

Directional Structurality: A Lightweight Proxy for Reasoning Data Quality

Ze Chen* Kaixuan Liu* Haomai Zhang
Jiaxuan Zou Zhuo Chen Xiaotian Tan[†]
China Agricultural University

Abstract

Reasoning-oriented language model training relies on traces that expose learnable dependencies, yet existing data-selection criteria use length, perplexity, difficulty, or judge-assigned quality. We argue that useful reasoning traces should also be sensitive to dependency order: intermediate steps are easier to encode in their natural premise-to-conclusion order but less predictable when dependency-revealing order is perturbed.

We introduce Local-DeltaAvg, a training-free data-selection metric. For each reasoning trace, we construct a semantically reversed counterpart that preserves problem content while perturbing dependency-revealing order. A fixed probe model scores both trajectories, and Local-DeltaAvg computes a paired local negative log-likelihood gap between the original and reversed contexts. The resulting probe-dependent score estimates the conditional encoding cost induced by order perturbation and identifies where this sensitivity is concentrated.

We evaluate Local-DeltaAvg as a relative ranking signal for reasoning-oriented continued-pretraining data selection, followed by an identical downstream SFT protocol. Under matched token budgets, it consistently outperforms length-, perplexity-, difficulty-, and judge-based selection across diverse reasoning benchmarks. Control analyses show that semantic reversal is essential and that high-gap regions align with annotated critical reasoning steps. These results suggest that dependency-order sensitivity is a lightweight and effective proxy for reasoning-data utility.

1 Introduction

Large language models increasingly acquire reasoning abilities from large collections of solution

traces. However, these corpora mix useful step-by-step derivations with redundant explanations, shallow shortcuts, and noisy rationales. Existing selection methods typically rely on length, perplexity, problem difficulty, or LLM-as-a-judge scores. While practical, these signals mostly measure surface complexity, fluency, input hardness, or external preference rather than internal logical structure. A trace can be long yet redundant, difficult for reasons unrelated to logic, or fluent while containing little reasoning progression.

To address this, we propose *directional structurality* as a training-relevant property of reasoning-data quality. High-value logical data should expose how conditions support intermediate conclusions and how constraints tighten across steps. A reasoning trace has stronger directional structurality when its dependencies are organized along a clear logical direction: they unfold coherently in the forward order, but become harder to recover when the trace is semantically reversed. This principle naturally extends from linear sequences to complex branching dependencies like reasoning trees and graphs.

To operationalize this property, we introduce LOCAL-DELTA AVG. For each trace, we construct a semantically reversed counterfactual that preserves the task content but reverses the explanatory order. A fixed probe model then scores both directions, computing a paired within-sample negative log-likelihood (NLL) difference inside local windows. Unlike absolute perplexity, this score measures how much predictability changes when the dependency direction is perturbed, naturally controlling for shared lexical content and domain familiarity. We use it as a relative ranking proxy within a fixed candidate pool, rather than as an absolute quality label. The local score forms a directional-gap curve, where peaks indicate reasoning *bottlenecks*, positions where dependencies are concentrated and least recoverable under reversal.

*Equal contribution.

[†]Corresponding author.

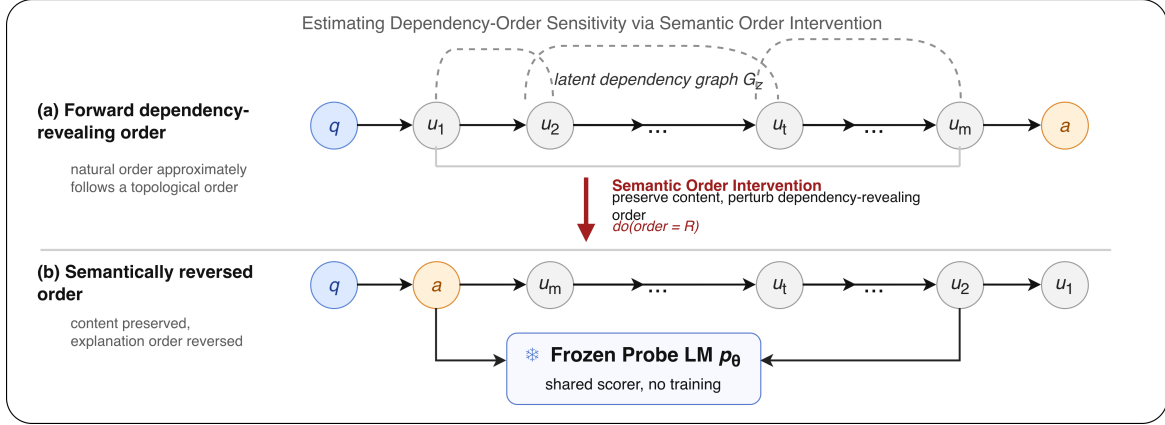


Figure 1: Illustration of dependency-order sensitivity under semantic order intervention.

We evaluate LOCAL-DELTA AVG under reasoning-oriented continued pretraining followed by an identical downstream SFT protocol. This setting cleanly separates structure learning from instruction-format imitation. Under matched token budgets across diverse benchmarks, data selected by our metric consistently improves over length, perplexity, difficulty, and judge-based baselines. Furthermore, the induced bottlenecks strongly align with annotated critical reasoning steps.

Our contributions are:

- We identify *directional structurality* as a training-relevant property of reasoning traces, grounded in directional dependency organization.
- We propose LOCAL-DELTA AVG, a paired within-sample NLL-gap proxy that yields local bottleneck curves without training.
- We validate its superiority over standard quality heuristics in reasoning-oriented continued pretraining and show its alignment with critical reasoning steps.

2 Related Work

Reasoning Data and Selection. Chain-of-thought prompting (Wei et al., 2022) and self-consistency (Wang et al., 2023) have established the value of intermediate reasoning traces. Recent pipelines rely on large trace corpora to internalize multi-step problem solving. To ensure data quality, common selection methods leverage small curated datasets (Zhou et al., 2023), LLM judges (Chen et al., 2024), or instruction-following difficulty (Li et al., 2024). Recent reasoning-specific methods like HES (Li et al., 2026) use token uncertainty to select long-CoT data. While

these signals are informative, they often measure verbosity, fluency, or judge preference (Zheng et al., 2023). Our work instead focuses on the internal logical structure of the trace, measuring its directional dependency organization.

Process Supervision and Verifiers. Process supervision provides fine-grained reasoning signals by training reward models or verifiers to score intermediate steps (Lightman et al., 2024; Cobbe et al., 2021; Uesato et al., 2022; Wang et al., 2024). While effective, these methods typically require step-level annotations or reward-model training. In contrast, LOCAL-DELTA AVG provides an unsupervised, probe-based proxy for reasoning structure. It can be used as a training-free signal for pretraining data selection, and its local gap scores inherently align with critical reasoning steps without supervised feedback.

3 Methodology

3.1 Problem Setup

LOCAL-DELTA AVG studies data selection for reasoning-oriented continued pretraining. Let $\mathcal{D} = \{x_i\}_{i=1}^N$ be a candidate pool of reasoning traces, where each trace is represented as $x = (q, u_1, \dots, u_m, a)$ with problem q , intermediate steps $u_1, \dots, u_m$, and final answer a , and is linearized into tokens $x = (s_1, \dots, s_T)$. We seek a score $S(x)$ that captures direction-sensitive dependency structure rather than marginal signals such as length, absolute perplexity, difficulty, or judge preference. We then select $\mathcal{D}_k = \text{TopK}_{x \in \mathcal{D}} S(x)$ for continued pretraining and evaluate the resulting model under the same downstream SFT protocol. The score is computed using a fixed probe language

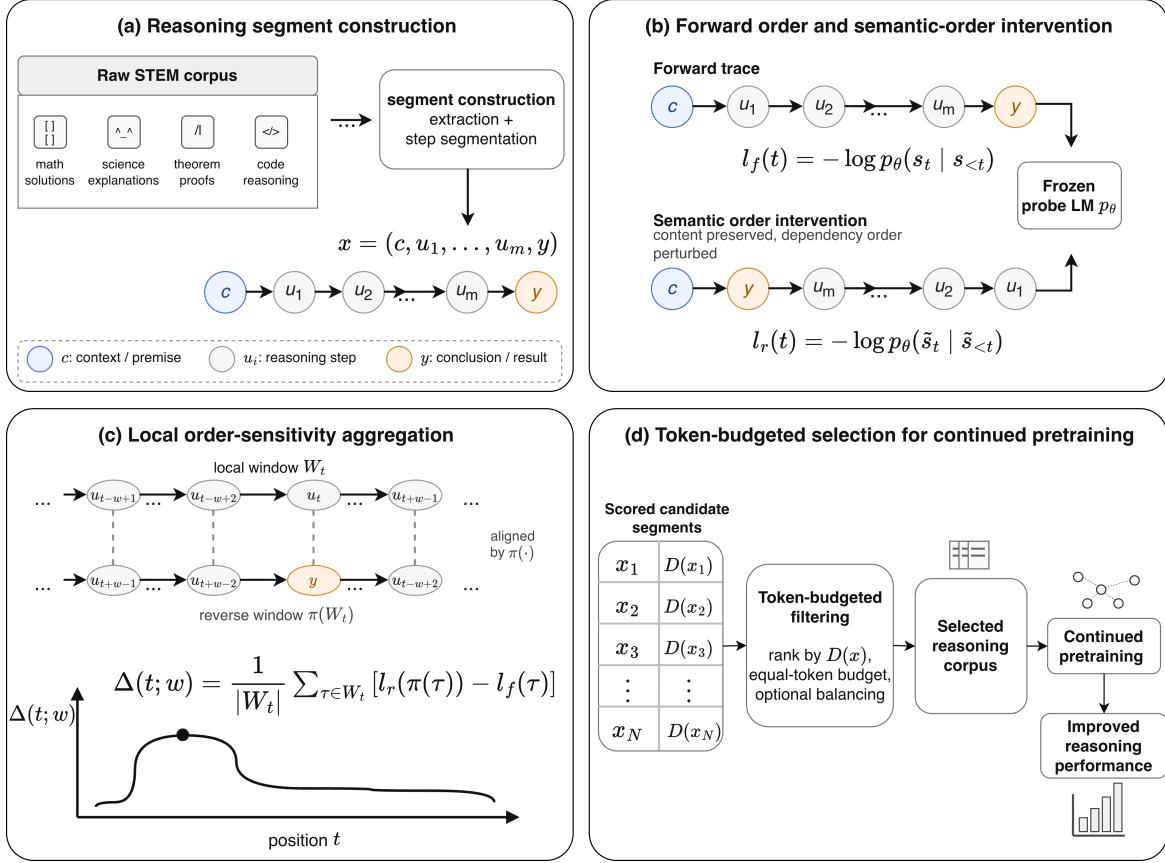


Figure 2: Main pipeline of LOCAL-DELTA AVG: segment construction, semantic-order intervention, local order-sensitivity aggregation, and token-budgeted selection for continued pretraining.

model p_θ and requires neither step-level supervision nor target-model feedback.

3.2 Order Intervention via Semantic Reversal

We view a reasoning trace as generated under a latent dependency graph whose forward order follows a topological order. This framework naturally captures both linear chains ($q \rightarrow u_1 \rightarrow u_2 \rightarrow \dots \rightarrow a$) and complex branching structures like reasoning trees or directed acyclic graphs (DAGs). Our central intervention perturbs the order in which dependencies are revealed, i.e., $do(\text{order}(x) = R)$. Concretely, we define a semantic reversal operator $\mathcal{R}$ that preserves the problem and answer content while reversing the explanatory order:

$$\tilde{x} = \mathcal{R}(x) = (q, a, u_m, u_{m-1}, \dots, u_1).$$

This construction is not intended to create ungrammatical text; it perturbs which dependencies are available before predicting each step while keeping much of the sample content fixed. Token-level and sentence-level reversals are used only as control variants to test whether the signal is driven by linguistic artifacts rather than dependency direction.

3.3 Local Directional Gap Curve

Let p_θ be a fixed probe language model. For the forward trace $x = (s_1, \dots, s_T)$, the token-level forward encoding cost is

$$\ell_f(t) = -\log p_\theta(s_t | C_t^{\text{fwd}}), \quad C_t^{\text{fwd}} = s_{<t}. \quad (1)$$

For the semantically reversed trace $\tilde{x} = (\tilde{s}_1, \dots, \tilde{s}_T)$, the reverse-direction encoding cost is

$$\ell_r(t) = -\log p_\theta(\tilde{s}_t | C_t^{\text{rev}}), \quad C_t^{\text{rev}} = \tilde{s}_{<t}. \quad (2)$$

Because the forward and reversed traces have different token orders, we use an alignment function $\pi(\cdot)$ to map a forward token position to its corresponding position in the semantically reversed trace. In practice, π can be obtained from step-level alignment or anchor-based alignment; unless otherwise specified, we use anchor-based alignment for efficiency. For a window size w , let $W_t = \{\tau : |\tau - t| \leq w/2, 1 \leq \tau \leq T\}$ denote the local window centered at position t . We define the local directional gap as

$$\Delta(t; w) = \frac{1}{|W_t|} \sum_{\tau \in W_t} [\ell_r(\pi(\tau)) - \ell_f(\tau)]. \quad (3)$$

Equivalently, with $p_\tau^{\text{fwd}} = p_\theta(s_\tau \mid C_\tau^{\text{fwd}})$ and $p_\tau^{\text{rev}} = p_\theta(\tilde{s}_{\pi(\tau)} \mid C_{\pi(\tau)}^{\text{rev}})$,

$$\Delta(t; w) = \frac{1}{|W_t|} \sum_{\tau \in W_t} \log \frac{p_\tau^{\text{fwd}}}{p_\tau^{\text{rev}}}. \quad (4)$$

The sequence $\{\Delta(t; w)\}_{t=1}^T$ forms a *local directional gap curve* over the reasoning trace.

This statistic is a paired local log-likelihood-ratio gap under an order intervention. If a region is easy to encode in the forward direction but difficult under semantic reversal, $\Delta(t; w)$ becomes large; such high-gap regions are directional bottlenecks where dependency information is concentrated. Using a paired difference rather than reverse loss alone helps control for shared lexical content, domain familiarity, and general fluency. Under the latent dependency graph view, if the probe model p_θ approximates the data distribution, the expected local gap satisfies

$$\mathbb{E}_x[\Delta_i] \approx H(U_i \mid C_i^{\text{rev}}) - H(U_i \mid C_i^{\text{fwd}}), \quad (5)$$

where C_i^{fwd} and C_i^{rev} denote the information available before predicting U_i under the natural and reversed dependency-revealing orders, respectively. Thus, LOCAL-DELTA AVG estimates dependency-order sensitivity rather than absolute perplexity. For the global variant, let $P_R = \mathcal{R}_\# P$ be the distribution induced by a bijective reversal $\mathcal{R}$. Since a bijection preserves entropy, the expected global gap can be written as

$$\mathbb{E}[\Delta_{\text{global}}] = \text{KL}(P_R \| p_\theta) - \text{KL}(P \| p_\theta). \quad (6)$$

Even when $p_\theta = P$, semantic reversal need not imply $P_R = P$, so the gap can remain nonzero. Accordingly, the score is a probe-dependent measure of sensitivity to the specified order intervention, not a distribution-free guarantee of logical validity.

3.4 Trace-level Scoring and Selection

To use the local curve for data selection, we aggregate it into a trace-level score. Our default score is the peak directional gap:

$$D(x) = \max_t \Delta(t; w). \quad (7)$$

High-gap positions on the local curve are interpreted as directional bottlenecks. They are defined by a high contrastive gap under the order intervention, rather than by high absolute loss alone. In the main experiments, we use $S(x) = D(x)$ as the default selection score. The final selected continued-pretraining subset is therefore obtained by ranking all candidates in $\mathcal{D}$ according to $S(x)$.

3.5 Multi-resolution Variants

We use LOCAL-DELTA AVG to denote the method family. When the local window spans the entire trace, it yields a *global variant* ($w = \infty$), which is cheaper but less localized and is mainly used for coarse pre-filtering. Our default setting is the *local variant* ($w = 64$), while Appendix discusses a structured extension for long traces where fixed local windows may miss non-local dependencies.

4 Experiments

4.1 Experimental Setup

We evaluate LOCAL-DELTA AVG as a data-selection signal for reasoning-oriented continued pretraining, followed by an identical downstream SFT protocol. Unless otherwise specified, the candidate pool is a 20B-token STEM reasoning corpus, from which each selection method chooses 2B tokens under the same training budget. We use Qwen2.5-7B-Base (Yang et al., 2024) as the main backbone and evaluate on GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021b; Lightman et al., 2024), AIME24, GPQA-D (Rein et al., 2023), and LiveCodeBench (Jain et al., 2025), along with HellaSwag (Zellers et al., 2019) and MMLU (Hendrycks et al., 2021a) as non-reasoning control benchmarks. Baselines include random STEM sampling, length filtering, forward PPL/loss filtering, LLM-as-a-judge scoring, and smaller-budget uncertainty/diversity controls.

Reproducibility and Setup Details. Corpus construction, MinHash deduplication, 13-gram benchmark decontamination, baseline prompts/models, and semantic-reversal implementation details are provided in Appendix B, Appendix B.2, and Appendix B.3. All main comparisons use the same candidate pool, token budget, CPT setup, and SFT protocol, isolating the effect of the selection score.

4.2 Intrinsic Validation of the Directional Signal

Before using LOCAL-DELTA AVG for data selection, we first test whether the score behaves like a reasoning-structure signal rather than a surface artifact.

Quality Sanity. We first ask whether the directional gap tracks externally meaningful data quality rather than generic fluency. As summarized in Table 1, the score increases monotonically from random web text to curated and reasoning-enriched

Validation	Key result	Interpretation
Quality tiers	Random web 0.061 → reasoning-enriched 0.385	Monotonic increase with external quality
Controlled regression	Local-DeltaAvg score $\beta = 0.35$, $t = 8.45$	Strongest predictor beyond length/PPL
Cross-dataset bins	Lowest bin 0.085 → highest bin 0.450	Same trend transfers across datasets

Table 1: Feasibility sanity checks for LOCAL-DELTA AVG. Full versions of validations 0.1/0.2/0.3 are reported in Appendix C.

data: the random web control remains close to zero (0.061), whereas the reasoning-enriched tier reaches the largest gap (0.385). The same compact feasibility chain also shows that the Local-DeltaAvg score remains the strongest predictor beyond length and forward PPL in a controlled regression, and that the monotonic trend transfers to independently binned datasets. Therefore, the score exhibits the basic sanity property expected of a reasoning-data quality signal.

Artifact Rejection. We next test whether the score reflects reasoning structure rather than superficial corruption artifacts. When we progressively degrade the CoT text while keeping the task family fixed, the score drops monotonically from full CoT to compressed CoT, direct answers, and shuffled rationales (Figure 3a), indicating sensitivity to structured dependencies rather than merely to length or absolute difficulty. We then compare reversal constructions to rule out the possibility that the gap is driven by ungrammatical token sequences. Figure 3b shows that mechanical token-level reversal produces the largest raw gap but almost no correlation with human reasoning quality, whereas semantic reversal yields the strongest intrinsic correlation. Semantic reversal also gives the best downstream CPT result under the same selection protocol; full numbers are reported in Appendix F.2. Therefore, the useful signal is tied to perturbing dependency-revealing order semantically, not to generic reversal noise or token-level gibberish.

Step Alignment and Robustness. Finally, we test whether the local peaks are interpretable and reasonably stable. Table 2 shows that high-gap regions substantially overlap with annotated critical reasoning steps, with Local $w = 64$ reaching 0.43 IoU / 0.71 R@3 on PRM800K and similar results on MATH-Shepherd; full results are in Appendix E.3.

Since LOCAL-DELTA AVG relies on a fixed probe model, we also check robustness to scorer

Dataset	Method	IoU	R@3
PRM800K	Random	0.08	0.21
PRM800K	Local $w = 64$	0.43	0.71
MATH-Shepherd	Random	0.07	0.19
MATH-Shepherd	Local $w = 64$	0.41	0.69

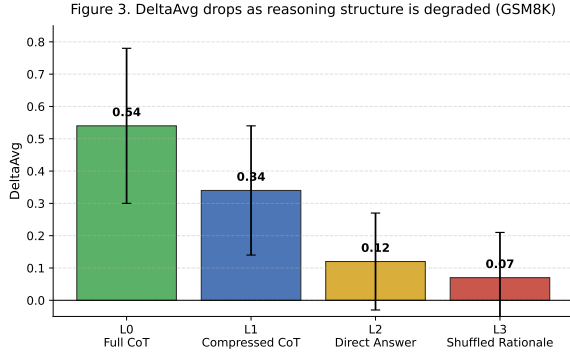
Table 2: Alignment between high-gap regions and annotated critical reasoning steps. Full metrics are reported in Appendix E.3.

choice. We find strong rank correlation among different probe architectures (Appendix G.1), suggesting that the ranking is not solely a probe-specific artifact. Additional confounder-controlled regressions against PPL, length, difficulty, and related variables are reported in Appendix F.3. Therefore, the local directional gap is not only interpretable at the step level, but also reasonably robust to probe choice and standard confounders.

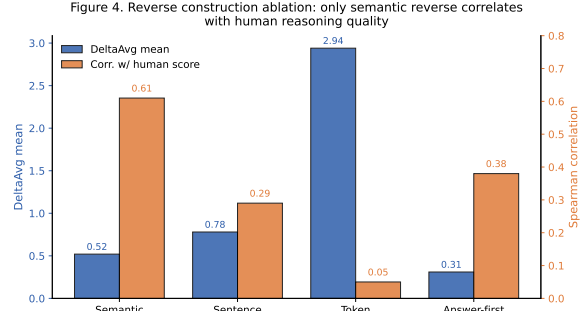
4.3 Main Results on Continued-Pretraining Data Selection

Statistical reporting. To assess training stochasticity at the main scale, we repeat the full 2B-token CPT comparison with three independent seeds and report the mean and standard deviation across runs. We separately report evaluation-level bootstrap confidence intervals over benchmark instances and use paired bootstrap tests when applicable; these intervals quantify evaluation uncertainty rather than CPT-run variance. We also retain the repeated 300M-token runs as a smaller-budget stability check.

We now evaluate whether the intrinsic directional signal translates into actual training utility under matched continued-pretraining and same-SFT protocols. Table 3 presents the main results. LOCAL-DELTA AVG achieves the best average performance, improving over Random-STEM by 2.6 points on average and over the LLM-judge baseline by 1.6 points on average. Importantly, on non-reasoning or less logic-intensive control tasks like HellaSwag (common-sense sentence completion) and MMLU (factual/knowledge retrieval), LOCAL-DELTA AVG yields flat or slightly lower performance. This selective behavior indicates that the score emphasizes directional structure rather than general linguistic fluency or domain familiarity. Table 4 complements the benchmark-wise results with the full-scale three-seed comparison, while the repeated 300M-token runs provide an additional smaller-budget stability check. The global



(a) Reasoning-structure degradation.



(b) Reversal construction ablation.

Figure 3: Intrinsic control experiments without training. (a) The score decreases as reasoning structure is compressed, removed, or shuffled. (b) Mechanical token reversal yields a large but uninformative gap, while semantic reversal best correlates with human reasoning quality.

variant already provides a strong cheap-coarse signal, while the local variant further improves performance by capturing localized directional bottlenecks. These results suggest that directional structurality provides a useful selection signal beyond standard surface-level heuristics.

4.4 Data Efficiency under Smaller Selection Budgets

We next test whether the selection advantage persists when the continued-pretraining budget is reduced. To focus strictly on the target reasoning capability, this and all subsequent ablation experiments report the average performance exclusively over the five primary reasoning benchmarks (excluding the two control tasks shown in Table 3). Table 5 shows that the local variant of LOCAL-DELTA AVG consistently outperforms Random-STEM, LLM-judge filtering, and the global variant from 500M to the main 2B-token setting. We include stronger uncertainty/diversity baselines at 500M due to compute constraints; their trends are consistent with the main-scale comparison. The gain is most pronounced at medium budgets, suggesting that directional structurality primarily improves data efficiency under fixed continued-pretraining compute. To estimate training stochasticity under feasible compute, we repeat the 300M-token setting with three CPT seeds. The ranking among methods remains stable. Additional component-level ablations are provided in Appendix E. Other ablations on cross-base transfer, confounder-controlled regressions, and an auxiliary SFT-only setting are reported in Appendix D.1, Appendix F.3, and Appendix D.3, respectively.

4.5 Directional Structurality Beyond Linear Chains

The main experiments instantiate LOCAL-DELTA AVG on linear reasoning traces, but the underlying notion of directional structurality is not restricted to a single chain. Many hard reasoning traces contain branching dependencies, where one premise supports multiple sub-claims or several steps reuse a shared prerequisite. In this setting, the same asymmetry principle should apply over dependency edges rather than only over local token windows. The core idea seamlessly extends from linear chains to branching trees and shared-dependency DAGs: useful reasoning traces reveal information more naturally along the forward dependency direction than under direction-perturbing alternatives.

To test this claim, we evaluate a structured extension on the Long-CoT, MATH-hard, and Olympiad-style subset of the 20B STEM candidate pool. As shown in Table 6, moving from the global variant to the local variant and then to the structured extension increases the after-Same-SFT performance in a 500M–1B-token targeted validation on the hard long-CoT subset. The gains are largest for the combined Local+TRCG/GRCG configuration, suggesting that edge-aware scoring captures cross-step directional structure that fixed local windows may miss. This is not evidence for a separate method family; rather, it shows that the same directional-structurality hypothesis remains useful when the latent dependency object is no longer well approximated by a chain.

We further test whether this gain is actually tied to branching structure rather than to scorer com-

Selection method	GSM8K	MATH500	AIME24	GPQA-D	LCB	HellaS.	MMLU	Avg.	Eval CI
Random-STEM	66.5	52.5	12.2	45.5	52.5	83.1	74.2	55.2	[54.3, 56.4]
Length filter	66.0	52.0	11.1	46.8	52.1	82.5	73.8	54.9	[53.8, 55.7]
Forward PPL / Loss	67.5	53.5	12.2	46.8	52.5	83.3	74.4	55.7	[54.6, 56.5]
LLM Judge / Quality Score	68.8	55.0	13.3	46.3	53.2	83.0	74.1	56.2	[55.3, 57.1]
LOCAL-DELTA AVG (global, $w = \infty$)	70.0	56.5	14.4	47.1	54.0	82.8	73.9	57.0	[56.1, 57.8]
LOCAL-DELTA AVG (local, $w = 64$)	71.5	58.5	16.7	47.3	54.0	82.6	73.8	57.8	[56.9, 58.9]

Table 3: Main continued-pretraining data-selection results on the primary reasoning and control benchmarks. LCB denotes LiveCodeBench, and HellaS. denotes HellaSwag. Each method selects the same number of tokens from the same STEM candidate pool and is trained with identical CPT and SFT settings. Benchmark-wise values are averaged across three independent CPT seeds; CIs are computed by paired bootstrap over evaluation examples. Table 4 reports the corresponding cross-seed mean and standard deviation after averaging each run over the five primary reasoning benchmarks.

Selection method	2B CPT, Mean $\pm$ Std.
Random-STEM	45.8 $\pm$ 0.2
Length filter	45.6 $\pm$ 0.2
Forward PPL / Loss	46.5 $\pm$ 0.2
LLM Judge / Quality Score	47.3 $\pm$ 0.2
LOCAL-DELTA AVG (global, $w = \infty$)	48.4 $\pm$ 0.3
LOCAL-DELTA AVG (local, $w = 64$)	49.6 $\pm$ 0.2

Table 4: Full-scale multi-seed validation under the 2B-token CPT budget. Values are the mean and standard deviation across three independent CPT seeds, with each run evaluated by its after-Same-SFT average over the five primary reasoning benchmarks.

plexity. Table 7 perturbs dependency edges before scoring and shows that both the structural score and downstream gain deteriorate as more edges are removed. Once all edges are dropped, performance falls back close to the disconnected step-level regime, indicating that the benefit comes from preserving directional dependencies across branches rather than from merely adding extra machinery.

Resolution choice. This analysis clarifies how the global, local, and structured variants should be used, rather than introducing separate competing methods. In practice, we use the local variant ($w = 64$) as the default for full-pool STEM selection: the global variant is useful for cheap pre-filtering, while the structured extension is reserved for long and complex traces. Detailed resolution trade-offs, including the short/medium/long-regime table and cost-quality analysis, are provided in Appendix E.4.

4.6 Robustness and Scaling

We additionally check whether the selection signal is robust to model architecture and confounder

controls. First, the gain transfers to other base models in targeted validation, including Llama-3.1-8B-Base and Qwen2.5-14B-Base. Second, rankings induced by the score are reasonably consistent across different probe models, suggesting that the metric is not tied to a single scorer. Third, confounder-controlled regressions show that the signal remains predictive after accounting for length, forward PPL, difficulty, and judge scores. Finally, the window-size stability check confirms $w = 64$ as a reliable default trade-off.

To keep the main evidence focused on the 20B-to-2B continued-pretraining setup, we report detailed cross-base results, cross-scorer robustness, confounder-controlled regressions, and window-size ablations in Appendix D.1, Appendix G.1, Appendix F.3, and Appendix E.

Beyond continued-pretraining data selection, we also briefly test whether the same directional signal can serve as an unsupervised inference-time reranking heuristic. As detailed in Appendix A, selecting the candidate with the largest score achieves performance competitive with PRM-based reranking, suggesting that the metric is also a lightweight unsupervised reasoning-quality estimator.

5 Conclusion

This work studies data selection for reasoning-oriented continued pretraining in language models. Instead of relying on surface-level signals such as length, perplexity, or problem difficulty, we ask whether the directional structure of a reasoning trace can reveal its training utility. Our central claim is that high-utility reasoning traces exhibit *directional irreversibility*: they incur lower conditional encoding cost when followed from premises to conclusions, but higher cost when semantically reversed from answers to supporting dependen-

Selection method	100M	300M (3 seeds)	500M	Role
Random-STEM	41.5	42.8 $\pm$ 0.4	44.2	random baseline
Length filter	41.2	42.6 $\pm$ 0.3	43.8	surface complexity
Forward PPL / Loss	42.0	43.4 $\pm$ 0.3	44.7	fluency / model familiarity
LLM Judge	42.4	43.8 $\pm$ 0.4	45.2	external quality preference
Token uncertainty (HES-style)	–	–	46.0	hard / uncertain instances
PPL + diversity	–	–	45.7	fluency w/ diverse coverage
LLM Judge + diversity	–	–	46.5	quality w/ diverse coverage
LOCAL-DELTA AVG (global, $w = \infty$)	43.8	45.3 $\pm$ 0.3	46.5	global directional score
LOCAL-DELTA AVG (local, $w = 64$)	44.6	46.1 $\pm$ 0.2	47.0	local bottleneck score
LOCAL-DELTA AVG (local, $w = 64$) + diversity	45.2	46.5 $\pm$ 0.2	47.6	avoid narrow high-gap selection

Table 5: Data efficiency under smaller selection budgets. By running at smaller pilot budgets (100M–500M tokens), we verify that the selection advantage holds across different compute scales and outperforms a wider range of standard heuristics. Stronger uncertainty/diversity baselines are included at 500M only due to compute constraints.

Selection method	Applied subset	Tokens	Cost	After Same-SFT Avg.	Role
Random-long	Long-CoT / Hard	500M–1B	0	39.0	Long-subset baseline
Local $w = \infty$	Long-CoT / Hard	500M–1B	Low	41.0	Sequence-level direction gap
Local $w = 64$	Long-CoT / Hard	500M–1B	Low+	43.0	Local bottleneck curve
TRCG/GRCG only	Long-CoT / Hard	500M–1B	Med	43.5	Structural edge gaps
Local + TRCG/GRCG	Long-CoT / Hard	500M–1B	Med	45.0	Recommended for hard traces

Table 6: Directional structurality in long-complex reasoning traces. This 500M–1B-token experiment on hard long-CoT traces validates that the same directional-structurality framework generalizes to branching dependencies. TRCG/GRCG extends LOCAL-DELTA AVG from local token bottlenecks to dependency-edge directional gaps.

Edge drop	TRCG mean	CPT Avg.
0%	0.62	41.8
10%	0.55	40.9
20%	0.46	39.5
50%	0.28	36.8
100%	–	41.1

Table 7: Compact edge-perturbation validation for the structured extension. As branching dependency edges are removed, both the structural score and downstream gain deteriorate, linking the advantage to preserved directional structure rather than to scorer complexity.

cies. Based on this claim, we introduce LOCAL-DELTA AVG, a probe-based metric that measures local forward–semantic-reversal NLL gaps and forms a local directional gap curve over each trace. The framework admits a cheaper global variant ($w = \infty$) and also extends naturally to branching dependency structures for long and complex traces. Within the evaluated reasoning-oriented continued-pretraining setting, LOCAL-DELTA AVG consistently outperforms standard selection baselines and aligns high-gap regions with critical reasoning steps; our additional experiments also support unsupervised best-of- N reranking. These results suggest that directional encoding asymmetry provides a lightweight, unsupervised signal for ranking candidate reasoning traces, with potential

utility both before training and at inference time.

Limitations

Our method has several limitations. First, LOCAL-DELTA AVG assumes that reasoning traces can be segmented into meaningful intermediate steps. This assumption generally holds for mathematical solutions, code reasoning, and structured chain-of-thought data, but it may be less reliable for implicit, poorly formatted, or highly compressed rationales. When step boundaries are ambiguous, semantic reversal may introduce noise because the answer-first, step-reversed construction no longer corresponds to a clean reversal of explanatory dependencies. Improving step segmentation and reversal construction for unstructured reasoning traces remains an important direction for future work.

Second, the score depends on a fixed probe language model. Although our cross-scorer experiments show that rankings are relatively stable across different probes, the absolute magnitude of the directional gap may still vary with tokenizer, model scale, domain familiarity, and language coverage. The local formulation also introduces design choices such as the window size and alignment strategy. We use $w = 64$ and anchor-based alignment as efficient defaults, but different domains or

reasoning granularities may require retuning.

Third, the tree-structured extension relies on reasoning dependency parsing. For long and complex traces, true dependencies may form directed acyclic graphs rather than trees, and parser errors can directly affect edge-level directional gaps. In addition, LOCAL-DELTA AVG is more computationally expensive than simple heuristics such as length or forward perplexity because it requires scoring both forward and semantically reversed traces and computing local-window gaps. While this cost is lower than training supervised verifiers or repeatedly querying strong LLM judges, large-scale deployment may still require batching, caching, or a two-stage filtering pipeline.

Finally, our experiments focus mainly on structured reasoning tasks, including mathematical, logical, coding, and knowledge-intensive reasoning. It remains unclear whether directional irreversibility is equally informative for open-ended generation, dialogue, creative writing, or multilingual reasoning. Moreover, selecting only high-gap examples may reduce data diversity if such examples concentrate in a narrow set of reasoning patterns. In practice, Local-DeltaAvg should therefore be combined with diversity constraints or domain balancing when constructing large SFT datasets.

References

- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. [Alpagasus: Training a better alpaca with fewer data](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. [Openthoughts: Data recipes for reasoning models](#). *CoRR*, abs/2506.04178.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. [Data-models: Understanding predictions with data and data with predictions](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. [Live-codebench: Holistic and contamination free evaluation of large language models for code](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. [From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7602–7635. Association for Computational Linguistics.
- Xiaoyuan Li, Yubo Ma, Chengpeng Li, Fengbin Zhu, Yiyao Yu, Keqin Bao, Wenjie Wang, Fuli Feng, and Dayiheng Liu. 2026. [Unified data selection for LLM reasoning](#). *CoRR*, abs/2605.22389.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Driani, Julian Michael, and Samuel R. Bowman. 2023. [GPQA: A graduate-level google-proof q&a benchmark](#). *CoRR*, abs/2311.12022.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- Gemma Team. 2024a. [Gemma 2: Improving open language models at a practical size](#). *CoRR*, abs/2408.00118.
- Llama Team. 2024b. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. [Solving math word problems with process- and outcome-based feedback](#). *CoRR*, abs/2211.14275.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. [Math-shepherd: Verify and reinforce llms step-by-step without human annotations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9426–9439. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. [Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [Hellaswag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A Inference-time Reranking

Although our main focus is continued-pretraining data selection, we briefly test whether the same directional signal can also serve as an unsupervised reranking heuristic. Table 8 shows that selecting the candidate with the largest $D(x)$ achieves performance close to PRM-based reranking, without training a process reward model. When the alignment score $R(x)$ is also used, the reranking result achieves performance competitive with PRM-based reranking in this setting.

B Additional Experimental Details

B.1 Candidate Pools and Preprocessing

Table 9 summarizes the main data sources used in our experiments. For continued-pretraining data selection, we use a 20B-token STEM reasoning candidate pool and select 2B-token subsets for reasoning-oriented continued pretraining. For sanity checks and interpretability analyses, we additionally use externally stratified data and process-supervision datasets. Dataset sources include OpenThoughts3 (Guha et al., 2025), Magpie (Xu et al., 2025), GSM8K (Cobbe et al., 2021), PRM800K (Lightman et al., 2024), and MATH-Shepherd (Wang et al., 2024).

B.2 Training and Evaluation Details

The main continued-pretraining experiment uses a 2B-token selected budget and the same down-

Reranking method	N	MATH500	AIME24
Greedy	1	74.5	17.8
Majority vote	8	78.5	23.3
PRM-based reranking	8	81.1	26.7
Local $D(x)$ reranking	8	80.8	25.6
Local $D(x) + R(x)$	8	81.5	27.8

Table 8: Inference-time best-of- N reranking. LOCAL-DELTA AVG provides an unsupervised reranking signal competitive with PRM-based reranking.

Dataset	Size	Usage	Purpose
STEM reasoning pool	20B tokens	Candidate pool	Continued-pretraining data selection
Nemotron-CC tiers	20K	Sanity check	External quality stratification
Magpie / OpenThoughts bins	20K	Sanity check	Cross-dataset validation
GSM8K train subset	5000	Control	Reasoning-structure degradation
Wikipedia control	5000	Control	Non-reasoning negative control
PRM800K	5000	Alignment	Critical-step localization
MATH-Shepherd	5000	Alignment	Cross-dataset step alignment
Long-CoT / hard subset	500M–1B tokens	Structured reasoning	TRCG/GRCG validation

Table 9: Summary of datasets used in the experiments. The main continued-pretraining candidate pool contains 20B STEM reasoning tokens; other datasets are used for sanity checks, controls, robustness, and interpretability.

stream SFT protocol for every selection method. The full-scale comparison uses three independent CPT seeds; Table 4 reports the mean and standard deviation across runs. Smaller-budget pilots and targeted validation runs are explicitly marked in their respective tables. The main full-scale backbone is Qwen2.5-7B-Base, with Llama-3.1-8B-Base and Qwen2.5-14B-Base used only for targeted cross-base validation. We evaluate on GSM8K, MATH500, AIME24, GPQA-D, and LiveCodeBench. CPT is run once per selected subset unless explicitly stated; after-Same-SFT evaluation is averaged over three seeds when applicable.

B.3 Scoring Implementation

For each reasoning trace, we construct a semantically reversed trace by moving the final answer before the explanation and reversing the order of reasoning steps while preserving the internal token order of each step. The same fixed probe language model scores both the forward and semantically reversed traces. The default sample score is the peak local directional gap $D(x)$.

C Additional Sanity Checks

C.1 External Quality Stratification

Table 12 reports the full version of the feasibility check summarized in the main text. We compute DeltaAvg independently for each example in externally stratified Nemotron-CC tiers. The score increases monotonically with external data qual-

ity, and adjacent tier differences are statistically significant under Welch’s two-sided t -test.

C.2 Length- and PPL-controlled Sanity Regression

Table 13 reports the second feasibility check: a regression analysis on the Nemotron-CC quality tiers. The target variable is the external quality tier. Even after controlling for length and forward perplexity, DeltaAvg remains the strongest predictor.

C.3 Cross-dataset Quality-bin Validation

Table 14 reports the third feasibility check, validating the same monotonic trend on an independent quality-binned dataset. The directional gap increases from the lowest-quality bin to the highest-quality bin.

C.4 Distribution across Data Types

Figure 4 visualizes the distribution of DeltaAvg across descriptive web text, curated data, and reasoning corpora. Descriptive text remains near zero, whereas reasoning-oriented data shows substantially higher directional gaps.

D Extended Data Selection Results

D.1 Targeted Cross-base Same-SFT Results

Table 15 provides a targeted cross-base validation for the continued-pretraining setup. To avoid turning the robustness check into a second full-scale benchmark suite, we compare only Random-STEM

Item	Value
Main base model	Qwen2.5-7B-Base
Probe model	Qwen2.5-1.5B-Base
Candidate pool	20B STEM reasoning tokens
Selected budget	2B tokens
Sequence length	4096
Global batch size	4M tokens
Training steps	500
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Peak learning rate	2×10^{-5}
Warmup steps	50
LR schedule	Cosine decay to 10% of peak
Precision	bfloat16
Hardware	8 $\times$ H100 80GB GPUs
SFT data	$\sim 50K$ samples ($\sim 20M$ tokens)
Evaluation decoding	Greedy (temperature = 0)
Evaluation seeds	3 (for SFT and evaluation)

Table 10: Main training configuration for continued pretraining and evaluation.

Component	Implementation
Forward trace	Original reasoning trace order
Semantic reversal	Answer-first, step-reversed construction
Step internal order	Preserved
Token-level reversal	Used only as a control variant
Sentence-level reversal	Used only as a control variant
Default window	$w = 64$
Default score	$D(x) = \max_t \Delta(t; w)$
Auxiliary features	$V(x), K(x), R(x)$

Table 11: Implementation details of semantic reversal and Local-DeltaAvg scoring.

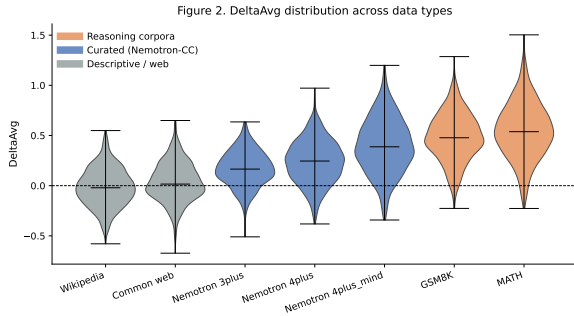


Figure 4: DeltaAvg distribution across data types. Reasoning corpora and reasoning-enriched curated data exhibit larger directional gaps than descriptive web text.

with the main LOCAL-DELTA AVG ($w = 64$) configuration on each additional base.

D.2 Full Main-result Visualization

Figure 5 visualizes the main data-filtering result on Qwen2.5-7B-Base. The main text reports the aggregate midtrain-only and after-SFT results under the 2B-token selection budget.

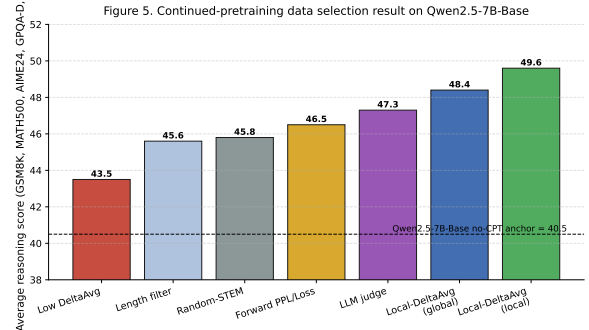


Figure 5: Visualization of the main continued-pretraining data filtering result on Qwen2.5-7B-Base.

D.3 SFT-only Auxiliary Results

Table 16 reports the auxiliary SFT-only setting. This experiment is not used as the main evidence for reasoning ability acquisition, but shows that the same signal can also transfer to instruction-tuning data selection.

E Additional Method Ablations

E.1 Curve Feature Ablation

Table 17 studies auxiliary features derived from the local directional gap curve. The default method

Data tier	N	Mean	Std.	95% CI	Avg. tokens	p -value
Random web	5000	0.061	0.200	[0.055, 0.067]	565	–
Nemotron 3plus	5000	0.155	0.220	[0.149, 0.161]	529	< 0.001
Nemotron 4plus	5000	0.240	0.250	[0.233, 0.247]	541	< 0.001
Nemotron 4plus_mind	5000	0.385	0.280	[0.377, 0.393]	618	< 0.001

Table 12: Feasibility validation 0.1: external quality stratification sanity check. The p -value is computed against the immediately lower-quality tier.

Predictor	Std. coef. β	t -stat.	p -value
Length	0.08	2.15	0.032
Forward PPL	-0.15	-3.82	< 0.001
DeltaAvg	0.35	8.45	< 0.001

Table 13: Feasibility validation 0.2: length- and PPL-controlled sanity regression. DeltaAvg remains the strongest independent predictor of external quality tier.

uses $D(x)$, while $V(x)$, $K(x)$, and $R(x)$ provide complementary information. This ablation is run as a 500M-token pilot rather than as additional full 2B-token CPT runs.

E.2 Alignment Method Ablation

Table 18 compares alignment strategies for mapping forward and semantically reversed traces into a shared coordinate system. Anchor alignment is used as the default because it offers the best cost-quality trade-off.

E.3 Critical-Step Alignment

Table 19 reports the full alignment analysis between high-gap regions and annotated critical reasoning steps.

E.4 Resolution Trade-off

Table 20 compares three resolutions of the same directional irreversibility framework. The best configuration depends on reasoning complexity.

E.5 Resolution Contribution and Cost Analysis

Table 21 decomposes the marginal contribution of stacking higher-resolution signals. Table 22 reports selection overlap between configurations, and Table 23 summarizes the cost-quality Pareto behavior. These analyses support the use of $w = 64$ as the main full-pool configuration and TRCG/GRCG as a targeted extension for long-structure subsets.

E.6 Window-size Ablation

For completeness, Table 24 reports the full window scan. This table is a 500M-token pilot used to

choose the default window size before the main 2B-token run.

F Control and Artifact Analyses

F.1 Reasoning Intensity Full Results

Table 25 shows that the directional gap decreases as the reasoning structure is compressed, removed, or shuffled. Table 26 further shows that after-Same-SFT performance follows the same trend.

F.2 Additional Reversal Construction Controls

Table 27 reports the downstream counterpart of the intrinsic reversal-construction analysis in the main text. Table 28 shows that semantic reversal is the only construction that yields a high correlation with human reasoning scores. Table 29 shows that the same construction also yields the strongest downstream gain under the same evaluation protocol.

F.3 Length, PPL, Difficulty, and Judge Confounders

Table 30 and Figure 6 show that the signal remains predictive after controlling for standard confounders. The dependent variable is midtrain gain under the same continued-pretraining and same-SFT protocol.

G Robustness and Generalization

G.1 Cross-scorer Robustness

We compute scores using several open-weight probe models, including Qwen2.5 (Yang et al., 2024), Llama-3.1 (Team, 2024b), Mistral-7B (Jiang et al., 2023), and Gemma2 (Team, 2024a). Table 31 reports pairwise correlations for

Quality bin	N	Mean	Std.	Avg. tokens
Q4, lowest quality	5000	0.085	0.210	450
Q3, medium-low	5000	0.160	0.225	480
Q2, medium-high	5000	0.285	0.245	510
Q1, highest quality	5000	0.450	0.280	540

Table 14: Feasibility validation 0.3: cross-dataset sanity validation on Magpie / OpenThoughts quality bins.

Base model	Tokens	Random-STEM	Local $w = 64$	Δ
Qwen2.5-7B-Base	2B	45.8	49.6	+3.8
Llama-3.1-8B-Base	2B	39.2	44.1	+4.9
Qwen2.5-14B-Base	2B	53.0	57.1	+4.1

Table 15: Targeted cross-base validation after the same SFT protocol. Only Random-STEM and the main Local-DeltaAvg configuration are compared on additional bases.

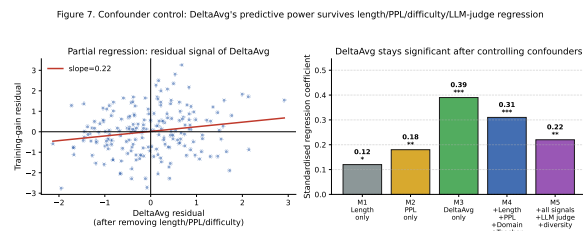


Figure 6: Partial-regression and confounder-control analysis.

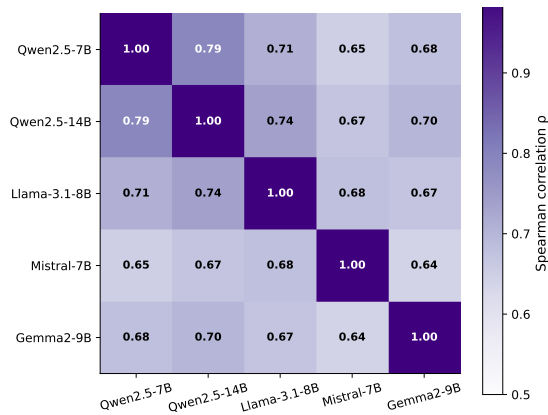


Figure 7: Spearman correlation of DeltaAvg rankings across scoring models ($N = 2000$).

the sequence-level degeneration. Table 32 reports the same analysis for the local peak score $D(x)$, and Table 33 reports agreement in bottleneck locations. Figure 7 visualizes scorer consistency for the local score.

G.2 Reverse-generator Robustness

Figure 8 reports agreement across different semantic-reversal generators. This analysis is relevant when semantic reversals are produced by LLM-based rewriting rather than purely rule-based

Figure 10: Spearman correlation of DeltaAvg across reverse generators ($N=200$)

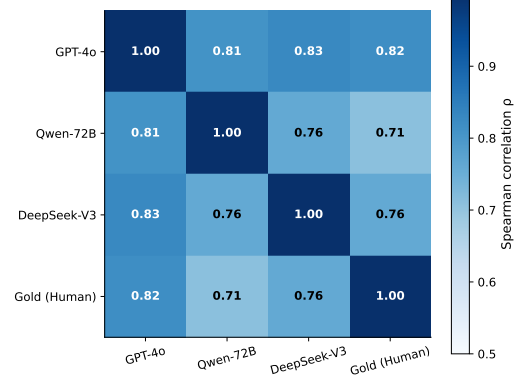


Figure 8: Spearman correlation of DeltaAvg rankings across semantic-reversal generators.

step reversal.

H Long and Structured Reasoning

H.1 Synthetic DAG Validation

Figure 9 reports synthetic DAG validation. DeltaAvg increases with directional strength and accumulates with chain length.

I Interpretability and Case Studies

I.1 Additional Critical-step Alignment Results

Table 34 reports the complete critical-step alignment results. Local-DeltaAvg at $w = 64$ achieves the strongest alignment.

I.2 Full Case Studies

Table 35 summarizes the case-study coverage, and Figure 10 shows the corresponding local directional gap curves with annotated critical-step spans.

Selection method	Candidate pool $\rightarrow$ SFT data	Base model	Budget	Eval Avg.
Random	50K $\rightarrow$ 10K samples	Qwen2.5-7B-Base	Equal tokens	42.8
Local $w = \infty$	50K $\rightarrow$ 10K samples	Qwen2.5-7B-Base	Equal tokens	47.0
Local $w = 64$	50K $\rightarrow$ 10K samples	Qwen2.5-7B-Base	Equal tokens	47.8

Table 16: Auxiliary SFT-only data-selection results.

Selection signal	Pilot tokens	After Same-SFT Avg.	p vs Rand.	p vs D
Random-STEM	500M	44.2	–	–
$D(x)$ only	500M	47.0	< 0.001	–
$D(x) + V(x)$	500M	47.4	< 0.001	0.24
$D(x) + K(x)$	500M	47.6	< 0.001	0.11
$D(x) + R(x)$	500M	47.9	< 0.001	0.041
$D(x) + V(x) + K(x) + R(x)$	500M	48.1	< 0.001	0.019

Table 17: Ablation of curve-derived features. The peak gap $D(x)$ is the main selection signal; auxiliary curve features provide additional gains. This table is a smaller-budget pilot.

I.3 Qualitative High/Low-score Schema

Figure 11 provides a qualitative schema contrasting high- and low-score samples.

J Inference-time Reranking and Additional Applications

J.1 Best-of- N Reranking Details

Table 36 reports best-of- N reranking results. Local-DeltaAvg provides an unsupervised reranking signal close to PRM-based reranking.

J.2 Additional Application Scenarios

Figure 12 summarizes three downstream scenarios, including continued-pretraining data selection, RL prompt selection (Shao et al., 2024), and inference-time reranking. We keep this figure in the appendix because the main paper focuses on continued-pretraining data selection and inference-time reranking.

K Additional Human and Causal Utility Validation

K.1 Human Annotation Validation

Figure 13 reports the correlation between DeltaAvg and human reasoning-quality annotations. The figure also shows that DeltaAvg adds predictive power beyond LLM-judge scores.

K.2 DataModels Marginal Contribution

Figure 14 reports DataModels-style evidence (Ilyas et al., 2022) that DeltaAvg predicts sample marginal contributions. High-score samples have larger training utility.

Alignment method	Avg. length	After Same-SFT Avg.	IoU	Relative cost
Anchor alignment	1,840	43.0	0.43	1.00×
Step-level alignment	1,840	43.5	0.48	1.18×
DTW alignment	1,840	43.2	0.46	1.52×
Naive sliding	1,840	41.0	0.31	0.92×

Table 18: Alignment-method ablation on the Long-CoT subset. Step-level alignment gives a small upper-bound gain, while anchor alignment is the most efficient default.

Dataset	Method	IoU	R@3	Pos. ρ
PRM800K	Random position	0.08	0.21	-0.02
PRM800K	Local $w = 32$	0.34	0.62	0.41
PRM800K	Local $w = 64$	0.43	0.71	0.49
PRM800K	Local $w = 128$	0.39	0.68	0.45
MATH-Shepherd	Random position	0.07	0.19	-0.01
MATH-Shepherd	Local $w = 64$	0.41	0.69	0.47

Table 19: Alignment between high-gap regions and critical reasoning steps. R@3 denotes recall using the top three high-gap regions.

Data type	Recommended config.	Local $w = \infty$	Local $w = 64$	TRCG/GRCG	Local+TRCG/GRCG	Comment
Short chain / GSM8K-like	Local $w = 64$	47.8	48.3	47.8	48.3	Graph adds little
Medium chain / MATH	Local $w = 64$	47.3	48.8	48.3	48.9	Local windows help
Long complex / Olympiad	Local+TRCG/GRCG	41.0	43.0	43.5	45.0	Edge gaps help
Mixed full STEM pool	Local $w = 64$	48.4	49.6	–	–	Default setting

Table 20: Resolution trade-off across data types. Sequence-level, local-window, and TRCG/GRCG scoring correspond to different resolutions of the same framework.

Stack configuration	Candidate pool $\rightarrow$ selected	Scorer cost	After Same-SFT	Δ	p -value
Random-STEM	20B $\rightarrow$ 2B	0	45.8	–	–
Local ($w = \infty$) only	20B $\rightarrow$ 2B	1.0×	48.4	+2.6	< 0.01
$w = \infty \rightarrow$ Local ($w = 64$)	20B $\rightarrow$ subset $\rightarrow$ 2B	1.05×	49.6	+1.2	< 0.01
Long-only Local + TRCG/GRCG	Long-CoT subset $\rightarrow$ 500M	1.35×	45.0	+1.5	< 0.01

Table 21: Marginal contribution breakdown under sequential stacking. The Δ is relative to the preceding row. Full-pool results stop at the main $w = 64$ configuration; TRCG/GRCG is evaluated only as a long-subset pilot.

Method A	Method B	Top-20% overlap	Complementarity
Local $w = \infty$	Length filter	28%	High
Local $w = \infty$	LLM Judge	42%	Medium
Local $w = \infty$	Local $w = 64$	62%	Low-Medium
Local $w = 64$	TRCG/GRCG	51%	Medium
Local $w = \infty$	TRCG/GRCG	36%	High

Table 22: Selection complementarity matrix. Different resolutions and baselines select largely distinct subsets.

Configuration	Cost	Recommended use case
Length filter	< 0.1×	Cheap coarse filtering for 10B+ pools
LLM Judge	100×	Expensive baseline for ≤ 50 K pools
Local $w = \infty$	1.0×	Coarse screening for 1M–1B samples
Local $w = 64$	1.05×	Main fine selection for 100K–1M samples
Local+TRCG/GRCG	1.35×	Structural filtering for Long-CoT subsets

Table 23: Cost-quality Pareto recommendations.

Window	Cost	After Same-SFT Avg.	IoU	R@3	p vs $w = \infty$
$w = 8$	$1.12\times$	45.1	0.28	0.54	0.19
$w = 16$	$1.08\times$	46.0	0.36	0.65	0.08
$w = 32$	$1.06\times$	46.8	0.41	0.69	0.041
$w = 64$	$1.05\times$	47.0	0.43	0.71	0.018
$w = 128$	$1.04\times$	46.9	0.39	0.68	0.055
$w = 256$	$1.02\times$	46.1	0.31	0.55	0.21
$w = \infty$	$1.00\times$	46.5	–	–	–

Table 24: Full window-size scan. The best overall trade-off is obtained at $w = 64$. All entries are from the 500M-token pilot budget.

Version	Construction	N	Avg _{fwd}	Avg _{rev}	DeltaAvg	Drop
L0 Full CoT	Complete step-by-step reasoning	5000	1.92	2.46	0.54	0%
L1 Compressed CoT	Keep key steps, remove explanation	5000	1.84	2.18	0.34	-37.0%
L2 Direct Answer	Keep only answer / short explanation	5000	1.71	1.83	0.12	-77.8%
L3 Shuffled Rationale	Shuffle reasoning steps	5000	2.21	2.28	0.07	-87.0%

Table 25: DeltaAvg response to reasoning-structure degradation.

Training version	DeltaAvg	After Same-SFT Avg.	Drop
L0 Full CoT	0.54	46.4	0
L1 Compressed CoT	0.34	44.5	-1.9
L2 Direct Answer	0.12	42.8	-3.6
L3 Shuffled Rationale	0.07	40.7	-5.7

Table 26: Downstream performance under different reasoning-trace completeness levels.

Reversal construction	Human corr.	CPT Avg.
Semantic reversal	0.32	49.6
Sentence reversal	0.18	47.5
Token reversal	0.05	46.2
Answer-first / Shuffle	-0.02	45.8

Table 27: Downstream CPT selection performance (Avg. over 5 benchmarks) for different reversal constructions, compared with their intrinsic human correlation.

Reversal method	Data	N	DeltaAvg	Human corr.	Pass rate	Main issue
Semantic reversal	GSM8K	1000	0.52	0.61	91%	Recommended
Sentence reversal	GSM8K	1000	0.78	0.29	–	Discourse artifact
Token reversal	GSM8K	1000	2.94	0.05	–	Syntax artifact
Shuffle rationale	GSM8K	1000	0.18	0.07	–	Randomized structure
Answer-first only	GSM8K	1000	0.31	0.38	94%	Insufficient reversal
Semantic reversal	Wikipedia	1000	-0.03	0.04	93%	Weak directionality
Sentence reversal	Wikipedia	1000	0.27	0.02	–	False positive

Table 28: Correlation between different reversal constructions and human reasoning scores. Token reversal creates the largest raw gap but does not correlate with reasoning quality.

Reversal method	After Same-SFT Avg.	Std.	p vs Rand.
Random baseline	42.8	0.41	–
Semantic reversal	46.7	0.35	< 0.001
Sentence reversal	43.6	0.47	0.062
Token reversal	42.2	0.52	0.51
Shuffle rationale	43.1	0.44	0.58

Table 29: Downstream gains under different reversal constructions.

Model	Controls	Core variable	Target	Std. coef.	<i>p</i> -value
M1	None	Length	Midtrain gain	0.08	0.035
M2	None	Avg _{fwd} / PPL	Midtrain gain	-0.12	0.002
M3	None	Local-DeltaAvg $w = \infty$	Midtrain gain	0.32	< 0.001
M4	Length + Avg _{fwd} + Domain + Teacher	Local-DeltaAvg $w = \infty$	Midtrain gain	0.25	< 0.001
M5	Length + PPL + difficulty + LLM judge	Local-DeltaAvg $w = \infty$	Midtrain gain	0.18	< 0.001

Table 30: Confounder-controlled regression. Local-DeltaAvg remains significant after controlling for length, PPL, difficulty, LLM-judge score, domain, and teacher source.

Scorer A	Scorer B	<i>N</i>	Pearson <i>r</i>	Spearman ρ	Top-20% overlap
Qwen2.5-7B-Base	Llama-3.1-8B-Base	2,000	0.74	0.71	58%
Qwen2.5-7B-Base	Mistral-7B-v0.3	2,000	0.68	0.65	51%
Llama-3.1-8B-Base	Gemma2-9B	2,000	0.70	0.67	54%
Qwen2.5-7B-Base	Qwen2.5-14B-Base	2,000	0.81	0.79	64%

Table 31: Cross-scorer robustness for Local-DeltaAvg at $w = \infty$.

Scorer A	Scorer B	<i>N</i>	Pearson <i>r</i>	Spearman ρ	Top-20% overlap
Qwen2.5-7B-Base	Llama-3.1-8B-Base	2,000	0.69	0.67	54%
Qwen2.5-7B-Base	Mistral-7B-v0.3	2,000	0.64	0.62	48%
Llama-3.1-8B-Base	Gemma2-9B	2,000	0.66	0.64	51%
Qwen2.5-7B-Base	Qwen2.5-14B-Base	2,000	0.78	0.76	61%

Table 32: Cross-scorer robustness for the Local-DeltaAvg peak score $D(x)$ at $w = 64$.

Scorer A	Scorer B	<i>N</i>	$\Delta(t)$ trace-level Spearman ρ	t^* Top-1 agreement
Qwen2.5-7B-Base	Llama-3.1-8B-Base	2,000	0.55	47%
Qwen2.5-7B-Base	Mistral-7B-v0.3	2,000	0.51	43%
Qwen2.5-7B-Base	Qwen2.5-14B-Base	2,000	0.66	56%

Table 33: Cross-scorer consistency of bottleneck curves and peak locations.

Figure 9. Validation on synthetic DAG: DeltaAvg tracks direction strength and length accumulation

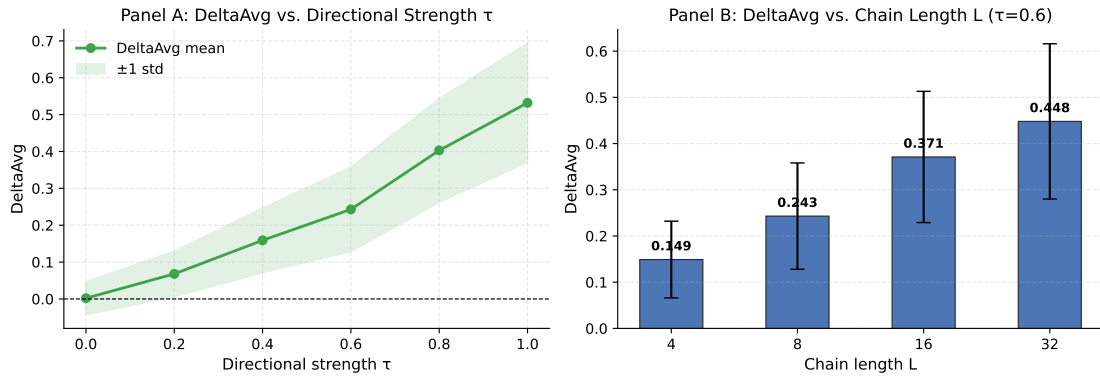


Figure 9: Synthetic DAG validation. Panel A varies directional strength τ ; Panel B varies chain length L .

Dataset	Method	<i>N</i>	IoU	R@3	Pos. ρ
PRM800K	Random t^*	5000	0.08	0.21	-0.02
PRM800K	Local $w = \infty$	5000	—	—	—
PRM800K	Local $w = 32$	5000	0.34	0.62	0.41
PRM800K	Local $w = 64$	5000	0.43	0.71	0.49
PRM800K	Local $w = 128$	5000	0.39	0.68	0.45
MATH-Shepherd	Random t^*	5000	0.07	0.19	-0.01
MATH-Shepherd	Local $w = 64$	5000	0.41	0.69	0.47

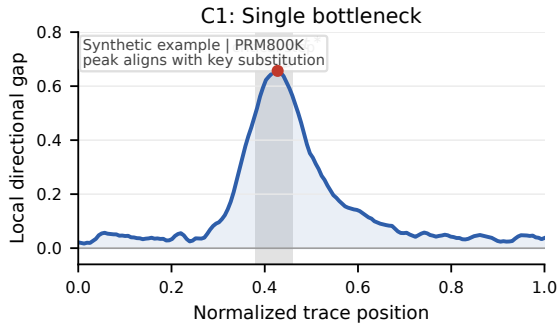
Table 34: Critical-step alignment results. R@3 denotes recall using the top three high-gap regions.

Case	Source	Type	Curve pattern	Critical-step relation
C1	PRM800K	High-quality multi-step math	Sharp single peak	Peak hits key substitution
C2	PRM800K	High-quality long reasoning	Multi-peak curve	Peaks align with multiple steps
C3	PRM800K	Short direct answer	Flat curve	No critical-step span
C4	MATH-Shepherd	Incorrect reasoning trace	Peak before error	Bottleneck marks divergence
C5	OpenThoughts3	High $D(x)$, low $R(x)$	Peak on filler span	Position filter rejects it
C6	OpenThoughts3	Low $D(x)$	Flat curve	Negative example
C7	GSM8K	Simple chain	Shallow broad peak	Simple local dependency
C8	OlympiadBench	Long complex logic	Multi-peak, high variance	Long-structure case

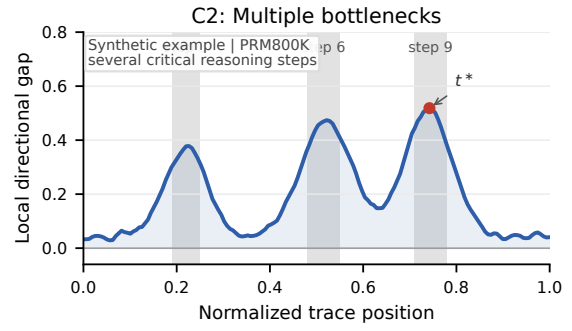
Table 35: Qualitative case-study schema.

Reranking method	N	MATH500	AIME24
Greedy	1	74.5	17.8
Majority vote	8	78.5	23.3
PRM-based reranking	8	81.1	26.7
Local $D(x)$ reranking	8	80.8	25.6
Local $D(x) + R(x)$	8	81.5	27.8

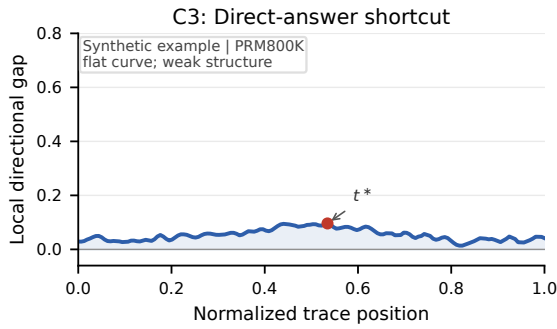
Table 36: Best-of- N reranking results. Local-DeltaAvg is competitive with PRM-based reranking without training a process reward model.



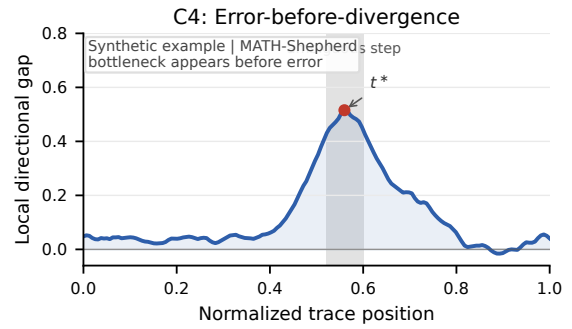
(a) C1: Correct long-chain reasoning.



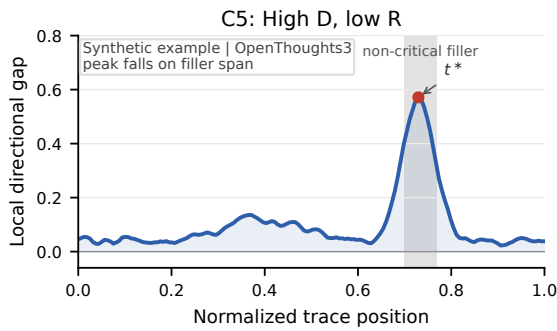
(b) C2: Redundant but structured reasoning.



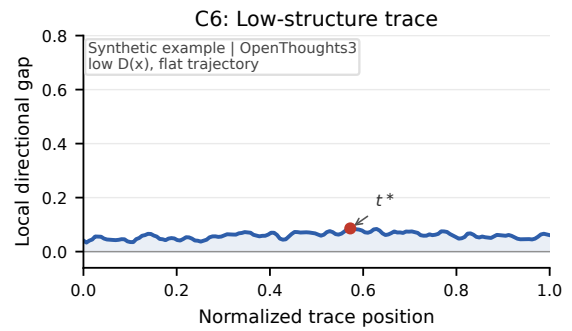
(c) C3: Direct-answer shortcut.



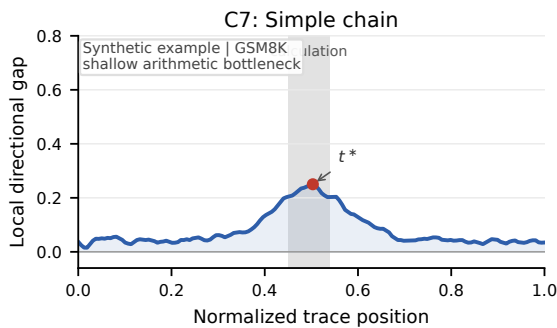
(d) C4: Key dependency introduction.



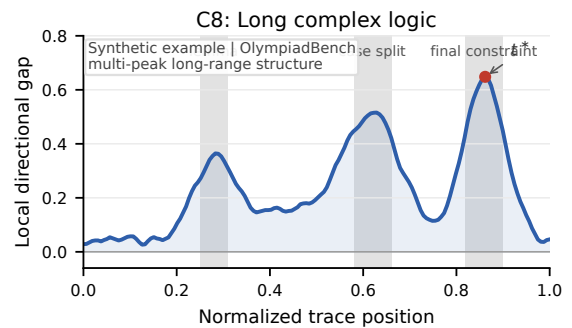
(e) C5: Shuffled or incoherent rationale.



(f) C6: Flat factual description.



(g) C7: Noisy reasoning trace.



(h) C8: Long complex structured trace.

Figure 10: Full case-study set. Each case visualizes the local directional gap curve and its relationship to critical reasoning steps.

Figure 8. Case Study Schema: High DeltaAvg (Left) vs Low DeltaAvg (Right)

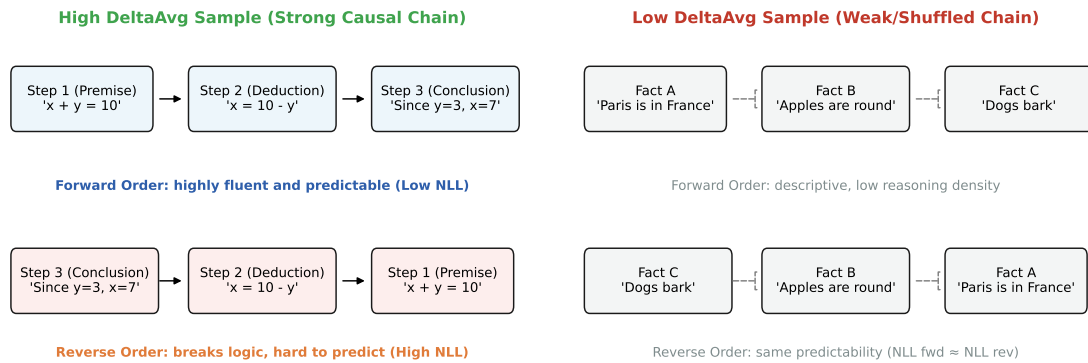


Figure 11: Qualitative schema for high- and low-DeltaAvg samples. This figure is used only as an illustrative example; the main interpretability evidence comes from the local curve and critical-step alignment results.

Figure 14. DeltaAvg consistently outperforms baselines in three downstream reasoning scenarios

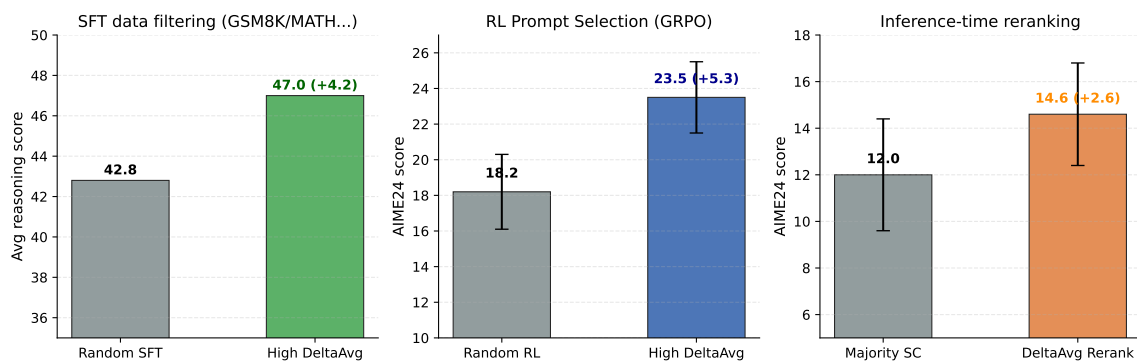


Figure 12: Additional downstream scenarios. The figure summarizes continued-pretraining data filtering, RL prompt selection, and inference-time reranking.

Figure 11. Human annotation validation: DeltaAvg correlates with human judgment and adds independent value

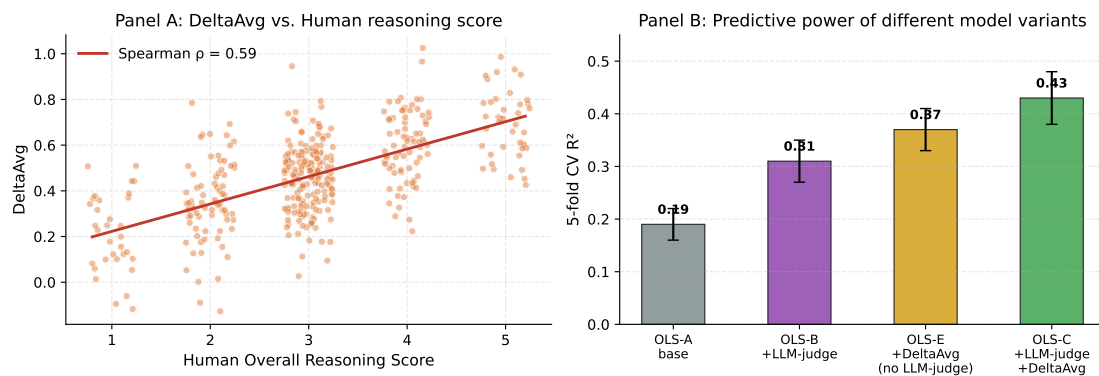


Figure 13: Human annotation validation. Panel A shows correlation with human reasoning scores. Panel B shows predictive power under different regression variants.

Figure 12. Causal evidence from DataModels: DeltaAvg predicts sample marginal contributions

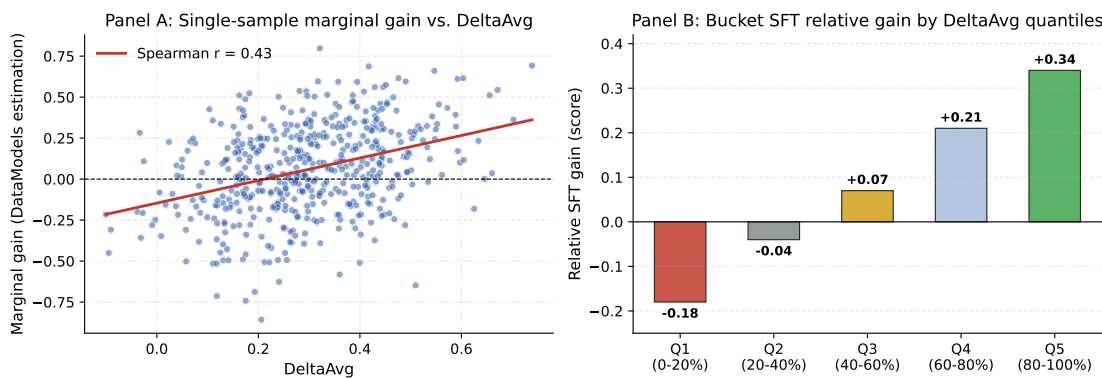


Figure 14: DataModels marginal contribution analysis. Panel A correlates DeltaAvg with single-sample marginal gain. Panel B groups samples by DeltaAvg quantiles and reports relative SFT gain.