CircuitSeer: Mining High-Quality Data by Probing Mathematical Reasoning Circuits in LLMs

Shaobo Wang[✉*,1,2] **Yongliang Miao**[*,1] **Yuancheng Liu**[*,1]
Qianli Ma[2] **Ning Liao**[2] **Dongrui Liu**[2] **Linfeng Zhang**[✉,1,2]

[1] EPIC Lab, SJTU [2] Shanghai Jiao Tong University
[*] Equal contribution [✉] Corresponding authors

Abstract

Large language models have demonstrated impressive reasoning capabilities, but scaling their performance often relies on massive reasoning datasets that are computationally expensive to train on. Existing data selection methods aim to curate smaller, high-quality subsets but often rely on costly external models or opaque heuristics. In this work, we shift the focus from external heuristics to the model's internal mechanisms. We find that complex reasoning tasks consistently activate a sparse, specialized subset of attention heads, forming core reasoning circuits. Building on this insight, we propose CircuitSeer, a novel data selection method that quantifies the reasoning complexity of data by measuring its influence on these crucial circuits. Extensive experiments on 4 models and 9 datasets demonstrate CircuitSeer's superiority. Notably, fine-tuning Qwen2.5-Math-7B on just 10% of data selected by our method achieves a 1.2-point gain in average Pass@1 over training on the full dataset, highlighting its efficiency and effectiveness.

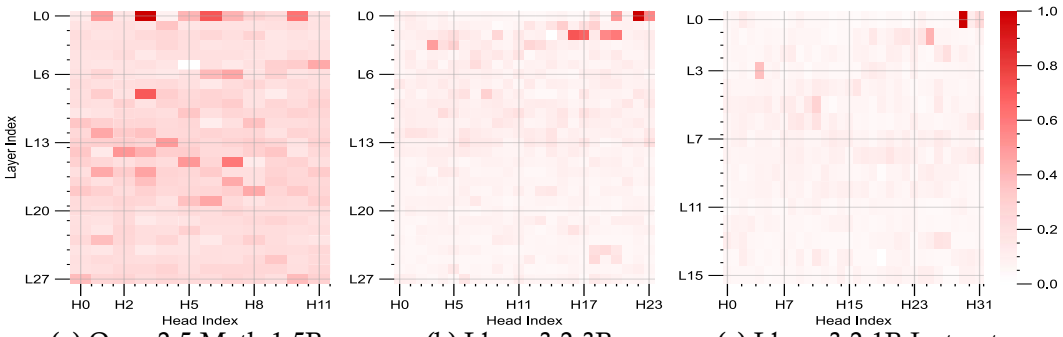
1 Introduction

Large language models (LLMs), like OpenAI-o1 [1] and DeepSeek-R1 [2], have achieved remarkable advances in complex reasoning tasks, significantly contributing to the development of massive reasoning datasets. While datasets like OpenR1 [3], Synthetic-1 [4], and AM [5] provide essential resources for supervised fine-tuning (SFT) to further enhance LLM reasoning capabilities, their massive scale often demands substantial training time and computational resources. This motivates methods that can identify compact, high-value subsets for reasoning-oriented fine-tuning.

Existing methods, such as LIMO [6], s1 [7], and the CoT Collection [8], have attempted to curate high-quality subsets from large-scale data. However, these approaches suffer from several critical limitations. They often rely on complex and resource-intensive pipelines, such as rejection sampling that requires a separate, more powerful model, or model-based quality assessments that are computationally expensive. Furthermore, their core logic is frequently built upon intricate data engineering and qualitative heuristics for quality, difficulty, and diversity. This reliance on external models and heuristic rules makes these methods not only costly but also opaque, difficult to reproduce, and lacking a principled, quantitative basis for determining data quality.

To overcome these limitations, we shift our focus from external heuristics to the model's internal mechanisms, seeking a more fundamental signal for data quality. Through extensive experimentation, we observed a striking and consistent phenomenon: during complex reasoning tasks, only a small, specific subset of attention head circuits remains highly active. Crucially, we find this pattern of sparse, specialized activation to be a consistent trend across the diverse large language models. This empirical finding aligns with the principles of Transformer circuits theory [9, 10], which posits that different circuits emerge within transformer models to handle different tasks. Figure 1 visually

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Qwen2.5-Math-1.5B (b) Llama3.2-3B (c) Llama3.2-1B-Instruct

Figure 1: Heatmaps of attention-head importance on probe reasoning dataset across 3 models: (a) Qwen2.5-Math-1.5B, (b) Llama3.2-3B, (c) Llama3.2-1B-Instruct. Brighter cells indicate higher contribution to reasoning; each model exhibits a sparse set of specialized reasoning heads, supporting our finding that only a subset of heads is for reasoning.

corroborates our observation, illustrating the sparse set of specialized attention heads that consistently activate during long reasoning, highlighting their pivotal role in this task. Building on this key insight, we propose CircuitSeer, a novel method that quantifies the complex reasoning attribute of data by harnessing these direct, model-internal signals. Unlike traditional data selection strategies that depend on external evaluations or coarse heuristics, CircuitSeer offers a principled and quantifiable approach rooted in the model's own processing. Our pipeline for identifying these specialized reasoning heads and data mining is illustrated in Figure 2. Across multiple models and benchmarks, CircuitSeer consistently selects compact subsets that yield stronger downstream reasoning performance, as summarized in Figure 3.

To validate our approach, we conducted comprehensive ablation studies exploring the critical role of these reasoning heads. These experiments confirmed that scores derived from the dominant reasoning heads substantially improve model performance over using all heads indiscriminately, underscoring their functional specialization. Additionally, the proposed soft-sampling strategy proved more effective than simple top-k selection in enhancing both data diversity and model generalization. Our contributions can be summarized as follows:

1. We identify a consistently emerging subset of reasoning-critical internal circuits (particularly attention heads) that underpin long-form mathematical reasoning, and we are the first to systematically harness these circuits to inform large-scale reasoning data selection.
2. We utilize these crucial internal circuits to calculate a model-internal score that reflects the reasoning attribute of data, enabling a data selection method that achieves state-of-the-art results across multiple models.
3. Through comprehensive experiments on 4 models and 9 test datasets, our method outperforms all baselines. Notably, on Qwen2.5-Math-7B, fine-tuning on just 10% of the data selected by CircuitSeer surpassed training on the full dataset by 1.2 points in average Pass@1.

2 Related works

Data Selection. Data selection plays a crucial role in supervised fine-tuning (SFT) of the large language model (LLM), with recent efforts mainly focusing on either task relevance, prioritizing data similar to the target distribution through clustering, in-context signals, or reasoning filters [11–15] or importance-based selection, which estimates data contribution via loss, gradients, or optimization-driven scoring [14–20]. However, fine-grained needs such as mathematical reasoning remain largely unexplored.

LLM Reasoning. Recent work on reasoning in LLMs has emphasized test-time scaling [7, 21–25], showing that increasing the number of generated tokens or verification steps can substantially improve reasoning accuracy. In parallel, open-source efforts have advanced the field by curating synthetic datasets with long reasoning trajectories [3–5] and distilling the reasoning skills of models such as DeepSeek-R1 [2] into more compact LLMs [26]. Despite these advances, when dealing with

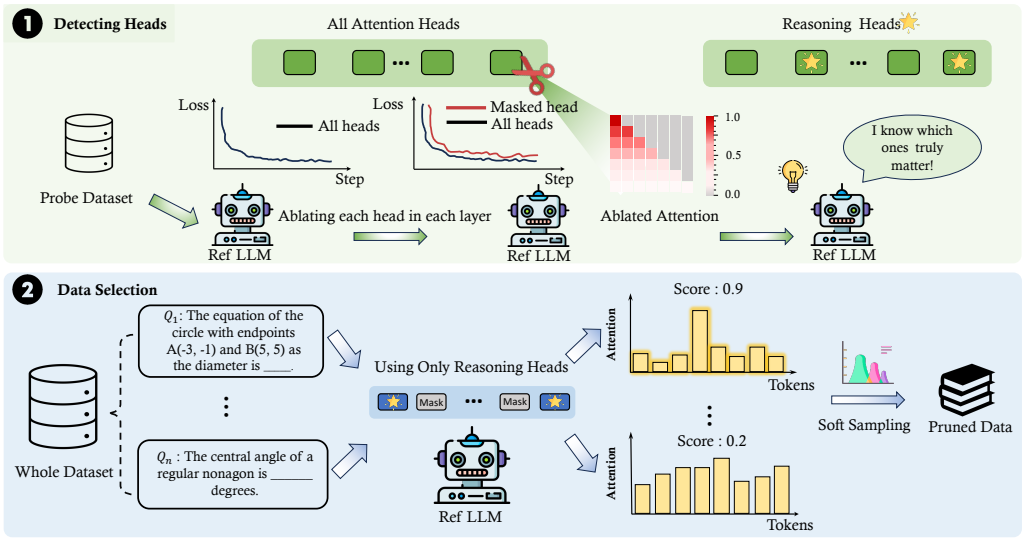


Figure 2: Overview of the CircuitSeer pipeline for reasoning-aware data selection. **Stage 1: Detecting Heads.** The reference LLM is evaluated on a probe set while ablating each attention head by flattening its causal attention distribution through small pre-softmax logit scaling; heads whose ablation yields a significant increase in loss are retained as reasoning heads. **Stage 2: Data Selection.** For each problem, we compute token-level attention distributions using only the detected heads and score samples with the formulation in Section 3.3 (Eq. (5)). Scores parameterize soft sampling to form a compact, diverse subset enriched with reasoning-heavy examples.

large-scale reasoning corpora, the community still largely lacks efficient and principled, inherently interpretable data selection methods.

Transformer circuits. Mechanistic interpretability aims to uncover human-understandable algorithms within neural networks [27–30]. Within this paradigm, Circuits Theory [9, 10, 27, 28, 30, 31] maps fine-grained connectivity patterns among attention heads and MLP neurons to algorithmic functions such as induction, composition, and copying.

Recent extensions, including attribution graphs [9, 10], trace causal information flows through these circuits, revealing components selectively activated by arithmetic or reasoning-related tasks [30]. CircuitSeer pinpoints reasoning-related circuits and further exploits them to guide targeted data selection.

3 Methodology

3.1 Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent a comprehensive dataset designed for mathematical reasoning, where N denotes the total number of samples. Each instance in this dataset is composed of an input problem x_i and a corresponding solution process y_i, which includes both the detailed reasoning steps and the final answer. The objective is to identify an optimal subset $\mathcal{D}^{'} \subset \mathcal{D}$ that is suited for SFT.

Given a pretrained language model $\mathcal{M}$, fine-tuning on a selected subset $\mathcal{D}'$ yields a task-specialized model $\mathcal{M}_{\mathcal{D}'}$. The goal is to select $\mathcal{D}'$ such that $\mathcal{M}_{\mathcal{D}'}$ achieves the highest accuracy on an unseen test set $\mathcal{D}_{\text{test}}$, while maintaining competitive performance on standard mathematical reasoning benchmarks compared with fine-tuning on the full dataset. Formally, the selection problem is:

$$\mathcal{D}' = \arg\max_{\mathcal{D}' \subseteq \mathcal{D}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}} \phi(\mathcal{M}_{\mathcal{D}'}(x), y)\,, \tag{1}$$

where y denotes the ground-truth answer and $\phi(\cdot,\cdot)$ represents the evaluation metric.

3.2 Detecting reasoning circuit

Transformer circuits [9, 10] show that neurons and attention heads in Transformer models play distinct roles. Yet in large language models (LLMs), how to systematically identify and understand the attention circuits responsible for reasoning remains largely unexplored. Our initial analysis involves an ablation experiment of attention heads in LLMs, evaluated on a mathematical reasoning probe dataset.

We denote $\mathcal{M}_{\text{ref}}$ as the reference model used for attention head circuits selection and data scoring, $\mathcal{D}_{\text{probe}}$ as a curated probe set for evaluating circuit importance, and let $\mathcal{H}$ be the set of all attention heads in $\mathcal{M}_{\text{ref}}$. For each circuit operationalized through a specific attention head $\tilde{h}$, we perform ablation to assess its contribution by employing undifferentiated attention [32], removing its selective attention pattern while preserving a coarse prefix-averaged value signal. Let $X \in \mathbb{R}^{n \times d_{\text{model}}}$ denote the hidden representations of an input sequence. For a candidate attention head $\tilde{h}$, we compute $Q = XW_q$, $K = XW_k$, and $V = XW_v$, where W_q, W_k, and W_v are the query, key, and value projection matrices of that head, and d_h denotes the head dimension. To ablate the head, we flatten its causal attention distribution by multiplying the pre-softmax logits by a small coefficient ϵ and applying the causal mask. This yields:

$$\tilde{h}(X) = \text{Softmax}\left(\epsilon \frac{QK^\top}{\sqrt{d_h}} + M_{\text{causal}}\right) V \stackrel{\text{def}}{=} \tilde{A}V, \tag{2}$$

where M_{causal} masks future tokens. When ϵ approaches zero, the resulting attention distribution approaches a uniform distribution over the visible prefix:

$$\tilde{A} = [\tilde{a}_{ij}], \quad \tilde{a}_{ij} = \begin{cases} \frac{1}{i}, & i \geq j, \\ 0, & i < j. \end{cases}$$

We do not scale the post-softmax attention probabilities directly. Instead, we scale the pre-softmax attention logits by a small coefficient ϵ, so that after the causal mask and softmax operation, the attention distribution collapses toward a uniform lower-triangular pattern.

The contribution of each candidate attention head is assessed by computing the expected increase in loss on the probe dataset induced by the intervention:

$$\Delta_{\tilde{h}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{probe}}}\left[\mathcal{L}(x, y; \mathcal{M}_{\text{ref}}^{-\tilde{h}}) - \mathcal{L}(x, y; \mathcal{M}_{\text{ref}})\right]. \tag{3}$$

where $\mathcal{L}(x, y; \mathcal{M}_{\text{ref}})$ denotes the loss of the reference model on (x, y), and $\mathcal{L}(x, y; \mathcal{M}_{\text{ref}}^{-\tilde{h}})$ denotes the loss after ablating the candidate head $\tilde{h}$. A larger loss increase indicates that the ablated head contributes more strongly to mathematical reasoning. These values are sorted in descending order, and the top-k heads are selected as the reasoning head set $\mathcal{H}_{\text{math}}$.

3.3 Using reasoning circuits to select data

Building upon the reasoning circuit heads identified in Section 3.2, we leverage them as data scorers to quantitatively evaluate the reasoning quality of each training sample $(x_i, y_i) \in \mathcal{D}$. For every sample, only the problem x_i is input to the reference model $\mathcal{M}_{\text{ref}}$, and scoring is performed exclusively using the selected reasoning circuits.

Let $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})$ denote the input sequence of n tokens. For each reasoning head $h \in \mathcal{H}_{\text{math}}$, the corresponding self-attention matrix $A^h \in \mathbb{R}^{n \times n}$ is extracted on model $\mathcal{M}_{ref}$, where $A^h_{j,k}$ represents the attention weight assigned by query position j to key position k. The mean attention on token k across all selected heads is:

$$\alpha_k = \frac{1}{|\mathcal{H}_{\text{math}}|} \sum_{h \in \mathcal{H}_{\text{math}}} \sum_{j=1}^{n} A^h_{j,k}, \tag{4}$$

where A^h is row-normalized such that $\sum_{k=1}^{n} A^h_{j,k} = 1$, and note that α_k denotes the received attention mass of token k.

The score for the input x_i, denoted as $S(x_i)$, is the variance of the mean attention across all tokens:

$$S(x_i) = \frac{1}{n}\sum_{k=1}^{n}\left(\alpha_k - \frac{1}{n}\sum_{m=1}^{n}\alpha_m\right)^2. \tag{5}$$

The score $S(x_i)$ measures how unevenly attention is received across tokens when only the detected reasoning heads are considered. For problems requiring multi-step reasoning, certain tokens often serve as key intermediate concepts, conditions, or operators, and therefore receive concentrated attention from many query positions. In contrast, simple or formulaic problems tend to induce more uniform received-attention patterns. A larger $S(x_i)$ therefore indicates that the sample contains richer multi-step reasoning structure, where conditions, intermediate concepts, and logical transitions are more explicitly organized, making it more likely to be useful for reasoning-oriented fine-tuning.

Once all samples have been assigned scores, these scores are normalized to induce a probability distribution over the dataset. Specifically, each sample x_i is assigned a selection probability proportional to its score $\frac{S(x_i)}{\sum_{(x_j,y_j)\in\mathcal{D}} S(x_j)}$. We perform soft sampling over $\mathcal{D}$ according to this probability distribution until the desired size $|\mathcal{D}'|$ is reached.

4 Experiments

4.1 Experimental setup

Training Details. We conducted experiments on the OpenR1-Math-220K dataset, synthesized by DeepSeek-R1 for long CoT training [2]. After filtering for samples with correct reasoning traces, we obtained about 196k high-quality examples.

As the reference model for our data scoring procedure (detailed in Sections 3.2 and 3.3), we employed Qwen2.5-Math-1.5B-Instruct [33], which complemented a weak-to-strong function.

We validated our method's effectiveness and generalizability by fine-tuning four diverse large language models: Qwen2.5-Math-7B-Base [33], Qwen2.5-Math-7B-Instruct [33], Llama-3.2-3B [34] and Llama-3.1-8B-Instruct [35]. For detailed training parameters, please refer to Appendix B.

Baselines. We compared our method with several representative data selection strategies, including: (i) Random: Randomly selected samples from the training data pool; (ii) Max Loss: Computed the training loss of each sample using Qwen2.5-Math-1.5B-Instruct and selected the samples with the highest loss; (iii) Quality: This baseline uses Qwen2.5-Math-1.5B-Instruct to evaluate data quality via prompt-based scoring and selects the top-rated samples; (iv) Diversity [7]: Data are uniformly sampled from different categories of mathematical problems to ensure diversity; (v) IFD [18]: Instruction-Following Difficulty is defined as the ratio between the conditioned answer loss (with instruction) and the answer loss (without instruction), and we selected the samples with the largest instruction-following difficulty.

Evaluation Settings. We used a rule-based evaluation framework [6], including metrics such as Pass@1 (P@1) [7] and Majority@16 (M@16).

We evaluated on nine benchmarks, including three competition-level benchmarks: AIME 2024, AIME 2025, and AMC 2023, and six datasets such as MATH500 [36], OlympiadBench [37], GAOKAO 2023 & 2024 [33], GAOKAO MATH [33], and KAOYAN [6]. The system prompt for evaluation was: *Please reason step by step, and put your final answer within \boxed{}*. For competition benchmarks, 16 solutions per problem were sampled. For other benchmarks, we used greedy decoding and sampled one solution per problem.

For every mathematical-results table, Avg. denotes the unweighted arithmetic mean of P@1 across the nine benchmarks; M@16 values do not enter this average.

4.2 Main results

CircuitSeer outperformed all baseline methods. As illustrated in Table 1, CircuitSeer consistently surpassed all baseline data selection techniques across a suite of nine benchmarks, showcasing reliable

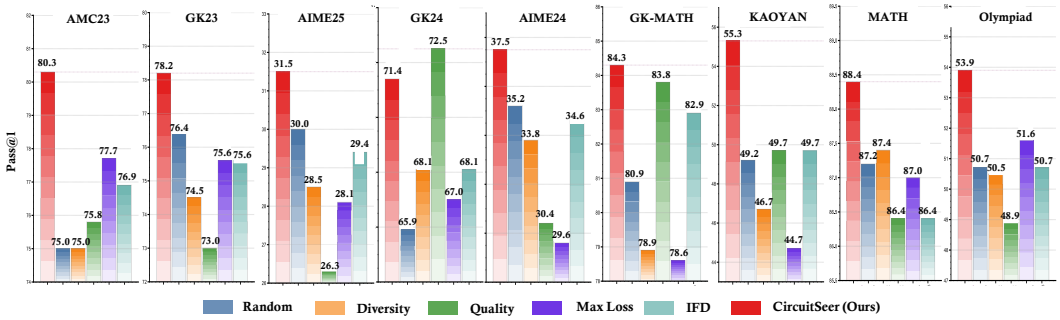


Figure 3: Pass@1 performance comparison of CircuitSeer against baseline methods across 9 benchmark datasets on Qwen2.5-Math-7B-Base. Our method consistently outperforms all baselines, demonstrating the effectiveness of leveraging internal reasoning circuits for data selection.

Table 1: Avg. is the unweighted mean of P@1 across nine benchmarks; M@16 is excluded. Main evaluation results across our full suite of mathematical reasoning benchmarks. We compare various data selection methods by fine-tuning four base models. For AIME and AMC, we report both Pass@1 (P@1) and Majority@16 (M@16) accuracy (%). For all other datasets, we report P@1. ↑ indicates an improvement over the Random Sampling baseline, while ↓ indicates a degradation. The best performance among selection methods for each setting is in **bold**, and the second best is underlined.

Method	AIME24		AIME25		AMC23		MATH	Olympiad	Kaoyan	GK 23	GK-Math	GK 24	Avg.
	P@1	M@16	P@1	M@16	P@1	M@16	P@1	P@1	P@1	P@1	P@1	P@1	P@1
Qwen2.5-Math-7B-Base													
Zero-shot	1.0	0.0	0.4	3.3	13.4	27.5	27.0	8.4	6.5	21.0	24.5	5.5	12.0
+ Random	35.2	53.3	30.0	43.3	75.0	**95.0**	87.2	50.7	49.2	76.4	80.9	65.9	61.2
+ Diversity	33.8↓1.4	53.3	28.5↓1.5	36.7↓6.6	75.0	90.0↓5.0	87.4↑0.2	50.5↓0.2	46.7↓2.5	74.5↓1.9	78.9↓2.0	68.1↑2.2	60.4↓0.8
+ Quality	30.4↓4.8	46.7↓6.6	26.3↓3.7	30.0↓13.3	75.8↑0.8	**95.0**	86.4↓0.8	48.9↓1.8	49.7↑0.5	73.0↓3.4	83.8↑2.9	**72.5**↑6.6	60.8↓0.4
+ Max Loss	29.6↓5.6	56.7↑3.4	28.1↓1.9	46.7↑3.4	77.7↑2.7	92.5↓2.5	87.0↓0.2	51.6↑0.9	44.7↓4.5	75.6↓0.8	78.6↓2.3	67.0↑1.1	60.0↓1.2
+ IFD	34.6↓0.6	53.3	29.4↓0.6	46.7↑3.4	76.9↑1.9	**95.0**	86.4↓0.8	50.7	49.7↑0.5	75.6↓0.8	82.9↑2.0	68.1↑2.2	61.6↑0.4
+ **CircuitSeer** (Ours)	**37.5**↑2.3	**63.3**↑10.0	**31.5**↑1.5	**50.0**↑6.7	**80.3**↑5.3	**95.0**	**88.4**↑1.2	**53.9**↑3.2	**55.3**↑6.1	**78.2**↑1.8	**84.3**↑3.4	71.4↑5.5	**64.5**↑3.3
Full-dataset	48.1	73.3	35.6	56.7	81.9	95.0	88.2	55.3	39.2	77.9	75.2	68.1	63.3
Qwen2.5-Math-7B-Instruct													
Zero-shot	11.7	20.0	10.8	20.0	56.7	72.5	84.0	40.7	44.7	64.4	75.5	61.5	50.0
+ Random	27.1	**50.0**	22.9	**36.7**	68.1	**92.5**	84.8	47.6	48.2	70.4	82.6	68.1	57.8
+ Diversity	24.2↓2.9	36.7↓13.3	21.5↓1.4	30.0↓6.7	66.9↓1.2	85.0↓7.5	84.4↓0.4	49.2↑1.6	49.2↑1.0	71.9↑1.5	80.9↓1.7	71.9↑3.8	57.8
+ Quality	23.5↓3.6	36.7↓13.3	19.2↓3.7	20.0↓16.7	68.9↑0.8	87.5↓5.0	83.4↓1.4	50.7↑3.1	48.7↑0.5	71.7↑1.3	82.3↓0.3	72.5↑4.4	57.9↑0.1
+ Max Loss	**27.3**↑0.2	46.7↓3.3	21.5↓1.4	33.3↓3.4	**70.0**↑1.9	**92.5**	85.0↑0.2	**51.0**↑3.4	41.7↓6.5	68.6↓1.8	79.5↓3.1	68.1	57.0↓0.8
+ IFD	25.8↓1.3	43.3↓6.7	22.3↓0.6	30.0↓6.7	68.3↑0.2	85.0↓7.5	**87.0**↑2.2	47.6	46.7↓1.5	71.9↑1.5	80.3↓2.3	67.0↓1.1	57.4↓0.4
+ **CircuitSeer** (Ours)	**27.3**↑0.2	46.7↓3.3	**23.1**↑0.2	**36.7**	68.1	**92.5**	**87.0**↑2.2	50.1↑2.5	**56.8**↑8.6	**74.0**↑3.6	**84.3**↑1.7	**73.6**↑5.5	**60.5**↑2.7
Full-dataset	44.2	63.3	35.0	50.0	79.2	95.0	89.0	51.4	36.2	80.5	77.5	62.6	61.7
Llama-3.1-8B-Instruct													
Zero-shot	4.4	13.3	0.2	0.0	20.2	35.0	45.0	3.6	5.5	34.0	8.8	11.0	14.7
+ Random	3.5	10.0	6.2	10.0	29.8	47.5	54.2	21.2	15.6	49.9	55.6	44.0	31.1
+ Diversity	3.7↑0.2	6.7↓3.3	**9.4**↑3.2	**20.0**↑10.0	29.8	40.0↓7.5	**58.4**↑4.2	23.6↑2.4	16.1↑0.5	51.9↑2.0	43.9↓11.7	45.1↑1.1	31.3↑0.2
+ Quality	1.7↓1.8	3.3↓6.7	4.0↓2.2	0.0↓10.0	21.1↓8.7	35.0↓12.5	46.4↓7.8	15.9↓5.3	3.5↓12.1	43.4↓6.5	27.1↓28.5	19.8↓24.2	20.3↓10.8
+ Max Loss	3.1↓0.4	6.7↓3.3	1.7↓4.5	3.3↓6.7	23.9↓5.9	45.0↓2.5	48.8↓5.4	18.5↓2.7	13.1↓2.5	42.3↓7.6	45.6↓10.0	22.0↓22.0	24.3↓6.8
+ IFD	2.5↓1.0	**13.3**↑3.3	3.5↓2.7	13.3↑3.3	29.4↓0.4	47.5	50.4↓3.8	21.6↑0.4	**17.1**↑1.5	48.8↓1.1	58.7↑3.1	**50.5**↑6.5	31.4↑0.3
+ **CircuitSeer** (Ours)	**4.2**↑0.7	**13.3**↑3.3	6.9↑0.7	13.3↑3.3	**31.4**↑1.6	**52.5**↑5.0	54.6↑0.4	**25.3**↑4.1	**17.1**↑1.5	**54.5**↑4.6	**62.4**↑6.8	44.0	**33.4**↑2.3
Full-dataset	0.0	0.0	0.2	0.0	2.0	5.0	1.0	0.9	2.5	2.9	15.1	4.4	3.2
Llama-3.2-3B													
Zero-shot	0.0	0.0	0.0	0.0	0.3	0.0	0.2	0.0	0.0	0.8	0.0	0.0	0.2
+ Random	0.2	0.0	**0.4**	0.0	9.4	22.5	25.6	6.5	3.0	**28.3**	16.8	13.2	11.5
+ Diversity	0.4↑0.2	0.0	**0.4**	**3.3**↑3.3	10.2↑0.8	20.0↓2.5	24.6↓1.0	7.0↑0.5	2.5↓0.5	25.7↓2.6	13.4↓3.4	4.4↓8.8	9.9↓1.6
+ Quality	0.0↓0.2	0.0	0.0↓0.4	0.0	6.6↓2.8	7.5↓15.0	22.4↓3.2	4.6↓1.9	**4.5**↑1.5	24.4↓3.9	**21.9**↑5.1	13.2	10.8↓0.7
+ Max Loss	0.2	**3.3**↑3.3	0.0↓0.4	0.0	6.1↓3.3	10.0↓12.5	23.4↓2.2	5.5↓1.0	4.0↑1.0	25.7↓2.6	18.5↑1.7	11.0↓2.2	10.5↓1.0
+ IFD	**1.0**↑0.8	0.0	0.2↓0.2	0.0	9.4	22.5	25.6	6.8↑0.3	0.5↓2.5	24.4↓3.9	13.1↓3.7	6.6↓6.6	9.7↓1.8
+ **CircuitSeer** (Ours)	**1.0**↑0.8	0.0	**0.4**	**3.3**↑3.3	**12.3**↑2.9	**27.5**↑5.0	**27.8**↑2.2	**8.6**↑2.1	**4.5**↑1.5	27.3↓1.0	17.1↑0.3	**15.4**↑2.2	**12.7**↑1.2
Full-dataset	5.0	13.3	9.0	30.0	36.4	52.5	54.2	24.3	2.5	36.4	38.7	24.2	25.6

and consistent enhancements. On the Qwen2.5-Math-7B-Base model, CircuitSeer achieved Pass@1 (P@1) scores of 37.5 on AIME24 and 31.5 on AIME25, notably outperforming Random Sampling and yielding substantial improvements of 10.0 and 6.7 points on the Majority@16 (M@16) metric. On AMC23, CircuitSeer excelled with a P@1 of 80.3, exceeding the best alternative by 2.6 points, while on Kaoyan, it reached 55.3, marking a 6.1-point advancement over Random. In the Qwen2.5-Math-7B-Instruct model, CircuitSeer achieved notable progress, with an impressive gain of 8.6 points on Kaoyan and a 5.5-point increase on GK24, underscoring its adaptability across model variations. Even when deployed on the smaller Llama-3.2-3B, CircuitSeer still delivered remarkable improvements, raising AMC23's M@16 metric to 27.5, a 5.0-point enhancement, and boosting MATH's P@1 to 27.8, reflecting a 2.2-point increase. These results collectively demonstrated CircuitSeer's improvements in mathematical reasoning across benchmarks.

Table 2: Ablation study on the use of reasoning heads for data selection. We compare selection based on the identified reasoning heads (Reasoning) and all attention heads (All) against random data selection (Random). The table reports Pass@1 (P@1) scores, with changes in percentage points relative to the random data selection baseline shown alongside the scores. All experiments were performed on Qwen2.5-Math-7B.

Strategy	AIME24	AIME25	AMC23	MATH	Olympiad	Kaoyan	GK23	GK-Math	GK24	Average
Random	35.2	30.0	75.0	87.2	50.7	49.2	76.4	80.9	65.9	61.2
All	33.8$_{\downarrow 1.4}$	26.0$_{\downarrow 4.0}$	75.6$_{\uparrow 0.6}$	87.6$_{\uparrow 0.4}$	51.1$_{\uparrow 0.4}$	42.2$_{\downarrow 7.0}$	**78.2**$_{\uparrow 1.8}$	80.9$_{\downarrow 0.0}$	64.9$_{\downarrow 1.0}$	60.0$_{\downarrow 1.2}$
Reasoning	**37.5**$_{\uparrow 2.3}$	**31.5**$_{\uparrow 1.5}$	**80.3**$_{\uparrow 5.3}$	**88.4**$_{\uparrow 1.2}$	**53.9**$_{\uparrow 3.2}$	**55.3**$_{\uparrow 6.1}$	**78.2**$_{\uparrow 1.8}$	**84.3**$_{\uparrow 3.4}$	**71.4**$_{\uparrow 5.5}$	**64.5**$_{\uparrow 3.3}$

Generalization across diverse models. CircuitSeer consistently demonstrated generalization across models of varying capacities. On the Qwen2.5-Math-7B-Instruct model, CircuitSeer excelled by achieving a P@1 of 27.3 on AIME24, surpassing other methods like Diversity and Quality, which achieved 24.2 and 23.5, respectively. On KAOYAN, CircuitSeer reached 56.8, outperforming the next best method, Diversity, which attained 49.2. Similarly, on GK24, CircuitSeer achieved 73.6, marking a significant improvement over the best baseline method, Quality, which scored 72.5. For the Qwen2.5-Math-7B-Base model, CircuitSeer yielded remarkable improvements, with P@1 on AIME25 rising to 31.5, compared to the next best method, Random, which scored 30.0. On MATH, CircuitSeer achieved 88.4, surpassing Diversity's 87.4, and on Kaoyan, it attained 55.3, which was significantly higher than Random's 49.2. Even with the smaller Llama-3.2-3B model, CircuitSeer delivered substantial improvements, such as AMC23 M@16 increasing to 27.5, outperforming Diversity, which achieved 20.0, and MATH P@1 reaching 27.8, compared to Random's 25.6.

4.3 Ablation studies

To validate the efficacy of our proposed components and understand their individual contributions, we conducted a series of ablation studies on the Qwen2.5-Math-7B. Our analysis was designed to answer four critical questions: (1) Was the identification of specialized reasoning heads essential? (2) Did our proposed score reflect the value of the reasoning data? (3) What was the impact of our soft sampling strategy compared to Top-K selection? (4) Were reasoning signals better captured from the input or the output of the sample?

Impact of head selection. A fundamental aspect of our approach was the identification of a specialized subset of attention heads crucial for mathematical reasoning, which offered a clearer and more pertinent signal for data selection compared to using all heads or a random subset. To validate this, we assessed two strategies for score computation:

1. **Reasoning Heads**: Utilized the reasoning heads identified through our intervention-based method detailed in Section 3.2.
2. **All Heads**: Calculated scores based on the attention weights from all heads in the model.

As presented in Table 2, the data subset selected with our identified reasoning heads yielded substantially superior performance. This result confirmed that these heads were indeed specialized for processing structural and logical information within reasoning traces. In contrast, using all heads introduced significant noise from heads dedicated to other, irrelevant functions (e.g., positional or syntactic processing), which diluted the quality signal and degraded selection performance. These results suggest that focusing on the identified reasoning heads provides a more informative signal for reasoning-oriented data selection than aggregating information across all heads. The improvement over random data selection further supports the effectiveness of our head-guided selection approach.

Effectiveness of CircuitSeer score. As shown in Table 1, soft sampling guided by high CircuitSeer scores consistently achieves state-of-the-art performance across multiple models. To validate these scores, we conducted a contrastive experiment comparing models trained on high-score, low-score, and randomly sampled data subsets (Table 3). Results reveal a clear trend: high-score samples involve problems requiring deeper multi-step reasoning, while low-score ones mostly capture simpler questions (see Appendix E). The superior performance of high-score selection over both low-score

Table 3: Ablation study on input score-based data selection. We compare selection based on high CircuitSeer scores (High) and low CircuitSeer scores (Low) against random data sampling (Random), which selects training examples without using these scores. All results are Pass@1 (P@1) accuracies. High-score selection outperforms both baselines, supporting the effectiveness of the CircuitSeer scoring mechanism.

Strategy	AIME24	AIME25	AMC23	MATH	Olympiad	Kaoyan	GK23	GK-Math	GK24	Average
Random	35.2	30.0	75.0	87.2	50.7	49.2	76.4	80.9	65.9	61.2
Low	29.8$_{\downarrow 5.4}$	25.9$_{\downarrow 4.1}$	70.2$_{\downarrow 4.8}$	80.1$_{\downarrow 7.1}$	45.2$_{\downarrow 5.5}$	46.2$_{\downarrow 3.0}$	73.4$_{\downarrow 3.0}$	76.8$_{\downarrow 4.1}$	65.1$_{\downarrow 0.8}$	57.0$_{\downarrow 4.2}$
High	**37.5**$_{\uparrow 2.3}$	**31.5**$_{\uparrow 1.5}$	**80.3**$_{\uparrow 5.3}$	**88.4**$_{\uparrow 1.2}$	**53.9**$_{\uparrow 3.2}$	**55.3**$_{\uparrow 6.1}$	**78.2**$_{\uparrow 1.8}$	**84.3**$_{\uparrow 3.4}$	**71.4**$_{\uparrow 5.5}$	**64.5**$_{\uparrow 3.3}$

Table 4: Ablation study comparing Soft Sampling and Top-k selection against a Random baseline. Performance changes for both methods are shown relative to Random. All scores are Pass@1 (P@1) accuracy. Soft Sampling demonstrates the most significant average improvement.

Strategy	AIME24	AIME25	AMC23	MATH	Olympiad	Kaoyan	GK23	GK-Math	GK24	Average
Random	35.2	30.0	75.0	87.2	50.7	49.2	76.4	80.9	65.9	61.2
Hard (Top-k)	33.1$_{\downarrow 2.1}$	30.8$_{\uparrow 0.8}$	77.5$_{\uparrow 2.5}$	87.6$_{\uparrow 0.4}$	49.5$_{\downarrow 1.2}$	50.8$_{\uparrow 1.6}$	75.3$_{\downarrow 1.1}$	**86.9**$_{\uparrow 6.0}$	**73.6**$_{\uparrow 7.7}$	62.8$_{\uparrow 1.6}$
Soft	**37.5**$_{\uparrow 2.3}$	**31.5**$_{\uparrow 1.5}$	**80.3**$_{\uparrow 5.3}$	**88.4**$_{\uparrow 1.2}$	**53.9**$_{\uparrow 3.2}$	**55.3**$_{\uparrow 6.1}$	**78.2**$_{\uparrow 1.8}$	84.3$_{\uparrow 3.4}$	71.4$_{\uparrow 5.5}$	**64.5**$_{\uparrow 3.3}$

selection and random data selection supports the effectiveness of the CircuitSeer scoring mechanism in identifying mathematical reasoning data that is more beneficial for training.

Soft sampling vs. top-k selection. We empirically examined our sampling strategy. Although Top-k selection is straightforward, it tends to over-concentrate on redundant high-score samples, reducing diversity. In contrast, our score-guided Soft Sampling maintains quality while introducing stochasticity for broader coverage. We compared it against a deterministic baseline that directly selected the highest-scoring samples.

As shown in Table 4, soft sampling achieved superior performance. We attributed this to its ability to foster a more diverse training subset by occasionally including lower-scoring-but-still-valuable samples. This stochasticity prevented the selection process from collapsing into a narrow region of the data distribution, which enhanced the model's generalization capabilities on unseen problems.

Input vs. output scoring. We examined whether reasoning signals were better captured from the problem statement or from the generated reasoning trajectory by comparing two variants: (1) output-based scoring, which assigned scores based on the solutions of the training samples, and (2) input-based scoring, which focused solely on the problem inputs of the training samples. As shown in Table 5, the input-based variant outperformed the output-based one across 5 out of 9 benchmarks, achieving an average Pass@1 (P@1) of 64.5% versus 62.8%. Input-based scoring captured reasoning complexity more effectively because attention patterns during the input phase reflected how the model parsed the problem, identified constraints, and constructed internal logical structures—revealing *what reasoning is required*. In contrast, activations during the output phase primarily reflected how reasoning is expressed, being shaped by stylistic factors such as solution format, generation length, and linguistic style rather than true reasoning depth.

4.4 Analysis

Soft sampling fosters superior data diversity. To understand why Soft Sampling consistently outperforms the deterministic Top-k baseline (Table 4), we analyze the source composition of the selected subsets. As shown in Figure 4, Top-k selection is strongly dominated by the `Cn_12` source, whereas Soft Sampling preserves a more balanced mixture across multiple sources such as `Olympiads`, `Cn_Contest`, and `Aops_Forum`. This difference reveals an important limitation of using the score as a hard ranking criterion: high-scoring examples are not necessarily independent or complementary, and the very top-ranked region can contain many near-redundant problems from the same distribution.

Table 5: Ablation study comparing outputs and inputs against a Random baseline. Performance changes for both are shown relative to Random. All scores are Pass@1 (P@1) accuracy.

Strategy	AIME24	AIME25	AMC23	MATH	Olympiad	Kaoyan	GK23	GK-Math	GK24	Average
Random	35.2	30.0	75.0	87.2	50.7	49.2	76.4	80.9	65.9	61.2
Output	35.6↑0.4	29.8↓0.2	73.9↓1.1	**89.0**↑1.8	**54.7**↑4.0	48.2↓1.0	**78.2**↑1.8	**85.8**↑4.9	70.3↑4.4	62.8↑1.6
Input	**37.5**↑2.3	**31.5**↑1.5	**80.3**↑5.3	88.4↑1.2	53.9↑3.2	**55.3**↑6.1	**78.2**↑1.8	84.3↑3.4	**71.4**↑5.5	**64.5**↑3.3

Table 6: Avg. is the unweighted mean of P@1 across nine benchmarks; M@16 is excluded. Evaluation of CircuitSeer with different reference models across mathematical benchmarks. The best performance among different reference models for each target model is in **bold**, and the second best is underlined.

Target Model	AIME24		AIME25		AMC23		MATH	Olympiad	Kaoyan	GK 23	GK-Math	GK 24	Avg.
	P@1	M@16	P@1	M@16	P@1	M@16	P@1	P@1	P@1	P@1	P@1	P@1	P@1
Qwen2.5-Math-7B													
Qwen3-0.6B	35.2	50.0	29.4	33.3	75.6	92.5	**89.6**	51.6	50.3	77.1	80.3	68.1	61.9
Gemma3-1B-Instruct	37.1	56.7	**32.5**	**50.0**	75.3	90.0	88.6	51.0	47.7	**78.7**	79.5	68.1	62.1
Qwen2.5-Math-1.5B-Instruct (Ours)	**37.5**	**63.3**	31.5	**50.0**	**80.3**	**95.0**	88.4	**53.9**	**55.3**	78.2	**84.3**	**71.4**	**64.5**
Llama-3.2-3B													
Qwen3-0.6B	0.2	**0.0**	0.0	0.0	9.4	12.5	24.6	7.1	3.5	27.0	16.8	12.1	11.2
Gemma3-1B-Instruct	0.2	**0.0**	0.0	0.0	6.9	12.5	25.0	6.8	3.5	25.7	**17.9**	11.0	10.8
Qwen2.5-Math-1.5B-Instruct (Ours)	**1.0**	**0.0**	**0.4**	**3.3**	**12.3**	**27.5**	**27.8**	**8.6**	**4.5**	**27.3**	17.1	**15.4**	**12.7**

In contrast, Soft Sampling treats the CircuitSeer score as a probability rather than a strict ordering. This allows the selected subset to remain enriched with reasoning-intensive examples while still covering less frequent problem families. The results in Table 4 support this interpretation: Top-k selection improves benchmarks that are closer to its overrepresented source distribution, but it is less robust on average; Soft Sampling achieves stronger overall generalization by balancing sample quality and distributional coverage.

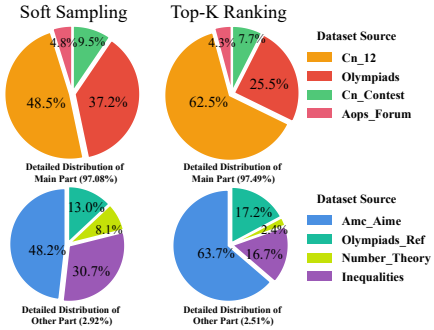


Figure 4: Detailed source distribution of subsets selected by Soft Sampling and Top-k selection. Infrequent categories are grouped as "Other part" for clarity.

Analysis of using different reference models. Table 6 evaluates reference models of different sizes and specializations. On Qwen2.5-Math-7B, Qwen3-0.6B and Gemma3-1B-Instruct achieve average P@1 scores of 61.9 and 62.1. Both exceed Random sampling (61.2), but remain below full-data SFT (63.3) and the default reasoning-oriented 1.5B reference (64.5). Thus, a smaller reference can provide a useful selection signal, while its capability limits the downstream gain.

On Llama-3.2-3B, the default reference reaches 12.7, compared with 11.2 and 10.8 for Qwen3-0.6B and Gemma3-1B-Instruct. These results show that the strength of the selection signal varies across reference models.

5 Conclusion

In this paper, we present CircuitSeer, a framework that identifies high-quality data by probing the internal reasoning circuits of large language models (LLMs). By quantifying activations of reasoning attention heads, CircuitSeer provides an interpretable, model-internal signal of reasoning complexity, effectively reducing the need for external heuristics. Experiments across diverse models and benchmarks show that CircuitSeer-selected subsets consistently outperform evaluated data-selection baselines, and in some settings even surpass full-data training, demonstrating both efficiency and robustness. Beyond empirical gains, this work highlights that reasoning circuits encode measurable structures of problem difficulty in supervised fine-tuning, suggesting that the internal dynamics of LLMs can serve as principled indicators for data quality. This mechanism-aware perspective enables more transparent, scalable, and domain-agnostic reasoning data curation, thereby bridging mechanistic interpretability with practical LLM training.

Acknowledgments and Disclosure of Funding

This work was supported by Alibaba Group through Alibaba Innovative Research Program and not associated with any competing interests.

References

[1] OpenAI. Learning to reason with LLMs, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[2] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948, 2025.

[3] Elie Bakouch, Leandro von Werra, and Lewis Tunstall. Open-R1: a fully open reproduction of DeepSeek-R1. Hugging Face Blog, 2025. URL https://huggingface.co/blog/open-r1.

[4] Justus Mattern, Sami Jaghouar, Manveer Basra, Jannik Straube, Matthew Di Ferrante, Felix Gabriel, Jack Min Ong, Vincent Weisser, and Johannes Hagemann. SYNTHETIC-1: Two Million Collaboratively Generated Reasoning Traces from Deepseek-R1, 2025. URL https://www.primeintellect.ai/blog/synthetic-1-release.

[5] Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 1.4 Million Open-Source Distilled Reasoning Dataset to Empower Large Language Model Training. arXiv preprint arXiv:2503.19633, 2025.

[6] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is More for Reasoning. In Second Conference on Language Modeling (COLM 2025), 2025.

[7] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025.

[8] Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[9] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit Tracing: Revealing Computational Graphs in Language Models. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

[10] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the Biology of a Large Language Model. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

[11] Shaobo Wang, Xuan Ouyang, Tianyi Xu, Yuzheng Hu, Jialin Liu, Guo Chen, Tianyu Zhang, Junhao Zheng, Kexin Yang, Xingzhang Ren, Dayiheng Liu, and Linfeng Zhang. OPUS: Towards Efficient and Principled Data Selection in Large Language Model Pre-training in Every Iteration. In Proceedings of the 43rd International Conference on Machine Learning, volume 306 of Proceedings of Machine Learning Research. PMLR, 2026.

[12] Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. Data Whisperer: Efficient Data Selection for Task-Specific LLM Fine-Tuning via Few-Shot In-Context Learning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025.

[13] Shaobo Wang, Jiaming Wang, Jiajun Zhang, Cong Wang, Yue Min, Zichen Wen, Xingzhang Ren, Fei Huang, Huiqiang Jiang, Junyang Lin, Dayiheng Liu, and Linfeng Zhang. Winning the Pruning Gamble: A Unified Approach to Joint Sample and Token Pruning for Efficient Supervised Fine-Tuning. arXiv preprint arXiv:2509.23873, 2025.

[14] Ming Shen. Rethinking Data Selection for Supervised Fine-Tuning. arXiv preprint arXiv:2402.06094, 2024.

[15] Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking Data Selection at Scale: Random Selection is Almost All You Need. In Findings of the Association for Computational Linguistics: EMNLP 2025, 2025.

[16] Kai Hua, Steven Wu, Ge Zhang, and Ke Shen. AttentionInfluence: Adopting Attention Head Influence for Weak-to-Strong Pretraining Data Selection. arXiv preprint arXiv:2505.07293, 2025.

[17] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data Selection for Language Models via Importance Resampling. In Advances in Neural Information Processing Systems, volume 36, 2023.

[18] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024.

[19] Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, and Furao Shen. A CLIP-Powered Framework for Robust and Generalizable Data Selection. In International Conference on Learning Representations, 2025.

[20] Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric Coreset Selection for High Pruning Rates. In The Eleventh International Conference on Learning Representations, 2023.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems, volume 35, 2022.

[22] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In International Conference on Learning Representations, 2024.

[23] Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. SELF-DISCOVER: Large Language Models Self-Compose Reasoning Structures. In Advances in Neural Information Processing Systems, volume 37, 2024.

[24] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In International Conference on Learning Representations, 2025.

[25] Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. Survey Certification.

[26] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling Reasoning Capabilities into Smaller Language Models. In Findings of the Association for Computational Linguistics: ACL 2023, 2023.

[27] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. Distill, 2020. doi: 10.23915/distill.00024.001.

[28] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[29] Jesse Vig and Yonatan Belinkov. Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2019.

[30] Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[31] David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled Transformers as a Laboratory for Interpretability. In Advances in Neural Information Processing Systems, volume 36, 2023.

[32] Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the Role of Attention Heads in Large Language Model Safety. In International Conference on Learning Representations, 2025.

[33] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv preprint arXiv:2409.12122, 2024.

[34] Meta AI. Llama 3.2. Hugging Face model collection, 2024. URL https://huggingface.co/collections/meta-llama/llama-32.

[35] Aaron Grattafiori et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024.

[36] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021.

[37] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[38] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing Transformer Training by Preventing Attention Entropy Collapse. In Proceedings of the 40th International Conference on Machine Learning, 2023.

[39] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive Attention Span in Transformers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[40] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In International Conference on Learning Representations, 2025.

[41] Mark Chen et al. Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374, 2021.

[42] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. In Advances in Neural Information Processing Systems, volume 36, 2023.

[43] Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. ClusterUCB: Efficient Gradient-Based Data Selection for Targeted Fine-Tuning of LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1026.

A Limitations

Although CircuitSeer demonstrates strong performance in filtering mathematical reasoning data through supervised fine-tuning (SFT), its current design requires direct access to the model's internal parameters and attention weights, which restricts applicability to closed-source, API-only large language models. In addition, CircuitSeer is not fully model-agnostic, as biases and architectural idiosyncrasies of the underlying models may influence the selection process and final outcomes. Future work will explore techniques for weight-free circuit attribution and bias-resilient scoring, thereby improving scalability, generalization, and real-world deployment.

B Detailed experimental settings

Training Details. All models are trained using the llamafactory framework with a learning rate of 5×10^{-5} for 3 epochs, utilizing fused AdamW as the optimizer. The total batch size is set to 32, with a warmup ratio of 0.1, and a cosine decay schedule towards zero. Both Flash Attention and DeepSpeed ZeRO-3 are adopted for training optimization.

For the Qwen-2.5-Math-7B model, we extend the model's context length from 4,096 to 16,384 via RoPE scaling, increasing the RoPE frequency from 10k to 300k. The max training sequence length is set to 16k. For Llama-3.2-3B, the max training sequence length is set to 8k. All experiments are conducted on 8 NVIDIA A100 GPUs. For Qwen2.5-Math-7B-Base, Qwen2.5-Math-7B-Instruct, and Llama-3.1-8B-Instruct, training on the full dataset takes approximately 54 hours, while training on the selected 10% of the data takes about 6 hours. For Llama-3.2-3B, full-dataset training requires roughly 27 hours, compared to approximately 3 hours for the selected 10% subset.

For the selection of reasoning heads, the probe dataset for the Detecting Heads stage is constructed using the top 300 samples from the OpenR1-Math-196k (After Filtered) dataset, sorted in descending order by their loss on the reference model. The number of selected top-k heads, k is set to 5% of the total number of attention heads for each model.

For Llama-3.1-8B-Instruct, full-dataset SFT was less stable than subset training, likely because this general-purpose instruction-tuned model is less aligned with long synthetic mathematical reasoning traces than the Qwen2.5-Math models, while its larger scale and longer full-corpus optimization make it more sensitive to distributional mismatch and noisy or redundant reasoning trajectories.

Computational cost. We outline the computational cost of CircuitSeer in Table 7, benchmarked with Qwen2.5-Math-7B. The pipeline begins with constructing the probe set by evaluating the OpenR1-Math-220K dataset with the reference model and selecting the 300 samples with the highest loss. Head detection takes only 12 minutes, and the subsequent data scoring requires 1 hour.

Table 7: Computational overhead of CircuitSeer across pipeline stages.

Stage	GPUs	Time	TFLOPs/h
Head detection	8×A100	12 min	3.68×10^{6}
Data scoring	8×A100	1 h	1.18×10^{5}
SFT on 10% data	8×A100	6 h	7.20×10^{6}
SFT on 100% data	8×A100	54 h	7.20×10^{6}

Overall, data selection introduces approximately 2.2% overhead compared with the 54-hour full-data SFT, while training on the selected 10% subset takes only 6 hours.

B.1 Licenses and terms of use

Tables 8–10 identify the releases discussed in this paper and the licenses declared by their official publishers. The linked model cards, dataset cards, and license files provide the governing terms. Public availability on Hugging Face or GitHub does not by itself imply unrestricted use or redistribution.

Table 8: Declared licenses of the training datasets and model releases. Different model families and datasets retain their respective license conditions.

Resource / release	Declared license
OpenR1-Math-220k	Apache-2.0
NuminaMath-CoT	Apache-2.0
KodCode-V1-SFT-R1	CC BY-NC 4.0
Qwen2.5-Math-1.5B-Instruct, Qwen2.5-Math-7B, and Qwen2.5-Math-7B-Instruct	Apache-2.0
Qwen3-0.6B	Apache-2.0
Qwen2.5-Coder-7B	Apache-2.0
Llama-3.1-8B-Instruct	Llama 3.1 Community License Agreement
Llama-3.2-3B and Llama-3.2-1B-Instruct	Llama 3.2 Community License Agreement
Gemma-3-1B-IT	Gemma Terms of Use

Conditions of reuse. Apache-2.0, MIT, and BSD-3-Clause require preservation of the applicable license and copyright notices when redistributing covered materials; Apache-2.0 also requires notices of modifications where applicable. KodCode-V1-SFT-R1 is released under CC BY-NC 4.0, which requires attribution and limits licensed reuse to noncommercial purposes. Llama releases are governed by their respective Community License Agreements and incorporated Acceptable Use Policies, including their attribution and redistribution conditions. Gemma is governed by its Terms of Use, including the use restrictions incorporated into those terms. These conditions apply to the corresponding resources rather than to every asset in the paper under a single common license.

Table 9: License declarations for evaluation resources. Software licenses are distinguished from the rights in underlying competition and examination questions; incomplete dataset-card declarations are reported as such.

Evaluation resource	Declared license / scope
MATH-500 split and MATH release	MIT in the upstream repositories; original problem rights remain distinct.
OlympiadBench	Apache-2.0 in the dataset card; MIT in the code repository.
AIME 2024, AIME 2025, and AMC 2023	Competition questions in the cited evaluation releases; no separate uniform license for the original question content is asserted here.
GAOKAO 2023, GAOKAO-MATH, GAOKAO 2024, and KAOYAN	Examination-derived benchmarks in the cited evaluation releases; original source rights and applicable terms are retained.
LiveCodeBench-v5	Dataset card declares "cc" without a specific CC variant; the evaluation code is MIT.
HumanEval	MIT
HumanEval+ / EvalPlus	Apache-2.0

Underlying question content. The mathematical evaluation releases are credited in Section 4.1 and through the LIMO and Qwen2.5-Math evaluation releases. A software repository's MIT license does not establish a blanket license for third-party competition or examination questions included in a benchmark. The corresponding source terms and original rights must therefore be considered separately for reuse or redistribution. Likewise, LiveCodeBench's generic "cc" label is not presented here as a verified grant under a particular Creative Commons variant.

Table 11: Avg. is the unweighted mean of P@1 across nine benchmarks; M@16 is excluded. Comparison between CircuitSeer and different attention score criteria across mathematical benchmarks. The best performance among different methods for each target model is in **bold**, and the second best is underlined.

Target Model	AIME24		AIME25		AMC23		MATH	Olympiad	Kaoyan	GK 23	GK-Math	GK 24	Avg.
	P@1	M@16	P@1	M@16	P@1	M@16	P@1	P@1	P@1	P@1	P@1	P@1	P@1
Llama-3.2-3B													
Attention Entropy	0.0	0.0	0.0	0.0	3.4	7.5	20.0	4.7	2.5	25.7	9.4	7.7	8.2
Attention Span	0.0	0.0	0.0	0.0	5.0	10.0	26.0	6.5	2.5	**27.8**	**17.7**	5.5	10.1
CircuitSeer (Ours)	**1.0**	0.0	**0.4**	**3.3**	**12.3**	**27.5**	**27.8**	**8.6**	**4.5**	27.3	17.1	**15.4**	**12.7**
Qwen2.5-Math-7B													
Attention Entropy	22.5	30.0	26.3	40.0	66.9	87.5	83.6	47.3	43.2	73.5	77.2	59.3	55.5
Attention Span	33.3	53.3	27.9	43.3	72.8	90.0	87.8	50.8	**56.8**	77.4	**84.3**	**74.7**	62.9
CircuitSeer (Ours)	**37.5**	**63.3**	**31.5**	**50.0**	**80.3**	**95.0**	**88.4**	**53.9**	55.3	**78.2**	**84.3**	71.4	**64.5**

Table 10: Licenses of the software components explicitly named in the experimental setup. Their linked license files specify the applicable notice and redistribution conditions.

Software resource	Declared license
LLaMA-Factory	Apache-2.0
DeepSpeed	Apache-2.0
FlashAttention	BSD-3-Clause

C Supplementary experiments

This section provides more detailed analysis of CircuitSeer.

C.1 Comparison of different scoring criterions

In addition to the variance-based score introduced in the main method, we also explore two alternative scoring criteria based on internal attention signals:

Attention Entropy.

Attention entropy [38] measures the dispersion or uncertainty of the attention distribution for each query position:

$$s_{\text{entropy}}(x) = \frac{1}{|\mathcal{H}|\,T} \sum_{(l,h)\in\mathcal{H}} \sum_{i=1}^{T} H\big(A_{i,:}^{(l,h)}(x)\big),$$

$$H(p) = -\sum_{j} p_j \log p_j,$$

where T denotes the number of tokens in the input sequence. $\mathcal{H}$ is the set of selected attention heads, indexed by layer l and head h. $A^{(l,h)}(x) \in \mathbb{R}^{T\times T}$ is the corresponding self-attention matrix, and the term $A_{i,:}^{(l,h)}(x)$ refers to the attention distribution vector for the i-th query token across all positions. p_j denotes the probability (attention weight) of the j-th token within the distribution vector.

Attention Span. Attention span [39] measures the strength of long-range dependency by computing the expected positional distance that each query token attends to:

$$s_{\text{span}}(x) = \frac{1}{|\mathcal{H}|\,T} \sum_{(l,h)\in\mathcal{H}} \sum_{i=1}^{T} \sum_{j=1}^{T} A_{i,j}^{(l,h)}(x)\,|i-j|,$$

where $A_{i,j}^{(l,h)}(x)$ represents the attention weight assigned from query token i to key token j.

Table 11 summarizes the results across multiple mathematical reasoning benchmarks, All experimental settings are kept identical to the main experiments. Our proposed variance-based score (CircuitSeer) achieves the best overall performance across both target models. This suggests that effective reasoning circuits are characterized not merely by the presence of attention, but by its structured concentration over semantically critical positions.

Table 12: Selection results under different probe set size. We report the selected attention heads under different probe set sizes. Each tuple (L, H) denotes the H-th head in layer L.

Probe Size	Selected Heads (Layer, Head)
300	**L0:** (0,1), (0,3), (0,6), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–14:** (12,1), (13,4), (14,2), (14,3) **L15–17:** (15,5), (15,7), (16,1), (16,3), (17,7) **L18–19:** (18,8), (19,6)
600	**L0:** (0,1), (0,3), (0,6), (0,7), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–14:** (12,1), (13,4), (14,2), (14,3) **L15–17:** (15,5), (15,7), (16,1), (16,3), (17,7) **L18–19:** (19,6)
900	**L0:** (0,1), (0,3), (0,6), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–15:** (12,1), (13,4), (14,2), (14,3), (15,5), (15,7) **L16–19:** (16,1), (16,3), (17,7), (18,8), (19,6)

C.2 Hyperparameter sensitivity analysis

A series of ablation studies is conducted to evaluate the sensitivity of the proposed method to key hyperparameters, including the proportion of selected attention heads, the probe set size, and the randomness in probe sampling.

Effect of top-k selection ratio. We examined three different selection ratios (1%, 5%, and 10% of all attention heads). The results are presented in Table 14.

It is observed that a selection ratio of 5% consistently yields the best overall performance across both models and most benchmarks. When too few heads are selected, important reasoning components may be omitted, leading to degraded performance. Conversely, when too many heads are included, less relevant or noisy attention patterns are incorporated, which negatively affects the quality of the selected data. These findings indicate that a balance between selectivity and coverage is required, and a moderate selection ratio is most effective for capturing reasoning-critical structures.

Sensitivity to probe set size. We further evaluated the robustness of the head detection method with respect to the probe set size. As shown in Table 12, the tested size of our probe set is varied from 300 to 900.

Strong stability is observed across different settings. The sets of selected heads obtained with probe sizes of 300 and 900 are identical, while the set obtained with 600 samples exhibits only minor differences. This suggests that CircuitSeer does not depend heavily on large probe sets and that a consistent set of core attention heads can be reliably identified even with relatively limited data.

Sensitivity to random seed. We examined the effect of stochasticity in probe sampling by repeating the head selection process with different random seeds. The corresponding results are reported in Table 13.

It is found that the selected head sets remain identical across all tested seeds, indicating complete invariance to random initialization. This behavior suggests that the identified heads correspond to intrinsic structural properties of the model rather than artifacts introduced by sampling noise.

C.3 Generalization beyond mathematical reasoning - code generation task

To further investigate the generalization ability of CircuitSeer, we conducted additional experiments on code generation tasks. For the training phase, we utilized the KodCode-V1-SFT-R1 dataset, employing CircuitSeer to select a 5% subset. For evaluation, we used three challenging benchmarks: LiveCodeBench-v5 (LCB (v5)) [40], HumanEval Base (HEB) [41], and HumanEval Plus (HEP) [42].

As detailed in Table 15, CircuitSeer delivers substantial quantitative improvements. Most notably, when applied to Qwen2.5-7b-coder, our method achieves Pass@1 scores of 6.0 on LCB (v5) and 79.2

Table 13: Selection results under different random seed.

Seed	Selected Heads (Layer, Head)
42	**L0:** (0,1), (0,3), (0,6), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–14:** (12,1), (13,4), (14,2), (14,3) **L15–17:** (15,5), (15,7), (16,1), (16,3), (17,7) **L18–19:** (18,8), (19,6)
62	**L0:** (0,1), (0,3), (0,6), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–14:** (12,1), (13,4), (14,2), (14,3) **L15–17:** (15,5), (15,7), (16,1), (16,3), (17,7) **L18–19:** (18,8), (19,6)
82	**L0:** (0,1), (0,3), (0,6), (0,10) **L5–8:** (5,11), (6,7), (8,3) **L12–14:** (12,1), (13,4), (14,2), (14,3) **L15–17:** (15,5), (15,7),(16,1), (16,3), (17,7) **L18–19:** (18,8), (19,6)

Table 14: Avg. is the unweighted mean of P@1 across nine benchmarks; M@16 is excluded. Comprehensive evaluation of Top-K selection strategies for mathematical benchmarks. The best performance among selection methods for each model is in **bold**, and the second best is underlined.

Model	AIME24		AIME25		AMC23		MATH	Olympiad	Kaoyan	GK 23	GK-Math	GK 24	Avg.
	P@1	M@16	P@1	M@16	P@1	M@16	P@1	P@1	P@1	P@1	P@1	P@1	P@1
Llama-3.2-3B													
Top 1%	0.8	**3.3**	**0.4**	0.0	11.3	15.0	21.0	7.1	4.0	**28.8**	**17.1**	11.0	11.3
Top 10%	0.0	0.0	0.0	0.0	**12.8**	25.0	21.4	5.3	3.5	28.3	13.7	9.9	10.5
CircuitSeer (5%)	**1.0**	0.0	**0.4**	**3.3**	12.3	**27.5**	**27.8**	**8.6**	**4.5**	27.3	**17.1**	**15.4**	**12.7**
Qwen2.5-Math-7B													
Top 1%	32.5	46.7	27.5	33.3	71.3	82.5	86.2	49.5	45.7	75.8	78.3	61.5	58.7
Top 10%	30.8	40.0	27.5	33.3	76.2	85.0	87.8	48.3	45.7	**79.7**	75.2	67.0	59.8
CircuitSeer (5%)	**37.5**	**63.3**	**31.5**	**50.0**	**80.3**	**95.0**	**88.4**	**53.9**	**55.3**	78.2	**84.3**	**71.4**	**64.5**

Table 15: Comparison of data selection strategies on code generation benchmarks. We report Pass@1 (P@1) accuracy on 3 benchmarks. The best performance among selection methods for each model is in **bold**, and the second best is underlined.

Model	LCB (v5) Avg (P@1)	HEB (P@1)	HEP (P@1)
Llama-3.2-3B			
Base	0	27.4	**23.8**
Quality	0	25.6	18.9
Diversity	0	25.0	22.5
Maxloss	0	26.8	21.3
Ifd	1.2	28.0	23.7
Random	0	27.4	22.5
CircuitSeer	1.2	**29.8**	21.9
Full data	**2.4**	29.3	19.5
Qwen2.5-7b-coder			
Base	2.4	81.7	76.2
Quality	4.2	80.4	75.6
Diversity	4.8	81.7	75.6
Maxloss	3.6	81.0	75.0
Ifd	3.0	83.5	77.4
Random	5.4	81.7	76.2
CircuitSeer	**6.0**	84.1	**79.2**
Full data	5.99	**86.6**	78.7

on HEP. This not only outperforms all existing selection baselines but also surpasses the performance of models trained on the entire dataset (5.99 and 78.7, respectively). Similarly, on the Llama-3.2-3B

model, CircuitSeer reaches a Pass@1 of 29.8 on HEB, exceeding both the baseline strategies and the full-data training result (29.3).

Achieving comparable or superior performance to full-dataset training using only a 5% subset demonstrates that CircuitSeer acts as a domain-agnostic method, successfully identifying high-quality reasoning data across diverse tasks.

C.4 Comparison with length-based selection

We compare CircuitSeer with a Longest baseline to test whether selecting the longest sequences alone accounts for its downstream performance. Longest ranks samples by sequence length and retains the longest sequences at the same selection budget. Both methods select a 10% subset for fine-tuning Llama-3.2-3B and Qwen2.5-Math-7B-Base.

Table 16: Comparison with length-based selection at a 10% data budget. All scores are percentages. The first three benchmarks report Pass@1 (P@1) / Majority@16 (M@16); the remaining columns report P@1, and Avg. is the average P@1 across the nine benchmarks. Bold indicates the best average within each comparison.

Method	AIME24 P@1 / M@16	AIME25 P@1 / M@16	AMC23 P@1 / M@16	MATH P@1	Olympiad P@1	Kaoyan P@1	GK23 P@1	GK-Math P@1	GK24 P@1	Avg. P@1
Llama-3.2-3B										
Longest	1.0 / 0.0	0.4 / 0.0	11.5 / 25.0	28.2	8.0	4.0	26.8	16.8	14.3	12.3
CircuitSeer (Ours)	1.0 / 0.0	0.4 / 3.3	12.3 / 27.5	27.8	8.6	4.5	27.3	17.1	15.4	**12.7**
Qwen2.5-Math-7B-Base										
Longest	36.8 / 60.0	32.2 / 53.3	78.5 / 92.5	88.0	54.2	51.5	77.5	83.2	69.2	63.5
CircuitSeer (Ours)	37.5 / 63.3	31.5 / 50.0	80.3 / 95.0	88.4	53.9	55.3	78.2	84.3	71.4	**64.5**

Table 16 shows that CircuitSeer achieves higher average P@1 than Longest on both models, although individual benchmarks exhibit mixed results. This supports the value of the received-attention concentration signal beyond ranking by length alone. It does not establish that the score is independent of sequence length. The shared generation pipeline of OpenR1-Math-220K reduces variation in instruction and reasoning styles, but this comparison does not separately isolate formatting effects or attention-sink behavior.

C.5 Comparison with additional data selection baselines

We further compare CircuitSeer with clusterUCB [43] and Data Selection with Importance Resampling (DSIR) [17]. The former uses gradient-based selection with clustering and an upper-confidence-bound strategy, while the latter selects data through importance resampling. Both additional baselines use NuminaMath-CoT as the reference dataset. We evaluate their selected subsets on the same two target models at a 10% data budget.

Table 17: Comparison with additional data selection baselines at a 10% data budget. All scores are percentages. The first three benchmarks report Pass@1 (P@1) / Majority@16 (M@16); the remaining columns report P@1, and Avg. is the average P@1 across the nine benchmarks. Bold indicates the best average within each comparison.

Method	AIME24 P@1 / M@16	AIME25 P@1 / M@16	AMC23 P@1 / M@16	MATH P@1	Olympiad P@1	Kaoyan P@1	GK23 P@1	GK-Math P@1	GK24 P@1	Avg. P@1
Llama-3.2-3B										
clusterUCB	0.0 / 0.0	0.0 / 0.0	8.8 / 17.5	23.8	4.8	3.5	28.0	16.8	12.1	10.9
DSIR	0.0 / 3.3	0.0 / 0.0	9.4 / 25.0	24.6	7.0	2.0	25.0	15.5	12.0	10.6
CircuitSeer (Ours)	1.0 / 0.0	0.4 / 3.3	12.3 / 27.5	27.8	8.6	4.5	27.3	17.1	15.4	**12.7**
Qwen2.5-Math-7B-Base										
clusterUCB	36.7 / 56.7	30.0 / 43.3	77.5 / 92.5	87.5	51.5	50.5	77.0	81.5	67.0	62.1
DSIR	31.1 / 46.7	26.3 / 36.7	73.4 / 92.5	85.4	47.6	50.5	76.9	81.2	65.2	59.7
CircuitSeer (Ours)	37.5 / 63.3	31.5 / 50.0	80.3 / 95.0	88.4	53.9	55.3	78.2	84.3	71.4	**64.5**

As shown in Table 17, CircuitSeer achieves the highest average P@1 on both target models. These comparisons extend the empirical evaluation to a gradient-based selection framework and

an importance-resampling method. The advantage is an aggregate result rather than a uniform improvement on every benchmark and metric.

C.6 Effect of data selection budget

We evaluate CircuitSeer against Random selection on Llama-3.2-3B at 5%, 10%, and 15% data budgets. This comparison tests whether the advantage over random selection persists under both stricter and looser data constraints.

Table 18: Effect of data selection budget on Llama-3.2-3B. All scores are percentages. The first three benchmarks report Pass@1 (P@1) / Majority@16 (M@16); the remaining columns report P@1, and Avg. is the average P@1 across the nine benchmarks. Bold indicates the best average within each comparison.

Method	AIME24 P@1 / M@16	AIME25 P@1 / M@16	AMC23 P@1 / M@16	MATH P@1	Olympiad P@1	Kaoyan P@1	GK23 P@1	GK-Math P@1	GK24 P@1	Avg. P@1
Llama-3.2-3B										
5% (Random)	0.2 / 0.0	0.0 / 0.0	6.1 / 15.0	9.6	1.5	2.0	12.2	9.4	5.5	5.17
5% (CircuitSeer)	0.2 / 0.0	0.4 / 0.0	6.6 / 20.0	10.2	4.6	2.0	14.3	13.1	11.0	**6.93**
10% (Random)	0.2 / 0.0	0.4 / 0.0	9.4 / 22.5	25.6	6.5	3.0	28.3	16.8	13.2	11.49
10% (CircuitSeer)	1.0 / 0.0	0.4 / 3.3	12.3 / 27.5	27.8	8.6	4.5	27.3	17.1	15.4	**12.72**
15% (Random)	0.2 / 0.0	0.4 / 3.3	13.4 / 27.5	27.0	8.4	4.0	24.4	16.8	15.4	12.2
15% (CircuitSeer)	1.0 / 0.0	0.4 / 3.3	15.0 / 25.0	28.0	8.6	5.0	25.9	16.5	16.5	**12.99**

Table 18 shows higher average P@1 for CircuitSeer at every evaluated budget, with a smaller advantage at larger budgets. This does not imply that subset training always outperforms full-data SFT: the latter remains stronger on average for Llama-3.2-3B (Table 1).

D Broader impacts

CircuitSeer is intended to improve the efficiency of selecting training data for mathematical reasoning. The potential benefits and risks below concern its use as a data-selection method; the experiments do not establish deployment safety or benefits outside the evaluated settings.

Potential benefits. Selecting a smaller training subset may lower fine-tuning time and computational cost, making reasoning-oriented adaptation more accessible to researchers with limited resources. The model-internal scoring signal may also help researchers examine data-selection behavior. These potential benefits depend on the target model and dataset, and must account for the additional cost of head detection and data scoring.

Potential risks. More efficient fine-tuning could also facilitate adaptation of language models for harmful purposes. Scores derived from a particular reference model may preserve or amplify its biases, and selection for mathematical benchmarks may underrepresent other domains or reasoning styles. Improved benchmark accuracy does not establish fairness, factual reliability, or safe behavior in downstream applications. CircuitSeer is not designed or evaluated as a content-safety filter.

Responsible use. Downstream users should respect the licenses and access conditions of the underlying data and models, examine the coverage of selected subsets, and evaluate performance and safety on the intended application. Independent safety screening and appropriate deployment controls remain necessary; attention-based selection alone does not provide these safeguards.

E Case study

For reference, we provide a few detailed case study examples here.

Example of a High-Scoring Sample (High CircuitSeer Score)

Input Problem (x):
At the 75th United Nations General Assembly, China pledged to adopt more effective policies and measures to strive to peak carbon dioxide emissions before 2030 and achieve carbon neutrality (referred to as the "dual carbon goals") before 2060. This demonstrates China's firm determination to address climate change, indicating that China's economic structure and socio-economic operational models will undergo profound transformations, greatly promoting the clean and green transition of its industrial chain. New energy vehicles, particularly electric vehicles, are key strategic emerging industries that play a crucial role in achieving the dual carbon goals.
To understand the sales of electric vehicles in a certain region, an organization obtained, via the least squares method, the linear regression equation for sales volume y (in units of 10,000 vehicles) with respect to year x:

$$\hat{y} = 4.7x - 9495.2$$

The sample variance of the sales volume y is $S_y^2 = 50$, and the sample variance of the year x is $S_x^2 = 2$.

Problem (1)
Calculate the correlation coefficient r between y and x, and based on this, determine the strength of the linear relationship between electric vehicle sales volume y and year x.

Problem (2)
The organization also surveyed 100 car owners in the region regarding gender and vehicle type purchased. The data are summarized in the table below:

	Non-electric Cars	Electric Cars	Total
Male	30	20	50
Female	15	35	50
Total	45	55	100

Can it be concluded, with 99% confidence, that the purchase of electric cars is associated with gender?

Problem (3)
Among the owners who purchased electric cars, a stratified sample of 11 individuals was selected according to gender. Then, 2 individuals were randomly selected from these 11. Let X denote the number of males among the 2 selected individuals. Find the probability distribution and the mathematical expectation of X.

Reference Formulas

(i) Linear regression equation: $\hat{y} = \hat{b}x + a$, where

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$
$$a = \overline{y} - \hat{b}\,\overline{x}.$$

(ii) Correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

If $r > 0.9$, it can be judged that y and x are strongly linearly correlated.

(iii) Chi-square test statistic for 2×2 contingency table:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$
where $n = a + b + c + d$.

Reasoning Trace (CoT, r):
(1) The correlation coefficient r is derived from the regression slope $\hat{b} = 4.7$ and the standard deviations $S_x = \sqrt{2}$, $S_y = \sqrt{50}$. Using the formula $r = \hat{b} \cdot \frac{S_x}{S_y}$, we compute:

$$r = 4.7 \cdot \frac{\sqrt{2}}{\sqrt{50}} = 4.7 \cdot \frac{1}{5} = 0.94.$$

Since $r > 0.9$, the linear relationship is strong.
(2) For the chi-square test, assign $a = 30, b = 20, c = 15, d = 35$. Compute:

$$\begin{aligned}\chi^2 &= \frac{100 \cdot (30 \cdot 35 - 20 \cdot 15)^2}{50 \cdot 50 \cdot 45 \cdot 55} \\ &= \frac{100 \cdot 750^2}{6{,}187{,}500} \\ &= \frac{56{,}250{,}000}{6{,}187{,}500} \approx 9.09.\end{aligned}$$

Since $9.09 > 6.635$ (critical value for $\alpha = 0.01$), we reject independence — gender and car type are associated at 99% confidence.
(3) Among 55 electric car owners (20M, 35F), a proportional stratified sample of 11 yields 4 males and 7 females. Selecting 2 from 11, $X \sim \text{Hypergeometric}(N = 11, K = 4, n = 2)$. Thus:

$$\begin{aligned}P(X = 0) &= \frac{\binom{7}{2}}{\binom{11}{2}} = \frac{21}{55}, \\ P(X = 1) &= \frac{\binom{4}{1}\binom{7}{1}}{\binom{11}{2}} = \frac{28}{55}, \\ P(X = 2) &= \frac{\binom{4}{2}}{\binom{11}{2}} = \frac{6}{55}.\end{aligned}$$

Expectation: $E[X] = 2 \cdot \frac{4}{11} = \frac{8}{11}$.

Final Answer (y):
(1) Correlation coefficient: $\boxed{0.94}$ — strong linear relationship.
(2) Chi-square statistic: $\boxed{9.09}$ — sufficient evidence at 99% confidence to conclude association between gender and electric car purchase.
(3) Probability distribution:

$$\begin{cases} P(X = 0) = \dfrac{21}{55}, \\ P(X = 1) = \dfrac{28}{55}, \\ P(X = 2) = \dfrac{6}{55}, \end{cases} \qquad \text{Expectation: } \boxed{\frac{8}{11}}.$$

Example: Integer Values of a Power Sum Expression (High CircuitSeer Score)

Input Problem (x):
How many positive integers n are there for which $\left(2022 + \frac{1}{2}\right)^n + \left(25 + \frac{1}{2}\right)^n$ is an integer?

Reasoning Trace (CoT, r):
First, express the numbers as fractions:

$$2022.5 = \frac{4045}{2}, \quad 25.5 = \frac{51}{2}$$

So the expression becomes:

$$\left(\frac{4045}{2}\right)^n + \left(\frac{51}{2}\right)^n = \frac{4045^n + 51^n}{2^n}$$

We require 2^n divides $4045^n + 51^n$.
Both 4045 and 51 are odd, so their powers are odd, and their sum is even (divisible by 2). But for higher n, we need divisibility by higher powers of 2.
Using the Lifting The Exponent Lemma (LTE) for $p = 2$: Since $4045 \equiv 1 \pmod 4$ and $51 \equiv 3 \pmod 4$, $4045 \equiv -51 \pmod 4$. For odd n, LTE gives $v_2(4045^n + 51^n) = v_2(4045 + 51) + v_2(n) = 12 + v_2(n)$, since $4045 + 51 = 4096 = 2^{12}$.
For 2^n to divide $4045^n + 51^n$, we need $12 + v_2(n) \geq n$.

- For $n = 1$, $v_2(1) = 0$, $12 \geq 1$: true.
- For $n = 3, 5, 7, 9, 11$, $v_2(n) = 0$, so $12 \geq n$.

So, all odd $n \leq 11$ are valid.
For even n, the sum modulo 4 is 2, not divisible by 4, so for $n \geq 2$ even, the sum is not divisible by 2^n.
Thus, the valid n are $1, 3, 5, 7, 9, 11$.

Final Answer (y):
There are $\boxed{6}$ positive integers n for which $\left(2022 + \frac{1}{2}\right)^n + \left(25 + \frac{1}{2}\right)^n$ is an integer.

Example: Only Even Prime Number (Low CircuitSeer Score)

Input Problem (x):
Find all even primes.

Reasoning Trace (CoT, r):
Prime numbers are defined as natural numbers greater than 1 with no positive divisors other than 1 and themselves. The only even prime number is 2, because:

- All even numbers (except 2) are divisible by 2, so they have at least three divisors: 1, 2, and themselves.
- 2 has exactly two divisors: 1 and 2.

For any even number $p > 2$, $p = 2k$ for $k > 1$, so p is divisible by 1, 2, and p, hence not prime.
Thus, 2 is the only even prime.

Final Answer (y):
The only even prime is $\boxed{2}$.

Example: Reciprocal of -1 (Low CircuitSeer Score)

Input Problem (x):
The reciprocal of -1 is_____.

Reasoning Trace (CoT, r):
The reciprocal of a number x is defined as the number y such that $x \cdot y = 1$. For $x = -1$:

$$(-1) \cdot y = 1 \implies y = \frac{1}{-1} = -1.$$

Verification: $(-1) \times (-1) = 1$, which satisfies the definition.
Thus, the reciprocal of -1 is -1.

Final Answer (y):
The reciprocal of -1 is $\boxed{-1}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contribution: identifying reasoning-critical attention heads and using their internal attention signals for reasoning-oriented data selection. The claims are supported by the experimental results in Section 4.2–Section 4.4, and the limitations are discussed in Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A discusses limitations, including the need for access to internal model parameters and attention weights, limited applicability to closed-source API-only models, and possible model-specific biases.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The method is empirically motivated and evaluated through experiments.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the data selection procedure, head detection method, scoring function, baselines, evaluation benchmarks, prompts, and training settings. Additional hyperparameters and implementation details are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:We provide the complete code required to reproduce the main experimental results in the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes the datasets, baselines, models, evaluation metrics, and decoding settings. Appendix B provides detailed training settings, including optimizer, learning rate, batch size, warmup ratio, sequence length, RoPE scaling, and hardware setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports extensive benchmark results and ablation studies, but does not include error bars, confidence intervals, or statistical significance tests due to the high computational cost of repeating large-scale SFT experiments.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the use of 8 NVIDIA A100 GPUs and provides runtime information for head detection, data scoring, 10% subset SFT, and full-dataset SFT in Section 4.4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: The research uses public datasets and publicly available models for data selection and evaluation, and does not involve human subjects, private data, or deceptive data collection. The work is intended to improve efficient and transparent model training.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:Appendix D discusses potential positive impacts, including improved efficiency, reduced training cost, and increased interpretability in reasoning-oriented data selection. It also discusses potential negative impacts, including misuse through more efficient LLM fine-tuning and the possibility of amplifying model-specific biases or benchmark-specific artifacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pre-trained model, image generator, scraped dataset, or other asset with a high risk of misuse. It proposes a data selection method evaluated on existing public datasets and models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the datasets, models, and benchmarks used in the experiments, we use publicly available assets and intend to respect their corresponding terms. We have cited the original owners of assets used in the paper properly.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, benchmark, model, or software package as a new asset. The contribution is a data selection method evaluated using existing assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, user studies, or crowdsourcing, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are central to the proposed method: the paper uses reference LLMs to identify reasoning-critical attention heads and to score data samples, and evaluates selected data through SFT of multiple target LLMs. These details are described in Sections 3 and 4.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.