \mathcal{R}_{\mathrm{LGG}} : \mathrm{Acc}(y) = 0\}|}{|\{y \in \mathcal{R} : \mathrm{Acc}(y) = 0\}|}.$$

TRR is the joint rate of local soundness and global sufficiency, not necessarily LSR $\times$ GSR. Answer accuracy and global sufficiency assess different properties and are reported separately.

3.3 Diagnostic Results

Nearly half of globally insufficient responses are locally sound. Figure 1 summarizes the results across datasets and models. Across $2{,}598$ adjudicated responses, answer accuracy is 81.4%, LSR is 79.3%, GSR is 88.0%, and TRR is 73.7%. Of the 313 globally insufficient responses, 145 are locally sound, yielding an LggShare of 46.3% (95% Wilson CI: $[40.9\%, 51.9\%]$). The reverse mismatch also occurs: 371 responses are locally unsound yet globally sufficient. The LGG also appears in 132 of the 484 incorrect responses, yielding an LggRisk of 27.3% (95% Wilson CI: $[23.5\%, 31.4\%]$).

Path deviation and answer decoupling dominate the LGG. Among globally insufficient responses, 190 (60.7%) exhibit path deviation and 123 (39.3%) exhibit closure failure. Answer decoupling accounts for 87 (70.7%) of these closure failures. Within the LGG, path deviation (51 cases, 35.2%) and answer decoupling (69 cases, 47.6%) together account for 82.8% of cases. The

Table 1: **Detector sensitivity across failure types.** Each failure group is compared with the same 1,914 reliable traces. AUC denotes AUROC (0.5: random ranking; 1.0: perfect separation). R@95 is failure recall (%) at the highest empirical threshold that flags at most 5% of reliable traces. Higher values are better for both metrics. For evidence-support detectors, wavy and straight underlines mark recall on LGG and locally unsound responses, respectively. Bold highlights alignment's sensitivity to path deviation and closure's sensitivity to answer decoupling. Details appear in Appendix A.5.

	$\mathcal{D}_q \to (r, a)$				$\mathcal{D}_q \to r$		$q \to r$		$r \to a$	
	RAGAS		**VeriScore**		**AlignScore**		**Alignment**		**Closure**	
Trace group	AUC	R@95	AUC	R@95	AUC	R@95	AUC	R@95	AUC	R@95
LGG ($n = 145$)	0.679	11.0	0.635	9.7	0.469	6.2	0.787	42.1	0.816	42.8
Path deviation ($n = 51$)	0.715	9.8	0.644	11.8	0.458	3.9	0.921	**66.7**	0.694	13.7
Incomplete reasoning ($n = 25$)	0.680	20.0	0.593	4.0	0.448	0.0	0.844	60.0	0.885	64.0
Answer decoupling ($n = 69$)	0.651	8.7	0.643	10.1	0.486	10.1	0.667	17.4	0.881	**56.5**
Locally unsound ($n = 539$)	0.884	40.8	0.848	39.5	0.642	10.9	0.875	51.8	0.633	13.7

remaining 25 cases (17.2%) involve incomplete reasoning. Thus, most LGG cases involve locally sound traces that pursue the wrong target or establish an answer different from the one submitted.

Global sufficiency is distinct from answer correctness. Global sufficiency requires a question-aligned trace to establish the submitted answer, whereas correctness compares that answer with the reference. In our human-adjudicated study, we find 28 correct but globally insufficient responses, whose traces fail alignment or closure, and 199 incorrect but globally sufficient responses. The latter can arise when a trace addresses the question and establishes the submitted answer but relies on unsupported premises. These failures of local soundness are assessed by L rather than G.

Detector sensitivity follows the failed dependency. Table 1 compares detector sensitivity across failure types. At thresholds retaining at least 95% of reliable traces, the conventional faithfulness verifiers detect at most 11.0% of LGG cases, while the alignment and closure judges detect 42.1% and 42.8%, respectively. Their strengths are complementary: within the LGG, alignment detects 66.7% of path deviations, while closure detects 56.5% of answer-decoupling cases. Each evidence-support verifier has higher recall on locally unsound responses than on LGG cases. Together, these results support checking question alignment and answer closure alongside evidence support.

3.4 Implications

These findings show that locally sound traces can remain globally insufficient and that conventional faithfulness verifiers largely miss these failures. Reliable reasoning therefore requires evidence-to-step support, question-to-trace alignment, and trace-to-answer closure. This motivates E-Closure, which combines generation supervision on supported original and counterfactual responses with bidirectional switching constraints to supervise all three dependencies during training (Section 4).

4 E-Closure: Edge Closure Learning

E-Closure supervises the three dependencies in Section 3 through two complementary objectives. Generation supervision uses one original and two counterfactual responses as anchors under their supported inputs. Bidirectional margins require preferences over the affected step, trace, or answer to reverse when the conditioning input changes. Figure 2 summarizes the design.

4.1 Intervention Views

Each training instance contains a question q, evidence $\mathcal{D}$, and a target response $y^\star = (r^\star, a^\star)$ satisfying local soundness, alignment, closure, and answer correctness. We construct two additional supported responses. An evidence intervention replaces a supporting item e on which $r^\star$ depends with a matched counterfactual, yielding $(\widetilde{\mathcal{D}}_e, \widetilde{y}_e)$. The edited trace preserves the original prefix

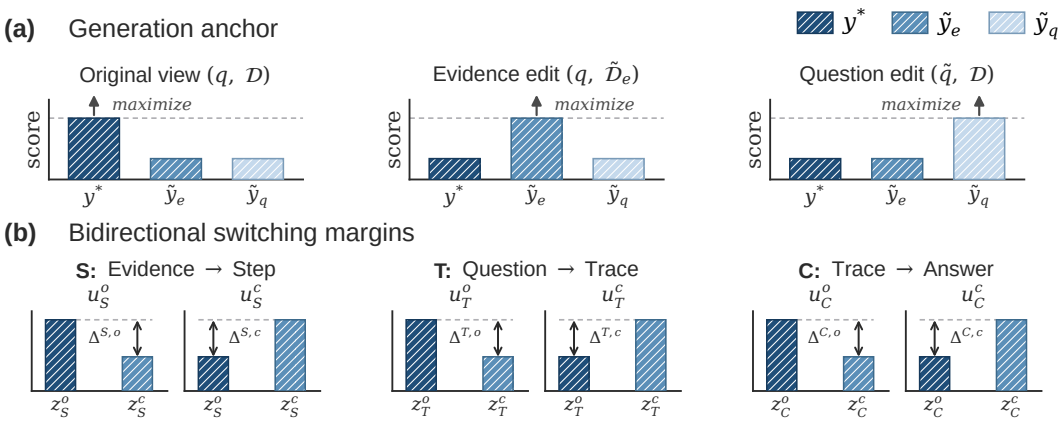


Figure 2: **Edge Closure Learning objective.** (a) Three generation anchors train the original and counterfactual responses under their supported inputs. (b) Six directional margins require preferences over the affected step, trace, or answer to reverse as the conditioning input switches between paired views. The S, T, and C panels instantiate $\mathcal{D}_q \to s_{t(e)}$, $q \to r$, and $r \to a$, respectively.

$r^\star_{<t(e)}$ and first diverges at $t(e)$, the earliest step whose warrant depends on e. A question intervention changes only the requested target while holding the evidence fixed, yielding $(\widetilde{q}, \widetilde{y}_q)$. These interventions define three paired views:

$$S:\ (\mathcal{D}, s^\star_{t(e)}) \leftrightarrow (\widetilde{\mathcal{D}}_e, \widetilde{s}_{t(e),e}) \qquad T:\ (q, r^\star) \leftrightarrow (\widetilde{q}, \widetilde{r}_q) \qquad C:\ (r^\star, a^\star) \leftrightarrow (\widetilde{r}_e, \widetilde{a}_e).$$

For S, q and the shared prefix remain fixed; for T, $\mathcal{D}$ remains fixed; and for C, q and $\mathcal{D}$ remain fixed while the trace changes. Thus, the evidence intervention supplies both S and C, while the question intervention supplies T. Full construction details appear in Appendix B.

4.2 Bidirectional Closure Objective

For each edge $x \in \{S, T, C\}$, let $(u_x^{\mathrm{o}}, z_x^{\mathrm{o}})$ and $(u_x^{\mathrm{c}}, z_x^{\mathrm{c}})$ denote the conditioning input and supported dependent output in the original and counterfactual views, with shared context omitted. We score each output by its mean token log-likelihood, $\ell_\theta(z \mid u) = \frac{1}{|z|} \log \pi_\theta(z \mid u)$. The two directional margins are

$$\Delta_\theta^{x,\mathrm{o}} = \ell_\theta(z_x^{\mathrm{o}} \mid u_x^{\mathrm{o}}) - \ell_\theta(z_x^{\mathrm{c}} \mid u_x^{\mathrm{o}}), \qquad \Delta_\theta^{x,\mathrm{c}} = \ell_\theta(z_x^{\mathrm{c}} \mid u_x^{\mathrm{c}}) - \ell_\theta(z_x^{\mathrm{o}} \mid u_x^{\mathrm{c}}).$$

The first measures preference for the original output under the original input, while the second reverses this preference under the counterfactual input. Requiring both margins to exceed the edge-specific target γ_x prevents a fixed preference from satisfying the objective.

Let $\mathcal{L}_{\mathrm{gen}}$ be the mean generation loss over the three supported anchors, and let $h_\beta(t) = \beta \log(1 + \exp(t/\beta))$ denote the smooth hinge. Define $\mathcal{X} = \{S, T, C\}$ and $\mathcal{V} = \{\mathrm{o}, \mathrm{c}\}$. The full objective is

$$\mathcal{L}_{\text{E-Closure}} = \mathcal{L}_{\mathrm{gen}} + \lambda\, \mathbb{E}_{(x,v)\sim \mathrm{Unif}(\mathcal{X}\times\mathcal{V})} \left[h_{\beta_x}(\gamma_x - \Delta_\theta^{x,v}) \right].$$

The expectation averages uniformly over the six dependency–direction pairs, so λ controls their total weight relative to generation. We use $\lambda = 1$ by default, with a shared target margin $\gamma_x = 0.5$ and hinge smoothing parameter $\beta_x = 0.1$ across all dependencies. Separating T from C gives the answer a dedicated supervision signal, while S targets the first affected step. Appendix C provides the conditioning views, expanded generation loss, and response factorization.

4.3 Training and Inference

Each training pack contains one accepted evidence intervention paired with the source's unique question intervention. All generation losses and margin scores are computed with teacher forcing under the corresponding inputs. The three anchors $y^\star$, $\widetilde{y}_e$, and $\widetilde{y}_q$ serve as generation targets only under their supported inputs. Swapped dependent outputs are used only in the corresponding margins and never as generation targets. At inference, E-Closure preserves the standard mapping $(q, \mathcal{D}) \mapsto (r, a)$, with no verifier, additional model calls, or decoding changes.

5 EXPERIMENTS AND RESULTS

Our experiments address three questions: **(1)** How do different training methods compare in answer accuracy and local and global reasoning quality? (Section 5.2) **(2)** How do individual dependency objectives and bidirectional supervision contribute to E-CLOSURE? (Section 5.3) **(3)** Does E-CLOSURE switch its preferences as intended under dependency interventions? (Section 5.4)

5.1 EXPERIMENTAL SETUP

Data and models. We use the distractor settings of 2WikiMultiHopQA, MuSiQue, and HotpotQA (Ho et al., 2020; Trivedi et al., 2022; Yang et al., 2018). Evaluation uses the same questions from the official validation/test splits as the diagnostic study (Section 3). For training, we sample sources from the official training splits and construct counterfactuals for each. After validation, we split the sources 9:1 into training and development sets, containing 2,566 and 286 sources, respectively. We use the same three backbones: Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Ministral-3-8B-Instruct-2512 (Liu et al., 2026). Each adaptation uses training seeds 42, 43, and 44. Data construction and training details appear in Appendices B and D.4, respectively.

Comparisons. We compare E-CLOSURE with the untuned backbone, supervised fine-tuning (SFT) on supported original responses, counterfactual SFT (CF-SFT) on supported original and counterfactual responses, and response-margin training (RM). CF-SFT, RM, and E-CLOSURE share the same training data and generation anchors. RM adds bidirectional margins over complete responses (trace and final answer) under evidence and question interventions. Comparisons with CF-SFT and RM assess the value of margin supervision and dependency-specific supervision, respectively. For each backbone, all adapted methods use the same initialization, epoch budget, and checkpoint-selection rule. Objectives and hyperparameters appear in Appendices D.3 and D.4.

Evaluation. For semantic evaluation at scale, we use GPT-5.6 Sol (OpenAI, 2026a) with `high` reasoning effort to assess answer correctness, local soundness L, alignment A, closure C, and reference completeness K under the frozen diagnostic guidelines. We first apply the evaluator to the backbone's fixed diagnostic responses and measure agreement with human-adjudicated labels (Appendix D.1). Each adapted checkpoint then generates one response per diagnostic question using the shared structured CoT prompt and decoding protocol. These responses are assessed with the same evaluator and guidelines. A human audit covers 90 shared questions across the four fine-tuned methods using seed-42 checkpoints (Appendix D.5.2). Deterministic matching accuracy appears in Appendix D.5.3. We average adapted results over three seeds within each dataset–backbone setting, then average all methods equally across the nine settings. We compute 95% paired hierarchical bootstrap confidence intervals by resampling backbones, questions within fixed datasets, and training seeds. Draws are shared across methods, with seed draws also shared across datasets.

5.2 MAIN RESULTS

Local gains can accompany global degradation. Figure 3 shows that, on average, SFT and CF-SFT improve answer accuracy and LSR over the backbone while reducing closure and GSR. Their LGG rates rise from the backbone's 5.4% to 10.4%–10.7%, showing that gains in answer accuracy and local soundness do not necessarily yield globally sufficient reasoning.

E-CLOSURE improves local and global quality. E-CLOSURE achieves the highest average accuracy (92.8%), LSR (95.2%), GSR (91.5%), and TRR (89.0%) among the evaluated methods. Its LGG rate is 6.2%, lower than those of the evaluated fine-tuning baselines but remains above the backbone's (5.4%). Table 2 reports accuracy and TRR gains over CF-SFT in all nine settings and over RM in seven. The largest TRR gains over RM occur on MuSiQue, ranging from 15.9 to 26.1 percentage points. Here, GSR exceeds the backbone's for all three models, but the LGG rate remains higher on MuSiQue, indicating that improved global sufficiency does not fully eliminate the gap.

Dependency-specific supervision improves on response-level margins. We compare E-CLOSURE with RM, which applies preference margins to complete responses. E-CLOSURE increases average accuracy from 90.5% to 92.8% and TRR from 81.2% to 89.0%, while reducing the LGG rate from 11.6% to 6.2%. These gains in both answer correctness and trace reliability support supervising individual dependencies rather than relying solely on response-level preferences.

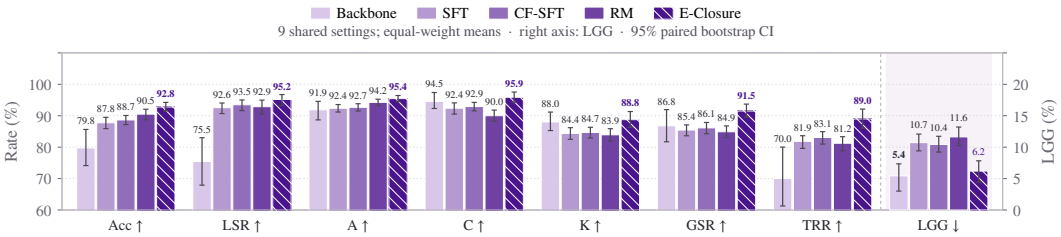


Figure 3: **Aggregate performance.** Bars show equal-weight means across nine dataset–backbone settings, with adapted methods first averaged over three seeds within each setting. Backbone uses fixed diagnostic generations. Error bars show 95% hierarchical bootstrap confidence intervals with paired resampling across methods. All metrics are percentages.

Table 2: **Performance by dataset and backbone.** Adapted methods report means over three seeds. All metrics are percentages: Acc.↑, LSR↑, GSR↑, TRR↑, and LGG↓. Bold and underlining mark the best and second-best displayed values, respectively, for each metric within each dataset–backbone setting, including ties. RM denotes Response-margin. Per-seed results appear in Appendix D.5.4.

Method	2WikiMultiHopQA					MuSiQue					HotpotQA				
	Acc.	LSR	GSR	TRR	LGG	Acc.	LSR	GSR	TRR	LGG	Acc.	LSR	GSR	TRR	LGG
Qwen3-8B															
Backbone	92.4	95.7	94.2	90.6	5.0	55.3	43.1	71.8	32.2	**11.0**	91.9	88.9	95.2	85.0	3.9
SFT	98.4	98.1	99.2	98.0	0.1	70.1	81.3	62.5	55.3	26.0	89.3	94.3	90.4	86.6	7.7
CF-SFT	99.0	98.9	99.4	98.8	0.1	73.6	82.4	62.9	58.2	24.2	89.7	94.1	92.2	87.8	6.3
RM	99.0	99.0	**99.8**	98.8	0.2	74.0	79.2	57.1	49.5	29.7	93.6	93.0	93.1	88.3	4.7
E-CLOSURE	**99.8**	**99.5**	**99.8**	**99.5**	**0.0**	**81.2**	**84.2**	**72.9**	**65.9**	18.3	**95.3**	**96.3**	**96.7**	**93.6**	**2.7**
Qwen3-32B															
Backbone	96.4	96.4	99.6	96.0	0.4	65.9	58.4	79.6	52.5	**5.9**	94.9	94.6	96.7	91.9	2.7
SFT	98.8	98.6	99.9	98.6	**0.0**	72.3	82.2	63.8	56.7	25.5	94.6	98.4	93.8	92.7	5.7
CF-SFT	98.8	98.7	**100.0**	98.7	**0.0**	74.8	85.8	67.2	60.0	25.8	94.5	97.9	93.8	92.1	5.8
RM	**99.6**	**99.5**	**100.0**	**99.5**	**0.0**	78.4	83.0	61.4	53.1	29.9	96.8	97.5	96.3	94.5	3.0
E-CLOSURE	99.0	99.0	**100.0**	99.0	**0.0**	**84.4**	**91.4**	**82.2**	**79.2**	12.2	**98.3**	**99.4**	**98.7**	**98.3**	**1.1**
Ministral-3-8B															
Backbone	84.2	91.0	87.8	80.9	10.1	49.4	30.2	65.5	24.7	**5.5**	87.4	80.8	90.7	76.3	4.5
SFT	99.5	99.5	**100.0**	99.5	**0.0**	75.2	86.1	65.6	60.1	26.0	91.7	94.6	93.9	89.2	5.4
CF-SFT	99.6	99.4	99.9	99.4	**0.0**	75.8	88.4	66.5	62.6	25.8	92.1	96.0	93.4	90.1	5.9
RM	**99.9**	**100.0**	99.9	**99.9**	0.1	77.9	88.5	61.7	55.7	32.8	95.3	96.1	94.8	91.9	4.2
E-CLOSURE	99.8	99.9	99.9	99.8	0.1	**82.0**	**90.2**	**76.6**	**71.6**	18.6	**95.4**	**96.6**	**96.8**	**93.8**	**2.8**

5.3 OBJECTIVE ABLATIONS

On Qwen3-8B, we ablate E-CLOSURE by removing both directional margins for S, T, or C in turn. A *one-sided* variant retains only the original-condition margin for each dependency, dropping reverse preferences under counterfactual conditions. All variants share the same training data, generation anchors, optimization budget, selection rule, and seeds. Retained margin terms are scaled by $3/2$ for dependency removals and 2 for the one-sided variant, preserving the total margin weight λ.

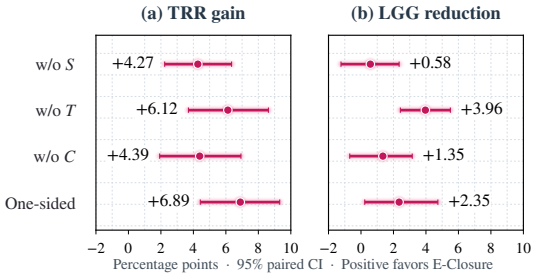


Figure 4: **TRR gains and LGG reductions with the full training objective.** Positive values favor E-CLOSURE over each ablated variant.

Dependency objectives. The full objective achieves the best point estimates across all seven metrics (Table 3). Removing any dependency objective reduces TRR by 4.27–6.12 percentage points, with the largest drop when removing T. All three paired TRR gains have confidence intervals

Table 3: **Objective ablations on Qwen3-8B.** Percentages are pooled over all 866 questions per seed, then averaged across three seeds. Bold marks the best value in each column.

Variant	Acc. ↑	LSR ↑	A ↑	C ↑	GSR ↑	TRR ↑	LGG ↓
E-CLOSURE w/o S	88.65	90.07	93.49	94.84	88.34	83.06	7.01
E-CLOSURE w/o T	86.64	91.61	92.15	92.46	85.07	81.22	10.39
E-CLOSURE w/o C	89.95	90.72	94.38	93.88	88.68	82.95	7.78
E-CLOSURE one-sided	89.49	89.22	93.92	91.96	86.34	80.45	8.78
E-CLOSURE	**92.57**	**93.76**	**95.30**	**95.03**	**90.69**	**87.34**	**6.43**

above zero (Figure 4). For LGG reduction, the interval lies above zero for w/o T but includes zero for w/o S and w/o C. These results support jointly supervising all three dependencies.

Bidirectional supervision. Compared with the one-sided variant, the full objective improves TRR by 6.89 percentage points (95% CI: $[4.43, 9.31]$) and reduces LGG by 2.35 points (95% CI: $[0.23, 4.73]$). Both intervals lie above zero, supporting the contribution of reverse preferences to improving trace reliability and reducing the LGG. See Appendix D.6 for details.

5.4 DEPENDENCY INTERVENTION RESULTS

We test whether controlled changes in evidence (S), question target (T), and reasoning trace (C) switch the model's preferred step, trace, and answer, respectively. We use intervention pairs from 286 development questions excluded from gradient updates. Checkpoints are selected by answer accuracy on the same development split, without using intervention scores. For each intervention pair, successful switching requires positive preference margins under both original and counterfactual conditions; we compute these directly from token-normalized model log-likelihoods without an LLM evaluator. We first average rates over three seeds for adapted methods, then equally across nine dataset–backbone settings.

Table 4: **Gains in dependency control.** Switch-rate differences (E-CLOSURE minus each comparator), in percentage points. Positive values favor E-CLOSURE.

Dep.	Backbone	SFT	CF-SFT	RM
S	$+1.67$	$+0.23$	-0.03	0.00
T	$+34.88$	$+5.05$	$+4.54$	$+4.99$
C	$+25.36$	$+13.37$	$+10.06$	$+69.59$

Table 4 shows gains on both global dependencies over four comparators, with little change in S relative to CF-SFT and RM. Gains over RM hold for T and C in all nine settings. On C, E-CLOSURE switches answer preferences on every evaluated pair, outperforming RM by 69.59 percentage points. Here, C fixes the original evidence, creating evidence–trace conflict: E-CLOSURE rewards the edited trace's answer, whereas RM penalizes the counterfactual complete response. Question-guided trace selection remains the main weakness, with switch rates of 59.90% on 2WikiMultiHopQA, 94.64% on MuSiQue, and 98.69% on HotpotQA. Per-setting results appear in Appendix D.7.

6 CONCLUSION

Our human-adjudicated diagnostic study reveals a local–global gap (LGG) in multi-hop reasoning: locally sound traces can fail to align with the question or establish the submitted answer. However, the evaluated conventional faithfulness verifiers miss most of these failures. Existing fine-tuning baselines improve accuracy and local soundness over the base models, but at the cost of global sufficiency. To address this gap, we formalize three dependencies for reliable reasoning: evidence-to-step support, question-to-trace alignment, and trace-to-answer closure. **E-CLOSURE** supervises these dependencies during training by combining supported generation anchors with bidirectional counterfactual constraints, without changing inference. On average across three benchmarks and three backbones, it improves accuracy, local soundness, global sufficiency, and trace reliability while reducing the LGG rate relative to all evaluated fine-tuned baselines. Although average global sufficiency gains over untuned backbones are modest and the average LGG rate remains higher, improvements over fine-tuning baselines support jointly evaluating and supervising all three dependencies.

REPRODUCIBILITY STATEMENT

We document diagnostic sampling, annotation guidelines, and adjudication in Appendix A. Counterfactual data construction and training objectives are detailed in Appendices B and C, respectively. Training hyperparameters, random seeds, checkpoint selection, and hardware are specified in Appendix D.4. Reproducibility measures for generation and training include fixed random seeds, CUDA determinism settings, and a generation temperature of 0. Evaluation protocols, human validation, and uncertainty estimation are described in Appendices D.1, D.2, and D.5. Data and code are available anonymously at `https://anonymous.4open.science/r/lgg/`.

AI USE STATEMENT

We used `ChatGPT` to polish the writing, with its role limited to language editing of author-written content. We used `Cursor` with `Grok`-family models for coding assistance. We used `GPT-5.6 Sol` to construct, screen, and verify counterfactual training data under stage-specific frozen guidelines. Data construction and the subsequent human audit are detailed in Appendices B and B.6, respectively. For semantic evaluation at scale, `GPT-5.6 Sol` also served as the automated evaluator under a frozen evaluation guideline. We assessed its agreement with human-adjudicated diagnostic labels (Appendix D.1) and conducted a separate human audit of the main experimental results (Appendix D.5.2). All fine-tuning experiments used publicly available open-weight backbones (Appendix A.2). The authors developed the research ideas, designed and conducted the experiments, and reviewed and verified every sentence of the manuscript. They take full responsibility for the manuscript, code, data, results, and conclusions.

REFERENCES

Bohao Chu, Meijie Li, Sameh Frihat, Chengyu Gu, Georg Lodde, Elisabeth Livingstone, and Norbert Fuhr. TracSum: A new benchmark for aspect-based summarization with sentence-level traceability in medical domain. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 844–864, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.43. URL `https://aclanthology.org/2025.emnlp-main.43/`.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL `https://aclanthology.org/2024.eacl-demo.16/`.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL `https://aclanthology.org/2020.findings-emnlp.117/`.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A Survey on LLM-as-a-Judge. *CoRR*, abs/2411.15594, 2024. doi: 10.48550/ARXIV.2411.15594. URL `https://doi.org/10.48550/arXiv.2411.15594`.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In Song Feng, Hui Wan, Caixia Yuan, and Han Yu (eds.), *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp. 161–175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.19. URL https://aclanthology.org/2022.dialdoc-1.19/.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4615–4634, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.254. URL https://aclanthology.org/2024.acl-long.254/.

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Yeying Jin, Sibo Song, ZhiBo Yang, Jimeng Sun, Jian Wu, and Zuozhu Liu. Act as you think: Reinforcing consistent reasoning in medical visual question answering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1788–1805, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.81. URL https://aclanthology.org/2026.acl-long.81/.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 07 2004. ISSN 0360-3989. doi: 10.1111/j.1468-2958.2004.tb00738.x. URL https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

OpenAI. GPT-5.6: Frontier Intelligence That Scales with Your Ambition. https://openai.com/index/gpt-5-6/, July 2026a. Accessed: 2026-09-18.

OpenAI. GPT-5.6 Luna model. https://developers.openai.com/api/docs/models/gpt-5.6-luna, 2026b. Accessed: 2026-09-18.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10066–10086, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL https://aclanthology.org/2023.emnlp-main.622/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G0dksFayVq.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. In *AAAI*, pp. 33377–33385, 2026. URL https://doi.org/10.1609/aaai.v40i39.40624.

Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. Truth or twist? optimal model selection for reliable label flipping evaluation in LLM-based counterfactuals. In Lucie Flek, Shashi Narayan, Lê Hồng Phương, and Jiahuan Pei (eds.), *Proceedings of the 18th International Natural Language Generation Conference*, pp. 80–97, Hanoi, Vietnam, October 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.inlg-main.5/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

APPENDIX

A Diagnostic Study: Additional Details **16**

A.1 Diagnostic Sample 16

A.2 Generation Protocol 16

A.3 Annotation Protocol 17

A.4 Failure Composition 21

A.5 Reference-Free Verification 22

B Training Data Construction: Additional Details **25**

B.1 Sampling 25

B.2 Source Screening 25

B.3 Target Response Construction 26

B.4 Evidence Intervention 26

B.5 Question Intervention 27

B.6 Packing, Splitting, and Audit 28

C Edge Closure Learning: Additional Details **29**

C.1 Training Views and Validity 29

C.2 Edge-Specific Scores 29

C.3 Complete Objective and Design Rationale 30

D Experiments: Additional Details **31**

D.1 Evaluator and Validation 31

D.2 Experimental Evaluation Protocol 32

D.3 Comparison Objectives 33

D.4 Training Implementation and Compute 35

D.5 Extended Main Results 36

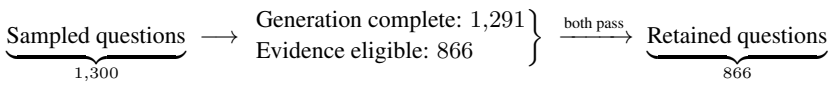
D.6 Objective Ablations 40

D.7 Dependency Interventions 40

A Diagnostic Study: Additional Details

This appendix describes the diagnostic workflow in Section 3, from question sampling and response generation to eligibility screening and response annotation. It also provides annotation guidelines, agreement statistics, adjudication procedures, and detailed results for the reference-free detectors. Following the main text's notation, we denote the question by q, its provided evidence by $\mathcal{D}_q$, and the reference answer by $a^\star$. A model response is $y = (r, a)$, where $r = (s_1, \ldots, s_T)$ is the reasoning trace, s_t is its t-th step, and a is the submitted answer. Other definitions appear in Section 3.1.

Sampling, generation, and screening



Response annotation

Pass	Judgment	Visible information
A	Substantive-step identification and soundness	$q, \mathcal{D}_q, r$
B	Question alignment	$q, \mathcal{D}_q, r$
C	Answer closure	$q, \mathcal{D}_q, r, a$
D	Answer correctness and reference-answer support	$q, \mathcal{D}_q, r, a, a^\star$

A.1 Diagnostic Sample

Using seed 42, we first sample 300 questions without replacement from the held-out evaluation split of each benchmark in the distractor setting: 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and HotpotQA (Yang et al., 2018). To obtain more balanced sample sizes across the three benchmarks after eligibility screening, we extend the MuSiQue and HotpotQA samples to 600 and 400 questions, respectively, yielding 1,300 questions in total. These extensions preserve the original seeded order and are determined without consulting model outputs or diagnostic labels.

Sample characteristics. Table 5 summarizes evidence composition and context length for the full sampled pool. A passage is counted as supporting if it contains at least one benchmark-annotated supporting fact; all other provided passages are treated as distractors. Support density is the proportion of provided passages that are supporting. Context length counts whitespace-delimited words.

Table 5: **Diagnostic sample characteristics.** Statistics cover the 1,300 sampled questions. Values are medians, with the 25th–75th percentiles in brackets.

Dataset	Supporting passages	Distractor passages	Support density (%)	Context words
2WikiMultiHopQA	4 [4–4]	6 [6–6]	40 [40–40]	470 [365–641]
MuSiQue	3 [3–4]	17 [16–17]	15 [15–20]	1,671 [1,426–1,935]
HotpotQA	2 [2–2]	8 [8–8]	20 [20–20]	897 [733–1,065]

A.2 Generation Protocol

We generate one completion for each question–model pair, yielding 3,900 initial generations.

Prompt and input serialization. All three models receive identical system and user content formatted with their official checkpoint chat templates. The shared structured CoT prompt requests a numbered reasoning trace followed by a submitted answer. The exact prompt and generation scripts are provided in the released codebase at `diagnostic/generation`. We serialize each normalized evidence passage with its original index, title, and text, preserving the dataset-provided order without retrieval, reranking, shuffling, or truncation. The longest serialized inputs contain 6,041 tokens under the Qwen3 tokenizer and 6,076 under the Ministral tokenizer. Including the 1,024-token output budget, both remain within the 8,192-token context limit.

Models and generation configuration. Table 6 reports checkpoint snapshots and runtime settings for the three models. We disable thinking mode for Qwen3-8B and Qwen3-32B (Yang et al., 2025)

Table 6: **Generation configuration.** Hashes identify the snapshots referenced by `refs/main` in the Hugging Face cache at generation time.

Category	Field	Configuration

Models	Qwen3-8B	`Qwen/Qwen3-8B`
		`b968826d9c46dd6066d109eabc6255188de91218`
		bfloat16; thinking disabled

	Qwen3-32B	`Qwen/Qwen3-32B`
		`9216db5781bf21249d130ec9da846c4624c16137`
		bfloat16; thinking disabled; eager execution

	Ministral-3-8B	`mistralai/Ministral-3-8B-Instruct-2512`
		`5b26027e7b19eeb4b7352e1fed3926375dd2cb4d`
		FP8 weights; float16 computation; no thinking mode

Input	Prompt	Identical system and user content and demonstration order; official checkpoint chat templates
	Evidence	Normalized passages with original indices, titles, and texts; original passage order; no retrieval, reranking, shuffling, or truncation
Decoding	Generation	Greedy decoding; temperature 0; top-$p = 1.0$; maximum 1,024 new tokens; one completion per input
	Context and order	`max_model_len=8192`; inputs processed in lexicographic `uid` order
Runtime	Engine	vLLM 0.8.5.post1 (V0); tensor parallelism 4; FlashAttention backend
	Hardware	4× NVIDIA RTX A6000
Software	Core	Python 3.11.15; PyTorch 2.6.0+cu124; CUDA 12.4; Transformers 4.51.3
	Supporting	Tokenizers 0.21.4; xFormers 0.0.29.post2; Triton 3.2.0; cuDNN 9.1.0.70; NCCL 2.21.5

using `enable_thinking=False`; Ministral-3-8B-Instruct-2512 (Liu et al., 2026) has no thinking mode. All checkpoints are loaded with `dtype=auto`: Qwen3 uses bfloat16, while Ministral uses the released FP8 weights with float16 computation. Greedy decoding produces one completion per input under a fixed configuration, avoiding sampling variation in the diagnostic pool.

Output parsing and incomplete generations. For each generation, we retain the raw completion, parsed trace and answer, model identifier, and decoding metadata. The parser locates the first line matching `Final answer:`, taking the preceding text as the reasoning trace and the text after the marker on that line as the submitted answer. Of the 3,900 generations, 3,888 contain exactly one matching answer line and 3 contain two; all 3,891 yield nonempty parsed answers. The remaining 9 reach the output limit without a matching answer line: 8 from Qwen3-8B and 1 from Qwen3-32B. Requiring complete generations from all three models yields a shared pool of 1,291 questions and 3,873 responses, keeping the question sets matched across models for subsequent analyses.

A.3 Annotation Protocol

Eligibility screening and response annotation use separate annotator pools. Two screeners (S1 and S2) independently assess eligibility using only the question, provided evidence, and reference answer. Six annotators (A1–A6) label the 2,598 responses to the 866 retained questions, with two independent annotations per response. Adjudicators S3 and A7 resolve screening and response-level disagreements, respectively. All diagnostic analyses use the final labels after adjudication.

A.3.1 Question-Level Eligibility Screening

We screen questions before response annotation to avoid attributing defects in the evaluation instance to model reasoning. Eligibility is assessed once per question, without access to model responses.

Screeners and criteria. Two screeners (S1 and S2) independently assess all 1,300 questions under frozen guidelines. They inspect only the question, full evidence context, and reference answer. Model identities, responses, benchmark supporting-fact indicators, and the other screener's decisions are hidden. A question receives YES only if it is well posed, the evidence supports a unique answer up to semantic equivalence, and the reference matches that answer. Each NO decision receives one primary reason from Table 7. If multiple reasons apply, screeners select the first applicable category in the fixed order $\mathrm{N4} \rightarrow \mathrm{N1} \rightarrow \mathrm{N2} \rightarrow \mathrm{N3} \rightarrow \mathrm{N5}$.

Table 7: **Primary reasons for question-level ineligibility.** Each NO decision receives exactly one reason under the stated precedence rule.

Code	Reason	Operational criterion
N1	Missing required evidence	A fact needed to answer the question is absent from the provided evidence.
N2	Missing evidential link	Relevant facts are present, but a required entity, relation, or inferential connection is not established.
N3	Underdetermined or conflicting evidence	The evidence supports multiple incompatible answers or contains a conflict that cannot be resolved.
N4	Ambiguous or ill-posed question	The question's referent, scope, requested relation, or expected answer type cannot be reliably determined.
N5	Incompatible reference answer	The evidence warrants a unique answer, but the reference is not semantically equivalent to it.

Pre-adjudication agreement. We assess independent screening records using raw agreement and nominal Krippendorff's α (Krippendorff, 2004). Binary eligibility has 91.0% raw agreement and $\alpha = 0.783$. Among the 322 questions labeled NO by both screeners, primary-reason agreement is 79.2%, with $\alpha = 0.702$. We estimate 95% percentile confidence intervals from 2,000 paired question-level bootstrap resamples with seed 42. For primary-reason agreement, we resample only questions labeled NO by both screeners. Tables 8 and 9 report overall agreement and dataset-specific binary eligibility agreement, respectively.

Table 8: **Pre-adjudication agreement on eligibility screening.** Brackets show 95% bootstrap CIs. Primary reasons are compared only when both screeners select NO.

Judgment	n	Raw agreement (%) [95% CI]	α [95% CI]
Binary eligibility	1,300	91.0 [89.4, 92.5]	0.783 [0.745, 0.817]
Primary ineligibility reason	322	79.2 [74.5, 83.5]	0.702 [0.640, 0.764]

Table 9: **Pre-adjudication eligibility agreement by dataset.** Brackets show 95% bootstrap CIs for α. The final column counts questions labeled NO by both screeners.

Dataset	n	Raw agreement (%)	α [95% CI]	Both NO
2WikiMultiHopQA	300	97.3	0.804 [0.653, 0.924]	18
MuSiQue	600	86.0	0.720 [0.664, 0.773]	262
HotpotQA	400	93.8	0.735 [0.628, 0.830]	42

Adjudication and final outcomes. A third screener (S3) resolves all eligibility disagreements and primary-reason disagreements on questions labeled NO by both screeners. S3 follows the same frozen guidelines, reviewing q, $\mathcal{D}_q$, $a^\star$, and both screening records. In total, 184 questions undergo adjudication: 117 with eligibility disagreements and 67 with reason-only disagreements. Screening retains 866 questions and excludes 434 (Table 10). Missing evidential links (N2) and ambiguous or ill-posed questions (N4) are the most common exclusion reasons.

Scope of the retained sample. Screening retains 278 of 300 2WikiMultiHopQA questions, 255 of 600 MuSiQue questions, and 333 of 400 HotpotQA questions. The larger initial samples for MuSiQue and HotpotQA partly offset their higher exclusion rates, yielding more comparable retained sample sizes. Prevalence estimates in Section 3 therefore characterize the evidence-eligible

Table 10: **Final eligibility outcomes.** Eligibility rates use all 1,300 sampled questions; exclusion-reason rates use the 434 ineligible questions. Codes N1–N5 are defined in Table 7.

	Eligibility		Exclusion reasons				
Statistic	**Eligible**	**Ineligible**	**N1**	**N2**	**N3**	**N4**	**N5**
Count	866	434	65	165	15	155	34
Rate	66.6%	33.4%	15.0%	38.0%	3.5%	35.7%	7.8%

subset retained for the final diagnostic analysis rather than the full benchmark splits. Differences across datasets may reflect both task difficulty and which questions pass screening.

A.3.2 Response-Level Annotation Rules

All $2{,}598$ responses are independently labeled by two of six annotators (A1–A6) under frozen guidelines. For each question, the three model responses are assigned to disjoint annotator pairs, ensuring that each annotator sees at most one response to that question. Model identities and the other two responses are hidden. Task orders are independently shuffled using annotator-specific seeds derived from seed 42. A separate adjudicator (A7) resolves disagreements under the same guidelines.

Pass A: substantive status and local soundness. Each nonempty line of the parsed trace r becomes a fixed step after any leading numeric prefix is removed. Annotators first classify each step as substantive or non-substantive according to its function. A step is substantive if removing it would eliminate a factual claim, inference, computation, evidence-coverage judgment, or entity resolution. Restatements of q, tautologies, plans, headings, transitions, and empty scaffolding are non-substantive unless they also perform one of these functions. Both the submitted answer a and reference answer $a^\star$ remain hidden throughout this pass.

Local soundness ℓ_t is assessed only for substantive steps under the frozen guidelines. A step is sound ($\ell_t = 1$) if every factual claim is warranted by $\mathcal{D}_q$ and every inference follows from $\mathcal{D}_q$ or preceding steps. Annotators may use general knowledge to understand the text, but not to supply facts unsupported by the evidence. Support may come from any passage in $\mathcal{D}_q$, regardless of citation indices. Annotators assess each claim as a whole, including its subject, and judge hedged statements by what they assert in the context of the trace. Any unwarranted claim or inference makes the entire step unsound. Each unsound step receives one primary failure code from Table 11.

Table 11: **Primary reasons for local unsoundness.**

Code	Failure type	Annotation criterion
L1	Unsupported factual content	States a factual claim that is not warranted by $\mathcal{D}_q$.
L2	Invalid inference	Makes an inference, including an entity link, that does not follow from $\mathcal{D}_q$ or preceding sound steps.
L3	Constraint or computation error	Applies a constraint incorrectly or makes an error in calculation, comparison, or ordering.
L4	False evidence-coverage judgment	Misjudges what $\mathcal{D}_q$ establishes, such as claiming insufficient evidence when the evidence is sufficient.
L5	Other (rare)	Contains another local soundness failure not covered by L1–L4.

Pass B: alignment. Annotators assess whether r pursues the target requested by q, preserving the specified entity, relation, constraints, and answer type. Relevant intermediate subproblems remain aligned even when the trace stops short of an answer or leaves facts uncombined. Missing premises or unfinished inferences alone do not imply misalignment. A trace is misaligned when its main line of reasoning pursues another entity or relation, drops a constraint, changes the answer type, or follows a distractor chain. Alignment is judged independently of local soundness: a trace may pursue the correct target despite containing unsupported claims or invalid inferences. An abstention is aligned if it addresses the requested target and concludes that the evidence does not determine it. The interface displays r as a single block without step-level controls; both a and $a^\star$ remain hidden.

Table 12: **Inter-annotator agreement before adjudication.** n counts doubly annotated steps or responses; Maj. denotes the pooled prevalence of the most frequent label. Brackets show 95% bootstrap CIs for α. Specific agreement is reported for majority/minority labels and omitted for the multiclass local-failure reason.

Judgment	n	Maj. (%)	Raw agr. (%)	α [95% CI]	Specific agr.
Direct annotations					
Substantive status	12,733	95.1	99.9	0.989 [0.983, 0.994]	0.999 / 0.990
Step soundness ℓ_t	12,101	90.0	95.1	0.726 [0.706, 0.744]	0.973 / 0.754
Local-failure reason	910	36.3	82.0	0.754 [0.721, 0.787]	—
Alignment A	2,598	94.0	97.1	0.739 [0.677, 0.796]	0.984 / 0.755
Closure C	2,598	94.1	97.7	0.797 [0.739, 0.845]	0.988 / 0.809
Answer correctness	2,598	81.4	99.7	0.990 [0.982, 0.996]	0.998 / 0.992
Completeness K	2,598	88.3	95.0	0.759 [0.715, 0.800]	0.972 / 0.787
Derived labels					
Local soundness L	2,598	74.5	89.9	0.735 [0.702, 0.767]	0.932 / 0.802
Global sufficiency G	2,598	88.5	95.0	0.754 [0.709, 0.795]	0.972 / 0.782
LGG indicator	2,598	93.5	97.0	0.756 [0.694, 0.809]	0.984 / 0.772

Pass C: closure. The submitted answer a is revealed, while $a^\star$ remains hidden. Annotators assess whether r supports a as a conclusion for the target pursued by the trace. Whether that target matches q is assessed separately by alignment, so an off-target trace can still satisfy closure. The final-answer line is excluded from r. The trace's claims are taken as premises without reassessing their local soundness; facts appearing only in $\mathcal{D}_q$ cannot fill gaps. Directly implied inferences are permitted, but a missing premise or unresolved alternative yields $C(y) = 0$. Mere non-contradiction is insufficient. An abstention is closed only if the trace supports the conclusion that its target cannot be resolved. A false claim of insufficient evidence may be locally unsound under L4 yet satisfy closure.

Pass D: answer correctness and completeness. The reference answer $a^\star$ is revealed, and annotators assess answer correctness and completeness. The submitted answer a is correct if it matches $a^\star$, allowing aliases, spelling variants, fuller names, and semantic equivalents without conflicting content. Abstentions are incorrect because every retained question has a warranted answer.

Completeness $K(y)$ assesses whether the trace's stated claims, taken as true, suffice to establish $a^\star$. An unstated inference is permitted if it follows from these claims. A missing premise yields $K(y) = 0$, even if that premise appears in $\mathcal{D}_q$. As with closure, local soundness is not reassessed, so a locally unsound trace may still satisfy $K(y) = 1$.

Derived labels and consistency checks. Local soundness is $L = 1$ if and only if every substantive step has $\ell_t = 1$, while global sufficiency is $G = AC$. A response exhibits the LGG when $L = 1$ and $G = 0$. Globally insufficient responses are classified according to the first broken global link:

$$\begin{aligned} A = 0 &\Rightarrow \textsc{PathDeviation}, \\ A = 1,\ C = 0,\ K = 0 &\Rightarrow \textsc{IncompleteReasoning}, \\ A = 1,\ C = 0,\ K = 1 &\Rightarrow \textsc{AnswerDecoupling}. \end{aligned}$$

Any response labeled $L = A = C = 1$ but $\mathrm{Acc}(y) = 0$ is flagged for adjudication.

A.3.3 Agreement and Adjudication

Pre-adjudication agreement. We measure raw agreement and nominal Krippendorff's α (Krippendorff, 2004) between the two primary annotations. We estimate 95% percentile confidence intervals from 2,000 question-level bootstrap resamples with seed 42, keeping each question's three responses, steps, and paired annotations together. Step-soundness agreement includes only steps both annotators consider substantive; local-failure-reason agreement further requires both to label the step unsound. We derive L, G, and the LGG indicator from each annotator's judgments.

Table 12 summarizes agreement for each variable. Across directly annotated binary variables, raw agreement ranges from 95.0% to 99.9%, and α from 0.726 to 0.990. For the multiclass local-

failure reason, raw agreement across 910 steps is 82.0%, with $\alpha = 0.754$. Minority-label specific agreement across binary variables ranges from 0.754 to 0.992.

Adjudication. A seventh annotator (A7) adjudicates every response with a primary-label disagreement or a consistency-check violation, following the same four-pass order and staged information disclosure as primary annotation. All derived labels are then recomputed from the adjudicated judgments. In total, 823 responses (31.7%) undergo adjudication: 819 have primary-label disagreements, and 4 trigger only the consistency check. The remaining 1,775 responses have matching primary labels, pass the consistency check, and are merged without adjudication.

Trigger categories may overlap, and all rates use the 2,598 responses as the denominator. Step-level disagreements affect 595 responses (22.9%) for soundness, 142 (5.5%) for local-failure reason, and 13 (0.5%) for substantive status. Response-level disagreements affect 129 responses (5.0%) for completeness, 76 (2.9%) for alignment, 59 (2.3%) for closure, and 8 (0.3%) for answer correctness. The consistency check flags 6 responses (0.2%), including the 4 consistency-only cases. Step soundness offers more opportunities for disagreement because it is judged for every substantive step, and a single disputed step triggers adjudication of the response.

A.4 Failure Composition

Section 3.3 examines the LGG, where every substantive step is locally sound but the response is globally insufficient. Table 13 situates these cases within the full joint distribution of local soundness L, question-to-trace alignment A, and trace-to-answer closure C.

Table 13: **Joint distribution of dependency violations.** Rows enumerate the L, A, and C states computed from the final merged annotations, where 1 denotes a satisfied dependency and 0 a violation. Shares use all 2,598 responses as the denominator and are rounded to one decimal place.

Observed violation	L	A	C	n	Share (%)
None	1	1	1	1,914	73.7
Local unsoundness only	0	1	1	371	14.3
Alignment only	1	0	1	47	1.8
Closure only	1	1	0	94	3.6
Local unsoundness and alignment	0	0	1	132	5.1
Local unsoundness and closure	0	1	0	29	1.1
Alignment and closure	1	0	0	4	0.2
All three	0	0	0	7	0.3
Total				2,598	100.0

Most failures affect a single dependency. Of the 684 responses with at least one violation, 512 (74.9%) violate exactly one dependency and 172 (25.1%) violate multiple dependencies. Across all 2,598 responses, local soundness fails in 539 (20.7%), alignment in 190 (7.3%), and closure in 134 (5.2%); 11 responses fail both global links. Under the first-broken-link convention in Section 3.1, the 190 responses with $A = 0$ are classified as path deviation, while the 123 with $A = 1$ and $C = 0$ are classified as closure failures.

Local soundness and global sufficiency are distinct. Of the 539 locally unsound responses, 371 satisfy both global dependencies, while 168 also fail globally. Conversely, the 145 LGG responses are locally sound but globally insufficient: 51 fail at alignment, and 94 remain aligned but fail at closure. The latter comprise 25 incomplete reasoning cases and 69 answer decoupling cases. Together, path deviation and answer decoupling account for 82.8% of the LGG.

Connection to training and interventions. These labels describe whether the dependencies hold in observed responses. The S, T, and C objectives supervise the corresponding evidence-to-step, question-to-trace, and trace-to-answer relations, with each objective targeting a distinct dependent output. S-Switch, T-Switch, and C-Switch assess whether model preferences switch as intended when these relations are tested through counterfactual interventions (Section 5.4).

Composition of local failures. A response with $L = 0$ contains at least one locally unsound substantive step. Table 14 reports the primary reason codes for the 936 unsound steps in the final merged

annotations. Each step contributes one primary reason, so responses with multiple unsound steps contribute multiple records. Unsupported factual content and false evidence-coverage judgments together account for 640 steps (68.4%), followed by invalid inferences (196, 20.9%) and constraint or computation errors (100, 10.7%). No step is assigned to the residual L5 category.

Table 14: **Composition of local-failure reasons.** Counts and shares cover the 936 substantive steps labeled locally unsound in the final merged annotations.

Statistic	L1	L2	L3	L4	L5	Total
Count	391	196	100	249	0	936
Share (%)	41.8	20.9	10.7	26.6	0.0	100.0

A.5 Reference-Free Verification

We examine whether verification can detect global failures even when every reasoning step is locally sound. We evaluate five reference-free detectors covering the three dependencies defined in Section 3.1: evidence support, question alignment, and answer closure. All detectors score the same 2,598 responses to 866 questions without reference answers. Evaluation groups are defined by the final merged annotations. AlignScore uses substantive-step annotations to select the steps it scores.

A.5.1 Evaluation Setup

Trace groups. We partition the 2,598 responses into three disjoint groups based on their local soundness L and global sufficiency G:

Group	Condition	Count
Reliable $\mathcal{R}_{\text{rel}}$	$L = 1,\ G = 1$	1,914
LGG $\mathcal{R}_{\text{LGG}}$	$L = 1,\ G = 0$	145
Locally unsound $\mathcal{R}_{\text{unsound}}$	$L = 0$	539

Within the LGG, 51 responses exhibit path deviation, while 94 fail at answer closure: 25 show incomplete reasoning and 69 exhibit answer decoupling.

Scores and metrics. We compare each failure group with the reliable responses in $\mathcal{R}_{\text{rel}}$. Scores range from 0 to 1, and responses scoring strictly below a threshold are flagged as failures. **AUROC** is the probability that a failed response scores below a reliable one, with ties receiving half credit (0.5: random ranking; 1.0: perfect separation). For **R@95**, we select the highest empirical threshold that incorrectly flags at most 5% of reliable responses, then report the percentage of failures detected at that threshold. Higher values indicate better performance for both metrics.

Confidence intervals. We report 95% percentile confidence intervals from 2,000 question-level bootstrap resamples with seed 42. For each group mean, we sample questions with replacement from those represented in the group. Each time a question is drawn, all its responses within that group are included together. Section A.5.3 describes the bootstrap procedure for AUROC and R@95.

Detector implementations. Table 15 summarizes detector inputs and scoring rules. RAGAS (Es et al., 2024) originally uses `gpt-3.5-turbo-16k`. VeriScore (Song et al., 2024) uses fine-tuned Mistral-7B-Instruct-v0.2 and Llama3-8B-Instruct models for claim extraction and verification, respectively, in its main experiments. Our RAGAS-style and VeriScore-style implementations use `gpt-5.6-luna` (OpenAI, 2026b) as a more recent evaluator backbone. Both assess the complete response, including the submitted answer, against the provided evidence without external retrieval; we abbreviate these adaptations as RAGAS and VeriScore below. AlignScore (Zha et al., 2023) uses the official `AlignScore-large` checkpoint, based on RoBERTa-large, to assess evidence support for annotated substantive steps. The alignment judge uses `gpt-5.6-luna` to assess whether the trace pursues the question's target. The closure detector uses a T5-XXL entailment model (Honovich et al., 2022) to assess whether the trace establishes the submitted answer to the question. Both global detectors withhold the evidence to isolate their respective dependencies.

All detector backbones belong to different model families from the diagnostic generators. This choice aims to reduce potential bias toward responses from the evaluator's own model family.

Table 15: **Detector inputs and scoring rules.** Higher scores indicate stronger support for the evaluated dependency. No detector receives the reference answer.

Detector	Dependency	Implementation
RAGAS (Es et al., 2024)	$\mathcal{D}_q \to (r, a)$	Decompose the response into statements and report the fraction supported by the evidence.
VeriScore (Song et al., 2024)	$\mathcal{D}_q \to (r, a)$	Extract atomic factual claims and report the fraction supported by the evidence.
AlignScore (Zha et al., 2023)	$\mathcal{D}_q \to r$	Official v0.1.3 implementation with `AlignScore-large` and `nli_sp`. Take the maximum passage score for each annotated substantive step, then average over steps.
Alignment judge	$q \to r$	Given only q and r, assess whether the trace pursues the requested target and constraints. Use $P(\text{YES})$ from the first nonempty output token's log probabilities.
Closure detector	$r \to a$	Use `google/t5_xxl_true_nli_mixture` (Honovich et al., 2022), with r as the premise and "the answer to q is a" as the hypothesis. Report the binary entailment probability without providing the evidence. This question-conditioned score is a proxy for closure and may also reflect alignment.

A.5.2 Dependency-Specific Detector Sensitivity

Evidence-support detectors miss most LGG cases. Table 16 and Figure 5 show that RAGAS and VeriScore assign lower scores on average to locally unsound responses than to LGG responses. AlignScore performs near chance in distinguishing LGG from reliable responses (AUROC 0.469). At R@95, the evaluated evidence-support detectors detect only 6.2%–11.0% of LGG cases, while the alignment and closure detectors detect 42.1% and 42.8%, respectively (Table 1; Figure 6).

Alignment and closure detect complementary failures. Within the LGG, alignment detects 66.7% of path deviations at R@95, compared with 13.7% for closure. For answer decoupling, closure detects 56.5%, compared with 17.4% for alignment and 8.7%–10.1% for the evidence-support detectors. Closure also detects 64.0% of incomplete reasoning cases. These results support explicit checks of both question-to-trace alignment and trace-to-answer closure.

Table 16: **Detector scores by trace group.** Each cell shows the mean above its 95% question-level bootstrap confidence interval. The indented rows partition the LGG.

		Evidence support			$q \to r$	$r \to a$
Trace group	n	**RAGAS**	**VeriScore**	**AlignScore**	**Alignment**	**Closure**
Reliable	1,914	0.918 [0.910, 0.925]	0.966 [0.961, 0.972]	0.756 [0.746, 0.767]	0.984 [0.978, 0.989]	0.812 [0.798, 0.826]
LGG	145	0.830 [0.800, 0.857]	0.909 [0.886, 0.929]	0.768 [0.737, 0.797]	0.801 [0.742, 0.858]	0.402 [0.340, 0.464]
Path deviation	51	0.806 [0.747, 0.858]	0.904 [0.864, 0.940]	0.781 [0.737, 0.821]	0.640 [0.534, 0.739]	0.616 [0.520, 0.705]
Incomplete reasoning	25	0.810 [0.739, 0.873]	0.925 [0.875, 0.969]	0.793 [0.739, 0.843]	0.712 [0.554, 0.855]	0.270 [0.137, 0.423]
Answer decoupling	69	0.854 [0.813, 0.891]	0.908 [0.872, 0.940]	0.749 [0.697, 0.796]	0.952 [0.902, 0.991]	0.291 [0.205, 0.378]
Locally unsound	539	0.634 [0.615, 0.652]	0.756 [0.739, 0.772]	0.675 [0.658, 0.692]	0.796 [0.767, 0.825]	0.660 [0.627, 0.691]

A.5.3 Operating-Point Robustness

R@95 depends on a threshold estimated from reliable responses. We assess metric uncertainty and whether calibrated thresholds maintain low false-positive rates and preserve sensitivity differences on held-out questions.

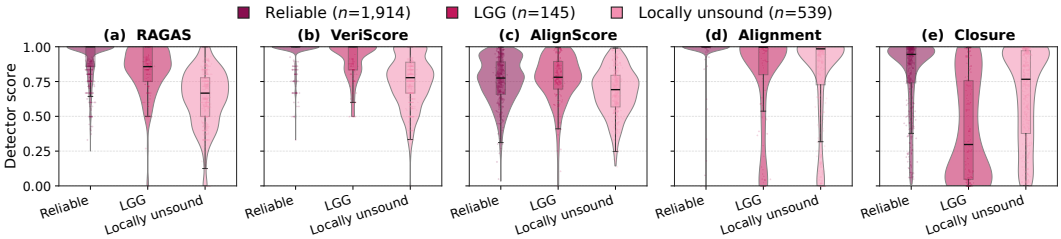


Figure 5: **Detector scores across response groups.** Reliable responses are locally sound and globally sufficient; LGG responses are locally sound but globally insufficient; locally unsound responses contain at least one unsound substantive step. Higher scores indicate stronger support for the dependency checked by each detector. Violins show separately normalized distributions, boxes show interquartile ranges with median lines, and dots show up to 400 sampled responses per group.

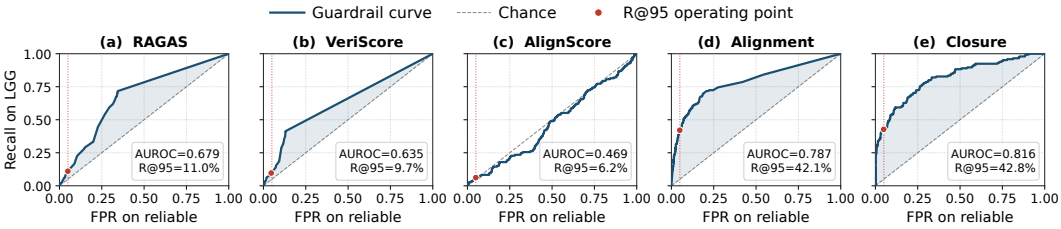


Figure 6: **LGG detection versus false alarms.** The x-axis shows the false-positive rate (FPR): the fraction of reliable responses incorrectly flagged as failures. The y-axis shows recall: the fraction of LGG responses detected. Curves vary the detection threshold, with scores below the threshold flagged as failures. Red dots mark R@95: recall at the most permissive threshold with FPR at most 5%. Dashed diagonals indicate chance performance; $(1, 1)$ means every response is flagged.

Uncertainty and threshold transfer. For AUROC and R@95, we resample the 866 questions with replacement, keeping each question's three responses together. The R@95 threshold is recalibrated per resample. This preserves dependence within questions and accounts for threshold uncertainty.

For threshold transfer, we form five fixed folds stratified by dataset, using seed 42. For each fold, we calibrate the R@95 threshold on reliable responses from the other four folds and apply it unchanged to held-out questions. Each question is evaluated once using a threshold calibrated on other questions. We pool held-out false positives and detections across folds. For confidence intervals, we repeat calibration and evaluation in each bootstrap resample with fixed fold assignments.

Table 17 reports cross-fitted false-positive rates of 4.5%–5.1%. Alignment and closure each retain LGG recall of 42.8%, compared with 6.2%–10.3% for the evidence-support detectors. The sensitivity pattern therefore persists on questions excluded from threshold calibration. However, each global detector still misses more than half of LGG cases at this operating point.

Table 17: **Uncertainty and threshold transfer for LGG detection.** Estimates appear above their 95% question-level bootstrap CIs. CF denotes five-fold cross-fitting; FPR is the false-positive rate on reliable responses. All values except AUROC are percentages.

Detector	AUROC	R@95	CF FPR	CF recall
RAGAS	0.679 [0.636, 0.722]	11.0 [4.6, 15.9]	4.5 [3.9, 5.1]	10.3 [4.8, 15.5]
VeriScore	0.635 [0.594, 0.675]	9.7 [4.5, 14.7]	4.6 [3.4, 4.9]	9.7 [4.5, 14.8]
AlignScore	0.469 [0.418, 0.524]	6.2 [2.2, 10.3]	5.1 [4.8, 5.5]	6.2 [2.2, 10.4]
Alignment	0.787 [0.739, 0.831]	42.1 [32.4, 51.0]	5.0 [4.7, 5.3]	42.8 [32.4, 51.1]
Closure	0.816 [0.774, 0.854]	42.8 [33.8, 51.6]	5.0 [4.8, 5.4]	42.8 [33.8, 51.3]

B TRAINING DATA CONSTRUCTION: ADDITIONAL DETAILS

Overview. This appendix describes the construction and validation of the shared data pool for E-CLOSURE and the fine-tuned baselines. Each family contains three generation anchors (Gold, Evidence-CF, and Question-CF) and their associated intervention pairs, as illustrated in Table 18. We retain 2,852 families; Table 19 reports retention by dataset. A deterministic, dataset-stratified $90/10$ split based on SHA-256 ranking yields 2,566 training families and 286 development families. Section B.6 details packing, splitting, and quality auditing.

$$\underbrace{\text{Sample} \rightarrow \text{Screen} \rightarrow y^\star \rightarrow \text{Interventions} \rightarrow \text{Pack} \rightarrow \text{Split} \rightarrow \text{Audit}}_{\text{training data construction}}$$

Table 18: **Example of training data construction.** The original response and two counterfactual responses form three generation anchors. The evidence intervention changes a supporting fact, while the question intervention changes the requested target.

View	Input	Warranted trace	Answer
Original	$(q, \mathcal{D})$	s_1: N was written by Ada. s_2: Ada was born in Paris, France. s_3: The requested country is France.	France
Evidence edit	$(q, \widetilde{\mathcal{D}}_e)$	s_1: N was written by Ada. s_2: Ada was born in Berlin, Germany. s_3: The requested country is Germany.	Germany
Question edit	$(\widetilde{q}, \mathcal{D})$	s_1: N was written by Ada. s_2: Ada was born in Paris, France. s_3: The requested city is Paris.	Paris

Constructors and verifiers. All construction, screening, and verification stages use GPT-5.6 Sol (OpenAI, 2026a) under separate frozen guidelines. Screens assess whether proposed inputs and targets are admissible without seeing generated traces. Each trace verifier returns the required semantic judgments as separate structured fields in a single call. The guidelines define each criterion independently, restrict the information that may justify it, and prohibit inferring one judgment from another. We conduct a final human audit of the constructed data, as detailed in Section B.6. The frozen guidelines are available in the released code at `construction/`. Sections B.1–B.6 detail the pipeline from source sampling to packing and human audit.

B.1 SAMPLING

We sample without replacement from the official training splits, using seed 42 for subsampling, and deduplicate questions by source identifier. The sampled pool contains 5,000 questions from 2WikiMultiHopQA (Ho et al., 2020), 15,000 from MuSiQue (Trivedi et al., 2022), and 10,000 from HotpotQA (Yang et al., 2018) (Table 19, *Sampled* column). We sample and process more MuSiQue and HotpotQA questions to offset their higher screening exclusion rates and improve balance in the retained data, consistent with the diagnostic screening results (Section A.3.1). Of these, 23,000 unique questions enter construction (*Processed* column): 2,000 from 2WikiMultiHopQA, 15,000 from MuSiQue, and 6,000 from HotpotQA. Eligibility is assessed only during source screening, which accounts for most exclusions in MuSiQue and HotpotQA (*Processed* to *Screened*).

B.2 SOURCE SCREENING

Seed screen. Source screening checks question validity and evidence sufficiency before trace or counterfactual construction, preventing later generation from masking defects in the original source. Let $z = (q, \mathcal{D}, a^\star)$ denote a processed source, where $a^\star$ is the reference answer. The SCREEN-SEED stage receives only z and selects a support set $E \subseteq \mathcal{D}$. Let $W(q, \mathcal{D}, a)$ indicate whether $\mathcal{D}$ warrants a as an answer to q, and $U(q, \mathcal{D}, a)$ whether it warrants a uniquely. We retain z only if

$$|E| \geq 2 \wedge W(q, E, a^\star) \wedge U(q, \mathcal{D}, a^\star) \wedge \neg W(q, \mathcal{D} \setminus E, a^\star).$$

The selected support set must contain at least two passages and establish the reference answer. The full context must warrant that answer uniquely, and the remaining passages must not independently

Table 19: **Training data construction.** Counts are unique source questions. Processed denotes entry into construction; Screened denotes passing the initial screen. The $y^\star$, Evidence, and Question columns require verified original responses, evidence interventions, and question interventions, respectively. Packed counts sources retained after semantic reconstruction and deterministic auditing.

Dataset	Sampled	Processed	Screened	$y^\star$	Evidence	Question	Packed
2WikiMultiHopQA	5,000	2,000	1,841	1,841	1,622	1,622	1,276
MuSiQue	15,000	15,000	2,827	2,827	1,685	946	559
HotpotQA	10,000	6,000	2,608	2,608	1,862	1,189	1,017
Total	30,000	23,000	7,276	7,276	5,169	3,757	2,852

establish it. We also reject sources with malformed or underdetermined questions, broken entity links, unsupported reference answers, or impractically long answers.

Retention. Source screening retains 7,276 of the 23,000 processed questions (31.6%): $1{,}841/2{,}000$ for 2WikiMultiHopQA, $2{,}827/15{,}000$ for MuSiQue, and $2{,}608/6{,}000$ for HotpotQA (Table 19, *Screened* column).

B.3 Target Response Construction

Target construction. Given $(q, \mathcal{D}, a^\star, E)$, the constructor generates a candidate response $\widehat{y} = (\widehat{r}, \widehat{a})$ and a completion indicator $Q^\star$. The trace has no fixed length. Each step states one locally supported substantive claim, and the trace pursues the target specified by q, making all required connections between dependent steps explicit. All evidence-dependent premises must be grounded in E, although the trace need not use every passage in E; no required premise may rely solely on $\mathcal{D} \setminus E$. The submitted answer $\widehat{a}$ must match $a^\star$ under the fixed normalization.

Target verification. The reference-blind VERIFY-TARGET verifier receives $(q, \mathcal{D}, E, \widehat{r}, \widehat{a})$, without access to $a^\star$, and returns four separately defined judgments: local soundness L, question–trace alignment A, trace–answer closure C, and evidence grounding R_E. Each judgment uses only its designated inputs: L uses $(q, \mathcal{D}, \widehat{r})$, A uses $(q, \widehat{r})$, and C uses $(\widehat{r}, \widehat{a})$. For closure, neither the question nor the evidence may supply missing trace premises. R_E requires a complete step-level support map and rejects any necessary premise supported only outside E. Semantic verification succeeds when all four criteria hold:

$$V^\star = L(q, \mathcal{D}, \widehat{r}) \wedge A(q, \widehat{r}) \wedge C(\widehat{r}, \widehat{a}) \wedge R_E(\widehat{r}).$$

A deterministic gate $H^\star$ checks normalized answer matching, support-identifier membership, schema validity, and structural well-formedness. We retain only candidates satisfying $V^\star \wedge Q^\star \wedge H^\star$, denoting each accepted response by $y^\star$.

Retention. All 7,276 screened sources yield an accepted target response: 1,841 from 2WikiMultiHopQA, 2,827 from MuSiQue, and 2,608 from HotpotQA (Table 19, $y^\star$ column).

B.4 Evidence Intervention

An evidence intervention changes one answer-determining claim while keeping the question fixed. The trace must first diverge where the claim is used and establish the new answer.

Evidence construction. Given $(q, \mathcal{D}, E, y^\star)$, the constructor replaces one passage $e \in E$ with a counterfactual passage $\widetilde{e}$, yielding

$$\widetilde{\mathcal{D}}_e = (\mathcal{D} \setminus \{e\}) \cup \{\widetilde{e}\}, \qquad \widetilde{y}_e = (\widetilde{r}_e, \widetilde{a}_e), \qquad \widetilde{a}_e \not\approx a^\star.$$

The edit changes exactly one semantic claim while preserving the passage identifier, title, and unrelated content. Any coreferential repetitions of that claim are updated consistently. The replacement value must preserve the answer type and yield a uniquely warranted answer.

Let t_e denote the first trace position affected by the edit. The constructor preserves the maximal prefix, with $\widetilde{s}_{j,e} = s_j^\star$ for all $j < t_e$, and rebuilds the trace from t_e using the edited context to establish the new answer. A completion indicator Q^e records whether construction succeeds.

Evidence screen. To assess the intervention without relying on the generated traces, SCREEN-EVIDENCE receives only $(q, \mathcal{D}, \widetilde{\mathcal{D}}_e, a^\star, \widetilde{a}_e)$. It checks that the edit is localized and coherent and that all unedited passages remain unchanged. The edited context must uniquely warrant the new answer and no longer warrant the original answer; the original context must not warrant the new answer:

$$U(q, \widetilde{\mathcal{D}}_e, \widetilde{a}_e) \wedge W(q, \widetilde{\mathcal{D}}_e, \widetilde{a}_e) \wedge \neg W(q, \widetilde{\mathcal{D}}_e, a^\star) \wedge \neg W(q, \mathcal{D}, \widetilde{a}_e).$$

Evidence verification. In a single structured call, VERIFY-EVIDENCE checks the edited response's local soundness, question–trace alignment, and trace–answer closure, together with step and closure switching. Let $J_{\mathcal{D}'}(s)$ indicate whether step s is warranted under context $\mathcal{D}'$. The shared prefix must remain valid under both contexts. At the first affected position t_e, each step must be warranted under its own context but not the other:

$$\forall j < t_e : \; J_{\mathcal{D}}(s^\star_j) = J_{\widetilde{\mathcal{D}}_e}(\widetilde{s}_{j,e}) = 1, \quad J_{\mathcal{D}}(s^\star_{t_e}) = J_{\widetilde{\mathcal{D}}_e}(\widetilde{s}_{t_e,e}) = 1, \quad J_{\mathcal{D}}(\widetilde{s}_{t_e,e}) = J_{\widetilde{\mathcal{D}}_e}(s^\star_{t_e}) = 0.$$

The edited response must satisfy

$$L(q, \widetilde{\mathcal{D}}_e, \widetilde{r}_e) \wedge A(q, \widetilde{r}_e) \wedge C(\widetilde{r}_e, \widetilde{a}_e),$$

while closure must hold for each matched trace–answer pair and fail for both crossed pairs:

$$C(r^\star, a^\star) \wedge C(\widetilde{r}_e, \widetilde{a}_e) \wedge \neg C(r^\star, \widetilde{a}_e) \wedge \neg C(\widetilde{r}_e, a^\star).$$

Let V^e denote the conjunction of the screening and verification criteria. A deterministic gate H^e checks schema validity, single-passage replacement, preservation of unedited content, exact prefix identity, first divergence at t_e, and normalized answer inequality. We retain only interventions satisfying $V^e \wedge Q^e \wedge H^e$. Matched evidence–response pairs serve as generation anchors; crossed trace–answer pairs are used only for margin scoring.

Retention. Of the 7,276 accepted targets, 5,169 (71.0%) yield a verified evidence intervention: 1,622 from 2WikiMultiHopQA, 1,685 from MuSiQue, and 1,862 from HotpotQA (Table 19, *Evidence* column).

B.5 QUESTION INTERVENTION

A question intervention changes the requested target while keeping the evidence fixed. The edited question must induce a distinct reasoning path, with a locally sound trace that aligns with the new target and establishes the new answer.

Question construction. For each source with an accepted evidence intervention, the constructor keeps $\mathcal{D}$ fixed and changes one target-determining component of q. It produces an edited question $\widetilde{q}$ and response $\widetilde{y}_q = (\widetilde{r}_q, \widetilde{a}_q)$, where $\widetilde{a}_q \not\simeq a^\star$. The edited question must be natural and unambiguous, require multiple reasoning steps, and avoid revealing the answer while preserving unrelated constraints. No prefix constraint is imposed because the new target may change the reasoning path from the first step. A completion indicator Q^q records whether construction succeeds.

Question screen. To assess the target change without relying on generated traces, SCREEN-QUESTION receives only $(q, \widetilde{q}, \mathcal{D}, a^\star, \widetilde{a}_q)$. It checks that the evidence remains unchanged, the requested target changes, and the edited question has a uniquely warranted answer. Neither answer may satisfy the other question:

$$U(\widetilde{q}, \mathcal{D}, \widetilde{a}_q) = 1, \qquad W(\widetilde{q}, \mathcal{D}, \widetilde{a}_q) = 1, \qquad W(\widetilde{q}, \mathcal{D}, a^\star) = 0, \qquad W(q, \mathcal{D}, \widetilde{a}_q) = 0.$$

Degenerate, tautological, ambiguous, single-step, or answer-leaking questions are rejected.

Question verification. In one structured call, VERIFY-QUESTION separately assesses the edited response's local soundness, question–trace alignment, and trace–answer closure, together with four T-switching judgments. Each alignment judgment uses only its question–trace pair; merely mentioning a relevant entity or relation does not establish alignment. The edited response must satisfy

$$L(\widetilde{q}, \mathcal{D}, \widetilde{r}_q) = 1, \qquad A(\widetilde{q}, \widetilde{r}_q) = 1, \qquad C(\widetilde{r}_q, \widetilde{a}_q) = 1.$$

Target switching requires alignment for both matched pairs and misalignment for both crossed pairs:

$$A(q, r^\star) = 1, \qquad A(\widetilde{q}, \widetilde{r}_q) = 1, \qquad A(q, \widetilde{r}_q) = 0, \qquad A(\widetilde{q}, r^\star) = 0.$$

Let V^q denote the conjunction of the screening and verification criteria. A deterministic gate H^q checks schema validity, identity of the shared evidence, normalized answer inequality, absence of literal answer leakage, support-identifier validity, and structural well-formedness. We retain only candidates satisfying $V^q \wedge Q^q \wedge H^q$. The matched question–response pair serves as a generation anchor, while crossed question–trace pairs are used only to score the two directional T margins. No closure-switching pair is constructed because the question and trace change together.

Retention. Of the $5{,}169$ sources with an accepted evidence intervention, $3{,}757$ (72.7%) also yield a verified question intervention: $1{,}622$ from 2WikiMultiHopQA, 946 from MuSiQue, and $1{,}189$ from HotpotQA (Table 19, *Question* column).

B.6 Packing, Splitting, and Audit

Acceptance and packing. Acceptance requires semantic approval V, successful construction Q, and deterministic checks H. For an intervention candidate i of type $x \in \{e, q\}$, we define $V_i^x = \mathrm{Screen}_i^x \wedge \mathrm{Verify}_i^x$ and $B_i^x = V_i^x \wedge Q_i^x \wedge H_i^x$. The target response follows the analogous rule $B^\star = V^\star \wedge Q^\star \wedge H^\star$. In both cases, $B = 1$ denotes an accepted candidate.

We pack a source and its views into a family only if it passes the seed screen and has an accepted target response, at least one accepted evidence intervention, and exactly one accepted question intervention. Deterministic checks cover schema validity, identifier consistency, normalized answer relations, and structural well-formedness. Evidence interventions additionally require exactly one edited passage, byte-identical unedited passages, exact prefix preservation, and first divergence at t_e. Question interventions require an unchanged context, a changed question, valid support identifiers, and no literal answer leakage. Semantic properties, including local soundness, alignment, closure, intervention isolation, and nondegeneracy, are assessed exclusively through V.

Each family stores the complete original and edited contexts, the Gold, Evidence-CF, and Question-CF views, and their proof graphs, traces, answers, support identifiers, intervention metadata, semantic judgments, and construction provenance. Only matched views serve as generation targets. Crossed evidence–step, question–trace, and trace–answer pairs are formed within each family solely for the corresponding S, T, and C margins.

Retention. Question verification yields $3{,}757$ candidate families. Removing one MuSiQue family with a duplicate original question leaves $3{,}756$. Reapplying the frozen construction and verification criteria retains $2{,}852$ final families: $1{,}276$ from 2WikiMultiHopQA, 559 from MuSiQue, and $1{,}017$ from HotpotQA (Table 19, *Packed* column).

Splitting. We split at the family level so that each source and all its derived views remain in the same partition. Within each dataset, we sort families by the SHA-256 hash of a fixed salt, dataset name, and source identifier. The first 10%, rounded to the nearest integer, form the development set; the remainder form the training set. This yields $2{,}566$ training families and 286 development families, with no source or derived-view overlap.

Human audit. After splitting, we draw a stratified random sample of 300 families using seed 42: 100 per dataset, comprising 90 training and 10 development families. Two annotators independently inspect each family's Gold response, one uniformly sampled Evidence-CF view, and its Question-CF view, without seeing automatic judgments, constructor rationales, or each other's decisions. The audit follows frozen guidelines shared by both annotators. They assess whether the Gold response is correct, supported, aligned, and closed; whether the evidence intervention is isolated and induces the required trace and answer changes; and whether the question intervention is valid and induces the required trace switch. A family passes only if all three inspected views pass.

Table 20 reports the adjudicated results. Dataset-level 95% confidence intervals use the Wilson method. Corpus-wide pass rates are weighted by each dataset's share of the final corpus, with 95% percentile intervals from $2{,}000$ dataset-stratified bootstrap resamples using seed 42. We resample families within each dataset, keeping their three view labels together and corpus weights fixed. Because pass labels are near ceiling, we report raw agreement alongside Gwet's AC_1. Before adjudication, annotators agree on $292/300$ family decisions (97.3%, $AC_1 = 0.972$); a third annotator resolves the eight disagreements under the same frozen guidelines. After adjudication, $290/300$ families pass (96.7% unweighted; 97.3% corpus-weighted).

Table 20: **Human audit results.** Adjudicated pass rates (%) with 95% confidence intervals. A family passes only if all three audited views pass.

Dataset	Audited	Gold	Evidence-CF	Question-CF	Family
2WikiMultiHopQA	100	100.0 [96.3,100.0]	99.0 [94.6,99.8]	100.0 [96.3,100.0]	99.0 [94.6,99.8]
MuSiQue	100	98.0 [93.0,99.4]	95.0 [88.8,97.8]	99.0 [94.6,99.8]	94.0 [87.5,97.2]
HotpotQA	100	99.0 [94.6,99.8]	97.0 [91.5,99.0]	100.0 [96.3,100.0]	97.0 [91.5,99.0]
Corpus-weighted	300	99.3 [98.3,100.0]	97.5 [95.6,99.0]	99.8 [99.4,100.0]	**97.3 [95.3,98.9]**

C Edge Closure Learning: Additional Details

This section explains how **E-Closure** converts validated intervention data into three edge-specific training signals. An admissible response has a locally sound trace that aligns with the question and establishes the submitted answer. Evidence and question interventions are drawn from $\mathcal{I}_e(z)$ and $\mathcal{I}_q(z)$, respectively. All targets and interventions pass the semantic, completion, and deterministic checks described in Appendix B, which also provides construction details and retention rates.

C.1 Training Views and Validity

Original view. A training source contains a question q, evidence $\mathcal{D}$, and an admissible target response $y^\star = (r^\star, a^\star)$ whose submitted answer matches the reference. This view anchors generation under the original input.

Evidence intervention. We replace a supporting passage $e \in E_{\mathrm{dep}}(r^\star)$, whose removal would leave a necessary premise of $r^\star$ unsupported, to obtain the corresponding edited evidence context $\widetilde{\mathcal{D}}_e$ and an admissible response $\widetilde{y}_e = (\widetilde{r}_e, \widetilde{a}_e)$. The traces share a prefix and first diverge at $t(e)$, the earliest step whose warrant changes under the edit. Validation requires each affected step to be warranted under its own evidence context but not the other, and each trace to establish its own answer but not the paired alternative. This intervention supplies the S edge from evidence to the first affected step and the C edge from trace to answer.

Question intervention. With $\mathcal{D}$ fixed, a localized target edit produces $\widetilde{q}$ and an admissible response $\widetilde{y}_q = (\widetilde{r}_q, \widetilde{a}_q)$. The edited trace must align with $\widetilde{q}$ but not q, while $r^\star$ must align with q but not $\widetilde{q}$. This intervention supplies the T edge, linking the requested target to the complete reasoning trace.

The resulting original and counterfactual pairs are

$$S:\ (\mathcal{D}, s^\star_{t(e)}) \leftrightarrow (\widetilde{\mathcal{D}}_e, \widetilde{s}_{t(e),e}) \qquad T:\ (q, r^\star) \leftrightarrow (\widetilde{q}, \widetilde{r}_q) \qquad C:\ (r^\star, a^\star) \leftrightarrow (\widetilde{r}_e, \widetilde{a}_e).$$

The notation omits inputs held fixed within each edge: q and the shared prefix for S, $\mathcal{D}$ for T, and the original $(q, \mathcal{D})$ for C. The evidence intervention supplies both S and C, while the question intervention supplies T; no separate response is constructed for C.

C.2 Edge-Specific Scores

Let $\pi_\theta(z \mid u)$ denote the probability of output sequence z given input u under the model with parameters θ. We instantiate the generic score $\ell_\theta(z \mid u)$ for three outputs: the first affected step, the complete trace, and the submitted answer:

$$F^S_{\theta,e}(s \mid \mathcal{D}') = \frac{1}{|s|} \log \pi_\theta\big(s \mid q, \mathcal{D}', s^\star_{<t(e)}\big),$$

$$F^T_\theta(r \mid q') = \frac{1}{|r|} \log \pi_\theta\big(r \mid q', \mathcal{D}\big),$$

$$F^C_\theta(a \mid r) = \frac{1}{|a|} \log \pi_\theta\big(a \mid q, \mathcal{D}, r\big).$$

Here, $|\cdot|$ denotes token count. Each score averages the output's autoregressive token log probabilities, accounting for differences in candidate length. All scores are computed with teacher forcing, without sampling.

Evidence to step. Both candidate steps are scored after the same validated prefix, so F^S isolates the first position where the evidence edit should change the trace. Restricting the score to this step separates the local dependency from later differences in the regenerated continuation.

Question to trace. The score F^T covers the complete trace without the final answer, giving the reasoning its own preference signal. Under q, the model should prefer $r^\star$ to $\widetilde{r}_q$; under $\widetilde{q}$, this preference should reverse.

Trace to answer. In F^C, q and $\mathcal{D}$ remain fixed while the trace changes. The model should prefer $a^\star$ after $r^\star$ and $\widetilde{a}_e$ after $\widetilde{r}_e$. Holding the other inputs fixed ties the required preference reversal to the change in trace. The edited trace under the original context is used only for answer scoring; $(q, \mathcal{D}, \widetilde{r}_e, \widetilde{a}_e)$ is never a generation anchor.

For each edge x, write F^x_θ for its score and $(u^{\mathrm{o}}_x, z^{\mathrm{o}}_x) \leftrightarrow (u^{\mathrm{c}}_x, z^{\mathrm{c}}_x)$ for the original and counterfactual pair. The two directional margins are

$$\Delta^{x,\mathrm{o}}_\theta = F^x_\theta(z^{\mathrm{o}}_x \mid u^{\mathrm{o}}_x) - F^x_\theta(z^{\mathrm{c}}_x \mid u^{\mathrm{o}}_x), \qquad \Delta^{x,\mathrm{c}}_\theta = F^x_\theta(z^{\mathrm{c}}_x \mid u^{\mathrm{c}}_x) - F^x_\theta(z^{\mathrm{o}}_x \mid u^{\mathrm{c}}_x).$$

Both margins must meet the positive target γ_x, so a fixed preference across the two inputs cannot satisfy both constraints. For each edge, $\gamma_x > 0$ sets the required separation, while $\beta_x > 0$ controls the smooth hinge transition.

C.3 Complete Objective and Design Rationale

Generation anchors. Each packed source z contains one accepted evidence intervention paired with the source's unique question intervention. The original response and the two counterfactual responses form three generation anchors, each paired with the input that warrants it.

Let $S^{\mathrm{rsp}}_\theta(y \mid q', \mathcal{D}')$ denote the mean log likelihood over the trace and answer tokens of a complete response, excluding fixed formatting delimiters. The generation loss averages over the three anchors:

$$\mathcal{L}_{\mathrm{gen}} = -\frac{1}{3}\left[S^{\mathrm{rsp}}_\theta(y^\star \mid q, \mathcal{D}) + S^{\mathrm{rsp}}_\theta(\widetilde{y}_e \mid q, \widetilde{\mathcal{D}}_e) + S^{\mathrm{rsp}}_\theta(\widetilde{y}_q \mid \widetilde{q}, \mathcal{D})\right].$$

Bidirectional supervision. The anchors directly encourage high likelihood for warranted responses. We complement them with bidirectional margins that supervise how output preferences should change under each intervention. Let $h_\beta(t) = \beta \log(1 + \exp(t/\beta))$ denote the smooth hinge, with $\mathcal{X} = \{S, T, C\}$ and $\mathcal{V} = \{\mathrm{o}, \mathrm{c}\}$. The complete per-source objective is

$$\mathcal{L}_{\text{E-Closure}} = \mathcal{L}_{\mathrm{gen}} + \lambda\, \mathbb{E}_{(x,v)\sim \mathrm{Unif}(\mathcal{X}\times\mathcal{V})}\left[h_{\beta_x}(\gamma_x - \Delta^{x,v}_\theta)\right].$$

The expectation averages uniformly over all six combinations of edge and direction, so λ controls the total margin weight relative to generation. The batch loss averages over sampled sources.

Design rationale. Let $\bar{\ell}_\theta$ denote mean token log-likelihood. For a fixed tokenization, split $y = (r, b)$, where b contains the answer delimiter, answer, and end-of-turn token. Then

$$S^{\mathrm{rsp}}_\theta(y \mid q, \mathcal{D}) = \frac{|r|}{|y|}\bar{\ell}_\theta(r \mid q, \mathcal{D}) + \frac{|b|}{|y|}\bar{\ell}_\theta(b \mid q, \mathcal{D}, r), \qquad |y| = |r| + |b|.$$

When $|r| \gg |b|$, trace changes can dominate response preferences without changing answer preferences. Separate T and C margins normalize their own scored spans, giving trace selection and answer closure dedicated supervision. The S margin similarly isolates the first step affected by the evidence intervention.

Training and inference. All terms are computed with teacher forcing and require no reference model. Additional views are used only during training. At inference, E-Closure retains the standard mapping $(q, \mathcal{D}) \mapsto (r, a)$, with no auxiliary verifier, additional model calls, or decoding changes.

Default hyperparameters. We use $\lambda = 1$, $\gamma_x = 0.5$, and $\beta_x = 0.1$ for all dependencies $x \in \{S, T, C\}$, with the same defaults across datasets and backbones. With token-normalized scores and an averaged margin loss, $\lambda = 1$ gives generation and margin supervision unit coefficients. The target $\gamma_x = 0.5$ encourages a positive preference gap in both directions, while $\beta_x = 0.1$ smooths the hinge transition on a scale smaller than this gap. For controlled comparisons, RM and all margin-based ablations share these values, with margin coefficients normalized to sum to λ. SFT and CF-SFT optimize generation alone, so the margin parameters do not apply.

D EXPERIMENTS: ADDITIONAL DETAILS

This appendix defines the evaluation protocol and validates the automatic judge used for behavioral outcomes. It then specifies the comparison objectives and training implementation before reporting full results, mechanism analyses, and robustness checks.

D.1 EVALUATOR AND VALIDATION

Evaluator. Human annotation across all methods, backbones, and training seeds is time-consuming. LLM-based judges offer a scalable alternative, but their reliability requires validation (Gu et al., 2024; Tan et al., 2025; Chu et al., 2025). We use GPT-5.6 Sol (OpenAI, 2026a) (`gpt-5.6-sol`) with `high` reasoning effort for all experimental results and evaluate `low` and `medium` as sensitivity conditions. We compare all three configurations with the final human adjudications of the $2{,}598$ diagnostic responses. Tables 21 and 22 report validation results for `high` and comparisons across reasoning efforts, respectively.

Judgment definitions. All judgments follow the frozen diagnostic guidelines (Section A.3). Answer correctness compares the submitted answer with the reference, allowing semantic equivalence. Local soundness L requires every substantive step to be warranted; alignment A requires the trace to pursue the requested target; and closure C requires it to establish the submitted answer. Reference completeness K asks whether the trace's stated premises suffice to establish the reference answer, even if the final inference remains implicit. The guidelines specify the information that may justify each judgment:

$$\mathcal{V}_L = (q, \mathcal{D}, r), \quad \mathcal{V}_A = (q, r), \quad \mathcal{V}_C = (r, a), \quad \mathcal{V}_K = (q, r, a^\star), \quad \mathcal{V}_{\mathrm{Acc}} = (q, a, a^\star).$$

For C and K, annotators and the evaluator treat the trace's claims as premises without reassessing their evidence support. Facts appearing only in the evidence cannot fill gaps in the trace. Answer correctness is assessed separately and does not determine alignment or closure.

Validation set and metrics. We evaluate all three configurations on the same $2{,}598$ diagnostic responses to 866 questions across three datasets and three backbones. Records are matched by question identifier and backbone, with identical inputs and responses across configurations. We compare predictions against the final human adjudications. All configurations yield complete binary labels without exclusions or imputation, and stored L and G labels match their recomputed values. The `low`, `medium`, and `high` configurations label 1, 2, and 1 responses, respectively, as reliable but incorrect; these cases remain included.

For each criterion, we report raw agreement, failure F1 ($F_{1,0}$), balanced accuracy, Cohen's κ, and the signed failure-rate difference:

$$\Delta p_0(x) = \Pr_{\mathrm{LLM}}(\widehat{x} = 0) - \Pr_{\mathrm{human}}(x = 0), \qquad x \in \{\mathrm{Acc}, L, A, C, K, G\}.$$

Failure F1 treats label 0 as positive; balanced accuracy averages recall across passing and failing labels. Positive Δp_0 indicates a higher automated failure rate. TRR and LGG differences use the same automated-minus-human convention, with all rate differences expressed in percentage points.

Validation uncertainty. We estimate 95% percentile confidence intervals using $2{,}000$ dataset-stratified question-level bootstrap resamples with seed 42. Each sampled question retains its three responses and their paired human and evaluator labels. Shared draws are used across criteria, configurations, and paired comparisons. Pooled validation estimates weight responses equally, while dataset-specific estimates use the corresponding stratum.

Agreement with human adjudication. Table 21 reports results for `high`, the configuration used throughout the experiments. Raw agreement ranges from 97.31% to 99.73% across the five directly evaluated criteria. Failure F1 reaches 0.993 for correctness, 0.937 for closure, and 0.972 for completeness. For local soundness and alignment, failure F1 is 0.938 and 0.971, respectively, with automated failure rates 2.16 and 0.12 percentage points above human estimates. Derived global sufficiency achieves 99.08% agreement and 0.962 failure F1. Overall, raw agreement is high, with the largest discrepancies concentrated in local soundness. All experimental methods are evaluated with the same evaluator, configuration, and frozen guidelines, providing a consistent basis for comparison. We further assess agreement with human judgments through a separate audit of the main experimental results (Appendix D.5.2).

Table 21: **Evaluator agreement with human adjudication.** Results use `high` on all 2,598 diagnostic responses. n_0 counts human failure labels. Agreement and balanced accuracy are percentages; $F_{1,0}$ and κ are unitless. Δp_0 is the automated minus human failure rate in percentage points. Brackets show 95% confidence intervals.

Criterion	n_0	**Agreement**	$F_{1,0}$	**BAcc.**	κ	Δp_0
Correctness	484	99.73 [99.50, 99.92]	0.993 [0.987, 0.998]	99.83 [99.70, 99.95]	0.991 [0.984, 0.997]	+0.27 [+0.08, +0.50]
Local soundness L	539	97.31 [96.57, 97.96]	0.938 [0.922, 0.953]	97.82 [97.18, 98.41]	0.921 [0.900, 0.940]	+2.16 [+1.46, +2.89]
Alignment A	190	99.58 [99.31, 99.81]	0.971 [0.953, 0.986]	98.80 [97.72, 99.65]	0.969 [0.949, 0.985]	+0.12 [-0.12, +0.38]
Closure C	134	99.35 [99.00, 99.65]	0.937 [0.904, 0.965]	97.19 [95.27, 98.85]	0.934 [0.899, 0.963]	+0.12 [-0.23, +0.46]
Completeness K	285	99.38 [99.00, 99.69]	0.972 [0.955, 0.986]	98.42 [97.21, 99.36]	0.968 [0.949, 0.984]	0.00 [-0.35, +0.35]
Global sufficiency G	313	99.08 [98.65, 99.42]	0.962 [0.944, 0.977]	98.37 [97.31, 99.24]	0.957 [0.937, 0.974]	+0.31 [-0.08, +0.69]

Sensitivity to reasoning effort. Table 22 compares `low`, `medium`, and `high` on the same responses. The `high` configuration achieves the highest mean agreement (99.07%) and mean failure F1 (0.962) across Acc, L, A, C, K, excluding the derived label G. Relative to `low`, mean agreement increases by 1.19 points (95% CI: [0.90, 1.50]), and mean failure F1 increases by 0.066 ([0.050, 0.082]). Across configurations, TRR is 1.54–2.23 points below the human estimate, while LGG rate differences range from -0.31 to $+0.12$ points.

Table 22: **Sensitivity to evaluator reasoning effort.** Mean agreement (%), mean failure F1, and maximum absolute failure-rate difference summarize Acc, L, A, C, K. Rate differences are automated minus human in percentage points. Brackets show 95% paired bootstrap confidence intervals.

Effort	**Mean agr.**	**Mean** $F_{1,0}$	$\max \lvert\Delta p_0\rvert$	ΔTRR	ΔLGG
`low`	97.88	0.897	1.54	-1.54 [-2.35, -0.81]	0.00 [-0.58, +0.58]
`medium`	97.96	0.902	2.12	-2.23 [-3.08, -1.46]	+0.12 [-0.46, +0.69]
`high`	99.07	0.962	2.16	-1.85 [-2.54, -1.23]	-0.31 [-0.73, +0.15]

Agreement on derived metrics. Under `high`, TRR is 71.82%, compared with 73.67% under human adjudication, a difference of -1.85 points (95% CI: $[-2.54, -1.23]$). LGG rates are closer: 5.27% versus 5.58%, a difference of -0.31 points ($[-0.73, 0.15]$), reflecting similar downward shifts in LSR and TRR. At the response level, `high` identifies 126 of the 145 human-labeled LGG cases and assigns 11 additional cases, yielding precision of 91.97%, recall of 86.90%, and F1 of 0.894. Corresponding LGG F1 values are 0.848 for `low` and 0.853 for `medium`.

Dataset-specific differences. Table 23 breaks down the metric differences by dataset. TRR differences are -0.12 points on 2WikiMultiHopQA, -1.70 on HotpotQA, and -3.92 on MuSiQue. LGG differences are $+0.12$, $+0.20$, and -1.44 points, respectively. These results characterize agreement on the diagnostic responses; the separate human audit of the main experiment assesses annotations of responses generated by the evaluated methods.

D.2 Experimental Evaluation Protocol

Evaluation set and generation. All methods are evaluated on the same 866 retained diagnostic questions: 278 from 2WikiMultiHopQA, 255 from MuSiQue, and 333 from HotpotQA. Each adapted checkpoint generates one response per question using the diagnostic study's structured CoT prompt. Backbone results use the fixed diagnostic generations. Generation uses greedy decoding with at most 1,024 new tokens and an 8,192-token sequence limit; explicit thinking mode is disabled where supported. All responses are assessed using the evaluator and `high` configuration described in Appendix D.1.

Table 23: **Dataset-specific metric differences for `high`.** n counts responses. Differences are automated minus human in percentage points, with 95% paired bootstrap confidence intervals. The pooled row weights responses equally.

Dataset	n	ΔTRR	ΔLGG
2WikiMultiHopQA	834	-0.12 [-0.48, 0.00]	+0.12 [0.00, +0.36]
HotpotQA	999	-1.70 [-2.80, -0.70]	+0.20 [-0.40, +0.90]
MuSiQue	765	-3.92 [-5.75, -2.35]	-1.44 [-2.75, -0.26]
All (pooled)	2,598	-1.85 [-2.54, -1.23]	-0.31 [-0.73, +0.15]

Metrics. We recompute $\widehat{L}$ from the substantive-step labels and derive $\widehat{G} = \widehat{A} \wedge \widehat{C}$. Across n responses, trace reliability and LGG rates are

$$\widehat{\mathrm{TRR}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}_i \widehat{G}_i, \qquad \widehat{\mathrm{LGG}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}_i (1 - \widehat{G}_i).$$

Answer accuracy is the mean correctness label; LSR and GSR are the mean $\widehat{L}$ and $\widehat{G}$ labels, respectively. Alignment and closure rates are computed analogously. We separately report a supplementary deterministic score based on normalized, bidirectional whole-token-span matching, which differs from both semantic evaluation and strict exact match (Appendix D.5). All rates are reported as percentages.

Aggregation. For adapted methods, we first compute each metric within each dataset–backbone–seed combination, then average over the three training seeds, $\{42, 43, 44\}$, within each dataset–backbone setting. Aggregate main results assign equal weight to the nine settings. Backbone results are aggregated equally across the same nine settings using their fixed generations. Thus, each dataset and backbone contributes equally to the aggregate main results, whereas the pooled evaluator-validation estimates in Appendix D.1 weight individual responses equally.

Uncertainty estimation for aggregate main results. We compute 95% paired hierarchical percentile confidence intervals from 5,000 bootstrap resamples with seed 2027. We resample backbones and, for each sampled backbone, resample questions within each fixed dataset and training seeds. Draws are shared across methods, and each training-seed draw is shared across datasets. Backbone results retain their fixed generations without training-seed resampling. Each replicate applies the same equal-weight aggregation as the point estimates, and the 2.5th and 97.5th percentiles form the confidence interval. The separate human audit, objective ablations, and dependency-intervention comparisons use the procedures specified in their respective appendix sections.

D.3 Comparison Objectives

Common setup. All adapted methods, including SFT, Counterfactual SFT (CF-SFT), Full-Response Margin (RM), E-Closure, and its ablations, start independently from the same pretrained backbone. They share the packed source partitions, source sampler, and context serialization. Methods using counterfactual supervision also share the accepted interventions and generation anchors. For each sampled source, one evidence intervention is selected uniformly and paired with the source's unique question intervention. All likelihoods are computed with teacher forcing, and minibatch averaging is implicit. For a complete response $y = (r, a)$ conditioned on input u, we define the token-normalized score

$$S_\theta^{\mathrm{rsp}}(y \mid u) = \frac{1}{|y|} \log \pi_\theta(y \mid u).$$

We write $u : z^+ \succ z^-$ to indicate a preference for z^+ over z^- under input u. Superscripts o and c denote original and counterfactual conditioning views, respectively. Both margin-based objectives use the smooth hinge

$$h_\beta(z) = \beta \log(1 + \exp(z/\beta)), \qquad \beta > 0.$$

SFT: original-response supervision. Given the original question q and evidence $\mathcal{D}$, standard SFT trains on the validated response $y^\star$ as its sole generation anchor:

$$\mathcal{L}_{\mathrm{SFT}} = -S_\theta^{\mathrm{rsp}}(y^\star \mid q, \mathcal{D}).$$

CF-SFT, RM, E-CLOSURE, and all controlled ablations retain this anchor. SFT thus isolates original-response supervision before adding counterfactual anchors or preference margins.

Counterfactual SFT: matched generation anchors. CF-SFT trains on the original response under $(q, \mathcal{D})$, the evidence-intervention response under $(q, \widetilde{\mathcal{D}}_e)$, and the question-intervention response under $(\widetilde{q}, \mathcal{D})$. The objective averages their generation losses:

$$\mathcal{L}_{\text{CF-SFT}} \equiv \mathcal{L}_{\mathrm{gen}} = -\frac{1}{3}\Big[S_\theta^{\mathrm{rsp}}(y^\star \mid q, \mathcal{D}) + S_\theta^{\mathrm{rsp}}(\widetilde{y}_e \mid q, \widetilde{\mathcal{D}}_e) + S_\theta^{\mathrm{rsp}}(\widetilde{y}_q \mid \widetilde{q}, \mathcal{D})\Big].$$

Each response is trained only under an input that warrants it. RM, E-CLOSURE, and all controlled ablations retain these three generation anchors. CF-SFT thus isolates counterfactual generation supervision before adding preference margins.

Full-Response Margin: preferences over complete responses. RM retains the three generation anchors in $\mathcal{L}_{\mathrm{gen}}$ and adds bidirectional margins over complete responses $y = (r, a)$. Each input should favor its matched response over the paired alternative, yielding four comparisons:

View	Conditioning	Preference
$\mathsf{R}_e^{\mathrm{o}}$	$(q, \mathcal{D})$	$y^\star \succ \widetilde{y}_e$
$\mathsf{R}_e^{\mathrm{c}}$	$(q, \widetilde{\mathcal{D}}_e)$	$\widetilde{y}_e \succ y^\star$
$\mathsf{R}_q^{\mathrm{o}}$	$(q, \mathcal{D})$	$y^\star \succ \widetilde{y}_q$
$\mathsf{R}_q^{\mathrm{c}}$	$(\widetilde{q}, \mathcal{D})$	$\widetilde{y}_q \succ y^\star$

For each row, let $u_{i,v}$ denote the conditioning input and $(y^+_{i,v}, y^-_{i,v})$ the preferred and dispreferred responses, with $i \in \{e, q\}$ and $v \in \mathcal{V} = \{\mathrm{o}, \mathrm{c}\}$. The preference margin and its penalty are

$$\Delta_\theta^{\mathrm{rsp},i,v} = S_\theta^{\mathrm{rsp}}(y^+_{i,v} \mid u_{i,v}) - S_\theta^{\mathrm{rsp}}(y^-_{i,v} \mid u_{i,v}), \qquad \Psi_{i,v} = h_{\beta_{\mathrm{rsp}}}\Big(\gamma_{\mathrm{rsp}} - \Delta_\theta^{\mathrm{rsp},i,v}\Big).$$

The full objective is

$$\mathcal{L}_{\mathrm{RM}} = \mathcal{L}_{\mathrm{gen}} + \frac{\lambda}{6}\sum_{v \in \mathcal{V}} (2\Psi_{e,v} + \Psi_{q,v}).$$

The four terms receive weights in the ratio $2 : 2 : 1 : 1$, allocating two thirds of the total margin weight to evidence interventions and one third to question interventions. This matches the allocation to S, C and T, respectively, in E-CLOSURE, while scoring complete responses. RM uses the same target margin and hinge smoothing as E-CLOSURE, with $\gamma_{\mathrm{rsp}} = 0.5$ and $\beta_{\mathrm{rsp}} = 0.1$, and requires no reference model. With generation anchors and margin allocation held fixed, this baseline tests the added value of dependency-specific supervision beyond preferences over complete responses.

E-CLOSURE: preferences on individual dependencies. E-CLOSURE retains the same generation anchors and adds bidirectional margins on three dependent outputs: the first affected step for S, the complete reasoning trace for T, and the answer for C. Let

$$p_e = s^\star_{<t(e)} = \widetilde{s}_{<t(e),e}$$

denote the prefix shared by the original and evidence-edited traces. The six comparisons are:

View	Conditioning	Preference
S^{o}	$(q, \mathcal{D}, p_e)$	$s^\star_{t(e)} \succ \widetilde{s}_{t(e),e}$
S^{c}	$(q, \widetilde{\mathcal{D}}_e, p_e)$	$\widetilde{s}_{t(e),e} \succ s^\star_{t(e)}$
T^{o}	$(q, \mathcal{D})$	$r^\star \succ \widetilde{r}_q$
T^{c}	$(\widetilde{q}, \mathcal{D})$	$\widetilde{r}_q \succ r^\star$
C^{o}	$(q, \mathcal{D}, r^\star)$	$a^\star \succ \widetilde{a}_e$
C^{c}	$(q, \mathcal{D}, \widetilde{r}_e)$	$\widetilde{a}_e \succ a^\star$

The S views change the evidence while keeping the question and shared prefix fixed. The T views change the question while keeping the evidence fixed. The C views change the trace while keeping both the question and evidence fixed. In C^{c}, the edited trace $\widetilde{r}_e$ serves only as conditioning for answer scoring; it is never a generation target under the original evidence.

Let $\mathcal{X} = \{S, T, C\}$. For each row, $\Delta_\theta^{x,v}$ is the difference between the preferred and dispreferred edge scores, with penalty

$$\Phi_{x,v} = h_{\beta_x}(\gamma_x - \Delta_\theta^{x,v}) .$$

The complete objective is

$$\mathcal{L}_{\text{E-CLOSURE}} = \mathcal{L}_{\text{gen}} + \frac{\lambda}{6} \sum_{x \in \mathcal{X}} \sum_{v \in \mathcal{V}} \Phi_{x,v}.$$

Each dependency receives one third of the total margin weight, split equally between its two directions. Appendix C.2 provides the full score definitions and conditioning views.

Controlled ablations. All ablations retain the same three generation anchors, including both counterfactual responses, and the same training protocol as E-CLOSURE. They vary which edges and directions receive margin supervision. Each retained edge contributes one margin per retained direction, as summarized below.

Variant	Edges	Directions	Weight per term
Full	$\{S, T, C\}$	$\{\mathrm{o}, \mathrm{c}\}$	$\lambda/6$
$-S$	$\{T, C\}$	$\{\mathrm{o}, \mathrm{c}\}$	$\lambda/4$
$-T$	$\{S, C\}$	$\{\mathrm{o}, \mathrm{c}\}$	$\lambda/4$
$-C$	$\{S, T\}$	$\{\mathrm{o}, \mathrm{c}\}$	$\lambda/4$
One-sided	$\{S, T, C\}$	$\{\mathrm{o}\}$	$\lambda/3$

Edge removal ($-S$, $-T$, $-C$). Removing edge x_0 drops both directional margins, $\Phi_{x_0,\mathrm{o}}$ and $\Phi_{x_0,\mathrm{c}}$, leaving four comparisons:

$$\mathcal{L}_{-x_0} = \mathcal{L}_{\text{gen}} + \frac{\lambda}{4} \sum_{x \in \mathcal{X} \setminus \{x_0\}} \sum_{v \in \mathcal{V}} \Phi_{x,v}.$$

The $-S$, $-T$, and $-C$ variants assess the contributions of evidence-to-step, question-to-trace, and trace-to-answer margins, respectively. The intervention data and generation anchors associated with the removed edge remain in training.

Direction removal (one-sided). This variant retains all three edges but only the original-condition margins, S^{o}, T^{o}, and C^{o}:

$$\mathcal{L}_{\text{orig}} = \mathcal{L}_{\text{gen}} + \frac{\lambda}{3} \sum_{x \in \mathcal{X}} \Phi_{x,\mathrm{o}}.$$

It assesses the contribution of enforcing preference reversal under counterfactual inputs, beyond original-condition margins and counterfactual generation supervision.

In every variant, the margin coefficients sum to λ. Relative to the full objective, each retained term is scaled by $3/2$ after edge removal and by 2 after direction removal. This preserves the total coefficient weight across different sets of active comparisons.

D.4 TRAINING IMPLEMENTATION AND COMPUTE

Training protocol. We pool sources across the three datasets and use a fixed split of $2{,}566$ training and 286 development sources. For each backbone and seed, all adapted methods start independently from the same pretrained checkpoint and share the source sampler, context serialization, and optimization settings. CF-SFT, RM, E-CLOSURE, and its ablations also use the same three generation anchors defined in Appendix D.3.

Number of trained models. We train 48 independent LoRA adapters, one per method–backbone–seed combination. The main comparison includes 36 adapters covering four objectives (SFT, CF-SFT, RM, and E-CLOSURE), three backbones, and three seeds, $\{42, 43, 44\}$. Four controlled ablations on Qwen3-8B add 12 adapters using the same seeds. This yields 24 adapters for Qwen3-8B

Table 24: **Training hyperparameters.** Settings are shared across objectives within each backbone. Margin parameters apply to RM, E-CLOSURE, and its ablations.

Setting	Value
LoRA targets	All linear layers in the language model
LoRA rank / scale / dropout	16 / 32 / 0
Precision / gradient checkpointing	BF16 / enabled
Batch size per GPU	1 source pack per microstep
Gradient accumulation	2 microsteps per optimizer update
Effective global batch size	8 source packs (4 GPUs $\times 1 \times 2$)
Scheduled training budget	5 epochs; 1,605 optimizer updates
Training seeds	$\{42, 43, 44\}$
Optimizer	AdamW
Peak learning rate	5×10^{-5} for Qwen3-8B and Qwen3-32B; 10^{-4} for Ministral-3-8B
Learning-rate schedule	Linear decay with 50 warmup optimizer updates
Weight decay / gradient clipping	0 / maximum gradient norm 1.0
Maximum sequence length	8,192 tokens, including input and scored output
Development evaluation	Every epoch (321 optimizer updates)
Margin weight	$\lambda = 1$
Margin target / temperature	$\gamma = 0.5$ / $\beta = 0.1$, shared across all margins

and 12 each for Qwen3-32B and Ministral-3-8B. Each adapter is trained on the pooled datasets, with epoch checkpoints serving as selection candidates from the same run.

Optimization. We train with data parallelism across four GPUs using the settings in Table 24. Batch sizes count source packs, with the number of likelihood evaluations per pack determined by the objective. A full training run comprises 3,210 microsteps per worker, yielding 1,605 optimizer updates with gradient accumulation.

Checkpoint selection. At the end of each epoch, we evaluate the model using greedy decoding with a maximum of 512 new tokens. For each run, we select the checkpoint with the highest weighted development answer accuracy:

$$A_{\text{sel}} = \tfrac{1}{2}A_{\text{o}} + \tfrac{1}{4}A_{\text{e}} + \tfrac{1}{4}A_{\text{q}},$$

where A_{o}, A_{e}, and A_{q} denote answer accuracy under the original, evidence-edited, and question-edited inputs, respectively. Answers are scored using deterministic normalization and matching against the target answer and available aliases. Ties favor the earlier checkpoint. We apply the same selection rule across all objectives, backbones, and seeds.

Hardware and compute. Main experiments with Qwen3-8B and Ministral-3-8B use four NVIDIA RTX A6000 GPUs with 256 GB of host memory; Qwen3-32B uses four NVIDIA H100 GPUs. For the Qwen3-8B ablations, seed 42 uses four RTX A6000 GPUs, while seeds 43 and 44 use four H100 GPUs because these GPUs were available.

Table 25 reports wall-clock time per complete training run, averaged over completed seeds. Timing starts at W&B initialization after model loading and ends at the last logged event. It includes training, development evaluation, and checkpoint I/O within this interval, but excludes data construction, separate test evaluation, and unlogged startup or shutdown overhead. GPU hours are computed as recorded wall-clock hours multiplied by four. Across all 48 recorded runs, logged usage totals 3,544.0 GPU hours across both GPU types. All methods share the scheduled source exposure and optimizer-update budget, while the number of likelihood evaluations depends on the objective. All adapted methods use the same single-pass generation procedure at inference.

D.5 EXTENDED MAIN RESULTS

This section provides evaluation details and analyses for the main results in Section 5.2.

Table 25: **Recorded time per training run (hours).** Values are means ± sample standard deviations across completed seeds. All runs use four GPUs. Main experiments use RTX A6000 GPUs for the 8B backbones and H100 GPUs for Qwen3-32B. Ablation statistics pool RTX A6000 timings for seed 42 and H100 timings for seeds 43 and 44.

Method	Qwen3-8B	Ministral-3-8B	Qwen3-32B
SFT	5.20 ± 0.13	4.89 ± 0.13	4.54 ± 0.02
CF-SFT	10.64 ± 0.18	10.09 ± 0.13	8.59 ± 0.04
RM	22.01 ± 0.32	20.93 ± 0.14	17.12 ± 0.06
E-CLOSURE	43.73 ± 0.05	42.05 ± 0.26	33.57 ± 0.04
E-CLOSURE w/o S	19.75 ± 11.76	—	—
E-CLOSURE w/o T	22.09 ± 9.17	—	—
E-CLOSURE w/o C	19.63 ± 11.81	—	—
E-CLOSURE one-sided	14.00 ± 11.60	—	—

D.5.1 EVALUATION SETUP

All methods are evaluated on the same 866 diagnostic questions: 278 from 2WikiMultiHopQA, 255 from MuSiQue, and 333 from HotpotQA. The main comparison includes 31,176 responses from 36 adapted checkpoints, plus 2,598 fixed diagnostic backbone responses. Adapted methods use three training seeds, $\{42, 43, 44\}$. For each metric, we compute rates within each dataset–backbone–seed setting, average across seeds, and then equally average the nine dataset–backbone settings. The backbone uses one fixed response per question and model, with the same averaging across settings. This gives each dataset equal weight despite differing question counts. Evaluator validation is reported separately in Appendix D.1.

D.5.2 HUMAN AUDIT OF EXPERIMENTAL RESULTS

Audit design. We audit SFT, CF-SFT, RM, and E-CLOSURE on the same 90 questions, with 10 questions per dataset–backbone setting and 360 responses in total. Each question is assigned to one backbone, and all audited checkpoints use training seed 42. Two reviewers assess every response: they agree on 335 responses, and a third reviewer adjudicates the remaining 25. We use the final human labels throughout. Automated and human judgments evaluate identical questions, evidence, traces, and submitted answers. Equal sample sizes across settings make pooled rates equivalent to equal-setting means.

Table 26: **Evaluator agreement on the human-audit sample.** Acc. denotes answer correctness. n_0 counts human-labeled failures among 360 responses. Failure F1 treats label 0 as positive; $G = A \wedge C$ is derived.

Label	n_0	Agreement (%)	$F_{1,0}$	κ
Acc.	31	99.72	0.984	0.983
L	35	95.83	0.737	0.715
A	14	99.17	0.903	0.899
C	32	99.44	0.969	0.966
K	63	98.33	0.951	0.941
G	44	98.61	0.945	0.937

Agreement with human adjudication. Table 26 reports agreement between automated and human labels. Across the five primary judgments, raw agreement ranges from 95.83% to 99.72%, with answer correctness agreeing on $359/360$ responses. Agreement for derived global sufficiency is 98.61%, with failure F1 of 0.945. Disagreement patterns differ by criterion: the evaluator identifies 31 of 32 human-labeled closure failures, misses one, and flags one additional response; for local soundness, it misses 14 of 35 human failures and flags one additional response.

Metric differences and uncertainty. Table 27 compares automated and human scores on the same responses. We define their difference as $\Delta M = M_{\text{auto}} - M_{\text{human}}$ and report paired differences in Table 28. All confidence intervals in this audit are 95% percentile intervals from $10{,}000$ paired bootstrap resamples using seed 42. Each resample draws 10 questions with replacement within each fixed dataset–backbone setting, preserving all four methods and both label sources. This procedure captures question-sampling uncertainty within the audited seed-42 runs.

Table 27: **Automated and human scores on the audit sample.** Each method contributes 90 responses. All scores are percentages: Acc.↑, LSR↑, GSR↑, TRR↑, and LGG↓. LGG additionally reports the number of responses with $L = 1$ and $G = 0$ in parentheses. Bold marks the best human score in each column.

Method	Labels	Acc.	LSR	GSR	TRR	LGG
SFT	Automated	91.11	95.56	85.56	83.33	12.22 (11)
	Human	91.11	**94.44**	85.56	82.22	12.22 (11)
CF-SFT	Automated	88.89	92.22	86.67	82.22	10.00 (9)
	Human	90.00	85.56	87.78	77.78	7.78 (7)
RM	Automated	92.22	93.33	85.56	82.22	11.11 (10)
	Human	**92.22**	91.11	86.67	81.11	10.00 (9)
E-CLOSURE	Automated	92.22	94.44	90.00	86.67	7.78 (7)
	Human	**92.22**	90.00	**91.11**	**84.44**	**5.56 (5)**

Across methods, automated and human TRR are 83.61% and 81.39%, respectively, giving a difference of $+2.22$ percentage points (95% CI: $[0.83,\ 3.61]$). Automated LGG is 10.28% (37 responses), versus 8.89% (32 responses) under human labels, a difference of $+1.39$ points ($[0.00,\ 2.78]$). For E-CLOSURE, the corresponding differences are $+2.22$ points for both TRR and LGG. The pooled TRR interval lies above zero, whereas the pooled LGG interval includes zero.

Table 28: **Automated minus human metric differences.** Values are percentage-point differences with 95% paired bootstrap confidence intervals. The pooled row averages all four methods.

Method	ΔTRR	ΔLGG
SFT	$+1.11$ [0.00, $+3.33$]	0.00 [0.00, 0.00]
CF-SFT	$+4.44$ [$+1.11$, $+8.89$]	$+2.22$ [0.00, $+5.56$]
RM	$+1.11$ [-2.22, $+4.44$]	$+1.11$ [-2.22, $+4.44$]
E-CLOSURE	$+2.22$ [0.00, $+5.56$]	$+2.22$ [0.00, $+5.56$]
All methods	$+2.22$ [$+0.83$, $+3.61$]	$+1.39$ [0.00, $+2.78$]

Method comparison under human labels. Under human labels, E-CLOSURE achieves a TRR of 84.44% and an LGG of 5.56% (5 responses), the highest and lowest point estimates, respectively. Its paired TRR gain is 6.67 points over CF-SFT (95% CI: $[1.11,\ 12.22]$) and 3.33 points over RM ($[-2.22,\ 10.00]$). The audit therefore supports a TRR improvement over CF-SFT, while the comparison with RM remains uncertain. The lower LGG point estimate relative to both baselines also persists under human review. These findings concern the audited responses from seed 42; variation across training seeds is assessed by the main experiments.

D.5.3 DETERMINISTIC MATCHING ACCURACY

Scoring rule. We apply the deterministic matcher used during training to the stored final answers on the diagnostic evaluation set. This evaluates the same predictions under an alternative scoring rule. Let $N(a)$ denote the token sequence obtained by removing an optional leading `Final answer:` label, lowercasing, replacing characters outside ASCII letters, digits, and whitespace with spaces, removing the articles `a`, `an`, and `the`, and collapsing whitespace. Write $u \preceq v$ when u is a contiguous whole-token span of v. The per-response score is

$$\mathrm{Match}(a, a^\star) = \mathbf{1}\begin{bmatrix} N(a) \neq \varnothing \wedge N(a^\star) \neq \varnothing \\ \wedge \big(N(a) \preceq N(a^\star) \vee N(a^\star) \preceq N(a)\big) \end{bmatrix}.$$

We denote its mean under the aggregation above by $\mathrm{Acc}_{\mathrm{match}}$. Only the stored prediction and reference answer are compared, without aliases or manual overrides. Although recorded as `acc_em` in the implementation, this bidirectional span-matching score differs from strict exact match. Development checkpoint selection uses the same normalizer and pairwise matcher but additionally accepts

available aliases and evaluates original and counterfactual inputs with a different generation length limit.

Sensitivity to the scoring rule. Table 29 compares semantic and deterministic accuracy. E-CLOSURE retains the highest aggregate point estimate under deterministic matching, reaching 88.44%, compared with 87.16% for RM and 84.96% for CF-SFT. Its gains are 1.27 and 3.48 percentage points, respectively, while its semantic-accuracy gain over RM is 2.28 points. Across individual settings, E-CLOSURE exceeds RM in five of nine settings and CF-SFT in eight, with one tie (Table 30). The aggregate accuracy advantage persists under deterministic matching, supporting its robustness to the scoring rule, while the magnitude and setting-level gains vary.

Deterministic matching can reject semantic equivalents or accept incomplete answers. In the diagnostic responses, `East Germany` versus `GDR` fails matching but passes semantic correctness, whereas `Latin` versus `Medieval Latin` passes matching but fails semantic correctness. The disagreement columns report predictions accepted by only one scoring rule, using the existing semantic labels.

Table 29: **Answer accuracy under two scoring rules.** Values are percentages, averaged over seeds and then equally across the nine settings. Match only and Semantic only report the proportions accepted exclusively by the corresponding rule, using the same weighting.

Method	$\mathrm{Acc}_{\mathrm{sem}}$	$\mathrm{Acc}_{\mathrm{match}}$	**Match only**	**Semantic only**
Backbone	79.75	74.73	0.45	5.47
SFT	87.76	84.21	0.51	4.06
CF-SFT	88.66	84.96	0.36	4.06
RM	90.51	87.16	0.47	3.81
E-CLOSURE	92.79	88.44	0.35	4.70

Table 30: **Deterministic answer accuracy by dataset and backbone (%).** Adapted methods report means $\pm$ sample standard deviations over seeds 42, 43, and 44. Backbone results use the fixed diagnostic predictions.

Model	**Method**	**2WikiMultiHopQA**	**MuSiQue**	**HotpotQA**
Qwen3-8B	Backbone	92.45	50.20	84.98
	SFT	98.44 ± 0.55	65.88 ± 0.68	81.98 ± 1.08
	CF-SFT	98.92 ± 0.95	69.02 ± 0.39	82.58 ± 0.79
	RM	98.56 ± 0.00	70.98 ± 3.59	87.49 ± 1.21
	E-CLOSURE	99.52 ± 0.21	76.60 ± 3.43	87.29 ± 0.63
Qwen3-32B	Backbone	96.40	57.65	86.49
	SFT	98.68 ± 0.21	68.63 ± 1.36	87.09 ± 1.04
	CF-SFT	98.80 ± 0.21	71.63 ± 1.48	86.79 ± 0.60
	RM	99.64 ± 0.36	75.42 ± 1.20	88.99 ± 0.92
	E-CLOSURE	99.04 ± 0.21	79.61 ± 3.86	89.29 ± 1.76
Ministral-3-8B	Backbone	84.17	41.57	78.68
	SFT	99.52 ± 0.21	72.29 ± 0.99	85.39 ± 2.00
	CF-SFT	99.28 ± 0.62	72.03 ± 2.61	85.59 ± 1.59
	RM	99.64 ± 0.36	74.77 ± 3.54	88.99 ± 0.17
	E-CLOSURE	99.28 ± 0.36	77.91 ± 1.48	87.39 ± 0.79

D.5.4 PER-SEED RESULTS

Tables 31–33 report both answer accuracy measures and the main reasoning metrics for each training seed. Here, $\mathrm{Acc}_{\mathrm{match}}$ uses deterministic matching, while $\mathrm{Acc}_{\mathrm{sem}}$ uses semantic correctness labels. All metrics are percentages; higher values are better except for LGG. The standard deviations in Table 30 summarize variation across training seeds, not uncertainty from question sampling.

Backbone results use the fixed diagnostic generations. All responses are scored by the same automated evaluator described in Appendix D.1. Generation uses greedy decoding with at most $1{,}024$ new tokens and an $8{,}192$-token sequence limit; explicit thinking mode is disabled where supported. Appendix D.4 documents checkpoint selection.

D.6 Objective Ablations

D.6.1 Ablation Setup

We conduct all ablations on Qwen3-8B using training seeds $\{42, 43, 44\}$. Removing S, T, or C drops both directional margins for that dependency. The one-sided variant retains all three dependencies but only their original-condition margins. All variants share the generation anchors, training data, backbone initialization, scheduled optimization budget, and checkpoint-selection rule.

To preserve the total margin coefficient λ, retained terms are scaled by $3/2$ after dependency removal and by 2 in the one-sided variant. The resulting objectives contain four and three margin terms, respectively, compared with six in the full objective.

D.6.2 Detailed Results

Following the main ablation analysis, we pool all 866 questions within each seed and then average over the three seeds. This weights questions equally, whereas the main comparison in Appendix D.5 weights datasets equally. For example, the full objective achieves pooled TRR of 87.34%, compared with 86.33% under equal dataset weights. Tables 34 and 35 report per-seed and dataset-specific results, respectively. Under deterministic answer matching, the full objective achieves 88.07% pooled accuracy, exceeding all four ablations; the closest variant is w/o C at 86.84%.

D.6.3 Failure Decomposition

We partition LGG responses using the same first-broken-link convention as the diagnostic study. Path deviation (PD) corresponds to $L = 1, A = 0$; incomplete reasoning (IR) to $L = 1, A = 1, C = 0, K = 0$; and answer decoupling (AD) to $L = 1, A = 1, C = 0, K = 1$. These categories are disjoint and together comprise $L = 1, G = 0$. Table 36 reports their rates over all evaluated responses, so they sum to the LGG rate before rounding.

Removing T increases all three components relative to the full objective: PD rises from 2.27% to 4.08%, IR from 3.66% to 4.70%, and AD from 0.50% to 1.62%. For the one-sided variant, the largest increase occurs in IR, from 3.66% to 5.62%. These descriptive patterns show that changes associated with an objective can span multiple failure types.

D.6.4 Paired Comparisons

Table 37 reports paired differences with 95% percentile confidence intervals from $10{,}000$ bootstrap resamples using seed 2031. Each resample draws questions from the pooled 866-question set and resamples training seeds, preserving pairing across variants. The same seed draw applies across all datasets.

The full objective improves TRR over every ablation, with all four confidence intervals above zero. LGG reduction intervals also exclude zero for w/o T and one-sided supervision, while those for w/o S and w/o C include zero. The closure gain over w/o C is 1.15 points, with an interval of $[-0.12, 2.35]$. Together, these comparisons support the full objective's contribution to joint trace reliability, while improvements in individual failure measures are less uniformly resolved.

D.7 Dependency Interventions

D.7.1 Evaluation Protocol

Development set. We evaluate 286 development sources: 128 from 2WikiMultiHopQA, 56 from MuSiQue, and 102 from HotpotQA. These sources are excluded from gradient updates but used for checkpoint selection based on weighted development answer accuracy. Intervention scores do not

Table 31: **Per-seed results for Qwen3-8B (%).** Dataset headings give the number of questions. Fixed denotes the untuned backbone's fixed diagnostic generations.

Method	Seed	$\mathrm{Acc}_{\mathrm{match}}$	$\mathrm{Acc}_{\mathrm{sem}}$	LSR	GSR	TRR	LGG
			2WikiMultiHopQA ($n = 278$)				
Backbone	Fixed	92.45	92.45	95.68	94.24	90.65	5.04
SFT	42	98.56	98.56	98.56	99.28	98.20	0.36
	43	98.92	98.92	98.20	99.28	98.20	0.00
	44	97.84	97.84	97.48	98.92	97.48	0.00
CF-SFT	42	98.56	98.56	98.56	99.64	98.56	0.00
	43	98.20	98.56	98.20	98.56	97.84	0.36
	44	100.00	100.00	100.00	100.00	100.00	0.00
RM	42	98.56	98.92	98.92	100.00	98.92	0.00
	43	98.56	99.28	99.28	100.00	99.28	0.00
	44	98.56	98.92	98.92	99.28	98.20	0.72
E-CLOSURE	42	99.64	99.64	99.28	99.64	99.28	0.00
	43	99.28	99.64	99.28	99.64	99.28	0.00
	44	99.64	100.00	100.00	100.00	100.00	0.00
			MuSiQue ($n = 255$)				
Backbone	Fixed	50.20	55.29	43.14	71.76	32.16	10.98
SFT	42	65.49	69.41	87.84	65.88	61.96	25.88
	43	65.49	69.80	81.57	58.04	50.59	30.98
	44	66.67	70.98	74.51	63.53	53.33	21.18
CF-SFT	42	68.63	73.73	87.84	68.24	65.88	21.96
	43	69.41	73.73	80.39	60.78	54.90	25.49
	44	69.02	73.33	78.82	59.61	53.73	25.10
RM	42	67.06	71.37	77.25	56.47	48.63	28.63
	43	74.12	75.69	78.82	57.25	50.59	28.24
	44	71.76	74.90	81.57	57.65	49.41	32.16
E-CLOSURE	42	75.69	80.78	88.24	76.86	70.59	17.65
	43	73.73	80.39	80.78	69.41	60.00	20.78
	44	80.39	82.35	83.53	72.55	67.06	16.47
			HotpotQA ($n = 333$)				
Backbone	Fixed	84.98	91.89	88.89	95.20	84.98	3.90
SFT	42	81.68	88.29	96.40	90.69	88.29	8.11
	43	81.08	89.19	93.39	89.49	84.98	8.41
	44	83.18	90.39	93.09	90.99	86.49	6.61
CF-SFT	42	82.88	90.09	95.50	93.39	89.79	5.71
	43	81.68	88.89	92.79	91.59	86.19	6.61
	44	83.18	90.09	93.99	91.59	87.39	6.61
RM	42	86.79	92.49	91.89	91.89	86.79	5.11
	43	88.89	95.20	94.29	93.39	89.79	4.50
	44	86.79	93.09	92.79	93.99	88.29	4.50
E-CLOSURE	42	86.79	95.20	97.00	97.30	94.59	2.40
	43	87.09	94.59	94.89	96.40	91.59	3.30
	44	87.99	96.10	97.00	96.40	94.59	2.40

Table 32: **Per-seed results for Qwen3-32B (%).** Dataset headings give the number of questions. Fixed denotes the untuned backbone's fixed diagnostic generations.

Method	Seed	$\mathrm{Acc}_{\mathrm{match}}$	$\mathrm{Acc}_{\mathrm{sem}}$	LSR	GSR	TRR	LGG
			2WikiMultiHopQA ($n = 278$)				
Backbone	Fixed	96.40	96.40	96.40	99.64	96.04	0.36
SFT	42	98.92	98.92	98.56	100.00	98.56	0.00
	43	98.56	98.56	98.20	99.64	98.20	0.00
	44	98.56	98.92	98.92	100.00	98.92	0.00
CF-SFT	42	98.56	98.56	98.56	100.00	98.56	0.00
	43	98.92	98.92	98.56	100.00	98.56	0.00
	44	98.92	98.92	98.92	100.00	98.92	0.00
RM	42	100.00	100.00	99.64	100.00	99.64	0.00
	43	99.28	99.28	99.28	100.00	99.28	0.00
	44	99.64	99.64	99.64	100.00	99.64	0.00
E-CLOSURE	42	98.92	98.92	98.92	100.00	98.92	0.00
	43	98.92	98.92	98.92	100.00	98.92	0.00
	44	99.28	99.28	99.28	100.00	99.28	0.00
			MuSiQue ($n = 255$)				
Backbone	Fixed	57.65	65.88	58.43	79.61	52.55	5.88
SFT	42	69.41	72.94	84.71	63.14	56.86	27.84
	43	69.41	72.55	80.00	64.31	56.47	23.53
	44	67.06	71.37	81.96	63.92	56.86	25.10
CF-SFT	42	70.59	74.51	87.45	66.27	60.00	27.45
	43	70.98	73.73	85.49	63.92	57.25	28.24
	44	73.33	76.08	84.31	71.37	62.75	21.57
RM	42	76.47	78.82	83.92	59.22	51.76	32.16
	43	75.69	79.61	82.35	63.92	53.73	28.63
	44	74.12	76.86	82.75	61.18	53.73	29.02
E-CLOSURE	42	75.29	80.39	91.37	78.43	76.08	15.29
	43	80.78	86.27	90.59	83.92	80.39	10.20
	44	82.75	86.67	92.16	84.31	81.18	10.98
			HotpotQA ($n = 333$)				
Backbone	Fixed	86.49	94.89	94.59	96.70	91.89	2.70
SFT	42	86.49	94.29	99.10	92.79	92.49	6.61
	43	88.29	95.50	98.50	94.29	93.39	5.11
	44	86.49	93.99	97.60	94.29	92.19	5.41
CF-SFT	42	86.19	93.99	98.20	93.09	91.59	6.61
	43	87.39	94.59	99.10	93.39	92.49	6.61
	44	86.79	94.89	96.40	94.89	92.19	4.20
RM	42	89.79	96.70	97.90	96.70	94.89	3.00
	43	87.99	96.70	97.30	95.20	93.39	3.90
	44	89.19	97.00	97.30	97.00	95.20	2.10
E-CLOSURE	42	87.99	97.90	99.10	98.50	97.90	1.20
	43	88.59	98.20	99.70	98.50	98.20	1.50
	44	91.29	98.80	99.40	99.10	98.80	0.60

Table 33: **Per-seed results for Ministral-3-8B (%).** Dataset headings give the number of questions. Fixed denotes the untuned backbone's fixed diagnostic generations.

Method	Seed	$\mathrm{Acc}_{\mathrm{match}}$	$\mathrm{Acc}_{\mathrm{sem}}$	LSR	GSR	TRR	LGG
			2WikiMultiHopQA ($n = 278$)				
Backbone	Fixed	84.17	84.17	91.01	87.77	80.94	10.07
SFT	42	99.28	99.28	99.28	100.00	99.28	0.00
	43	99.64	99.64	99.64	100.00	99.64	0.00
	44	99.64	99.64	99.64	100.00	99.64	0.00
CF-SFT	42	99.64	100.00	99.64	99.64	99.64	0.00
	43	99.64	100.00	100.00	100.00	100.00	0.00
	44	98.56	98.92	98.56	100.00	98.56	0.00
RM	42	99.64	100.00	100.00	100.00	100.00	0.00
	43	99.28	99.64	100.00	99.64	99.64	0.36
	44	100.00	100.00	100.00	100.00	100.00	0.00
E-CLOSURE	42	98.92	99.64	100.00	99.64	99.64	0.36
	43	99.64	100.00	100.00	100.00	100.00	0.00
	44	99.28	99.64	99.64	100.00	99.64	0.00
			MuSiQue ($n = 255$)				
Backbone	Fixed	41.57	49.41	30.20	65.49	24.71	5.49
SFT	42	73.33	76.47	82.75	67.84	60.39	22.35
	43	71.37	74.12	89.02	68.63	64.31	24.71
	44	72.16	74.90	86.67	60.39	55.69	30.98
CF-SFT	42	73.73	76.86	85.49	66.67	61.57	23.92
	43	69.02	74.51	91.37	69.41	67.45	23.92
	44	73.33	76.08	88.24	63.53	58.82	29.41
RM	42	74.51	78.82	85.88	69.02	60.78	25.10
	43	71.37	74.90	87.06	58.82	52.55	34.51
	44	78.43	80.00	92.55	57.25	53.73	38.82
E-CLOSURE	42	76.86	82.35	89.41	78.43	72.55	16.86
	43	79.61	82.75	90.20	78.82	74.90	15.29
	44	77.25	80.78	90.98	72.55	67.45	23.53
			HotpotQA ($n = 333$)				
Backbone	Fixed	78.68	87.39	80.78	90.69	76.28	4.50
SFT	42	85.89	92.79	94.29	93.99	89.19	5.11
	43	83.18	89.79	93.39	93.09	87.09	6.31
	44	87.09	92.49	96.10	94.59	91.29	4.80
CF-SFT	42	86.79	93.69	96.10	95.20	91.89	4.20
	43	86.19	92.49	95.80	93.69	90.09	5.71
	44	83.78	90.09	96.10	91.29	88.29	7.81
RM	42	89.19	94.59	95.50	95.80	92.19	3.30
	43	88.89	96.40	95.80	95.50	92.49	3.30
	44	88.89	94.89	97.00	93.09	90.99	6.01
E-CLOSURE	42	87.09	95.20	96.10	96.40	93.09	3.00
	43	86.79	94.89	95.80	96.70	93.09	2.70
	44	88.29	96.10	97.90	97.30	95.20	2.70

Table 34: **Per-seed objective ablations on Qwen3-8B.** All values are percentages over the 866 questions pooled within each seed. $\mathrm{Acc}_{\mathrm{match}}$ uses deterministic matching, while $\mathrm{Acc}_{\mathrm{sem}}$ uses semantic correctness labels.

Variant	Seed	$\mathrm{Acc}_{\mathrm{match}}$	$\mathrm{Acc}_{\mathrm{sem}}$	LSR	A	C	GSR	TRR	LGG
w/o S	42	84.64	87.99	90.30	92.73	95.27	87.99	83.49	6.81
	43	84.06	88.11	88.45	93.19	94.80	87.99	81.87	6.58
	44	86.03	89.84	91.45	94.57	94.46	89.03	83.83	7.62
w/o T	42	82.33	87.07	91.22	91.34	92.84	84.64	81.18	10.05
	43	81.76	85.57	91.69	92.03	92.38	84.99	80.60	11.09
	44	83.26	87.30	91.92	93.07	92.15	85.57	81.87	10.05
w/o C	42	87.30	90.42	90.88	94.69	93.88	89.03	82.56	8.31
	43	87.76	90.30	89.38	94.46	93.88	88.57	82.45	6.93
	44	85.45	89.15	91.92	94.00	93.88	88.45	83.83	8.08
One-sided	42	85.91	90.42	88.11	93.65	92.84	86.72	80.48	7.62
	43	84.06	88.57	88.11	93.65	92.84	86.72	79.79	8.31
	44	85.68	89.49	91.45	94.46	90.18	85.57	81.06	10.39
E-CLOSURE	42	87.64	92.38	95.15	95.96	95.61	92.03	89.03	6.12
	43	87.07	92.03	92.15	94.92	94.46	89.49	84.76	7.39
	44	89.49	93.30	94.00	95.03	95.03	90.53	88.22	5.77

Table 35: **Dataset-specific ablations on Qwen3-8B.** Values are three-seed mean percentages. Each dataset reports deterministic matching accuracy, TRR, and LGG.

Variant	2WikiMultiHopQA			MuSiQue			HotpotQA		
	$\mathrm{Acc}_{\mathrm{match}}$	TRR	LGG	$\mathrm{Acc}_{\mathrm{match}}$	TRR	LGG	$\mathrm{Acc}_{\mathrm{match}}$	TRR	LGG
w/o S	99.04	98.56	0.12	69.28	57.39	18.56	85.09	89.79	3.90
w/o T	98.80	98.92	0.00	64.84	54.12	26.80	82.28	87.19	6.51
w/o C	98.56	98.80	0.00	73.99	56.08	21.83	86.89	90.29	3.50
One-sided	98.56	98.68	0.00	71.11	51.11	24.05	84.88	87.69	4.40
E-CLOSURE	99.52	99.52	0.00	76.60	65.88	18.30	87.29	93.59	2.70

enter selection, so we report this analysis as development-set diagnostics. Three fixed backbones and 36 adapted checkpoints yield 39 scoring runs on the same sources.

Controlled inputs. Each source provides a validated original response, evidence and question counterfactuals, and the first affected step under the evidence edit. When multiple evidence edits are available, we select the first under a fixed ordering by passage identifier and serialized edit content. The selected edit is shared across all checkpoints.

For S, we change the evidence while holding the question and shared trace prefix fixed, comparing the original and counterfactual affected steps. For T, we change the question while holding the evidence fixed, comparing the original and question-counterfactual traces without their final answers. For C, we change the supplied trace while holding the original question and evidence fixed, comparing the original and evidence-counterfactual answers. Thus, C-switching measures whether answer preferences follow the supplied trace with external evidence held fixed.

Scoring. For a candidate completion y under serialized input x, we compute its token-normalized log likelihood with teacher forcing:

$$f_\theta(y \mid x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}).$$

For each dependency $d \in \{S, T, C\}$, let x_o, x_c denote the original and counterfactual inputs, and y_o, y_c their corresponding preferred candidates. We define

$$\Delta_o^d = f_\theta(y_o \mid x_o) - f_\theta(y_c \mid x_o), \quad \Delta_c^d = f_\theta(y_c \mid x_c) - f_\theta(y_o \mid x_c), \quad \mathrm{Switch}_d = \mathbf{1}\left[\Delta_o^d > 0 \wedge \Delta_c^d > 0\right].$$

Table 36: **Deterministic accuracy and LGG decomposition.** Values are percentages pooled over 866 questions within each seed, then averaged over three seeds. PD, IR, and AD use all responses as the denominator and sum to LGG before rounding.

Variant	$\mathrm{Acc}_{\mathrm{match}}$	PD	IR	AD	LGG
w/o S	84.91	2.69	3.66	0.65	7.01
w/o T	82.45	4.08	4.70	1.62	10.39
w/o C	86.84	2.77	4.20	0.81	7.78
One-sided	85.22	2.42	5.62	0.73	8.78
E-CLOSURE	88.07	2.27	3.66	0.50	6.43

Table 37: **Paired gains from the full objective.** Cells show gains in percentage points above their 95% bootstrap confidence intervals. Metric and TRR gains are full minus ablated; LGG reduction is ablated minus full. Positive values favor E-CLOSURE.

Comparator	Metric	Metric gain	TRR gain	LGG reduction
w/o S	LSR	$+3.70$ $[+1.96, +5.54]$	$+4.27$ $[+2.23, +6.35]$	$+0.58$ $[-1.23, +2.35]$
w/o T	A	$+3.16$ $[+1.39, +5.08]$	$+6.12$ $[+3.70, +8.62]$	$+3.96$ $[+2.42, +5.50]$
w/o C	C	$+1.15$ $[-0.12, +2.35]$	$+4.39$ $[+1.92, +6.93]$	$+1.35$ $[-0.69, +3.16]$
One-sided	$\mathrm{Acc}_{\mathrm{sem}}$	$+3.08$ $[+1.50, +4.70]$	$+6.89$ $[+4.43, +9.31]$	$+2.35$ $[+0.23, +4.73]$

Successful switching requires positive margins in both directions; ties count as failures. Evaluation uses a zero threshold, distinct from the training target margin $\gamma = 0.5$. Twelve candidate scores determine six directional margins without response generation or LLM judging. For adapted methods, rates are averaged over three seeds per dataset–backbone setting, then equally across nine settings.

D.7.2 DIRECTIONAL RESULTS

Table 38 reports success in each direction and the joint switch rate. For C, RM prefers the original answer under the original trace on 100% of pairs but prefers the counterfactual answer under the edited trace on only 30.41%. Its 69.59-point switch-rate gap relative to E-CLOSURE therefore arises entirely in the counterfactual direction. For T, E-CLOSURE achieves 86.02% original-direction success, 98.39% counterfactual-direction success, and 84.41% joint switching. Its remaining T-switch failures are concentrated in the original-conditioning direction.

Table 38: **Directional preference success and joint switching (%).** O and CF denote positive original and counterfactual margins; SW requires both on the same pair. Rates are equally averaged across nine settings after seed averaging for adapted methods.

Method	S			T			C		
	O	CF	SW	O	CF	SW	O	CF	SW
Backbone	99.60	98.44	98.24	66.90	82.64	49.54	94.90	79.73	74.64
SFT	100.00	99.68	99.68	85.00	94.36	79.36	97.51	89.13	86.63
CF-SFT	100.00	99.94	99.94	83.53	96.34	79.87	97.70	92.24	89.94
RM	100.00	99.91	99.91	83.68	95.75	79.42	100.00	30.41	30.41
E-CLOSURE	100.00	99.91	99.91	86.02	98.39	84.41	100.00	100.00	100.00

D.7.3 PER-SETTING RESULTS

Table 39 reports switch rates for each dataset–backbone setting. E-CLOSURE achieves 100% C-switching across all 2,574 source–checkpoint evaluations, comprising the same 286 sources eval-

uated under nine backbone–seed checkpoints. Mean T-switch rates are 59.90% on 2WikiMultiHopQA, 94.64% on MuSiQue, and 98.69% on HotpotQA. Question-guided trace switching therefore remains less consistent on 2WikiMultiHopQA, while S-switching is already near ceiling across the fine-tuned methods.

Table 39: **Dependency switch rates by dataset and backbone (%).** Adapted methods report three-seed means; Backbone uses one fixed run. Datasets contain 128, 56, and 102 sources, respectively.

Method	2WikiMultiHopQA			MuSiQue			HotpotQA		
	S	T	C	S	T	C	S	T	C
Qwen3-8B									
Backbone	99.22	37.50	84.38	96.43	46.43	94.64	99.02	65.69	75.49
SFT	99.74	62.50	94.01	99.40	80.95	92.26	100.00	93.79	86.93
CF-SFT	100.00	55.73	94.79	100.00	86.31	96.43	100.00	94.44	89.22
RM	100.00	55.99	98.96	100.00	85.12	0.60	100.00	97.06	6.54
E-CLOSURE	100.00	56.77	100.00	100.00	95.24	100.00	100.00	99.02	100.00
Qwen3-32B									
Backbone	100.00	39.84	59.38	98.21	42.86	69.64	100.00	67.65	55.88
SFT	100.00	65.10	87.50	100.00	82.74	75.00	100.00	94.44	69.93
CF-SFT	100.00	59.38	91.15	100.00	88.69	76.79	100.00	94.77	76.80
RM	99.48	57.55	68.23	100.00	86.31	11.90	100.00	93.14	10.13
E-CLOSURE	100.00	58.33	100.00	100.00	94.64	100.00	100.00	98.04	100.00
Ministral-3-8B									
Backbone	97.66	41.41	85.94	94.64	44.64	76.79	99.02	59.80	69.61
SFT	99.22	58.85	99.22	99.40	82.74	89.88	99.35	93.14	84.97
CF-SFT	99.48	59.64	99.22	100.00	85.12	92.26	100.00	94.77	92.81
RM	99.74	60.94	62.76	100.00	83.93	8.33	100.00	94.77	6.21
E-CLOSURE	99.22	64.58	100.00	100.00	94.05	100.00	100.00	99.02	100.00

D.7.4 PAIRED COMPARISONS

Table 40 compares E-CLOSURE with CF-SFT using 95% percentile confidence intervals from $10{,}000$ bootstrap resamples with seed 20260915. Questions are resampled within datasets, with shared draws across methods and backbones. Training seeds are resampled within each backbone, with shared draws across datasets and methods. Dataset–backbone settings remain fixed.

Table 40: **Paired switch gains over CF-SFT.** Gains are computed as E-CLOSURE minus CF-SFT and reported in percentage points, with 95% paired bootstrap confidence intervals.

Dependency	Gain	95% CI
S	-0.03	$[-0.17, +0.00]$
T	$+4.54$	$[+1.93, +7.33]$
C	$+10.06$	$[+7.51, +12.81]$

The full objective improves T-switching by 4.54 percentage points and C-switching by 10.06 points, with both intervals above zero. The S-switch difference is -0.03 points, with an interval that includes zero. Complete per-setting and per-seed directional rates are provided in the supplementary CSV. These scores measure preference reversal between fixed candidates; perfect C-switching on these development sources does not imply perfect closure in freely generated responses.