CoT-Core: Accelerating LLM Evaluation via CoT-Aware Coreset Selection

Anonymous authors
Paper under double-blind review

Abstract

Evaluating Large Language Models (LLMs) incurs substantial computational overhead during continuous development, since a fixed benchmark is rerun for many checkpoints. Coreset selection reduces this workload, but existing methods either require historical response logs that new benchmarks lack, as in Item Response Theory, or inherit the lexical bias of question-only embeddings. We propose CoT-Core, a training-free core question selection framework. Because lexically disparate questions can require related reasoning operations, CoT-Core prompts an LLM to unroll a zero-shot Chain-of-Thought (CoT) trajectory for each question and clusters the resulting question-trajectory representations to choose representative items. The subset is constructed once, without reference answers or historical target-model responses, and reused across evaluations. On GSM8K, MMLU, MMLU-Pro, and GPQA, CoT-Core attains the strongest history-free score estimation in most dataset-budget settings. Rerunning the full evaluation pipeline for eleven target models on MMLU-Pro further shows that a 5% subset reduces output tokens by 94.77%, while its estimated scores deviate from full-benchmark scores by only 0.0065 on average. CoT-Core therefore provides a practical option for repeated evaluation when historical responses are unavailable.

1 Introduction

The rapid advancement of large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) has been accompanied by increasingly comprehensive evaluation frameworks and benchmarks, such as MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and HELM (Liang et al., 2022). Evaluating a modern LLM across such suites can require predictions for tens of thousands of test items (Srivastava et al., 2023). Moreover, evaluation is rarely a one-time event: throughout continuous pre-training, alignment, and fine-tuning, the same benchmark is rerun for many model checkpoints to monitor progress and steer development (Biderman et al., 2023; Groeneveld et al., 2024), multiplying this overhead. The practical objective is therefore not merely to answer fewer questions, but to obtain reliable estimates of full-benchmark performance while substantially reducing repeated generation.

Recent literature highlights that this exhaustive paradigm is highly inefficient, because large evaluation suites contain substantial redundancy (Wang et al., 2025; Zhang et al., 2025). This creates a need for *coresets*: small subsets of items whose scores reconstruct full-benchmark results. While various sampling heuristics exist, the most systematic approaches align with two paradigms. Psychometric methods score questions with Item Response Theory (IRT) (Polo et al., 2024), calibrated on historical model correctness, and are therefore bottlenecked by their reliance on massive evaluation logs. Consequently, they face a "cold start" problem: for newly constructed benchmarks or private evaluation sets, the response records needed to calibrate item parameters simply do not exist. Geometric methods, such as the widely adopted k-center greedy algorithm (Sener & Savarese, 2017), require no such logs and operate directly on pretrained question embeddings. Dense encoders, however, are known to rely heavily on surface lexical overlap (Sinha et al., 2021; Thakur et al., 2021), so proximity in these spaces reflects shared wording or domain rather than a shared solution procedure. Random sampling and question-only selection nonetheless remain strong low-overhead baselines against which any added preprocessing must be justified.

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.

To overcome both the dependence on historical logs and the lexical bias of question-only features, we propose CoT-Core, a training-free core question selection framework. Our approach is motivated by a simple observation: lexically disparate questions, which appear unrelated in standard text embedding spaces, can require the same underlying operations, whereas similarly worded questions can require entirely different ones. To expose this distinction, CoT-Core first prompts an LLM to unroll a zero-shot Chain-of-Thought (CoT) trajectory for each question, making candidate laws, intermediate quantities, and distinctions between answer options explicit in the text. Embedding each question together with its trajectory and applying standard clustering then yields representative items grouped by attempted solution procedures rather than surface form alone. The trajectory serves as an auxiliary feature, not a reference solution: selection uses neither answer keys nor target-model responses. Because the trajectories and the resulting subset are constructed once per benchmark and configuration, this preprocessing is paid a single time and amortized over all subsequent evaluations.

We assess this design along three axes. First, on GSM8K, MMLU, MMLU-Pro, and GPQA, we measure how faithfully the selected subsets reconstruct full-benchmark scores and model rankings. Second, rerunning full and subset evaluation for eleven target models on MMLU-Pro quantifies the generation a reusable subset actually demands: at a 5% budget, output tokens fall by 94.77% while the estimated scores deviate from full-benchmark scores by 0.65 percentage points on average. Third, replacing each trajectory with equally long but unrelated text tests whether the gains stem from the solution content or merely from appending more tokens.

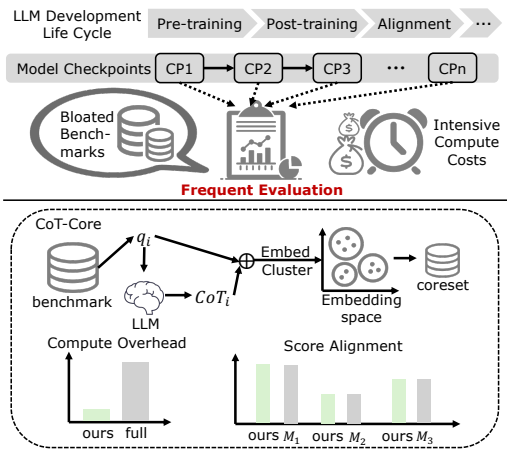

Figure 1: Motivation and overview of CoT-Core: construct a subset once and reuse it across model evaluations. The cost and score-alignment sketches are schematic, not measured runtime or accuracy curves.

Our main contributions are summarized as follows: 1) We propose CoT-Core, a training-free coreset selection framework that circumvents surface-level lexical traps by clustering question-trajectory features, using neither reference answers nor historical target-model responses. 2) Methodologically, we decouple spatial coreset selection from downstream proxy estimation: centroid-based sampling supports Standard Aggregation (SA) for a strictly history-free pipeline and remains compatible with estimators such as GP-IRT, while a conditional reconstruction bound relates representation geometry, response consistency, and aggregation weights. 3) Extensive experiments show high-fidelity score and ranking estimation, the strongest among history-free selectors in most dataset-budget settings, and length-matched placebo controls and ablations delineate the boundary conditions of reasoning-aware pruning.

2 Related Work

LLM Benchmark Evaluation and Redundancy As the demand for rigorous assessment grows, evaluating models across massive suites has become computationally bottlenecked. Studies such as EssenceBench (Wang et al., 2025) have systematically quantified sample redundancy within prominent evaluation suites, revealing significant overlap at both the textual level and the ranking level. Similarly, Zhang et al. (2025) demonstrated that repeated testing of similar capabilities yields diminishing returns. Identifying a representative coreset is therefore an urgent necessity for sustainable AI development.

Geometric and Training-free Methods. One way to remove this redundancy is to seek coverage directly in a representation space. Geometric algorithms such as k-center greedy (Sener & Savarese, 2017) do so without any historical model responses, which makes them immediately applicable to new benchmarks. Their selection quality, however, is inherited from that space: applied to pretrained embeddings of *raw problem text*, distances track wording and domain, so two questions can share vocabulary while requiring different solution procedures. CoT-Core retains this history-free property

but changes the feature rather than the search, augmenting the question representation with evidence about the procedures a question calls for.

Performance-Based and Psychometric Methods. A second branch infers item informativeness from model behavior instead of text geometry. Dataset cartography uses training dynamics (Swayamdipta et al., 2020), while TinyBenchmarks (Polo et al., 2024) and ATLAS (Li et al., 2025) leverage Item Response Theory. These methods therefore exploit exactly the evidence that geometric selectors discard, namely how models actually answer each item, and they remain the stronger choice when such responses are plentiful. The same dependence, however, is a cold-start constraint on newly constructed or private benchmarks, where no response matrix exists yet. CoT-Core removes that constraint at the selection stage only: its subsets can still be paired with GP-IRT, but that estimator continues to require historical data.

CoT-Aware Representations for Coreset Selection Chain-of-Thought prompting (Wei et al., 2022) offers a source of exactly this missing evidence: it elicits explicit intermediate solution text, making the laws, quantities, and case distinctions a question calls for visible in the text itself, even though that text need not be correct or faithful to the model's internal computation. Clustering and CoT have already been combined, but for a different purpose. Auto-CoT (Zhang et al., 2022) clusters question embeddings to assemble diverse demonstrations, so the trajectories are the output of selection rather than its input. Our setting reverses that relationship: to group evaluation items by solution procedure, the trajectory must enter the feature that selection operates on. State-aware analyses of CoT (Yu et al., 2025) support treating trajectory semantics as informative in this way, while stopping short of establishing that ordinary embedding distance measures logical equivalence.

CoT-Core therefore occupies the remaining position: it uses questions and their trajectories as a unified feature space that exposes **solution-related operations and concepts**, while keeping the history-free applicability of geometric selection. Because the generator and encoder are pretrained and the selector fits nothing to target-model responses, the pipeline is *training-free* in the sense of requiring no evaluation history, not in the sense of being independent of pretraining data or generator behavior.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{N}$ denote an LLM evaluation benchmark comprising N test instances, where q_i is the question prompt and a_i is the reference answer. For a target model $\mathcal{M}$, the full-benchmark score under a fixed evaluation protocol is the item-weighted mean:

$$S_{\mathcal{D}}(\mathcal{M}) = \frac{1}{N} \sum_{i=1}^{N} m(\mathcal{M}(q_i), a_i) \tag{1}$$

where $\mathcal{M}(q_i)$ is the model's output and $m(\cdot, \cdot) \in [0, 1]$ is the item-level scoring function, such as exact-match accuracy.

Coreset evaluation separates **Subset Selection** from **Performance Estimation**. A selector identifies $B \ll N$ representative items with indices $\mathcal{S} = \{s_1, \ldots, s_B\}$, and an estimator f_{est} predicts the full-benchmark score from the target model's responses to them:

$$\hat{S}_{\mathcal{D}}(\mathcal{M}) = f_{est}(\mathcal{M}, \mathcal{S}) \approx S_{\mathcal{D}}(\mathcal{M}). \tag{2}$$

Keeping the two stages separate makes each reusable: CoT-Core supplies the indices, after which Standard Aggregation (SA) reads the responses directly while a calibrated estimator may additionally exploit historical ones. The budget B is a user-specified input, and we report fidelity across budgets rather than claiming one minimal subset size.

Throughout, two questions *share a solution procedure* when answering them invokes the same operations, formulations, or domain-specific laws, irrespective of their surface entities and values. Since procedural overlap is not annotated in these benchmarks, we evaluate this notion through its consequences: score fidelity under a fixed budget and sensitivity to replacing the trajectory content.

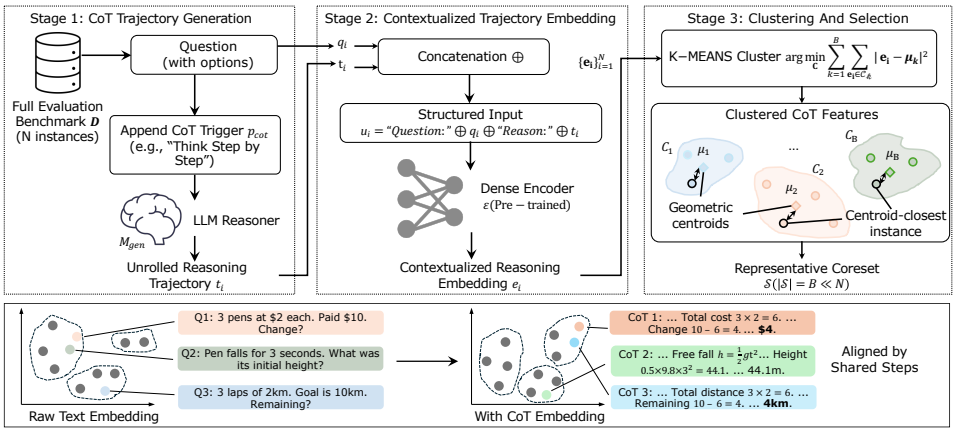

Figure 2: **Overview of CoT-Core.** The generator receives the question and candidate options, not the reference answer. CoT-Core embeds the resulting question-trajectory pairs and selects cluster representatives. The illustrative Q1/Q3 pair shares arithmetic operations despite different contexts, whereas Q1/Q2 shares wording but requires different operations: the cases a question-only representation is prone to confuse.

3.2 CoT-Core Framework Overview

Figure 2 implements construction in three stages: (1) zero-shot trajectory generation, (2) contextualized trajectory embedding, and (3) clustering and representative selection. All three precede the evaluation of any target model and run once per benchmark, feature configuration, and budget, after which the cached trajectories, embeddings, and selected item IDs are reused and only a new model's responses to those items are needed. Construction cost is therefore paid once and amortized over every subsequent evaluation.

3.3 Stage 1: CoT Trajectory Generation

The question text q_i leaves the operations needed for a solution implicit. Our first stage makes them explicit: an instruction-tuned generator $\mathcal{M}_{gen}$ produces a trajectory $t_i = \mathcal{M}_{gen}(q_i \oplus p_{cot})$ from the question and a fixed CoT instruction p_{cot}, such as *"Let's think step by step"*, where $\oplus$ denotes concatenation. The prompt contains only the question and its candidate options; withholding reference answers, correctness labels, demonstrations, and historical responses is what keeps the pipeline applicable to benchmarks with no evaluation history. We retain every trajectory regardless of final-answer correctness, since a partially incorrect attempt still surfaces the laws and quantities a question involves, and trajectories serve selection only, never reaching the target model at evaluation time.

3.4 Stage 2: Contextualized Trajectory Embedding

Embedding a trajectory in isolation would discard the premise it was derived from, so we keep both. Each instance is serialized with explicit role markers as *"Question:"* q_i *"Reason:"* t_i, and a pretrained dense encoder $\mathcal{E}$ maps the sequence into $\mathbb{R}^d$. We L2-normalize the resulting vectors, so that clustering compares directions rather than magnitudes, and stack them into the feature matrix $\mathbf{E} \in \mathbb{R}^{N \times d}$. These question-trajectory features expose the operations a solution attempt performs, which question-only embeddings leave implicit; Section 4.4 tests directly whether that added content or merely its length drives the observed gains.

3.5 Stage 3: CoT-Aware Clustering and Selection

We apply standard K-Means to partition the trajectory embeddings into B nonempty clusters, minimizing within-cluster squared Euclidean distance. For cluster C_k with arithmetic centroid μ_k, we

select the centroid-nearest item within that cluster:

$$\min_{\{C_k,\mu_k\}_{k=1}^{B}} \sum_{k=1}^{B} \sum_{i \in C_k} \|\mathbf{e}_i - \mu_k\|_2^2, \qquad s_k \in \arg\min_{i \in C_k} \|\mathbf{e}_i - \mu_k\|_2. \tag{3}$$

Under subject-specific quotas the rule applies within each subject, with quotas summing to the total budget, and drawing one item per cluster guarantees B distinct questions. On unit vectors, squared Euclidean distance is twice the cosine distance, yet ordinary K-Means still differs from its spherical variant because centroids need not have unit norm; Appendix D shows the two perform almost identically here, so we keep the simpler algorithm.

For a fixed target model, write $y_i = m(\mathcal{M}(q_i), a_i) \in [0, 1]$. Standard Aggregation with normalized nonnegative weights w_k is

$$\widehat{S}_w = \sum_{k=1}^{B} w_k y_{s_k}, \qquad \sum_{k=1}^{B} w_k = 1. \tag{4}$$

We use uniform weights $w_k = 1/B$ by default and state the weighting whenever selectors are compared, since cluster-mass weights $w_k = |C_k|/N$ define a different estimator. SA consumes only the selected responses, so the pipeline runs without evaluation history; the same subsets also feed GP-IRT, which calibrates item parameters from historical responses before inferring target-model ability. One construction thus serves both regimes. Appendix E bounds the reconstruction error, showing that a compact representation helps only when items grouped together also elicit similar responses.

4 Experiments

4.1 Experimental Setup

Datasets & Implementation Details. We evaluate on GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024), and GPQA Diamond (Rein et al., 2024), which together span grade-school arithmetic, broad multi-subject knowledge, and expert-level science. The default trajectory generator is Phi-4 (Abdin et al., 2024), used through OpenCompass (OpenCompass Contributors, 2023); the default encoder is BGE-M3 (Chen et al., 2024). Raw question-trajectory pairs are the selection features unless a transformation or representation ablation is explicitly identified. Generator and encoder choices concern coreset construction, not the models whose benchmark performance is being estimated.

Baselines & Model Roles. History-free selectors are **Random**, **Question-Emb K-Means**, k-**Center Greedy** (Sener & Savarese, 2017), and CoT-Core. **Correctness K-Means** and **IRT** (Polo et al., 2024) use historical responses for selection. To test generalization to unseen models, we hold out a benchmark-specific pool of 100 target models, with logs sourced from TinyBenchmarks (Polo et al., 2024) and the Open LLM Leaderboard (Beeching et al., 2023). History-dependent selectors and estimators are calibrated on a strictly disjoint set of models; we write H for its size, reserving N for items and M for target models. The main comparisons use $H = 25$ to reflect a realistic low-resource setting, with larger pools in Appendix C. The supplementary token analysis reruns full and subset evaluation on MMLU-Pro for its own target models, eleven of which have complete paired token and score records (Appendix G).

Evaluation Protocol. Coresets are extracted at nominal budgets of $1\%, 5\%, 10\%, 15\%$, and 20%, with integer rounding and subject allocation determining the realized item counts. To separate selection quality from downstream aggregation, every subset is evaluated under two decoupled estimators: **SA** (Eq. (4)) and historically calibrated **GP-IRT** (Polo et al., 2024). To control initialization variance, we report means over five independent rounds for all algorithmic methods and twenty trials for Random, with each round reusing the cached features rather than regenerating trajectories. Appendix A.4 identifies the main comparisons, ablations, length-matched controls, and token analysis as separate result blocks.

Evaluation Metrics. Both score reconstruction and ranking are evaluated across target models. For one selection round and M models, let S_m and $\widehat{S}_m$ be the full-benchmark and estimated scores in

Table 1: Performance comparison using SA. Historical selectors use $H = 25$ calibration models; history-free selectors do not. Boldface highlights reported best values among history-free selectors, not statistical significance. GPQA results at the nominal 1% budget are omitted, since that ratio leaves too few of its 198 items for a meaningful subset.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
GSM8K	*History-Free Selectors*					
	Random	0.0890 / 0.9044	0.0382 / **0.9747**	0.0246 / 0.9851	0.0187 / **0.9897**	0.0142 / 0.9920
	Question-Emb K-Means	0.0875 / 0.8951	**0.0321** / 0.9743	0.0241 / 0.9797	**0.0175** / 0.9871	**0.0133** / **0.9931**
	k-Center Greedy	**0.0683** / **0.9118**	0.0349 / 0.9738	0.0283 / 0.9847	0.0218 / **0.9897**	0.0175 / 0.9925
	CoT-Core (Ours)	0.0755 / 0.8884	0.0331 / 0.9696	**0.0209** / **0.9853**	0.0197 / 0.9884	0.0154 / 0.9918
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0596 / 0.9129	0.0347 / 0.9700	0.0259 / 0.9770	0.0231 / 0.9838	0.0198 / 0.9886
	IRT	0.0664 / 0.9172	0.0308 / 0.9649	0.0196 / 0.9817	0.0175 / 0.9852	0.0145 / 0.9879
MMLU	*History-Free Selectors*					
	Random	0.0330 / **0.9489**	0.0149 / 0.9880	0.0102 / 0.9935	**0.0079** / 0.9957	0.0063 / 0.9969
	Question-Emb K-Means	**0.0298** / 0.9484	0.0215 / 0.9877	0.0154 / 0.9923	0.0203 / 0.9943	0.0252 / 0.9958
	k-Center Greedy	0.0477 / 0.9461	0.0530 / 0.9867	0.0393 / 0.9909	0.0360 / 0.9942	0.0317 / 0.9952
	CoT-Core (Ours)	0.0324 / 0.9353	**0.0118** / **0.9881**	**0.0081** / **0.9941**	0.0081 / **0.9964**	**0.0060** / **0.9971**
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0304 / 0.9616	0.0203 / 0.9754	0.0181 / 0.9773	0.0169 / 0.9781	0.0160 / 0.9785
	IRT	0.0253 / 0.9589	0.0112 / 0.9895	0.0081 / 0.9938	0.0068 / 0.9966	0.0056 / 0.9969
MMLU-Pro	*History-Free Selectors*					
	Random	0.0311 / 0.9556	0.0132 / **0.9899**	0.0093 / 0.9941	**0.0070** / 0.9956	0.0058 / 0.9967
	Question-Emb K-Means	0.0300 / 0.9579	0.0133 / 0.9890	0.0111 / **0.9947**	0.0116 / 0.9957	0.0100 / **0.9969**
	k-Center Greedy	0.0369 / 0.9524	0.0432 / 0.9857	0.0334 / 0.9923	0.0306 / 0.9936	0.0272 / 0.9957
	CoT-Core (Ours)	**0.0296** / **0.9682**	**0.0124** / 0.9893	**0.0082** / 0.9941	0.0072 / **0.9965**	**0.0055** / **0.9969**
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0257 / 0.9789	0.0144 / 0.9902	0.0123 / 0.9929	0.0107 / 0.9936	0.0095 / 0.9941
	IRT	0.0379 / 0.9540	0.0151 / 0.9891	0.0111 / 0.9930	0.0092 / 0.9943	0.0073 / 0.9962
GPQA	*History-Free Selectors*					
	Random	- / -	0.1252 / 0.2851	0.0797 / 0.4480	0.0635 / 0.5579	0.0536 / 0.6022
	Question-Emb K-Means	- / -	0.1098 / 0.2856	0.0859 / 0.3656	0.0615 / 0.5338	0.0522 / **0.6263**
	k-Center Greedy	- / -	0.1232 / 0.2197	0.0821 / 0.4036	0.0608 / 0.5355	0.0565 / 0.6151
	CoT-Core (Ours)	- / -	**0.1044** / **0.4334**	**0.0731** / **0.5839**	**0.0601** / **0.5646**	**0.0516** / 0.6013
	History-Dependent Selectors (H=25)					
	Correctness K-Means	- / -	0.0936 / 0.3125	0.0743 / 0.3435	0.0660 / 0.4498	0.0570 / 0.4770
	IRT	- / -	0.1054 / 0.2576	0.0723 / 0.3806	0.0552 / 0.6081	0.0513 / 0.6176

$[0, 1]$. We use

$$\mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M} |\widehat{S}_m - S_m|, \qquad \mathrm{Sim} = \mathrm{Corr}\big(R(S), R(\widehat{S})\big), \tag{5}$$

where R gives descending ordinal ranks and Corr is Pearson correlation, so MAE measures score reconstruction and Sim global model-order agreement; a difference of 0.001 in MAE corresponds to 0.1 percentage points (pp). Where MAE carries a $\pm$, the reported spread is the pooled standard deviation of individual absolute errors rather than a confidence interval, so method differences are read as effect sizes and point estimates. Because Sim summarizes the whole ranking, Appendix A.3 additionally reports a top-weighted ranking metric; Appendix A.2 collects the remaining tie-handling and aggregation conventions.

4.2 Main Results

We evaluate the selected subsets along two axes that matter for repeated evaluation: how closely their scores reconstruct full-benchmark results, and how well they preserve the ranking among candidate models. Table 1 reports the history-free comparison under SA and Table 2 the same subsets under calibrated GP-IRT, each giving MAE for reconstruction and Sim for ranking.

Score Reconstruction Across Datasets and Budgets. Under SA (Table 1), CoT-Core attains the lowest MAE among history-free selectors in 12 of the 19 reported dataset-budget cells, including all four GPQA budgets. The margins are largest where question text alone is least informative: on GPQA at 5%, MAE falls from 0.1252 for Random to 0.1044, a reduction of 2.08 pp, while Sim rises

Table 2: Performance comparison using GP-IRT with $H = 25$ historical models. Every estimator in this table uses historical calibration; the row groups describe selection-stage dependencies only. Boldface marks the best history-free value in each column. GPQA 1% values remain omitted, with actual budgets determined by rounding and subject quotas.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0827 / **0.8846**	0.0297 / 0.9701	0.0202 / 0.9843	**0.0158** / 0.9893	0.0131 / 0.9919
	Question-Emb K-Means	0.0975 / 0.8443	0.0296 / **0.9715**	0.0206 / 0.9776	**0.0158** / 0.9866	**0.0123** / 0.9928
	k-Center Greedy	0.1016 / 0.8597	**0.0289** / 0.9707	0.0216 / 0.9830	0.0166 / **0.9895**	0.0136 / **0.9929**
	CoT-Core (Ours)	**0.0800** / 0.8700	0.0290 / 0.9642	**0.0197** / **0.9844**	0.0166 / 0.9882	0.0135 / 0.9914
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0830 / 0.8608	0.0281 / 0.9691	0.0183 / 0.9830	0.0156 / 0.9886	0.0121 / 0.9940
	IRT	0.0789 / 0.8957	0.0308 / 0.9659	0.0177 / 0.9838	0.0158 / 0.9877	0.0134 / 0.9909
	History-Free Selectors					
MMLU	Random	0.0212 / 0.9639	0.0116 / 0.9897	0.0089 / 0.9938	0.0072 / 0.9957	0.0060 / 0.9968
	Question-Emb K-Means	0.0225 / 0.9606	0.0120 / 0.9891	0.0105 / 0.9933	0.0130 / 0.9947	0.0165 / 0.9961
	k-Center Greedy	**0.0198** / **0.9695**	0.0190 / 0.9893	0.0191 / 0.9929	0.0209 / 0.9951	0.0205 / 0.9957
	CoT-Core (Ours)	0.0223 / 0.9577	**0.0111** / **0.9901**	**0.0076** / **0.9948**	**0.0068** / **0.9962**	**0.0059** / **0.9969**
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0222 / 0.9648	0.0142 / 0.9856	0.0121 / 0.9864	0.0113 / 0.9850	0.0110 / 0.9836
	IRT	0.0209 / 0.9651	0.0131 / 0.9883	0.0095 / 0.9932	0.0073 / 0.9961	0.0066 / 0.9968
	History-Free Selectors					
MMLU-Pro	Random	0.0250 / 0.9598	0.0104 / **0.9904**	0.0074 / 0.9944	0.0060 / 0.9955	0.0051 / 0.9966
	Question-Emb K-Means	0.0244 / 0.9547	0.0103 / 0.9895	0.0081 / **0.9949**	0.0083 / 0.9955	0.0074 / 0.9968
	k-Center Greedy	0.0240 / 0.9670	0.0146 / 0.9896	0.0149 / 0.9935	0.0162 / 0.9943	0.0162 / 0.9955
	CoT-Core (Ours)	**0.0240** / **0.9724**	**0.0102** / 0.9901	**0.0069** / 0.9943	**0.0058** / **0.9963**	**0.0047** / **0.9968**
	History-Dependent Selectors (H=25)					
	Correctness K-Means	0.0234 / 0.9747	0.0091 / 0.9925	0.0081 / 0.9948	0.0075 / 0.9957	0.0070 / 0.9958
	IRT	0.0257 / 0.9549	0.0109 / 0.9903	0.0095 / 0.9943	0.0085 / 0.9960	0.0075 / 0.9971
	History-Free Selectors					
GPQA	Random	- / -	0.1019 / 0.2807	0.0679 / 0.4782	0.0568 / **0.5842**	0.0497 / 0.6248
	Question-Emb K-Means	- / -	**0.0887** / 0.2759	0.0703 / 0.3900	0.0552 / 0.5704	**0.0482** / **0.6388**
	k-Center Greedy	- / -	0.0973 / 0.2351	0.0701 / 0.4334	0.0551 / 0.5547	0.0519 / 0.6335
	CoT-Core (Ours)	- / -	0.0922 / **0.5123**	**0.0649** / **0.6206**	**0.0545** / 0.5812	0.0482 / 0.6318
	History-Dependent Selectors (H=25)					
	Correctness K-Means	- / -	0.0728 / 0.3970	0.0597 / 0.4130	0.0559 / 0.5022	0.0503 / 0.5293
	IRT	- / -	0.0887 / 0.2545	0.0622 / 0.4248	0.0500 / 0.6402	0.0473 / 0.6475

from 0.2851 to **0.4334**. On MMLU-Pro the subsets remain the most accurate at 1%, 5%, 10%, and 20%, reaching 0.0082 against 0.0093 for Random at 10%. GSM8K is the consistent exception: its arithmetic questions already state the operations a solution needs, so an unrolled trajectory adds little, and Random or Question-Emb selection is equally effective. Such cases bound where reasoning-aware features help rather than undermine the mechanism, an interpretation we examine directly in Section 4.3.

Compatibility with Historical Estimation. Pairing the same subsets with GP-IRT (Table 2) widens the advantage to 14 of 19 cells, confirming that the selected items remain informative under a calibrated estimator. The stronger result concerns the history-dependent selectors: at budgets of 5% and above on MMLU, and 10% and above on MMLU-Pro, CoT-Core matches or improves on both IRT and Correctness K-Means, for instance 0.0069 against 0.0081 on MMLU-Pro at 10%. A selector that reads no evaluation history therefore reaches the accuracy of selectors built from response matrices, in the regimes where those matrices are most expensive to collect. On GSM8K, and on GPQA where only 198 items are available for calibration and selection alike, the history-dependent methods retain an edge.

Model Ranking Under Heavy Compression. Score fidelity alone does not guarantee that a subset ranks models correctly, and ranking is what most development decisions rely on. The gap is widest under heavy compression on the hardest benchmark: with a 5% GPQA budget under GP-IRT, CoT-Core preserves a Sim of **0.5123**, close to twice the 0.2759 retained by question-only clustering. The lead holds at 5% and 10% under both estimators, the regimes where compression is most aggressive, and narrows as budgets grow. Appendix A.3 reports a top-weighted ranking metric that gives the same ordering at these budgets.

From Score Fidelity to Generation Demand. Fidelity is only useful if it comes with a real reduction in work, so we measure both on the same subsets. Rerunning full and subset evaluation for eleven target models on MMLU-Pro (Table 16) cuts aggregate output tokens by 94.77% at a 5% budget, while the estimated scores stay within 0.65 pp of the full-benchmark values on average. Because the construction cost is paid once and the subset is reused, this saving accrues to every subsequent model. Section 4.4 reports the per-model figures and Appendix G the aggregation; token counts quantify generation demand rather than wall-clock or GPU time.

4.3 Analysis

Where the Added Information Helps. The gains track how much a trajectory can add to the question text. GSM8K states its arithmetic operations directly, and at 5% CoT-Core's MAE of 0.0331 is within 0.1 pp of the 0.0321 obtained from question-only clustering: unrolling adds length without adding discriminative content. On GPQA, where a solution must first identify the governing law or weigh distractors, the same budget lowers MAE from 0.1252 to 0.1044. We read this as a statement about incremental information rather than benchmark difficulty as such, since these comparisons do not separate difficulty from the informativeness of the question text. Low budgets likewise do not degrade uniformly: MMLU-Pro attains its lowest MAE at 1% (0.0296), and 1% of MMLU's 14,042 items is about 140 items spread over 57 subjects, so subject coverage rather than nominal budget governs whether a subject is missed.

What the Trajectories Change in the Representation. Regrouping items by solution procedure requires distances to move in two directions, and both are visible on GPQA. Appendix B reports the full problem statements, the generated trajectories, and a t-SNE view of the same embeddings.

Case 1: Connecting Related Solution Procedures. A CERN bubble-chamber question and a spherical-detector question describe disjoint experimental settings, and their question-only cosine distance of 0.5461 places them in separate clusters, where each consumes a representative of its own. Both generated trajectories invoke relativistic kinematics and time dilation, and the distance falls by about 61% to 0.2128, so one item can stand in for the pair and release budget for a genuinely different procedure. Both attempts also contain arithmetic errors, reproduced verbatim in the appendix: the shift follows the concepts an attempt invokes rather than the correctness of its result.

Case 2: Separating Different Solution Procedures. Conversely, two questions about a 6000K star with dark spots share almost all of their wording while asking for different quantities, atomic state populations and macroscopic photometric variation. Their question-only distance of 0.1957 makes them candidates for the same cluster, in which one would be pruned as redundant. Once one trajectory invokes the Boltzmann distribution and the other a T^4 flux relation, the distance nearly doubles to 0.3890, keeping a separate representative for each procedure. The two pairs therefore move in opposite directions, each tracking the solution procedure rather than the surface wording.

4.4 Ablation Studies

Generator Scale and the Verbosity Hypothesis. A practical question is how strong the generator must be. We vary it from 1B to 72B on MMLU-Pro with the encoder fixed to BGE-M3; this changes the model producing selection features, not the target models being evaluated. Cleaning, encoder, and clustering sensitivity is reported in Appendix D.

Scale buys little here. In Table 3 the 72B generator has the lowest MAE at 10%, yet a 1B generator is best at 1% and Phi-4 at 20%, so the ranking never becomes monotonic in size. A modest generator is therefore sufficient to produce useful selection features, which keeps construction cost low, and the practical criterion is the quality of the resulting coreset rather than parameter count. One explanation is a **verbosity penalty**: conversational filler dilutes the operations that distinguish one solution attempt from another. The claim concerns the share of non-discriminative text, not length as such, since

Table 3: Generator-scale ablation on MMLU-Pro (MAE ↓). Pooled individual-model error standard deviations range from ± 0.0043 to ± 0.0241 and are not seed-level uncertainty.

Generator	1%	10%	20%
LLaMA-3-1B	**0.0280**	0.0109	0.0057
Mistral-8B	0.0316	0.0081	0.0066
Phi-4	0.0295	0.0082	**0.0054**
Qwen-72B	0.0305	**0.0077**	0.0065

appending a trajectory lengthens the input and still lowers MAE relative to question-only features. Size, training data, capability, and output length move together across these four models, so we test the content question directly in the next ablation, which holds token count fixed.

Length-Matched Placebo Controls. Appending a trajectory lengthens the text that the encoder sees, so the decisive test is whether its content or its bulk produces the gain. We replace each Phi-4 trajectory with synthetic text from Python's Faker library or an unrelated Wikipedia passage, truncated to the token count of the trajectory it displaces, and hold every other stage of the pipeline fixed so that the appended content is the only variable.

Under tight budgets, content rather than length reproduces the benefit. Question-relevant trajectories give the lowest mean MAE at both budgets (Table 4), improving on the strongest placebo by 0.219 pp at 5% and 0.013 pp at 10%. The 5% column carries this evidence: there the placebos are the only configurations that fall behind plain Random, so padding a question with irrelevant text is worse than not clustering on it at all. This is the verbosity penalty seen from the opposite direction, now with length held constant, which places the effect in what the appended text says rather than in how much of it there is. The margin is budget-dependent: by 10% the three variants lie within 0.13 pp of one another, and Appendix H shows the ordering reversing at 1% on MMLU-Pro and at 20% on GPQA. The two placebo sources also swap order across budgets, so the comparison isolates the content source rather than any property of coherent versus synthetic filler.

Table 4: Length-matched placebo controls on MMLU-Pro with SA (MAE ↓). The lower three rows append text of identical token count to the same question; Random needs no representation. Pooled individual-model error standard deviations range from ± 0.0061 to ± 0.0128.

Appended text	**5%**	**10%**
Random	0.01315	0.00929
Unrelated Wiki	0.01457	0.00852
Faker text	0.01729	0.00837
CoT (ours)	**0.01238**	**0.00824**

Output-Token Use and Score Fidelity. The remaining question is what the subsets save. We rerun full and subset evaluation for eleven target models on MMLU-Pro at a 5% budget and record both generation and score (Table 16). Full evaluation consumes 69,263,070 output tokens against 3,621,929.2 for the subsets, a **94.77% reduction**, while the mean absolute gap between each full score and its trial-averaged subset score is **0.65 percentage points**. The reduction is therefore close to the nominal budget, and it does not come at the price of the score estimate.

This saving recurs for every model that reuses the subset, whereas generation, embedding, and clustering are paid once per benchmark and configuration. Because output tokens measure generation demand rather than per-token compute, we report the reduction and fidelity without converting them into wall-clock or GPU-hour savings.

5 Conclusion

CoT-Core builds reusable benchmark subsets by augmenting questions with generated reasoning trajectories before clustering, using neither reference answers nor historical target-model responses, so the same items support SA and, with calibration data, GP-IRT. Across four benchmarks it gives the strongest history-free score and ranking fidelity in most dataset-budget settings, and on MMLU-Pro an eleven-model study cuts output tokens by 94.77% at a mean score gap of 0.65 percentage points. Length-matched placebos trace the tight-budget gain to what the appended text says rather than how much of it there is, and a modest generator suffices to produce it, keeping the one-time construction cost low. CoT-Core is therefore a practical option whenever a benchmark will be rerun and no evaluation history exists, with its value set by the required estimation quality and the number of models that reuse the subset.

AI USE STATEMENT

We use generative models to run the experiments, including the reasoning trajectories and transformed text described in the method and ablations, and to polish the writing of this paper. The authors are responsible for the final claims, reported results, and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 specifies the selection and estimation procedures, and Appendix E gives the full proof of the reconstruction bound. Appendix A provides model settings, transformation prompts, and the distinction between result blocks. Appendix G defines the supplementary token and mean-score accounting. Reproducing each numerical comparison additionally requires its exact model and item manifests, input configuration, selected indices and weights, and per-model, per-round predictions.

ETHICS STATEMENT

Compressed evaluation can miss subgroups or change the ordering of closely matched models. Coreset estimates should therefore not be interpreted as exhaustive assessments of model capabilities. Generated trajectories are fallible selection features, not reference solutions. Any release of benchmark items, outputs, or code must respect the applicable licenses and access conditions.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

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

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2025.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*, 2024.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. Unnatural language inference. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers)*, pp. 7329–7346, 2021.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, et al. Rethinking llm evaluation: Can we evaluate llms with 200x less data? *arXiv preprint arXiv:2510.10457*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. *arXiv preprint arXiv:2509.00190*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. Redundancy principles for mllms benchmarks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12492–12504, 2025.

A Implementation Details

This section records the implementation settings and transformation prompts associated with the experiments.

A.1 Model Configurations

The default generator is Phi-4 and the default encoder is BGE-M3. The reported generation settings use temperature $T = 0.7$ and a maximum output length of 1024 tokens. Raw trajectories are used for the main selector; cleaning and abstraction are separate interventions. The generator receives a question, its candidate options where applicable, and a zero-shot request for intermediate solution steps. This is CoT prompting and does not require a model-specific thinking-mode switch. Generator, encoder, and target-model roles are distinct: generated text is used to construct selection features, not to provide extra solution context to the target models.

A.2 Metric Definitions and Aggregation

Ties and averaging. For the reported Sim results, a shared model permutation orders tied scores before ordinal ranks are assigned, so Sim is distinct from average-tie Spearman and is not a correlation between question rankings. Repeated-selection MAE averages absolute errors over models and rounds, and the accompanying $\pm$ is the pooled standard deviation of those individual absolute errors, not the standard deviation of round-level MAE or a confidence interval. Ranking statistics, in contrast, are computed once per round before aggregation. The token study reports a third quantity, $|\overline{\widehat{S}}_m - S_m|$, the gap of a trial-averaged score, which need not equal the average of per-trial absolute errors; Appendix G defines its aggregation.

A.3 Top-Weighted Ranking Results

Sim in Eq. (5) weights the whole model ordering equally, so we also measure agreement concentrated at the top of the ranking. Let π order all candidate models by estimated score and π^* order them by full-benchmark score, which supplies nonnegative linear relevance gains:

$$\mathrm{DCG@}k(\pi) = \sum_{r=1}^{k} \frac{S_{\pi(r)}}{\log_2(r+1)}, \qquad \mathrm{NDCG@}k = \frac{\mathrm{DCG@}k(\pi)}{\mathrm{DCG@}k(\pi^*)}. \tag{6}$$

We set $k = 20$ for the 100-model GPQA pool and average over tied estimated-score orders. The entire candidate pool is ranked before truncation at k.

Table 5 reports GPQA NDCG@20 under SA. CoT-Core gives the highest point estimate at 5%, 10%, and 15%, matching the Sim ordering in Section 4.2, and trails at 20%, where every method has ample items. Because relevance gains use full scores, NDCG can be insensitive to swaps among close-scoring models, so these values describe ranking agreement rather than preservation of every top-model pair.

Table 5: Additional ranking evaluation on GPQA with SA: NDCG@20 (↑). The candidate pool is ranked by estimated scores, with full-benchmark scores as relevance. Boldface marks the largest value in each column.

Method	5%	10%	15%	20%
Random	0.8402	0.8729	0.9010	0.9088
Question-Emb	0.8202	0.8556	0.9063	0.9286
k-Center Greedy	0.8239	0.8835	0.9047	**0.9314**
CoT-Core	**0.8474**	**0.8941**	**0.9090**	0.9163

A.4 Result Blocks and Provenance

Main and calibrated comparisons. Tables 1 and 2 compare selectors within their original benchmark response pools. SA consumes selected-item scores and their specified weights; GP-IRT ad-

ditionally uses historical calibration. Because a history-free selector consumes no calibration responses, its SA rows are by construction identical across the historical pool sizes tabulated in Appendix C; only the calibrated estimators respond to H. Throughout, Sim denotes the ordinal-rank correlation of Eq. (5) and NDCG@20 is reported under its own name.

Encoder and clustering blocks. In Table 14, the BGE-M3 rows report the same selections and scores as the MMLU-Pro block of Table 1, so the default configuration is directly comparable across the two tables; the Jina-v3 and Qwen3-8B rows come from additional runs that differ only in the encoder. Table 15 likewise varies the clustering rule alone and carries its own K-Means reference row. Each block therefore isolates one configuration choice, and we compare within a block rather than across them.

Placebo and output-token blocks. Within the placebo study, we held the selection and evaluation settings fixed and truncated each Faker or Wiki replacement to the token count of the corresponding Phi-4 trajectory. Thus the compared representations differ in the appended text rather than its intended token budget. The token analysis in Appendix G instead reruns full and subset evaluation on MMLU-Pro and uses the eleven complete model-level usage records shown there. Its scope and trial-averaged score statistic are therefore separate from the main 100-model protocol.

Reproducibility records. Each result block requires its benchmark version and item order, target and historical model IDs, checkpoint revisions, generation prompt and sampling settings, tokenization used for placebo length matching, encoder pooling and truncation, selected indices and weights, and evaluation-round identifiers. These records determine whether two nominally identical ratios or model names identify the same experiment. Source-level result names such as `macro` do not by themselves define subject weighting; the score formulas and actual saved weights determine the estimand.

A.5 Prompt Templates for Hierarchical Cleaning

The generation and transformation experiments use Phi-4 with different instructions. The initial generation uses only the question/options and the fixed CoT trigger p_{cot}, without reference answers or few-shot demonstrations. The cleaning and abstraction prompts below are retained verbatim. Using the same generator controls its identity, but does not isolate a single text property: these prompts can change length, domain detail, and solution content together. The second generation pass also contributes to preprocessing cost.

System Prompt: Level-1 Denoised (Cleaned CoT)

You are a strict reasoning text cleaner. Your task is to extract only the valid, successful logical steps from the provided reasoning process that lead to the final answer. Strictly remove any self-corrections, dead-ends, hesitations, or redundant thoughts. Retain all necessary calculations, entities, and the core logical trajectory. Output ONLY the cleaned reasoning text without any introductory, conversational, or concluding remarks.

System Prompt: Level-2 Skeleton (Abstract CoT)

You are an expert in mathematical and logical abstraction. Your task is to extract the 'Logical Skeleton' (Granularity 2) from the provided Question and Reasoning.
Strictly follow these rules:
1. Abstract away ALL specific numbers, values, and concrete entities (e.g., replace '3 apples' with 'initial quantity', '18 km/h' with 'velocity', 'John' with 'agent').
2. Summarize the problem setup and the step-by-step reasoning process as a sequence of generalized meta-operations (e.g., 'Identify variables → Set up system of linear equations → Solve by substitution').
3. Preserve the exact logical trajectory, sequence of steps, and causal relationships.
4. Output ONLY the abstracted step-by-step sequence. Do not include any conversational filler, introductory, or concluding remarks.

B Case Study: Full Problem Descriptions and Reasoning Trajectories

This appendix provides the four GPQA problem statements and Phi-4-generated trajectories used in the qualitative analysis in Section 4.3, together with a global view of the same embeddings. Candidate option lists are omitted from these excerpts, although the generated outputs refer to option letters. The trajectories are reproduced verbatim, including their arithmetic and reasoning errors, to show what the encoder actually receives. These two pairs illustrate the mechanism; the benchmark-level evidence is the score fidelity reported in Section 4.2.

Table 6 records the pairwise cosine distances quoted in Section 4.3, measured with BGE-M3 on the question-only and question-trajectory representations.

Table 6: Cosine distances between the two case-study pairs under both representations, computed with BGE-M3 on L2-normalized embeddings.

Pair	Relation	Question-only	Question-CoT
Case 1 (A, B)	Related procedures, different wording	0.5461	0.2128
Case 2 (C, D)	Different procedures, shared wording	0.1957	0.3890

B.1 Global View of the Embedding Space

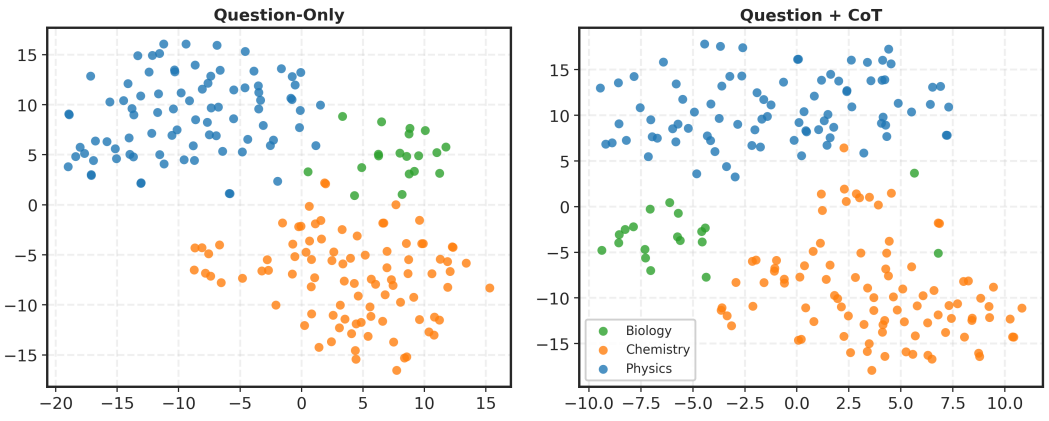

Figure 3: **t-SNE of GPQA embeddings.** (**Left**) Question-only features. (**Right**) Question-CoT features. Colors denote domains. The projection illustrates how neighborhoods change; two-dimensional distances are qualitative, and the estimation results quantify selection quality.

In the question-only projection (Figure 3, left), Physics and Chemistry form two large adjacent masses with the smaller Biology group wedged along their boundary, a layout that their disjoint vocabularies alone would produce. Once trajectories are added (right), Biology detaches into its own region and Physics spreads into a wider band, while a few items of each domain come to sit among another domain's neighbors. Both movements are consistent with grouping items by the operations a solution performs, since one domain can contain calculation-heavy and recall-heavy questions alike.

B.2 Case 1: Related Solution Concepts with Different Wording

These questions place relativistic kinematics, time dilation, and survival probabilities in different experimental contexts. Their generated attempts share some of the required concepts but do not correctly implement every step.

Question A (Experimental Resolution): In the CERN Bubble Chamber a decay occurs, $X^0 \rightarrow Y^+Z^-$ in $\tau_0 = 8 \times 10^{-16}$ s, i.e. the proper lifetime of X^0. What minimum resolution is needed to observe at least 30% of the decays? Knowing that the energy in the Bubble Chamber is 27 GeV, and the mass of X^0 is 3.41 GeV.

Unrolled CoT Trajectory (Question A)

To solve this problem, we need to determine the minimum resolution required to observe at least 30% of the decays of the particle X^0. This involves understanding the relationship between the proper lifetime, the energy of the particle, and the distance it travels before decaying.

Step 1: Calculate the Lorentz factor (γ)

The Lorentz factor γ is given by:

$$\gamma = \frac{E}{mc^2}$$

where:

- $E = 27$ GeV is the total energy,
- $m = 3.41$ GeV$/c^2$ is the rest mass of X^0,
- c is the speed of light.

Substituting the values:

$$\gamma = \frac{27 \text{ GeV}}{3.41 \text{ GeV}} \approx 7.92$$

Step 2: Calculate the dilated lifetime (τ)

The dilated lifetime τ is given by:

$$\tau = \gamma\tau_0$$

where $\tau_0 = 8 \times 10^{-16}$ s is the proper lifetime.

$$\tau = 7.92 \times 8 \times 10^{-16} \text{ s} \approx 6.336 \times 10^{-15} \text{ s}$$

Step 3: Calculate the average distance traveled before decay

The average distance d traveled by the particle before decaying is:

$$d = v\tau$$

where $v \approx c$ for high-energy particles. Thus:

$$d \approx c\tau = (3 \times 10^8 \text{ m/s}) \times 6.336 \times 10^{-15} \text{ s} \approx 1.9008 \times 10^{-6} \text{ m}$$

Step 4: Determine the resolution needed to observe 30% of decays

To observe at least 30% of the decays, the resolution Δx should be such that:

$$\Delta x \leq \frac{d}{\ln(1/0.3)}$$

Calculating $\ln(1/0.3)$:

$$\ln(1/0.3) \approx 1.204$$

Thus:

$$\Delta x \leq \frac{1.9008 \times 10^{-6} \text{ m}}{1.204} \approx 1.578 \times 10^{-6} \text{ m}$$

Step 5: Compare with given options
The closest option that satisfies $\Delta x \leq 1.578 \times 10^{-6}$ m is: C) 2.08×10^{-6} m
Thus, the minimum resolution needed is approximately 2.08×10^{-6} m.
ANSWER: C

Question B (Kinematic State): Particles are collided at the center of a spherical detector producing new type of particles that travel uninterrupted at ultra-relativistic velocities highly centered around Lorentz factor of ~ 20. On average, one third of these fast-decaying particles reaches the detector inner walls. The radius of the detector is 30 meters. What Lorentz factor is needed in order to have about two thirds of these particles reaching the detector inner walls?

Unrolled CoT Trajectory (Question B)

To solve this problem, we need to understand the relationship between the Lorentz factor (γ) and the distance traveled by the particles before they decay. The Lorentz factor is given by:

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

where v is the velocity of the particle and c is the speed of light.
The mean lifetime of a particle in its rest frame is τ_0. In the lab frame, due to time dilation, the mean lifetime is $\tau = \gamma\tau_0$.
The average distance d a particle travels before decaying is given by:

$$d = v\tau = v\gamma\tau_0$$

For ultra-relativistic particles, $v \approx c$, so:

$$d \approx c\gamma\tau_0$$

We know that with $\gamma \approx 20$, one third of the particles reach the detector's inner walls, which are 30 meters away. This implies that the average distance traveled by these particles is approximately 30 meters.
Now, we want two thirds of the particles to reach the detector, meaning the average distance should be approximately 60 meters (since two thirds reaching implies the average distance is twice as much as when one third reaches).
Using the proportionality:

$$\frac{d_2}{d_1} = \frac{\gamma_2}{\gamma_1}$$

where $d_1 = 30$ meters and $d_2 = 60$ meters, and $\gamma_1 = 20$, we solve for γ_2:

$$\frac{60}{30} = \frac{\gamma_2}{20}$$

$$2 = \frac{\gamma_2}{20}$$

$$\gamma_2 = 40$$

Thus, the Lorentz factor needed for about two thirds of the particles to reach the detector's inner walls is 40.
ANSWER: C

Analysis of the Shared Concepts: Question A concerns experimental resolution in a bubble chamber, whereas Question B concerns particle survival to a spherical detector. The shared solution ingredients can be less explicit in these question descriptions than in the generated attempts:

1. **Relativistic Kinematics:** Calculating or using the Lorentz factor (γ).
2. **Time Dilation:** Linking proper lifetime to lab-frame lifetime ($\tau = \gamma\tau_0$).
3. **Mean Decay Length:** Relating distance and lifetime ($d \approx c\gamma\tau_0$).
4. **Threshold Probability:** Relating distance to a survival fraction, a step required by the problems but handled incorrectly in the generated outputs.

The shared mathematical vocabulary offers a plausible explanation for the recorded distance reduction, consistent with a semantic-equalizer effect. It is not proof of logical isomorphism or equal difficulty. In particular, Question B replaces exponential survival with a linear proportionality argument, and Question A's stated threshold is inconsistent with the selected option. These substantive errors are preserved in the quoted text, showing that useful solution concepts and incorrect reasoning can coexist in the same trajectory.

B.3 Case 2: Different Procedures with Overlapping Wording

These questions share terms about spotted stars and effective temperatures but ask for different quantities: atomic state populations and macroscopic brightness variations.

Question C (Quantum States): Astronomers are studying a star with a 1.5 solar radius and 1.1 solar masses. When the star's surface is not covered by dark spots, its T_{eff} is 6000K. However, when 40% of its surface is covered by spots, the overall photospheric effective temperature decreases to 5500K. In the stellar photosphere, when examining the ratio of the number of neutral atoms of Ti in two energetic levels (level 1 and level 2), astronomers have observed that this ratio decreases when the star has spots. What is the factor by which this ratio changes when the star does not have spots compared to when it has spots? Note that the transition between the energy levels under consideration corresponds to a wavelength of approximately 1448 Å. Assume that the stellar photosphere is in LTE.

Unrolled CoT Trajectory (Question C)

To solve this problem, we need to understand how the ratio of the number of neutral titanium (Ti) atoms in two energy levels changes with temperature. This is governed by the Boltzmann distribution, which describes the population of atoms in different energy states at thermal equilibrium.
The Boltzmann distribution is given by:

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} \exp\left(-\frac{\Delta E}{kT}\right)$$

where: - n_2 and n_1 are the number of atoms in energy levels 2 and 1, respectively. - g_2 and g_1 are the statistical weights (degeneracies) of the energy levels. - ΔE is the energy difference between the two levels. - k is the Boltzmann constant. - T is the temperature.
The energy difference ΔE can be related to the wavelength λ of the transition using the formula:

$$\Delta E = \frac{hc}{\lambda}$$

where: - h is Planck's constant. - c is the speed of light.
Given: - $\lambda = 1448\,\text{Å} = 1448 \times 10^{-10}\,\text{m}$ - $T_1 = 6000\,\text{K}$ (without spots) - $T_2 = 5500\,\text{K}$ (with spots)
We need to find the factor by which the ratio $\frac{n_2}{n_1}$ changes from T_1 to T_2.
The ratio of the populations at two different temperatures is:

$$\frac{\left(\frac{n_2}{n_1}\right)_{T_1}}{\left(\frac{n_2}{n_1}\right)_{T_2}} = \frac{\exp\left(-\frac{\Delta E}{kT_1}\right)}{\exp\left(-\frac{\Delta E}{kT_2}\right)} = \exp\left(\frac{\Delta E}{k}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right)$$

Substitute $\Delta E = \frac{hc}{\lambda}$:

$$\frac{\left(\frac{n_2}{n_1}\right)_{T_1}}{\left(\frac{n_2}{n_1}\right)_{T_2}} = \exp\left(\frac{hc}{\lambda k}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right)$$

Now, calculate the numerical value:
- $h = 6.626 \times 10^{-34}$ J s - $c = 3 \times 10^8$ m/s - $k = 1.381 \times 10^{-23}$ J/K

$$\frac{hc}{\lambda k} = \frac{(6.626 \times 10^{-34})(3 \times 10^8)}{1448 \times 10^{-10} \times 1.381 \times 10^{-23}}$$

$$= \frac{1.9878 \times 10^{-25}}{1.999688 \times 10^{-26}} \approx 9.94 \times 10^0$$

Now, calculate the temperature difference term:

$$\frac{1}{T_2} - \frac{1}{T_1} = \frac{1}{5500} - \frac{1}{6000} = \frac{6000 - 5500}{5500 \times 6000} = \frac{500}{33000000} \approx 1.515 \times 10^{-5}$$

Now, calculate the exponential:

$$\exp\left(9.94 \times 1.515 \times 10^{-5}\right) = \exp(0.0001506) \approx 1.1507$$

Thus, the factor by which the ratio changes is approximately 1.15, which is closest to option A.
ANSWER: A

Question D (Photometric Flux): Astronomers are currently observing a star with a radius equal to that of the Sun. One hemisphere of the star is covered in dark spots with a filling factor of 20%. The star has an effective temperature (T_{eff}) of 6000K, and the spots exhibit a temperature difference of 1000K. As only one hemisphere is spotty, photometric time-series observations will reveal periodic variations in brightness due to rotational modulation. Interestingly, this situation can closely resemble the presence of an exoplanet. To produce the same amplitude signal in the star's light curve (if the star was not covered by spots!), what should be the radius of a hypothetical exoplanet relative to the radius of the host star (i.e. $R_{\text{pl}}/R_{\text{star}}$)?

Unrolled CoT Trajectory (Question D)

To solve this problem, we need to determine the radius of a hypothetical exoplanet that would produce the same amplitude of brightness variation in the star's light curve as the dark spots on one hemisphere of the star.
Step-by-step Solution:
1. **Understand the Spot Coverage:** - The star has a filling factor of 20% for the dark spots on one hemisphere. This means 20% of the visible hemisphere is covered by spots.
2. **Calculate the Area of the Spots:** - The visible hemisphere of the star has an area of $2\pi R_{\text{star}}^2$. - The area covered by spots is $0.2 \times 2\pi R_{\text{star}}^2 = 0.4\pi R_{\text{star}}^2$.
3. **Brightness Reduction Due to Spots:** - The effective temperature of the star is $T_{\text{eff}} = 6000\,\text{K}$. - The temperature of the spots is $T_{\text{spot}} = 6000\,\text{K} - 1000\,\text{K} = 5000\,\text{K}$.
4. **Calculate the Flux Reduction:** - The flux from a blackbody is proportional to T^4. - The flux reduction factor for the spots is $\left(\frac{5000}{6000}\right)^4 = \left(\frac{5}{6}\right)^4$.
5. **Calculate the Total Flux Reduction:** - The total flux reduction due to the spots is the product of the area covered by spots and the flux reduction factor:

$$\Delta F_{\text{spots}} = 0.4\pi R_{\text{star}}^2 \times \left(\frac{5}{6}\right)^4$$

6. **Calculate the Equivalent Exoplanet Radius:** - The flux reduction due to a transiting exoplanet is given by the ratio of the area of the planet to the area of the star:

$$\Delta F_{\text{planet}} = \pi R_{\text{planet}}^2 / \pi R_{\text{star}}^2 = (R_{\text{planet}}/R_{\text{star}})^2$$

- Set $\Delta F_{\text{spots}} = \Delta F_{\text{planet}}$:

$$0.4 \times \left(\frac{5}{6}\right)^4 = (R_{\text{planet}}/R_{\text{star}})^2$$

7. **Solve for $R_{\text{planet}}/R_{\text{star}}$:** - Calculate $\left(\frac{5}{6}\right)^4$:

$$\left(\frac{5}{6}\right)^4 = \frac{625}{1296} \approx 0.4823$$

- Substitute back:

$$0.4 \times 0.4823 \approx 0.1929$$

- Solve for $R_{\text{planet}}/R_{\text{star}}$:

$$(R_{\text{planet}}/R_{\text{star}})^2 = 0.1929$$

$$R_{\text{planet}}/R_{\text{star}} = \sqrt{0.1929} \approx 0.439$$

8. **Choose the Closest Answer:** - The closest answer to 0.439 is 0.39.
ANSWER: D

Analysis of the Different Procedures: Question C and Question D share terms such as "star," "dark spots," "effective temperature," and "6000K," but ask about different physical quantities. Their generated attempts make the following distinction explicit:

- **Question C (Quantum Statistical Mechanics):** The trajectory invokes the *Boltzmann distribution* and the wavelength-dependent energy difference $\Delta E = hc/\lambda$ to compare atomic level populations at two temperatures.

- **Question D (Macroscopic Photometry):** The trajectory invokes a T^4 flux relation and geometric area ratios to compare brightness modulation with a transit signal.

These different solution concepts offer a plausible explanation for the recorded distance increase, consistent with a semantic-repeller effect. They do not imply that question-only embeddings inevitably assign the pair to one cluster, or that the generated calculations are correct. Question C mishandles powers of ten in $hc/(\lambda k)$ and the subsequent exponential. Question D confuses the flux remaining from a spotted region with the loss relative to an unspotted region, in addition to its assumptions about projected area. These are substantive errors, and we retain them precisely because they show that the representation can separate the concepts a question invokes from the correctness of their execution.

C Extended Main Results

In the main text (Section 4.2), the history-dependent comparisons use $H = 25$ calibration models. Here, we retain the full results for $H \in \{50, 100, 200\}$. History-free SA rows repeated across historical pool sizes reuse the same evaluations; they are not additional independent trials. All GP-IRT estimates, including those using history-free selectors, require historical calibration.

Table 7: Performance comparison at H=50 using the SA estimator. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0890 / 0.9044	0.0382 / **0.9747**	0.0246 / 0.9851	0.0187 / **0.9897**	0.0142 / 0.9920
	Question-Emb K-Means	0.0875 / 0.8951	**0.0321** / 0.9743	0.0241 / 0.9797	**0.0175** / 0.9871	**0.0133** / **0.9931**
	k-Center Greedy	**0.0683** / **0.9118**	0.0349 / 0.9738	0.0283 / 0.9847	0.0218 / **0.9897**	0.0175 / 0.9925
	CoT-Core (Ours)	0.0755 / 0.8884	0.0331 / 0.9696	**0.0209** / **0.9853**	0.0197 / 0.9884	0.0154 / 0.9918
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0623 / 0.9130	0.0339 / 0.9748	0.0256 / 0.9816	0.0230 / 0.9848	0.0195 / 0.9887
	IRT	0.0670 / 0.9053	0.0329 / 0.9754	0.0241 / 0.9833	0.0169 / 0.9908	0.0143 / 0.9907
	History-Free Selectors					
MMLU	Random	0.0330 / **0.9489**	0.0149 / 0.9880	0.0102 / 0.9935	**0.0079** / 0.9957	0.0063 / 0.9969
	Question-Emb K-Means	**0.0298** / 0.9484	0.0215 / 0.9877	0.0154 / 0.9923	0.0203 / 0.9943	0.0252 / 0.9958
	k-Center Greedy	0.0477 / 0.9461	0.0530 / 0.9867	0.0393 / 0.9909	0.0360 / 0.9942	0.0317 / 0.9952
	CoT-Core (Ours)	0.0324 / 0.9353	**0.0118** / **0.9881**	**0.0081** / **0.9941**	0.0081 / **0.9964**	**0.0060** / **0.9971**
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0272 / 0.9638	0.0191 / 0.9789	0.0161 / 0.9825	0.0152 / 0.9838	0.0144 / 0.9851
	IRT	0.0225 / 0.9588	0.0113 / 0.9896	0.0083 / 0.9938	0.0069 / 0.9960	0.0065 / 0.9968
	History-Free Selectors					
MMLU-Pro	Random	0.0311 / 0.9556	0.0132 / **0.9899**	0.0093 / 0.9941	**0.0070** / 0.9956	0.0058 / 0.9967
	Question-Emb K-Means	0.0300 / 0.9579	0.0133 / 0.9890	0.0111 / **0.9947**	0.0116 / 0.9957	0.0100 / **0.9969**
	k-Center Greedy	0.0369 / 0.9524	0.0432 / 0.9857	0.0334 / 0.9923	0.0306 / 0.9936	0.0272 / 0.9957
	CoT-Core (Ours)	**0.0296** / **0.9682**	**0.0124** / 0.9893	**0.0082** / 0.9941	0.0072 / **0.9965**	**0.0055** / **0.9969**
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0283 / 0.9769	0.0172 / 0.9903	0.0152 / 0.9933	0.0131 / 0.9942	0.0113 / 0.9947
	IRT	0.0412 / 0.9642	0.0208 / 0.9902	0.0151 / 0.9938	0.0122 / 0.9955	0.0102 / 0.9963
	History-Free Selectors					
GPQA	Random	- / -	0.1252 / 0.2851	0.0797 / 0.4480	0.0635 / 0.5579	0.0536 / 0.6022
	Question-Emb K-Means	- / -	0.1098 / 0.2856	0.0859 / 0.3656	0.0615 / 0.5338	0.0522 / **0.6263**
	k-Center Greedy	- / -	0.1232 / 0.2197	0.0821 / 0.4036	0.0608 / 0.5355	0.0565 / 0.6151
	CoT-Core (Ours)	- / -	**0.1044** / **0.4334**	**0.0731** / **0.5839**	**0.0601** / **0.5646**	**0.0516** / 0.6013
	History-Dependent Selectors (H=50)					
	Correctness K-Means	- / -	0.1269 / 0.2524	0.0942 / 0.3297	0.0771 / 0.4398	0.0666 / 0.4601
	IRT	- / -	0.1047 / 0.3878	0.0780 / 0.4071	0.0656 / 0.5104	0.0540 / 0.5748

Table 8: Performance comparison at H=50 using GP-IRT. All rows use historical calibration for estimation. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0933 / **0.8809**	0.0325 / 0.9670	0.0203 / 0.9839	**0.0157** / 0.9894	0.0129 / 0.9917
	Question-Emb K-Means	**0.0781** / 0.8796	0.0310 / **0.9679**	0.0204 / 0.9815	0.0161 / 0.9871	**0.0123** / **0.9930**
	k-Center Greedy	0.0972 / 0.8676	0.0333 / 0.9658	0.0214 / 0.9833	0.0167 / **0.9899**	0.0143 / 0.9929
	CoT-Core (Ours)	0.0849 / 0.8713	**0.0297** / 0.9592	**0.0192** / **0.9856**	0.0163 / 0.9893	0.0133 / 0.9918
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0698 / 0.9103	0.0292 / 0.9718	0.0183 / 0.9855	0.0152 / 0.9887	0.0140 / 0.9926
	IRT	0.0805 / 0.8858	0.0278 / 0.9642	0.0191 / 0.9844	0.0149 / 0.9910	0.0128 / 0.9924
	History-Free Selectors					
MMLU	Random	**0.0207** / **0.9658**	0.0106 / **0.9908**	0.0082 / 0.9945	0.0067 / 0.9961	**0.0056** / 0.9971
	Question-Emb K-Means	0.0220 / 0.9597	0.0123 / 0.9899	0.0108 / 0.9934	0.0146 / 0.9948	0.0192 / 0.9961
	k-Center Greedy	0.0208 / 0.9651	0.0239 / 0.9892	0.0244 / 0.9925	0.0257 / 0.9945	0.0244 / 0.9955
	CoT-Core (Ours)	0.0233 / 0.9535	**0.0100** / 0.9900	**0.0068** / **0.9949**	**0.0066** / **0.9965**	0.0056 / **0.9971**
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0207 / 0.9723	0.0141 / 0.9868	0.0129 / 0.9887	0.0125 / 0.9892	0.0121 / 0.9894
	IRT	0.0213 / 0.9655	0.0105 / 0.9911	0.0077 / 0.9949	0.0066 / 0.9964	0.0065 / 0.9971
	History-Free Selectors					
MMLU-Pro	Random	0.0239 / 0.9624	**0.0097** / **0.9906**	0.0072 / 0.9943	**0.0058** / 0.9957	0.0049 / 0.9967
	Question-Emb K-Means	0.0238 / 0.9593	0.0098 / 0.9899	0.0075 / **0.9944**	0.0073 / 0.9957	0.0069 / **0.9968**
	k-Center Greedy	0.0249 / 0.9582	0.0146 / 0.9895	0.0149 / 0.9936	0.0163 / 0.9947	0.0163 / 0.9956
	CoT-Core (Ours)	**0.0226** / **0.9691**	0.0105 / 0.9899	**0.0070** / 0.9939	0.0059 / **0.9959**	**0.0047** / 0.9966
	History-Dependent Selectors (H=50)					
	Correctness K-Means	0.0226 / 0.9668	0.0102 / 0.9909	0.0103 / 0.9946	0.0101 / 0.9951	0.0093 / 0.9954
	IRT	0.0278 / 0.9518	0.0100 / 0.9914	0.0083 / 0.9947	0.0079 / 0.9958	0.0076 / 0.9966
	History-Free Selectors					
GPQA	Random	- / -	0.1281 / 0.2834	**0.0596** / 0.4492	0.0414 / **0.6098**	**0.0352** / **0.6768**
	Question-Emb K-Means	- / -	0.1370 / 0.2460	0.0601 / 0.4870	**0.0379** / 0.6004	0.0362 / 0.6651
	k-Center Greedy	- / -	**0.1172** / 0.3289	0.0612 / 0.4532	0.0434 / 0.5656	0.0364 / 0.6267
	CoT-Core (Ours)	- / -	0.1242 / **0.3571**	0.0612 / **0.5275**	0.0412 / 0.5904	0.0362 / 0.6616
	History-Dependent Selectors (H=50)					
	Correctness K-Means	- / -	0.1028 / 0.3905	0.0491 / 0.4433	0.0382 / 0.5559	0.0373 / 0.6144
	IRT	- / -	0.1117 / 0.3240	0.0505 / 0.4854	0.0411 / 0.6455	0.0352 / 0.6663

Table 9: Performance comparison at H=100 using the SA estimator. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0890 / 0.9044	0.0382 / **0.9747**	0.0246 / 0.9851	0.0187 / **0.9897**	0.0142 / 0.9920
	Question-Emb K-Means	0.0875 / 0.8951	**0.0321** / 0.9743	0.0241 / 0.9797	**0.0175** / 0.9871	**0.0133** / **0.9931**
	k-Center Greedy	**0.0683** / **0.9118**	0.0349 / 0.9738	0.0283 / 0.9847	0.0218 / **0.9897**	0.0175 / 0.9925
	CoT-Core (Ours)	0.0755 / 0.8884	0.0331 / 0.9696	**0.0209** / **0.9853**	0.0197 / 0.9884	0.0154 / 0.9918
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0522 / 0.9226	0.0344 / 0.9699	0.0299 / 0.9843	0.0266 / 0.9865	0.0233 / 0.9903
	IRT	0.0650 / 0.9049	0.0295 / 0.9734	0.0216 / 0.9797	0.0168 / 0.9880	0.0141 / 0.9893
	History-Free Selectors					
MMLU	Random	0.0330 / **0.9489**	0.0149 / 0.9880	0.0102 / 0.9935	**0.0079** / 0.9957	0.0063 / 0.9969
	Question-Emb K-Means	**0.0298** / 0.9484	0.0215 / 0.9877	0.0154 / 0.9923	0.0203 / 0.9943	0.0252 / 0.9958
	k-Center Greedy	0.0477 / 0.9461	0.0530 / 0.9867	0.0393 / 0.9909	0.0360 / 0.9942	0.0317 / 0.9952
	CoT-Core (Ours)	0.0324 / 0.9353	**0.0118** / **0.9881**	**0.0081** / **0.9941**	0.0081 / **0.9964**	**0.0060** / **0.9971**
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0184 / 0.9762	0.0111 / 0.9906	0.0100 / 0.9917	0.0091 / 0.9927	0.0082 / 0.9936
	IRT	0.0222 / 0.9565	0.0105 / 0.9895	0.0089 / 0.9932	0.0071 / 0.9950	0.0063 / 0.9962
	History-Free Selectors					
MMLU-Pro	Random	0.0311 / 0.9556	0.0132 / **0.9899**	0.0093 / 0.9941	**0.0070** / 0.9956	0.0058 / 0.9967
	Question-Emb K-Means	0.0300 / 0.9579	0.0133 / 0.9890	0.0111 / **0.9947**	0.0116 / 0.9957	0.0100 / **0.9969**
	k-Center Greedy	0.0369 / 0.9524	0.0432 / 0.9857	0.0334 / 0.9923	0.0306 / 0.9936	0.0272 / 0.9957
	CoT-Core (Ours)	**0.0296** / **0.9682**	**0.0124** / 0.9893	**0.0082** / 0.9941	0.0072 / **0.9965**	**0.0055** / **0.9969**
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0289 / 0.9815	0.0219 / 0.9919	0.0192 / 0.9937	0.0175 / 0.9940	0.0157 / 0.9948
	IRT	0.0274 / 0.9561	0.0111 / 0.9923	0.0085 / 0.9938	0.0064 / 0.9953	0.0058 / 0.9961
	History-Free Selectors					
GPQA	Random	- / -	0.1252 / 0.2851	0.0797 / 0.4480	0.0635 / 0.5579	0.0536 / 0.6022
	Question-Emb K-Means	- / -	0.1098 / 0.2856	0.0859 / 0.3656	0.0615 / 0.5338	0.0522 / **0.6263**
	k-Center Greedy	- / -	0.1232 / 0.2197	0.0821 / 0.4036	0.0608 / 0.5355	0.0565 / 0.6151
	CoT-Core (Ours)	- / -	**0.1044** / **0.4334**	**0.0731** / **0.5839**	**0.0601** / **0.5646**	**0.0516** / 0.6013
	History-Dependent Selectors (H=100)					
	Correctness K-Means	- / -	0.1124 / 0.4882	0.0925 / 0.4670	0.0792 / 0.5425	0.0742 / 0.5639
	IRT	- / -	0.1125 / 0.3425	0.0713 / 0.4955	0.0511 / 0.6022	0.0466 / 0.6193

Table 10: Performance comparison at H=100 using GP-IRT. All rows use historical calibration for estimation. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0881 / **0.8824**	0.0292 / **0.9732**	0.0200 / 0.9856	0.0161 / **0.9899**	0.0130 / 0.9921
	Question-Emb K-Means	0.0899 / 0.8723	**0.0276** / 0.9723	0.0210 / 0.9804	**0.0158** / 0.9877	**0.0125** / **0.9931**
	k-Center Greedy	0.0831 / 0.8703	0.0279 / 0.9725	0.0216 / 0.9832	0.0177 / 0.9897	0.0149 / 0.9925
	CoT-Core (Ours)	**0.0818** / 0.8751	0.0285 / 0.9664	**0.0188** / **0.9860**	0.0162 / 0.9889	0.0135 / 0.9917
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0741 / 0.9099	0.0263 / 0.9727	0.0205 / 0.9893	0.0195 / 0.9910	0.0177 / 0.9934
	IRT	0.0651 / 0.8984	0.0260 / 0.9748	0.0195 / 0.9809	0.0152 / 0.9894	0.0128 / 0.9902
	History-Free Selectors					
MMLU	Random	**0.0199** / 0.9662	0.0098 / 0.9909	0.0076 / 0.9944	0.0062 / 0.9960	**0.0053** / 0.9970
	Question-Emb K-Means	0.0206 / 0.9626	0.0107 / 0.9895	0.0086 / 0.9940	0.0103 / 0.9949	0.0132 / 0.9962
	k-Center Greedy	0.0209 / **0.9684**	0.0129 / 0.9902	0.0141 / 0.9935	0.0161 / 0.9953	0.0164 / 0.9959
	CoT-Core (Ours)	0.0218 / 0.9631	**0.0096** / **0.9911**	**0.0069** / **0.9952**	**0.0061** / **0.9964**	0.0054 / **0.9971**
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0190 / 0.9715	0.0100 / 0.9906	0.0083 / 0.9932	0.0075 / 0.9941	0.0066 / 0.9948
	IRT	0.0188 / 0.9684	0.0099 / 0.9913	0.0076 / 0.9945	0.0064 / 0.9957	0.0057 / 0.9966
	History-Free Selectors					
MMLU-Pro	Random	0.0229 / 0.9629	0.0094 / **0.9907**	0.0066 / **0.9947**	0.0053 / 0.9960	0.0045 / 0.9969
	Question-Emb K-Means	0.0227 / 0.9581	**0.0093** / 0.9899	0.0069 / 0.9946	0.0061 / 0.9959	0.0054 / 0.9970
	k-Center Greedy	0.0229 / 0.9640	0.0102 / 0.9900	0.0086 / 0.9941	0.0096 / 0.9949	0.0099 / 0.9958
	CoT-Core (Ours)	**0.0209** / **0.9706**	0.0099 / 0.9903	**0.0065** / 0.9944	**0.0051** / **0.9965**	**0.0043** / **0.9972**
	History-Dependent Selectors (H=100)					
	Correctness K-Means	0.0295 / 0.9550	0.0103 / 0.9927	0.0091 / 0.9956	0.0085 / 0.9970	0.0086 / 0.9970
	IRT	0.0255 / 0.9520	0.0088 / 0.9929	0.0066 / 0.9948	0.0051 / 0.9962	0.0046 / 0.9971
	History-Free Selectors					
GPQA	Random	- / -	0.1022 / 0.2447	0.0549 / 0.4685	0.0415 / 0.5990	0.0377 / 0.6439
	Question-Emb K-Means	- / -	0.0990 / 0.2167	0.0557 / 0.3998	0.0412 / 0.5598	0.0376 / **0.6586**
	k-Center Greedy	- / -	0.1040 / 0.1707	0.0565 / 0.3994	0.0446 / 0.5275	0.0411 / 0.5964
	CoT-Core (Ours)	- / -	**0.0836** / **0.3995**	**0.0482** / **0.6417**	**0.0380** / **0.6166**	**0.0369** / 0.6505
	History-Dependent Selectors (H=100)					
	Correctness K-Means	- / -	0.0653 / 0.4379	0.0443 / 0.5324	0.0415 / 0.5921	0.0431 / 0.6339
	IRT	- / -	0.0944 / 0.2877	0.0490 / 0.5302	0.0360 / 0.6203	0.0330 / 0.6608

Table 11: Performance comparison at H=200 using the SA estimator. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0890 / 0.9044	0.0382 / **0.9747**	0.0246 / 0.9851	0.0187 / **0.9897**	0.0142 / 0.9920
	Question-Emb K-Means	0.0875 / 0.8951	**0.0321** / 0.9743	0.0241 / 0.9797	**0.0175** / 0.9871	**0.0133** / **0.9931**
	k-Center Greedy	**0.0683** / **0.9118**	0.0349 / 0.9738	0.0283 / 0.9847	0.0218 / **0.9897**	0.0175 / 0.9925
	CoT-Core (Ours)	0.0755 / 0.8884	0.0331 / 0.9696	**0.0209** / **0.9853**	0.0197 / 0.9884	0.0154 / 0.9918
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0546 / 0.9239	0.0348 / 0.9694	0.0303 / 0.9800	0.0277 / 0.9817	0.0259 / 0.9870
	IRT	0.0723 / 0.9048	0.0283 / 0.9757	0.0205 / 0.9841	0.0172 / 0.9866	0.0148 / 0.9895
	History-Free Selectors					
MMLU	Random	0.0330 / **0.9489**	0.0149 / 0.9880	0.0102 / 0.9935	**0.0079** / 0.9957	0.0063 / 0.9969
	Question-Emb K-Means	**0.0298** / 0.9484	0.0215 / 0.9877	0.0154 / 0.9923	0.0203 / 0.9943	0.0252 / 0.9958
	k-Center Greedy	0.0477 / 0.9461	0.0530 / 0.9867	0.0393 / 0.9909	0.0360 / 0.9942	0.0317 / 0.9952
	CoT-Core (Ours)	0.0324 / 0.9353	**0.0118** / **0.9881**	**0.0081** / **0.9941**	0.0081 / **0.9964**	**0.0060** / **0.9971**
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0211 / 0.9699	0.0139 / 0.9885	0.0112 / 0.9898	0.0099 / 0.9911	0.0085 / 0.9935
	IRT	0.0223 / 0.9558	0.0097 / 0.9901	0.0072 / 0.9928	0.0060 / 0.9940	0.0055 / 0.9956
	History-Free Selectors					
MMLU-Pro	Random	0.0311 / 0.9556	0.0132 / **0.9899**	0.0093 / 0.9941	**0.0070** / 0.9956	0.0058 / 0.9967
	Question-Emb K-Means	0.0300 / 0.9579	0.0133 / 0.9890	0.0111 / **0.9947**	0.0116 / 0.9957	0.0100 / **0.9969**
	k-Center Greedy	0.0369 / 0.9524	0.0432 / 0.9857	0.0334 / 0.9923	0.0306 / 0.9936	0.0272 / 0.9957
	CoT-Core (Ours)	**0.0296** / **0.9682**	**0.0124** / 0.9893	**0.0082** / 0.9941	0.0072 / **0.9965**	**0.0055** / **0.9969**
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0332 / 0.9847	0.0244 / 0.9918	0.0222 / 0.9927	0.0194 / 0.9937	0.0174 / 0.9940
	IRT	0.0233 / 0.9582	0.0123 / 0.9901	0.0080 / 0.9939	0.0061 / 0.9957	0.0054 / 0.9963
	History-Free Selectors					
GPQA	Random	- / -	0.1252 / 0.2851	0.0797 / 0.4480	0.0635 / 0.5579	0.0536 / 0.6022
	Question-Emb K-Means	- / -	0.1098 / 0.2856	0.0859 / 0.3656	0.0615 / 0.5338	0.0522 / **0.6263**
	k-Center Greedy	- / -	0.1232 / 0.2197	0.0821 / 0.4036	0.0608 / 0.5355	0.0565 / 0.6151
	CoT-Core (Ours)	- / -	**0.1044** / **0.4334**	**0.0731** / **0.5839**	**0.0601** / **0.5646**	**0.0516** / 0.6013
	History-Dependent Selectors (H=200)					
	Correctness K-Means	- / -	0.1261 / 0.2994	0.1072 / 0.3593	0.0915 / 0.4311	0.0839 / 0.4810
	IRT	- / -	0.1055 / 0.2956	0.0776 / 0.4353	0.0610 / 0.5367	0.0529 / 0.5793

Table 12: Performance comparison at H=200 using GP-IRT. All rows use historical calibration for estimation. Boldface marks the best history-free value in each column; history-dependent selectors are listed separately.

Benchmark	Method	Sampling Ratio				
		1%	5%	10%	15%	20%
		MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)	MAE (↓) / Sim (↑)
	History-Free Selectors					
GSM8K	Random	0.0881 / 0.8761	0.0287 / **0.9744**	0.0192 / 0.9859	**0.0152** / **0.9898**	0.0125 / 0.9921
	Question-Emb K-Means	0.0869 / 0.8594	0.0279 / 0.9729	0.0203 / 0.9811	0.0153 / 0.9874	**0.0121** / **0.9931**
	k-Center Greedy	0.0879 / 0.8655	**0.0265** / 0.9731	0.0192 / 0.9817	0.0158 / 0.9891	0.0135 / 0.9923
	CoT-Core (Ours)	**0.0738** / **0.8793**	0.0278 / 0.9693	**0.0187** / **0.9860**	0.0157 / 0.9886	0.0131 / 0.9914
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0705 / 0.9242	0.0260 / 0.9762	0.0198 / 0.9875	0.0175 / 0.9890	0.0163 / 0.9918
	IRT	0.0706 / 0.9023	0.0252 / 0.9758	0.0181 / 0.9857	0.0142 / 0.9892	0.0130 / 0.9916
	History-Free Selectors					
MMLU	Random	**0.0186** / 0.9710	0.0093 / **0.9919**	0.0074 / 0.9948	0.0062 / 0.9963	0.0053 / 0.9972
	Question-Emb K-Means	0.0201 / 0.9657	0.0104 / 0.9906	0.0090 / 0.9942	0.0112 / 0.9953	0.0150 / 0.9964
	k-Center Greedy	0.0188 / **0.9720**	0.0160 / 0.9914	0.0173 / 0.9937	0.0192 / 0.9955	0.0190 / 0.9959
	CoT-Core (Ours)	0.0216 / 0.9617	**0.0085** / **0.9919**	**0.0063** / **0.9956**	**0.0059** / **0.9966**	**0.0052** / **0.9974**
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0186 / 0.9722	0.0094 / 0.9925	0.0080 / 0.9938	0.0073 / 0.9944	0.0064 / 0.9956
	IRT	0.0184 / 0.9714	0.0086 / 0.9921	0.0067 / 0.9938	0.0055 / 0.9954	0.0049 / 0.9964
	History-Free Selectors					
MMLU-Pro	Random	0.0225 / 0.9646	0.0094 / **0.9908**	0.0069 / 0.9942	**0.0055** / **0.9955**	0.0048 / 0.9963
	Question-Emb K-Means	0.0228 / 0.9611	**0.0094** / 0.9904	0.0074 / **0.9943**	0.0069 / 0.9952	0.0062 / **0.9967**
	k-Center Greedy	0.0228 / 0.9673	0.0117 / 0.9901	0.0111 / 0.9940	0.0127 / 0.9949	0.0127 / 0.9958
	CoT-Core (Ours)	**0.0216** / **0.9747**	0.0098 / 0.9907	**0.0069** / 0.9937	0.0056 / 0.9954	**0.0047** / 0.9962
	History-Dependent Selectors (H=200)					
	Correctness K-Means	0.0234 / 0.9685	0.0097 / 0.9933	0.0092 / 0.9954	0.0087 / 0.9966	0.0090 / 0.9968
	IRT	0.0221 / 0.9619	0.0095 / 0.9908	0.0072 / 0.9940	0.0052 / 0.9958	0.0046 / 0.9963
	History-Free Selectors					
GPQA	Random	- / -	0.1108 / 0.2931	0.0875 / 0.4409	0.0578 / 0.5951	0.0423 / **0.6625**
	Question-Emb K-Means	- / -	0.1111 / 0.2699	0.0826 / 0.4144	0.0521 / **0.6110**	0.0448 / 0.6466
	k-Center Greedy	- / -	0.1170 / 0.2745	0.0863 / 0.4452	0.0623 / 0.5391	0.0487 / 0.5753
	CoT-Core (Ours)	- / -	**0.1031** / **0.3926**	**0.0737** / **0.5764**	**0.0492** / 0.6092	**0.0423** / 0.6502
	History-Dependent Selectors (H=200)					
	Correctness K-Means	- / -	0.0805 / 0.3920	0.0605 / 0.4511	0.0449 / 0.4943	0.0406 / 0.5771
	IRT	- / -	0.0837 / 0.3378	0.0586 / 0.4497	0.0484 / 0.5412	0.0374 / 0.6340

D Configuration Sensitivity

These ablations vary the preprocessing and clustering stages around the default configuration of Section 4.1, and each block carries its own reference row.

D.1 Trajectory Transformation

To examine sensitivity to trajectory transformation, we compare three levels on GPQA using Phi-4 (prompts in Appendix A.5): *Raw CoT* (unaltered), *Cleaned CoT* (a prompted removal of boilerplate and failed paths while retaining entities), and *Abstract CoT* (a compressed operation-level description).

Table 13: Hierarchical-cleaning ablation on GPQA. Cleaned CoT improves the reported MAE, with a ranking trade-off at 10%.

Abstraction Level	GPQA Budget (MAE ↓ / Sim ↑)		
	5%	10%	20%
Raw CoT	0.1044 / 0.4334	0.0731 / **0.5839**	0.0516 / 0.6013
Cleaned CoT	**0.1007** / **0.5041**	**0.0666** / 0.4970	**0.0442** / **0.6438**
Abstract CoT	0.1206 / 0.3561	0.0831 / 0.4506	0.0518 / 0.5714

Cleaned CoT has lower MAE at all three reported budgets, with a lower Sim than Raw CoT at 10% (Table 13). *Abstract CoT* generally loses accuracy relative to Raw. A possible **semantic-starvation**

explanation is that removing domain-specific anchors discards useful selection information. However, the prompts jointly change length, domain detail, failed paths, and self-corrections; this experiment does not isolate abstraction or verbosity alone. The length-matched placebo study addresses a different comparison, namely the content of appended text at matched token counts, rather than confirming that explanation. We keep Raw CoT as the default because it needs a single generation pass, and report cleaning as an optional refinement that buys lower error with a second pass.

D.2 Embedding Model

To test whether the effect depends on BGE-M3, we compare question-only and question-CoT features under BGE-M3, Jina-v3, and Qwen3-8B on MMLU-Pro; the first two are encoder-based embedding models, while the Qwen configuration uses a decoder-based representation. CoT lowers MAE for all three encoders at every listed budget (Table 14), though the margin varies with the encoder. Sim moves less uniformly, with one exception per encoder (BGE-M3 at 10%, Jina-v3 at 1%, Qwen3-8B at 20%), so the MAE gain is the property that transfers across encoders here.

Table 14: Embedding-model ablation on MMLU-Pro with SA. Entries are MAE (↓) / Sim (↑) at nominal budgets. The BGE-M3 rows are the main-table configuration; the Jina-v3 and Qwen3-8B rows come from the additional encoder runs.

Encoder	Representation	1%	10%	20%
BGE-M3	Question-Emb	0.0300 / 0.9579	0.0111 / 0.9947	0.0100 / 0.9969
BGE-M3	CoT-Core	0.0296 / 0.9682	0.0082 / 0.9941	0.0055 / 0.9969
Jina-v3	Question-Emb	0.0319 / 0.9658	0.0090 / 0.9945	0.0064 / 0.9965
Jina-v3	CoT-Core	0.0305 / 0.9558	0.0088 / 0.9945	0.0061 / 0.9971
Qwen3-8B	Question-Emb	0.0357 / 0.9650	0.0125 / 0.9939	0.0118 / 0.9969
Qwen3-8B	CoT-Core	0.0345 / 0.9734	0.0089 / 0.9942	0.0062 / 0.9965

D.3 Clustering Algorithm

We additionally compare ordinary K-Means on normalized embeddings, explicit Spherical K-Means, and K-Medoids on MMLU-Pro. For unit vectors u, v, $\|u - v\|_2^2 = 2(1 - u^\top v)$, yet the two variants remain distinct because only Spherical K-Means constrains its centroids to the sphere. In Table 15 they nonetheless give nearly identical point estimates, so we keep the simpler ordinary K-Means. K-Medoids has higher error at 1% and 10% and lower error at 20%, so no single rule dominates across budgets and the reported behavior is not an artifact of one clustering algorithm.

Table 15: Clustering ablation on MMLU-Pro with SA. Entries are MAE (↓) / Sim (↑). Each algorithm is evaluated in the clustering-ablation block, including its own K-Means reference; these entries do not replace the main-table configuration.

Clustering method	1%	10%	20%
K-Means	0.0278 / 0.9685	0.0080 / 0.9945	0.0062 / 0.9969
Spherical K-Means	0.0278 / 0.9685	0.0081 / 0.9945	0.0059 / 0.9970
K-Medoids	0.0333 / 0.9631	0.0089 / 0.9938	0.0052 / 0.9970

E Conditional Reconstruction Bound

This appendix states and proves the bound that relates the clustering geometry of Section 3.5 to score-reconstruction error. Define the normalized quantization error $Q = N^{-1} \sum_k \sum_{i \in C_k} \|\mathbf{e}_i - \mu_k\|_2^2$ and the cluster masses $\pi_k = |C_k|/N$.

Proposition 1 (conditional reconstruction bound). For a fixed target model, suppose that scores within each cluster satisfy

$$|y_i - y_j| \le L_{\mathcal{M}} \|\mathbf{e}_i - \mathbf{e}_j\|_2 + \eta_{\mathcal{M}}, \qquad i, j \in C_k, \tag{7}$$

for constants $L_{\mathcal{M}}, \eta_{\mathcal{M}} \ge 0$. With the arithmetic centroids and within-cluster representatives of Eq. (3), the Standard Aggregation estimate of Eq. (4) satisfies

$$\left|\widehat{S}_w - S_{\mathcal{D}}(\mathcal{M})\right| \le L_{\mathcal{M}} \sqrt{2Q} + \eta_{\mathcal{M}} + \frac{1}{2} \sum_{k=1}^{B} |w_k - \pi_k|. \tag{8}$$

The three terms separate quantization error, the within-cluster response residual, and the discrepancy between the reported weights and the cluster masses; the last term vanishes under cluster-mass weighting. Consequently, a compact representation alone does not control estimation error: items placed in one cluster must also elicit similar responses.

Setup. We prove Proposition 1 for a fixed target model and a fixed partition into nonempty clusters. Write $n_k = |C_k|$, $q_k = n_k^{-1} \sum_{i \in C_k} \|\mathbf{e}_i - \mu_k\|_2^2$, and $\pi_k = n_k/N$, so that $Q = \sum_k \pi_k q_k$. The centroid μ_k is the arithmetic mean of the vectors in C_k and s_k is a nearest member of C_k to that mean.

Representative quantization error. By the choice of s_k, its squared distance to the centroid cannot exceed the average squared distance:

$$\|\mathbf{e}_{s_k} - \mu_k\|_2^2 \le q_k. \tag{9}$$

Since $\sum_{i \in C_k} (\mathbf{e}_i - \mu_k) = 0$, the centroid identity yields

$$\frac{1}{n_k} \sum_{i \in C_k} \|\mathbf{e}_i - \mathbf{e}_{s_k}\|_2^2 = q_k + \|\mathbf{e}_{s_k} - \mu_k\|_2^2 \le 2q_k. \tag{10}$$

Summing with weights π_k and applying Cauchy–Schwarz gives

$$\frac{1}{N} \sum_k \sum_{i \in C_k} \|\mathbf{e}_i - \mathbf{e}_{s_k}\|_2 \le \left(\frac{1}{N} \sum_k \sum_{i \in C_k} \|\mathbf{e}_i - \mathbf{e}_{s_k}\|_2^2 \right)^{1/2} \le \sqrt{2Q}. \tag{11}$$

Cluster-mass reconstruction. For $\widehat{S}_\pi = \sum_k \pi_k y_{s_k}$, the triangle inequality and the response-smoothness assumption in Eq. (7) imply

$$\left|\widehat{S}_\pi - S_{\mathcal{D}}(\mathcal{M})\right| = \left| \frac{1}{N} \sum_k \sum_{i \in C_k} (y_{s_k} - y_i) \right| \tag{12}$$

$$\le \frac{1}{N} \sum_k \sum_{i \in C_k} |y_{s_k} - y_i| \tag{13}$$

$$\le \frac{L_{\mathcal{M}}}{N} \sum_k \sum_{i \in C_k} \|\mathbf{e}_{s_k} - \mathbf{e}_i\|_2 + \eta_{\mathcal{M}} \tag{14}$$

$$\le L_{\mathcal{M}} \sqrt{2Q} + \eta_{\mathcal{M}}. \tag{15}$$

The first inequality also supplies a finite-sample response-consistency bound without any smoothness assumption. Its right-hand side can be measured after selection, but using full target-model responses to optimize it would change the information available to the selector.

Changing aggregation weights. Let $a_k = w_k - \pi_k$. Because $\sum_k a_k = 0$ and $0 \le y_{s_k} \le 1$, the positive and negative masses of a both equal $\frac{1}{2} \sum_k |a_k|$. Consequently,

$$|\widehat{S}_w - \widehat{S}_\pi| = \left| \sum_k a_k y_{s_k} \right| \le \frac{1}{2} \sum_k |w_k - \pi_k|. \tag{16}$$

Combining this inequality with the cluster-mass reconstruction bound proves Eq. (8). □

Scope of the result. The proof does not require a global optimum of K-Means and applies to the union of subject-wise partitions. It does require arithmetic centroids and within-cluster nearest representatives. The constants in the response-smoothness assumption are representation- and model-dependent; choosing a large residual, such as $\eta_{\mathcal{M}} = 1$, makes the result vacuous for scores in $[0, 1]$. A useful bound therefore needs small, independently supported constants together with a small weighting discrepancy. The proposition accordingly specifies what a selection representation must deliver for geometric coverage to control estimation error, and Section 4 supplies the corresponding comparison between question-trajectory and question-only features.

F Limitations and Practical Use

Computational trade-offs. The cost of generating trajectories over the entire benchmark is incurred once, before any subset is reused. We therefore report this construction term separately from the recurring downstream cost, and Section 4.4 shows that a 1B generator already yields competitive subsets, so the term can be kept small. How many reuses are needed to amortize it depends on the generator, target model, hardware, and output lengths.

Generator dependence and representation validity. A generated trajectory describes the generator's attempted solution, so arithmetic mistakes, incorrect methods, truncation, and stylistic boilerplate all enter the representation; how often such errors help or hurt selection is not measured here, and downstream estimation error remains our primary evidence. The placebo study matches token counts but not domain content, style, or relevance, so it compares auxiliary text sources rather than the presence or absence of logic. The resulting orderings vary with budget and dataset, including budgets where a placebo or Random outperforms CoT, and the generator, encoder, and clustering ablations each cover a limited set of configurations.

Evaluation scope. The current comparisons concern particular benchmark versions, response logs, and model pools. Benefits are not uniform across datasets, budgets, or metrics. Global rank correlation can mask errors among closely matched models, and high NDCG with similar relevance scores need not preserve their exact ordering. Transfer to newer, stronger, or different model families requires corresponding target-model evaluations; changing the trajectory generator alone does not provide that evidence. Finally, matching solution operations does not guarantee matching item difficulty or target-model success, which is why downstream score reconstruction remains our reported measure.

G Output-Token Use and Score Fidelity

Table 16 joins downstream output-token use with full- versus subset-score fidelity on MMLU-Pro at the nominal 5% budget. This study reruns the full evaluation pipeline for its own eleven target models, so its scope is separate from the main 100-model response pools. Phi-4 appears only as the trajectory generator, with its full-evaluation token count listed for reference and excluded from the summary row. Model names and absolute score gaps are retained as reported.

Let P_m and $\overline{\widehat{P}}_m$ denote full-evaluation and mean subset output tokens, and let S_m and $\overline{\widehat{S}}_m$ be the corresponding scores in $[0, 1]$. With $M = 11$, the summaries are

$$r_{\text{out}} = 1 - \frac{\sum_{m=1}^{M} \overline{\widehat{P}}_m}{\sum_{m=1}^{M} P_m}, \qquad \overline{\Delta}_{\text{pp}} = \frac{100}{M} \sum_{m=1}^{M} |\overline{\widehat{S}}_m - S_m|. \tag{17}$$

Thus token reduction is a ratio of totals, not an unweighted average of model-wise reductions; score gaps weight the models equally. The gap of a trial-averaged score is not the trial-average absolute error, so it does not quantify single-subset reliability. Eight of the eleven reported mean-score gaps are at most one percentage point.

Table 16: Output-token use and score fidelity at a 5% sampling budget. Scores are percentages; absolute gaps are in percentage points (pp). Token reduction excludes input tokens and preprocessing. The summary row sums token counts and averages the reported absolute gaps across the eleven target models; it does not pool their task accuracies. Phi-4 is listed only for its full-evaluation token count, as the trajectory generator rather than a target model of this study.

Target model	Full output tokens	Mean 5% output tokens	Reduction (%)	Full score (%)	Mean 5% score (%)	Absolute gap (pp)
Gemma-3-1B-Instruct	6,489,650	338,505.6	94.78	14.89	13.57	1.31
Gemma-3-4B-Instruct	6,401,580	336,711.4	94.74	42.95	42.70	0.25
LLaMA-3.2-1B-Instruct	5,863,418	310,567.8	94.70	21.61	21.79	0.18
LLaMA-3.2-3B-Instruct	6,461,163	338,832.6	94.76	39.26	38.16	1.10
Mistral-3-3B-Instruct	9,190,233	472,823.2	94.86	54.28	54.11	0.17
Mistral-3-8B-Instruct	8,265,916	439,329.6	94.69	63.09	63.27	0.18
Phi-3.5-MoE	3,623,012	191,211.8	94.72	58.16	57.05	1.11
Phi-mini-MoE	4,521,428	236,958.4	94.76	46.63	46.94	0.31
Phi-tiny-MoE	3,906,998	201,222.0	94.85	34.88	35.86	0.99
Qwen-3-1.7B-Instruct	6,508,844	336,700.8	94.83	40.15	41.10	0.95
Qwen-3-8B-Instruct	8,030,828	419,066.0	94.78	56.75	57.39	0.63
11-model summary	**69,263,070**	**3,621,929.2**	**94.77**	—	—	**0.65**
Phi-4 (generator)	5,832,703	—	—	—	—	—

Precision and protocol. Scores and gaps retain the precision of the underlying records, so three reported gaps differ by 0.01 pp from subtracting the displayed scores. The eleven-model mean gap is computed from the gap column itself rather than from the rounded scores.

Connection to one-time construction. A subset built for a fixed configuration is reused across models, so this table isolates the recurring downstream cost and excludes the one-time construction term. Phi-4 is listed with its full-evaluation count alone, since it serves here as the trajectory generator. The reported quantities are output volume and mean score gaps; converting them into wall-clock or GPU savings depends on the serving setup.

H Additional Length-Matched Placebo Results

The following tables complement the MMLU-Pro comparison in Table 4. Both controls retain the original question and replace its Phi-4 trajectory with token-length-matched Faker synthetic text or an unrelated Wiki passage. They test auxiliary text sources, not verified absence versus presence of reasoning. All entries report mean MAE $\pm$ pooled individual-model error standard deviation, not seed-level uncertainty; boldface denotes the lowest mean among the listed methods.

Remaining MMLU-Pro budgets. Table 17 reports the 1% and 20% budgets alongside the 5% and 10% results in the main text. Wiki has the lowest mean at 1% and CoT at 20%, confirming the budget dependence noted in Section 4.4. At 1% the subject quotas leave roughly 120 of the 12,032 items, about nine per subject, so every selector operates near the resolution limit of this benchmark.

Table 17: Additional length-matched placebo results on MMLU-Pro with SA, complementing the 5% and 10% budgets in Table 4.

Selection features / baseline	1% budget	20% budget
Random	0.03113 ± 0.02420	0.00584 ± 0.00454
Question + unrelated Wiki	$\mathbf{0.02857} \pm 0.02278$	0.00640 ± 0.00449
Question + Faker text	0.02996 ± 0.02163	0.00606 ± 0.00487
Question + CoT (CoT-Core)	0.02957 ± 0.02414	$\mathbf{0.00549} \pm 0.00437$

Separate GPQA comparison. Table 18 retains the GPQA Diamond results as a distinct dataset block. CoT has the lowest mean at 5% and 10%, with absolute reductions of 0.323 and 0.090 percentage points relative to the stronger Faker control. At 20%, both placebo variants outperform CoT, which reinforces the dataset- and budget-dependent nature of the effect.

Table 18: Length-matched placebo results on GPQA Diamond with SA. Each cell shows mean MAE above its pooled error standard deviation. These results are separate from the MMLU-Pro comparisons.

Selection features / baseline	5%	10%	20%
Random	0.12523 $\pm$0.09043	0.07970 $\pm$0.05732	0.05359 $\pm$0.03963
Question + unrelated Wiki	0.11285 $\pm$0.08399	0.07617 $\pm$0.05755	0.04783 $\pm$0.03763
Question + Faker text	0.10765 $\pm$0.07934	0.07402 $\pm$0.05728	**0.04358** $\pm$0.03377
Question + CoT (CoT-Core)	**0.10442** $\pm$0.07620	**0.07312** $\pm$0.05705	0.05160 $\pm$0.03853

Taken together, these tables locate the benefit of question-related auxiliary text in the tight-budget regime, where the coreset must cover the benchmark with the fewest items.