What Makes a Good Curriculum? Disentangling the Effects of Data Ordering on LLM Mathematical Reasoning

Anonymous ACL submission

Abstract

Curriculum learning (CL)—ordering training data from easy to hard—has become a popular strategy for improving reasoning in large language models (LLMs). Yet prior work employs disparate difficulty metrics and training setups, leaving open fundamental questions: *When does curriculum help? Which direction—forward or reverse—is better? And does the answer depend on what we measure?* We address these questions through a unified offline evaluation framework that decomposes curriculum difficulty into five complementary dimensions: *Problem Difficulty*, *Model Surprisal*, *Confidence Margin*, *Predictive Uncertainty*, and *Decision Variability*. Through controlled post-training experiments on mathematical reasoning benchmarks with *Llama3.1-8B*, *Mistral-7B*, and *Gemma3-4B*, we find that: *(i)* no curriculum strategy dominates universally—the relative effectiveness of forward versus reverse CL depends jointly on model capability and task complexity; *(ii)* even within a single metric, samples at different difficulty levels produce distinct gains depending on task demands; and *(iii) Task-aligned* curricula focus on shaping the model's final representations and generalization, whereas *inner-state* curricula modulate internal states such as confidence and uncertainty. Our findings challenge the notion of a universal curriculum strategy and offer actionable guidance across model and task regimes, with some metrics indicating that prioritizing decision-uncertain samples can further enhance learning outcomes.

1 Introduction

Curriculum Learning (CL), which organizes training data by difficulty, has become an effective approach for improving the reasoning ability of large language models (LLMs) (Bengio et al., 2009; Kumar et al., 2010; Jiang et al., 2015; Gan et al., 2021; Feng et al., 2025). By presenting examples in a structured progression, CL has demonstrated improvements across diverse reasoning domains, including code generation (Naïr et al., 2024; Pouransari et al., 2024), instruction following (Lee et al., 2024; Pang et al., 2024), and commonsense reasoning (Maharana and Bansal, 2022). Yet despite its widespread adoption, a fundamental question remains unresolved: *what makes a curriculum strategy effective?*

The answer is far from straightforward. Existing work employs a heterogeneous landscape of difficulty metrics that fall into two broad families. *Problem-side metrics* quantify intrinsic task complexity—such as reasoning depth (Jung and Jung, 2025), linguistic complexity (Liu, 2008), symbolic density (Polu et al., 2022; Meadows et al., 2024), or model-independent success rates like Acc@K (Tong et al., 2024). These metrics enable curricula that progress from structurally simple to complex problems, often improving both efficiency and accuracy (Jung and Jung, 2025; Ding et al., 2024; Parashar et al., 2025). *Model-side metrics*, in contrast, capture difficulty from the model's perspective—measuring uncertainty (Zhou et al., 2020; Peng et al., 2023), confidence (Pattnaik et al., 2024), or surprisal (Pattnaik et al., 2024; Mohiuddin et al., 2022). By adaptively focusing on samples that challenge the model, these metrics can better expose knowledge gaps and enhance generalization in high-difficulty regimes.

However, prior work varies substantially in training protocols, evaluation setups, and even the direction of curriculum ordering (forward vs. reverse), making systematic comparison difficult. Many studies adopt online or reinforcement-learning–based curricula that dynamically adjust sampling (Chen et al., 2025; Li et al., 2025; Zhang et al., 2025a), confounding the effect of data ordering with optimization dynamics or reward shaping. Moreover, the choice of difficulty metric is often treated as a design decision rather than an empirical question, leaving practitioners without clear

guidance on which metric to use, when forward or reverse ordering is preferable, or how these choices interact with model capability and task complexity.

We address these gaps through a controlled, multidimensional evaluation of curriculum strategies in mathematical reasoning. We decompose curriculum difficulty into five complementary metric families: *Problem Difficulty*, *Model Surprisal*, *Confidence Margin*, *Predictive Uncertainty*, and *Decision Variability*, and systematically construct both forward (easy to hard) and reverse (hard to easy) curricula for each. Crucially, we focus on *supervised fine-tuning (SFT)* under an *offline* setting, where training data are fixed but reordered before learning. This design isolates the effect of data ordering by holding objectives, budgets, and optimization procedures constant, providing a clean lens to study curriculum effects without confounds from adaptive sampling or reward engineering.

We conduct controlled post-training experiments on standard mathematical reasoning benchmarks using three representative open-source base models—*Llama-8B*, *Mistral-7B*, and *Gemma-4B*—spanning different scales and architectural families. This enables us to examine not only *whether* curricula help, but *when*, *why*, and *for whom* they are most effective. Our findings reveal four key insights that challenge conventional lessons about curriculum learning:

① **Data ordering is a first-class learning signal.** Even when the training budget and data composition are fixed, curriculum structure alone can substantially alter learning dynamics and reasoning outcomes, highlighting that *how* we present data matters as much as *what* we present.

② **No curriculum dominates universally.** The relative advantage of forward versus reverse CL depends jointly on model reasoning capacity, task complexity, and the chosen difficulty metric. Stronger models benefit from forward CL on simpler tasks, while weaker models or harder tasks often favor reverse ordering—a pattern that holds across multiple metric families.

③ **Curricula reshape optimization trajectories.** *Inner-state* curricula primarily influence learning dynamics by adjusting convergence speed and stability, whereas *task-aligned* curricula affect the final convergence point, shaping the model's representations and generalization, suggesting that curricula can steer not only *how* models learn, but also *where* they ultimately converge.

④ **Curricula modulate internal model states.** Beyond reasoning accuracy, CL shapes predictive uncertainty and confidence: Forward CL yields more cautious, uncertainty-aware models, while Reverse CL preserves decisiveness with higher confidence, highlighting that curricula influence not only *what* models learn, but also *how* they calibrate and express their internal properties.

Together, these results establish a unified framework for analyzing curriculum effects in LLMs and provide actionable guidance across different scenarios. For example, prioritizing decision-uncertain or high-confidence problems during training can further enhance a model's reasoning capabilities.

2 Method

We design a metric-driven offline curriculum framework to systematically evaluate how different difficulty metrics and ordering strategies affect LLM reasoning, as shown in Fig 1. Our approach consists of three stages: *(i)* quantifying difficulty through complementary problem-side and model-side metrics, *(ii)* constructing curricula by reordering training data according to these metrics, and *(iii)* evaluating the resulting models on both in-distribution and out-of-distribution benchmarks. Figure 1 illustrates the complete pipeline.

2.1 Difficulty Metrics: Two Complementary Views

To capture the multifaceted nature of difficulty, we decompose it into five metric families spanning two perspectives: *problem-side metrics* that quantify intrinsic task complexity, and *model-side metrics* that reflect the model's internal state when processing each example.

2.1.1 Problem-Side Metrics

Problem-side metrics assess difficulty independently of model behavior, either through human annotation or automated analysis. We collectively refer to these as *Problem Difficulty* and consider four instantiations:

Reasoning Step (RS) counts the number of distinct logical operations required to reach a solution, capturing the depth of the reasoning chain.

Symbol Complexity (SC) quantifies notational richness on a 1–5 scale, where 1 denotes basic arithmetic and 5 denotes advanced constructs (integrals, limits, set theory).

Comprehension Difficulty (CD) measures how hard a problem is to *understand* rather than solve,

rated 1–5 based on ambiguity, context requirements, and prerequisite knowledge.

Acc@K (ACC) quantifies empirical difficulty by sampling K independent generations and computing the proportion of correct solutions:

$$\mathrm{ACC} = \frac{1}{K}\sum_{s=1}^{K} z_s = \hat{p}, \tag{1}$$

where $z_s \in 0, 1$ denotes the correctness of the s-th generation (1 if correct, 0 otherwise).

We estimate RS, SC, and CD using **Qwen2.5-Math-72B-Instruct**, while ACC is computed with respect to the target model under evaluation.

2.1.2 Model-Side Metrics

Model-side metrics capture difficulty from the model's perspective by measuring internal properties during generation. To account for stochasticity, we decode with temperature 0.7 and average each metric over 20 independent completions per problem.

Notation Given a problem x, model M_θ generates response $y_{1:T}$ token-by-token with distribution $p_\theta(y_t \mid y_{<t}, x)$. Let $\ell_t = \log p_\theta(y_t \mid y_{<t}, x)$ denote the log-probability of the emitted token. For the top-k candidates at position t, we define the truncated distribution $q_{t,i} = e^{\ell_{t,i}} / \sum_{j=1}^{k} e^{\ell_{t,j}}$ (with $k = 5$). We denote sorted log-probabilities as $\ell_{t,(1)} \geq \ell_{t,(2)} \geq \cdots$.

Model Surprisal quantifies how unexpected a sequence is to the model via perplexity. We measure this at two granularities:

Sequence-Level Perplexity (SLP) provides a holistic measure over the entire sequence:

$$\mathrm{SLP} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \ell_t\right). \tag{2}$$

Token-Level Perplexity (TLP) captures local uncertainty by averaging token-wise entropy:

$$\mathrm{TLP} = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\left(-\sum_{i=1}^{k} q_{t,i}\log q_{t,i}\right)\right). \tag{3}$$

Confidence Margin assesses decisiveness by measuring the separation between top predictions. The *Logit Gap (LG)* averages this margin across positions where at least two candidates exist:

$$\mathrm{LG} = \frac{1}{U}\sum_{t\in\mathcal{G}}\left(\ell_{t,(1)} - \ell_{t,(2)}\right), \tag{4}$$

where $\mathcal{G} = \{t : m_t \geq 2\}$ and $U = |\mathcal{G}|$.

Predictive Uncertainty quantifies distributional dispersion via entropy. We compute this at both sequence and token levels:

Sequence-Level Entropy (SLE) sums token-wise entropies:

$$\mathrm{SLE} = \sum_{t=1}^{T}\left(-\sum_{i=1}^{k} q_{t,i}\,\log_2 q_{t,i}\right). \tag{5}$$

Token-Level Entropy (TLE) measures local uncertainty:

$$\mathrm{TLE} = -\sum_{i=1}^{k} q_{t,i}\,\log_2 q_{t,i}. \tag{6}$$

Decision Variability measures the stability of a model's predictions across repeated generations of the same problem. It is quantified by the *Variance of Acc@K* (VACC), which reflects fluctuations in correctness across K trials. Given binary correctness indicators $z_s \in \{0, 1\}$, we define

$$\mathrm{VACC} \;=\; \frac{1}{K}\sum_{s=1}^{K}\left(z_s - \hat{p}\right)^2 \;=\; \hat{p}\left(1 - \hat{p}\right), \tag{7}$$

where $\hat{p}$ is the empirical accuracy from Eq. 1.

2.2 Curriculum Construction

Given a difficulty metric, we construct three primary curriculum strategies by reordering the training data:

Forward Curriculum Learning (FCL) orders examples in ascending difficulty (easy-to-hard), allowing models to build foundational knowledge before tackling complex instances.

Reverse Curriculum Learning (RCL) orders examples in descending difficulty (hard-to-easy), exposing models to challenging cases early to potentially accelerate learning.

Single Group Curriculum (SGC) isolates a specific difficulty tier (Low, Medium, or High) by training exclusively on shuffled examples from that group, enabling controlled analysis of tier-specific effects.

We also explore two group based variants, **Group Forward Curriculum (GFC)** and **Group Reverse Curriculum (GRC)**, which partition the data into three tiers, shuffle samples within each tier, and then arrange the tiers in forward or reverse order. These strategies balance global difficulty progression with robustness to fine grained ranking noise. Full implementation details and additional experimental results are provided in Appendix D.

2.3 Metric-driven Curriculum Framework

For each combination of metric and curriculum strategy, we fine-tune a base model and evaluate it across multiple benchmarks to examine three questions: *(i)* whether curriculum-based ordering improves over random shuffling, *(ii)* which direction of progression (forward or reverse) is more effective, and *(iii)* how these effects vary with model capacity and task complexity. In this offline setup, where the data remain fixed but their order is rearranged, we isolate the impact of curriculum structure while keeping the training objective, budget, and optimization procedure constant. This controlled design enables clear attribution of performance differences to data ordering alone.

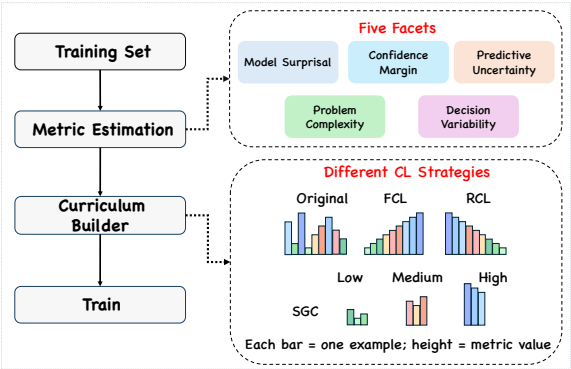

Figure 1: Overview of the metric-driven curriculum framework. **Left:** From the training set, we estimate difficulty metrics, construct curricula, and train models under controlled settings. **Right (top):** Five metric families spanning problem-side and model-side perspectives. **Right (bottom):** Curriculum schedules, including *Original* (random order), *FCL* (easy→hard), *RCL* (hard→easy), and *SGC* (single-tier focus). Each bar denotes a training instance, with height reflecting its metric value.

3 Experiment

3.1 Setup

Datasets During training, we use MetaMathQA-40K (Yu et al., 2023) as our training dataset. It consists of 40K automatically generated, high-quality mathematical question–answer pairs specifically designed to evaluate and enhance advanced reasoning in large language models. The full MetaMathQA corpus is first shuffled, and 20K samples are randomly selected as the training set. To better evaluate generalization, we assess model performance on the in-distribution *MetaMathQA (MMQA)* and four out-of-distribution mathematical reasoning benchmarks: *ASDiv* (Miao et al., 2020), *GSM8K* (Cobbe et al., 2021), *MATH* (Hendrycks et al., 2021), and *MathBench (MBench)* (Liu et al., 2024a). For *MBench*, we focus on application-style tasks from the *Arithmetic*, *Primary*, and *Middle* difficulty levels.

Models We employ three open-source large language models in our experiments: Gemma3-4B-pt (Team et al., 2025), Mistral-7B-v0.3 (Jiang et al., 2023), and Llama3.1-8B (Dubey et al., 2024). For brevity, we hereafter denote them as *Gemma3-4B*, *Llama3-8B*, and *Mistral-7B*, respectively. All three are base (non-instruction-tuned) models, allowing us to isolate the effects of CL from instruction tuning or other external factors. Experimental configurations are described in Section C, with implementation code released at `https://anonymous.4open.science/r/Curriculum_evaluation-A472`.

3.2 FCL vs. RCL with Multidimensional Metrics

We evaluate Gemma3-4B, Llama3-8B, and Mistral-7B under different curriculum learning strategies on five mathematical reasoning benchmarks: four out-of-distribution datasets (ASDiv, GSM8K, MATH, and MBench) and one in-distribution dataset (MMQA). The in-distribution setting measures how curriculum ordering affects performance within the training domain, while the out-of-distribution setting evaluates generalization under distribution shift. Each dataset contains 512 randomly sampled problems. Results are reported across five metric dimensions: *problem difficulty*, *model surprisal*, *confidence margin*, *predictive uncertainty*, and *decision variability*.

3.2.1 Problem Difficulty

Problem difficulty metrics include reasoning steps (RS), symbol complexity (SC), comprehension difficulty (CD), and Acc@K (ACC). Results on the five mathematical reasoning benchmarks are shown in Table 1.

Overall, three consistent patterns emerge. **(i)** Curricula based on problem difficulty metrics consistently outperform the randomly shuffled baseline, with gains of up to 3–6% on easier datasets such as *ASDiv* and *GSM8K* (e.g., *Llama-8B*: +4.7% on *ASDiv* under $\text{RS}_{\downarrow}$), confirming that difficulty-aware ordering enhances reasoning performance. **(ii)** The benefit diminishes as task complexity increases: on

MBench and *MATH*, improvements drop below 2% or even reverse (e.g., *Mistral-7B*: 18.8→16.4% on *MATH*), suggesting that complex reasoning leaves less room for curriculum effects. **(iii)** Under distribution shift, Reverse CL generally excels on simpler datasets, whereas Forward CL performs better on harder ones (e.g., *Gemma-4B*: 19.5→24.8% on *MATH* under $SC_{\uparrow}$). Notably, *Mistral-7B* consistently favors Reverse CL across all benchmarks, and reasoning-step and symbol-complexity metrics yield the most reliable gains.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$RS_{\uparrow}$	54.49	75.78	54.30	35.77	25.78
	$RS_{\downarrow}$	53.91	75.39	56.05	30.27	19.53
	$SC_{\uparrow}$	52.73	76.37	55.86	37.50	24.80
	$SC_{\downarrow}$	51.37	77.34	53.52	30.27	24.02
	$CD_{\uparrow}$	48.63	76.17	52.34	33.01	21.88
	$CD_{\downarrow}$	50.59	76.56	54.10	32.42	20.90
	$ACC_{\uparrow}$	58.59	75.20	56.25	28.32	24.61
	$ACC_{\downarrow}$	56.64	77.15	57.42	36.52	23.63
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$RS_{\uparrow}$	50.59	74.22	59.77	36.91	16.41
	$RS_{\downarrow}$	60.16	73.24	65.62	38.09	17.58
	$SC_{\uparrow}$	51.37	71.68	55.66	39.45	16.41
	$SC_{\downarrow}$	50.59	72.85	59.57	37.89	16.99
	$CD_{\uparrow}$	52.15	72.85	57.81	36.72	16.21
	$CD_{\downarrow}$	53.76	73.63	62.30	38.12	16.80
	$ACC_{\uparrow}$	47.27	63.87	54.30	35.55	14.65
	$ACC_{\downarrow}$	52.15	71.48	60.35	36.52	16.80
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$RS_{\uparrow}$	54.49	77.34	60.35	35.74	21.88
	$RS_{\downarrow}$	63.67	77.73	70.90	38.67	20.31
	$SC_{\uparrow}$	61.72	75.78	64.95	38.48	20.90
	$SC_{\downarrow}$	59.79	74.61	64.85	37.70	17.77
	$CD_{\uparrow}$	49.61	71.09	57.33	31.64	19.74
	$CD_{\downarrow}$	52.93	76.95	62.89	36.33	19.13
	$ACC_{\uparrow}$	56.25	76.17	60.16	36.91	21.02
	$ACC_{\downarrow}$	55.86	78.71	65.04	36.52	20.51

Table 1: Accuracy (%) on five mathematical reasoning benchmarks for curricula constructed from **problem difficulty** metrics (RS, SC, CD, ACC). The **baseline** uses randomly shuffled data; ↑ and ↓ indicate ascending (Forward CL) and descending (Reverse CL) ordering, respectively. See Sec. 2.1 for metric definitions.

3.2.2 Model Surprisal

Model surprisal is quantified using sequence-level perplexity (SLP) and token-level perplexity (TLP). Results are reported in Table 2.

We can observe that perplexity serves as an effective yet partially unstable curriculum signal for mathematical reasoning. **(i)** In most cases, Forward CL (low to high perplexity) outperforms Reverse CL, suggesting that starting from low-perplexity samples provides a more stable optimization signal, consistent with prior findings (Zhang et al., 2025c). **(ii)** However, the directionality is not universal. In some cases, Reverse CL achieves higher gains (e.g., *Mistral-7B*: +2.7% on *GSM8K* under $SLP_{\downarrow}$), implying that high-perplexity samples can help models adapt to semantic noise and improve robustness (Zhang et al., 2024), while Forward CL may overfit early confident patterns. **(iii)** SLP provides a more stable signal for curriculum direction, favoring Forward CL in most cases (11/15, including all for *Gemma-4B* and *Llama-8B*). In contrast, TLP exhibits less stability but often achieves stronger gains, with RCL frequently surpassing FCL on harder datasets (e.g., *Llama-8B*: 21.09→24.02 on *MATH*), suggesting its heightened sensitivity to local variations can occasionally yield larger improvements.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$SLP_{\uparrow}$	56.84	75.20	56.84	32.42	24.41
	$SLP_{\downarrow}$	48.63	73.44	52.15	31.45	24.02
	$TLP_{\uparrow}$	53.12	75.59	54.30	38.48	22.85
	$TLP_{\downarrow}$	46.09	71.88	49.02	33.79	20.70
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$SLP_{\uparrow}$	51.17	72.27	57.81	40.23	17.38
	$SLP_{\downarrow}$	52.73	73.24	63.09	40.43	17.19
	$TLP_{\uparrow}$	53.32	70.70	62.70	40.04	18.55
	$TLP_{\downarrow}$	54.49	73.24	60.55	39.43	16.21
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$SLP_{\uparrow}$	57.23	76.76	65.43	42.04	20.90
	$SLP_{\downarrow}$	50.98	72.66	60.16	40.04	18.36
	$TLP_{\uparrow}$	58.40	77.15	67.77	33.20	21.09
	$TLP_{\downarrow}$	58.98	75.39	66.02	36.72	24.02

Table 2: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed with **model surprisal** (SLP, TLP).

3.2.3 Confidence Margin

Confidence margin, measured by the logit gap (LG), quantifies how confidently a model prefers its top prediction over alternatives. Larger gaps indicate stronger confidence in token-level decisions.

Table 3 shows that the logit gap (LG) serves as an effective and stable curriculum signal for mathematical reasoning. **(i)** Across models, LG-based curricula consistently outperform the shuffled baseline, particularly on easier datasets such as *ASDiv* (e.g., *Llama-8B*: 74.02→77.93%), following a consistent trend (RCL > Baseline > FCL). This confirms that LG provides a reliable measure of task difficulty and model confidence. **(ii)** Reverse CL (high-confidence first) strengthens model confidence by exposing it to high-margin examples early in training, allowing the model to form more decisive predictions and improving reasoning performance (e.g., *Gemma-4B*: +3.5% on *ASDiv*, +3.0% on *MATH*). Overall, LG offers a clear

and direction-sensitive curriculum signal, with Reverse CL providing the most consistent gains across benchmarks.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$LG_{\uparrow}$	52.73	75.20	53.52	35.55	24.61
	$LG_{\downarrow}$	54.30	77.15	57.03	38.09	24.72
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$LG_{\uparrow}$	55.27	74.22	60.16	40.04	16.21
	$LG_{\downarrow}$	52.24	70.12	59.96	40.23	17.97
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$LG_{\uparrow}$	56.25	77.93	69.92	34.77	19.73
	$LG_{\downarrow}$	58.79	75.39	66.80	38.48	21.68

Table 3: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed with **confidence margin** (LG).

3.2.4 Predictive Uncertainty

The predictive uncertainty of a model is measured by entropy-based metrics, including entropy at the sentence level (SLE) and entropy at the token level (TLE). These metrics reflect how dispersed or peaked the model's predictive distributions are, with higher entropy indicating greater uncertainty in its predictions.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$SLE_{\uparrow}$	53.71	75.59	55.86	34.57	24.22
	$SLE_{\downarrow}$	53.52	78.52	55.47	33.40	23.63
	$TLE_{\uparrow}$	53.52	75.78	56.05	35.16	22.85
	$TLE_{\downarrow}$	50.59	75.59	53.91	31.45	25.79
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$SLE_{\uparrow}$	54.49	72.27	59.57	38.87	14.84
	$SLE_{\downarrow}$	52.73	74.02	59.96	39.94	18.75
	$TLE_{\uparrow}$	54.66	70.70	60.16	40.21	16.80
	$TLE_{\downarrow}$	50.00	73.24	60.74	40.43	16.98
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$SLE_{\uparrow}$	58.98	79.49	61.91	39.45	21.29
	$SLE_{\downarrow}$	56.25	74.02	62.89	34.96	21.88
	$TLE_{\uparrow}$	58.01	72.85	70.12	35.94	23.83
	$TLE_{\downarrow}$	55.47	76.95	56.84	36.72	22.46

Table 4: Accuracy (%) on five mathematical reasoning benchmarks with curricula derived from **predictive uncertainty** (entropy-based metrics, SLE and TLE).

Table 4 shows that entropy-based curricula provide informative yet unstable difficulty signals. **(i)** Both SLE and TLE effectively capture predictive uncertainty, especially on easier datasets such as *ASDiv* (*Llama-8B*: 79.49% under $SLE_{\uparrow}$). However, their stability is lower than that of logit-gap or perplexity-based metrics, and performance occasionally declines on harder tasks (*Mistral-7B*: –3.9% on *MATH*). **(ii)** For in-distribution evaluation, forward CL consistently outperforms reverse CL across models, indicating that ordering by confidence (low-entropy first) facilitates more stable optimization. **(iii)** Between metrics, TLE is more noise-prone yet occasionally yields larger gains (*Llama-8B*: +7.6% on *MBench* under $TLE_{\uparrow}$), consistent with findings that token-level entropy is often dominated by noisy high-entropy outliers rather than stable uncertainty patterns (Wang et al., 2025). In contrast, SLE aggregates uncertainty more coherently, offering steadier, model-agnostic improvements. Overall, entropy-based curricula provide informative but unstable signals.

3.2.5 Decision Variability

Decision variability, measured as the Variance of Acc@K (VACC). It captures how "blurred" a problem appears to the model by quantifying the variability of correctness across K independent generations for the same problem.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$VACC_{\uparrow}$	56.25	76.56	57.42	31.45	26.17
	$VACC_{\downarrow}$	60.16	75.39	58.40	36.72	24.41
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$VACC_{\uparrow}$	47.27	63.87	54.30	35.55	14.65
	$VACC_{\downarrow}$	55.15	71.48	60.67	36.92	16.80
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$VACC_{\uparrow}$	56.05	76.76	63.28	35.16	19.53
	$VACC_{\downarrow}$	63.09	78.91	60.97	41.02	21.48

Table 5: Accuracy (%) on five mathematical reasoning benchmarks with curricula derived from **decision variability** (VACC).

Table 5 shows that variability-based curricula exhibit clear directional asymmetries. **(i)** Reverse CL, which begins with high-variability (less stable) problems, consistently outperforms Forward CL across most settings—winning 12 of 15 comparisons—with notable gains on easier datasets (*MMQA*: +5.9%, *ASDiv*: +4.9% for *Llama-8B*). **(ii)** Forward CL is competitive only in isolated cases, such as *Gemma-4B* on *MATH* (+1.8%), where uncertainty is already limited. **(iii)** Across models, the performance gap between directions narrows on more difficult benchmarks (*MBench*, *MATH*), suggesting that once reasoning difficulty dominates, the influence of variability ordering diminishes. Overall, these results indicate that initi-

ating learning with high-variability examples (Reverse CL) promotes faster adaptation on simpler reasoning tasks, whereas both strategies converge as task complexity increases.

3.3 Metric-Tiered Curricula

Building on the analysis of how different CL strategies influence model reasoning in the previous section, we further investigate *how* samples with varying metric levels affect learning across tasks of different difficulty. To this end, we conduct **metric tiered training** on *Llama3 8B*. For each metric, the training set is divided into three equal sized tiers, namely **Low**, **Medium**, and **High**, based on metric values. Within each tier, samples are shuffled, and the model is trained on one tier at a time under a fixed data budget to isolate the effect of each difficulty band. Evaluation is performed on four benchmarks of increasing complexity, including *ASDiv*, *GSM8K*, *MBench*, and *MATH*, which are further grouped into *easy* and *hard* categories to analyze tier effects across task difficulty. The five metrics (Section 2.1) represent complementary dimensions: problem difficulty (ACC), model surprisal (TLP), predictive uncertainty (TLE), confidence margin (LG), and decision variability (VACC). A **Baseline** trained on a randomly shuffled one third subset of the data serves as a size matched control.

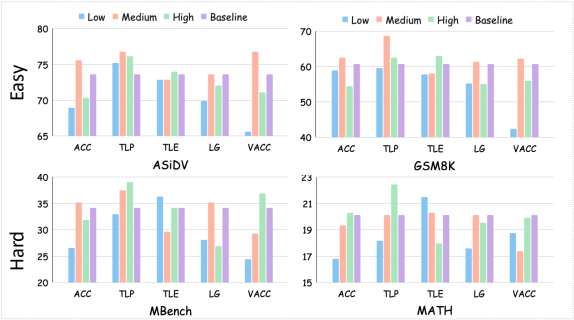

Figure 2: Results of metric-tiered curricula on Llama3-8B for four benchmarks. For each metric, the training set is split into three equal-size tiers (Low, Medium, High), each comprising one third of the data and shuffled; models are trained on one tier at a time. Bars show accuracy (%).

The results in Figure 2 reveal two consistent trends. **(i)** For entropy-based metrics, the most effective tier depends on task difficulty. On easier benchmarks (*ASDiv*, *GSM8K*), training on the **high-entropy** tier yields stronger gains, suggesting that uncertain samples provide valuable learning signals for simpler tasks. In contrast, on harder benchmarks (*MBench*, *MATH*), the **low-entropy** tier performs better, indicating that lower-uncertainty examples become more useful as task complexity increases. **(ii)** For the remaining metrics (ACC, TLP, LG, VACC), improvements are largely concentrated in the **medium** tier. Task difficulty further modulates this effect: easier tasks benefit more from the low-to-medium range, while harder ones favor the medium-to-high range. This pattern suggests that low-surprisal, low-confidence, and low-variability examples support simpler reasoning, whereas challenging reasoning tasks rely more on informative, high-difficulty samples.

3.4 Effect of CL Strategies on Internal Metrics

Beyond external performance, we further examine *how curricula influence internal model states*, including confidence and uncertainty, and *how the direction of the curriculum affects these properties*. We fine tune *Llama3 8B* using three training schedules: Forward CL (FCL), Reverse CL (RCL), and a randomly shuffled baseline (SHUF). Each schedule is constructed using one of five metrics that represent different dimensions of difficulty: TLP, SLP, TLE, SLE, and LG (see Section 2.1). For each metric, the training data are reordered based on its values, and the resulting models are evaluated on five benchmarks *using the same metric that guided training*. From each dataset, we sample 200 problems and report the average metric values.

The results are shown in Figure 3. Across all five internal metrics, curriculum direction induces clear and consistent shifts in model internal states. Both FCL and RCL systematically move the model away from the randomized baseline (SHUF), confirming that curriculum ordering fundamentally reshapes the model's internal confidence and uncertainty distributions. Specifically, for TLP, SLP, and LG, we observe a consistent trend of FCL $<$ RCL, suggesting that training on low-perplexity or low-confidence samples first yields more conservative and stable internal representations. In contrast, SLE shows the opposite behavior, where prioritizing low-certainty data increases model uncertainty, indicating a trade-off between calibration and robustness under different curriculum directions.

3.5 Impact of CL Strategies on Training Dynamics

In previous experiments, we demonstrated that CL strategies substantially influence both the model's

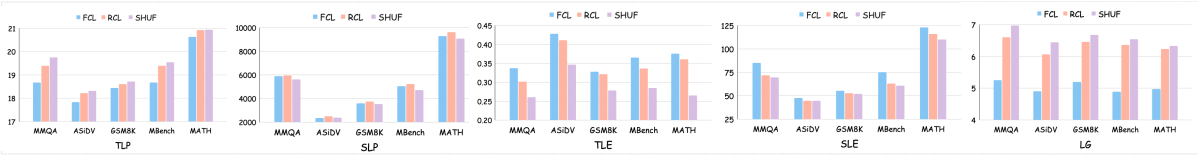

Figure 3: Internal-state metrics across different curriculum strategies. We present TLP, SLP, TLE, SLE, and LG (from left to right, defined in Section 2.1) for *Llama3-8B* trained under Forward CL (FCL), Reverse CL (RCL), and the shuffled baseline (SHUF) across five benchmarks.

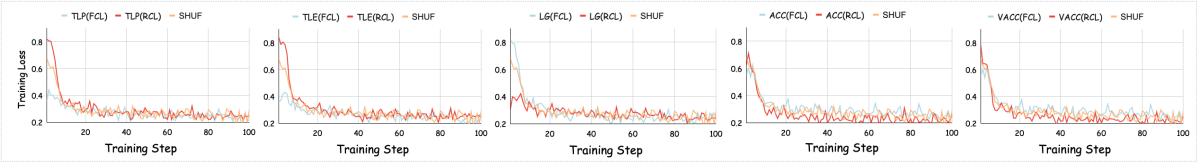

Figure 4: Training loss dynamics over 100 training steps under different curriculum learning strategies. We compare FCL, RCL, and the shuffled baseline (SHUF) across five metric dimensions (from left to right): TLP, TLE, LG, ACC, and VACC, showing how loss evolves over training under different CL strategies.

external performance and internal states. Here, we further investigate *how different CL strategies shape the model's training dynamics*. We consider five representative metrics spanning distinct learning dimensions: *problem complexity (ACC)*, *model surprisal (TLP)*, *model uncertainty (TLE)*, *confidence margin (LG)*, and *decision variability (VACC)* (see Section 2.1). Training loss is tracked under two curriculum schedules: Forward CL (FCL) and Reverse CL (RCL), with a randomly shuffled curriculum (SHUF) serving as the baseline. This setup enables a systematic examination of how curriculum organization shapes model optimization behavior over time.

We present results in Figure 4. These results reveal three distinct mechanisms through which curriculum strategies influence training dynamics. ***(i) Internal-state metrics*** (TLP, TLE, LG) capture difficulty from the model's own predictive behavior and primarily regulate the convergence dynamics during training: starting with less-perplexed, more-certain, or less-confident samples produces higher initial loss but faster convergence, while beginning with high-perplexity or high-confidence samples yields slower convergence yet similar final loss. Such metrics mainly shape the model's internal uncertainty and confidence distributions, indirectly enhancing reasoning ability. ***(ii) Task-aligned metrics*** (ACC, VACC) measure difficulty using task-level outcomes that reflect alignment with the learning objective, having limited effect on convergence speed but substantial impact on the final loss and generalization: training on easier or higher-variability samples leads to lower steady-state loss, indicating improved data fitting. Overall, curriculum learning influences training through two complementary pathways: *internal-state metrics shape model properties such as confidence and uncertainty, while task-aligned metrics affect data fitting and generalization. Together, these mechanisms explain why CL strategies can effectively enhance reasoning performance.*

Conclusion

We propose a multidimensional offline curriculum learning framework to systematically analyze how forward and reverse curricula affect LLMs in mathematical reasoning across three open-source base models: *Gemma 4B*, *Mistral 7B*, and *Llama 8B*. Our results show that well-chosen CL strategies can significantly enhance reasoning performance, with the optimal direction depending jointly on model capability and task complexity. Across different difficulty metrics such as logit gap, entropy, and perplexity, the effectiveness of training tiers varies with task difficulty: harder tasks benefit from high-difficulty and high-uncertainty samples, whereas easier tasks favor low-difficulty and high-confidence ones. Moreover, CL influences not only external performance but also internal calibration and optimization dynamics. *Internal-state metrics* influence model confidence and uncertainty, while *task-aligned metrics* affect data fitting and generalization. Overall, these results highlight the key role of data organization in shaping learning dynamics and provide practical insights for designing effective curricula for LLM post-training.

Limitations

While this work systematically investigates curriculum learning (CL) strategies for LLMs, several limitations remain. **(i)** We focus on *offline* CL, where data ordering is predefined before training. Although this enables controlled comparisons, it does not capture the adaptivity of *online* or self-paced curricula that evolve dynamically during training; we leave this for future work. **(ii)** Our analysis mainly examines solution accuracy and internal model states. While these metrics reveal how CL influences learning dynamics, we have not yet explored the model's full reasoning trajectories or intermediate solution steps, which we plan to investigate in future research.

Ethical Considerations

All experiments presented in this study were conducted using publicly available datasets and models licensed for academic research purposes. To the best of our knowledge, this work does not present any ethical concerns.

References

Shuang Ao, Stefan Rueger, and Advaith Siddharthan. 2023. Confidence-aware calibration and scoring functions for curriculum learning. In *International Conference on Machine Vision*. Available: https://arxiv.org/pdf/2301.12589.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Annual International Conference on Machine Learning*. Available: https://dl.acm.org/doi/pdf/10.1145/1553374.1553380.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023. Skill-it! a data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2307.14430.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 2025. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*. Available: https://arxiv.org/pdf/2505.14970.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. Available: https://arxiv.org/pdf/2110.14168.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Animashree Anandkumar, and 1 others. 2024. Easy2hard-bench: Standardized difficulty labels for profiling llm performance and generalization. *Advances in Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2409.18433.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Available: https://arxiv.org/pdf/2407.21783.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*. Available: https://www.sciencedirect.com/science/article/pii/0010027793900584.

Tianye Fang, Xuanshu Luo, and Martin Werner. 2025. Entropy-driven curriculum for multi-task training in human mobility prediction. *arXiv preprint arXiv:2509.01613*. Available: https://arxiv.org/pdf/2509.01613.

Qi Feng, Yihong Liu, and Hinrich Schütze. 2025. Your pretrained model tells the difficulty itself: A self-adaptive curriculum learning paradigm for natural language understanding. *arXiv preprint arXiv:2507.09758*. Available: https://aclanthology.org/2025.acl-srw.15.pdf.

Tao Feng, Zifeng Wang, and Jimeng Sun. 2023. Citing: Large language models create curriculum for instruction tuning. *arXiv preprint arXiv:2310.02527*. Available: https://arxiv.org/pdf/2310.02527.

Zifa Gan, Hongfei Xu, and Hongying Zan. 2021. Self-supervised curriculum learning for spelling error correction. In *Conference on Empirical Methods in Natural Language Processing*. Available: https://aclanthology.org/2021.emnlp-main.281.pdf.

Hasan Abed Al Kader Hammoud, Kumail Alhamoud, Abed Hammoud, Elie Bou-Zeid, Marzyeh Ghassemi, and Bernard Ghanem. 2025. Train long, think short: Curriculum learning for efficient reasoning. *arXiv preprint arXiv:2508.08940*. Available: https://arxiv.org/pdf/2508.08940.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*. Available: https://arxiv.org/pdf/2103.03874.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. Available: https://arxiv.org/pdf/2106.09685v1/1000.

Yufei Huang and Deyi Xiong. 2024. It2acl learning easy-to-hard instructions via 2-phase automated curriculum learning for large language models. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Available: https://aclanthology.org/2024.lrec-main.822.pdf.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active curriculum learning. In *First Workshop on Interactive Learning for Natural Language Processing in ACL*. Available: https://aclanthology.org/2021.internlp-1.6.pdf.

Albert Jiang, Timothée Balesni, Farima Hatami, Thomas Joseph, Thibaut Lavril, Gaël Leroux, Antoine Lheritier, Raphaël Lostis, Walid Moufarek, Alexandre Rochette, Lucile Saulnier, Roman Tang, Pablo Villegas, Pierre Wajsburt, and Guillaume Lample. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*. Available: https://arxiv.org/pdf/2310.06825.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. 2015. Self-paced curriculum learning. In *AAAI Conference on Artificial Intelligence*. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9608/9467.

Jeesu Jung and Sangkeun Jung. 2025. Reasoning steps as curriculum: Using depth of thought as a difficulty signal for tuning llms. *arXiv preprint arXiv:2508.18279*. Available: https://arxiv.org/pdf/2508.18279.

Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint arXiv:2405.07490*. Available: https://arxiv.org/pdf/2405.07490.

M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*. Available: https://ai.stanford.edu/~bpacker/selfPacedLVM.pdf.

Bruce W. Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Available: https://aclanthology.org/2024.findings-naacl.82.pdf.

Mengdi Li, Jiaye Lin, Xufeng Zhao, Wenhao Lu, Peilin Zhao, Stefan Wermter, and Di Wang. 2025. Curriculum-rlaif: Curriculum alignment with reinforcement learning from ai feedback. *arXiv preprint arXiv:2505.20075*. Available: https://arxiv.org/pdf/2505.20075.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *International Conference on Learning Representations*. Available: https://arxiv.org/pdf/2305.20050.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*. Available: https://web.archive.org/web/20170809174544id_/http://lingviko.net/jcs.pdf.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024a. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics ACL 2024*. Available: https://arxiv.org/pdf/2405.12209.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. 2024b. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*. Available: https://arxiv.org/pdf/2402.10738.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Available: https://aclanthology.org/2022.naacl-main.72.pdf.

Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2024. A symbolic framework for evaluating mathematical reasoning and generalisation with transformers. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Available: https://aclanthology.org/2024.naacl-long.84.pdf.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In *Annual Meeting of the Association for Computational Linguistics*. Available: https://aclanthology.org/2020.acl-main.92.pdf.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale, and Shafiq Joty. 2022. Data selection curriculum for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Available: https://aclanthology.org/2022.findings-emnlp.113.pdf.

Marwa Naïr, Kamel Yamani, Lynda Lhadj, and Riyadh Baghdadi. 2024. Curriculum learning

for small code language models. In *Student Research Workshop at Annual Meeting of the Association for Computational Linguistics*. Available: https://aclanthology.org/2024.acl-srw.44.pdf.

Wei Pang, Chuan Zhou, Xiao-Hua Zhou, and Xiaojie Wang. 2024. Phased instruction fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Available: https://aclanthology.org/2024.findings-acl.341.pdf.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and 1 others. 2025. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*. Available: https://arxiv.org/pdf/2506.06632.

Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. Enhancing alignment using curriculum learning & ranked preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Available: https://aclanthology.org/2024.findings-emnlp.754.pdf.

Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. 2023. Token-level self-evolution training for sequence-to-sequence learning. In *Annual Meeting of the Association for Computational Linguistics*. Available: https://aclanthology.org/2023.acl-short.73.pdf.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*. Available: https://arxiv.org/pdf/2202.01344.

Hadi Pouransari, Chun-Liang Li, Jen-Hao Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. 2024. Dataset decomposition: Faster llm training with variable sequence length curriculum. In *Advances in Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2405.13226.

Jun Rao, Xuebo Liu, Lian Lian, Shengjun Cheng, Yunjie Liao, and Min Zhang. 2024. Commonit: Commonality-aware instruction tuning for large language models via data partitions. In *Conference on Empirical Methods in Natural Language Processing*. Available: https://aclanthology.org/2024.emnlp-main.561.pdf.

Linxin Song, Jieyu Zhang, Tianxiang Yang, and Masayuki Goto. 2022. Adaptive ranking-based sample selection for weakly supervised class-imbalanced text classification. *arXiv preprint arXiv:2210.03092*. Available: https://arxiv.org/pdf/2210.03092.

Kv Aditya Srivatsa and Ekaterina Kochmar. 2024. What makes math word problems challenging for LLMs? In *Findings of the Association for Computational Linguistics: NAACL 2024*. Available: https://aclanthology.org/2024.findings-naacl.72.pdf.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*. Available: https://arxiv.org/pdf/2503.19786.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. In *Advances in Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2407.13690.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*. Available: https://arxiv.org/pdf/2506.01939.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*. Available: https://arxiv.org/pdf/2203.11171.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, and 1 others. 2024. Training large language models for reasoning through reverse curriculum reinforcement learning. *arXiv preprint arXiv:2402.05808*. Available: https://arxiv.org/pdf/2402.05808.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*. Available: https://arxiv.org/pdf/2309.12284.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2024. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *International Conference on Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2203.14465.

Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, Hongxin Wei, and 1 others. 2024. On the noise robustness of in-context learning for text generation. In *Advances in Neural Information Processing Systems*. Available: https://arxiv.org/pdf/2405.17264.

Ruiqi Zhang, Daman Arora, Song Mei, and Andrea Zanette. 2025a. Speed-rl: Faster training of reasoning models via online curriculum learning. *arXiv preprint arXiv:2506.09016*. Available: https://arxiv.org/pdf/2506.09016.

Xuemiao Zhang, Xu Liangyu, Feiyu Duan, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025b. Preference curriculum: LLMs should always be pretrained on their preferred data. In *Findings of the Association for Computational Linguistics: ACL 2025*. Available: https://aclanthology.org/2025.findings-acl.1091.pdf.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. 2025c. Beyond random sampling: Efficient language model pretraining via curriculum learning. *arXiv preprint arXiv:2506.11300*. Available: https://arxiv.org/pdf/2506.11300.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*. Available: https://arxiv.org/pdf/2205.10625.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*. Available: https://aclanthology.org/2020.acl-main.620.pdf.

Contents of Appendix

A Related Work **13**
- A.1 Curriculum Learning for Large Language Models 13
- A.2 Curriculum Metrics for Mathematical Reasoning 13

B Additional Curriculum Strategies **14**

C Experimental Details **14**
- C.1 Hardware and Software Environment 14
- C.2 Training Settings 14
- C.3 Parameter Setting 15

D Additional Results **15**
- D.1 Problem Difficulty 15
- D.2 Model Surprisal 15
- D.3 Confidence Margin 15
- D.4 Predictive Uncertainty 16
- D.5 Decision Variability 16
- D.6 Result Analysis of FCL/RCL and GFC/GRC 16

A Related Work

A.1 Curriculum Learning for Large Language Models

Curriculum learning (CL), proposed in machine learning and cognitive science—trains models on examples arranged from easy to hard to stabilize optimization and improve efficiency (Bengio et al., 2009; Elman, 1993; Kumar et al., 2010). This principle has recently been scaled to large language models (LLMs). Several works incorporate CL directly into LLM training: for instance, Zelikman et al. construct an implicit easy-to-hard curriculum via self-training on verified chains of thought, boosting mathematical and symbolic reasoning; Huang and Xiong design a two-phase automated curriculum across tasks and within-task instructions, yielding stronger performance and better zero-shot generalization; and other researchers develop CL schemes driven by diverse difficulty metrics (Chen et al., 2023; Rao et al., 2024; Xi et al., 2024; Parashar et al., 2025). Complementing training-time CL, Zhou et al. apply CL at test time by decomposing complex problems into ranked subproblems and solving them from easy to hard, markedly enhancing reasoning on challenging benchmarks. A parallel line of work applies CL before or alongside pretraining by reordering or sampling data to improve efficiency and downstream reasoning (Zhang et al., 2025b; Kim and Lee, 2024; Feng et al., 2023). For example, Zhang et al. cast preference-aware data selection as a pretraining-time curriculum, prioritizing high-utility examples and phasing in lower-utility data, which improves sample efficiency and downstream performance relative to uniform or domain-agnostic mixtures.

A.2 Curriculum Metrics for Mathematical Reasoning

Mathematical reasoning tasks naturally exhibit graded difficulty, making them well-suited for curriculum learning approaches that progress from easier to harder problems. Existing work on curriculum metrics for this domain can be broadly divided into two categories: problem-side metrics and model-conditional metrics. Problem-side metrics are derived from an instance's structure and meaning, not from model behavior, and may be computed analytically or via annotation. For example, Jung and Jung order instances by reasoning depth, using shallow-to-deep pacing to guide

training, while Hammoud et al. employ a length-based curriculum that begins with generous chain-of-thought budgets and gradually anneals them to encourage concise reasoning. Other studies define difficulty through symbolic complexity (Tong et al., 2024), linguistic complexity (Srivatsa and Kochmar, 2024), or verifiability signals such as self-consistency and stepwise correctness (Wang et al., 2022; Lightman et al., 2023), leveraging these properties to schedule learning. Model-conditional metrics instead adapt curricula based on model behavior. Perplexity has been used as a difficulty signal to sequence training data and in-context examples, yielding faster convergence and improved in-context learning performance (Zhou et al., 2020; Liu et al., 2024b). Entropy-based uncertainty has also been proposed as a pacing signal, where prioritizing high-information samples leads to more stable training and stronger downstream results, including higher BLEU scores (Zhou et al., 2020; Fang et al., 2025). Beyond these, logit gap signals (Jafarpour et al., 2021), calibrated confidence estimates (Ao et al., 2023), and adaptive verification metrics (Wang et al., 2022; Song et al., 2022) have been explored to construct easy-to-hard schedules that accelerate training efficiency and boost generalization.

Despite growing interest in curriculum learning, there is limited work that systematically compares curriculum metrics and examines how CL strategies relate to task type and model reasoning capabilities. To address this gap, we introduce a multidimensional evaluation benchmark for head-to-head comparisons of CL strategies. To control for confounds, we focus on offline curricula and conduct experiments on base language models (i.e., non–instruction-tuned language models).

B Additional Curriculum Strategies

In addition to the Forward CL and Reverse CL settings introduced in Section 2.2, we explore two group-based curriculum variants that aim to reduce the sensitivity of training to fine-grained ranking noise.

Group Forward Curriculum (GFC). The dataset is partitioned into three tiers—*Low*, *Medium*, and *High*—based on metric values using thresholding or quantile splits. Samples within each tier are randomly shuffled to remove intra-group ordering effects. Training proceeds from easier to harder groups (*Low* $\rightarrow$ *Medium* $\rightarrow$ *High*), preserving the forward progression of difficulty.

Group Reverse Curriculum (GRC). Using the same grouping procedure, training is instead conducted in reverse order (*High* $\rightarrow$ *Medium* $\rightarrow$ *Low*). This variant emphasizes difficult samples early while maintaining stability against local ranking variance.

C Experimental Details

C.1 Hardware and Software Environment

All experiments were performed on a node equipped with eight NVIDIA RTX A6000 GPUs (48 GB each), running CUDA 12.8 with driver version 570.86.16. The software environment was set up with Python 3.10, PyTorch 2.2.1, and the HuggingFace Transformers library to ensure reproducibility.

C.2 Training Settings

All fine-tuning experiments were conducted using parameter-efficient LoRA training (Hu et al., 2022) on Llama3-8B and Gemma-3-4B, and Mistral-7B backbones. We adopted the PEFT framework with a low-rank adaptation configuration of rank 64, scaling factor 128, and dropout 0.05. Each model was fine-tuned under mixed precision (`bfloat16`) using the HuggingFace `Transformers` and `Accelerate` libraries with `DeepSpeed ZeRO Stage 2` for distributed optimization. To improve memory efficiency, we enabled `FlashAttention-2` and non-reentrant gradient checkpointing throughout training.

We selected the final checkpoint based on the lowest validation loss, which was evaluated every 25 training steps.

Table 6: Key hyperparameters used in all fine-tuning experiments.

Category	Setting / Value
Dataset	Training samples: 20,000 Validation samples: 500 Test samples: 500
Optimization	Optimizer: AdamW Weight decay: 0.01 Learning rate: 5×10^{-5} Scheduler: Cosine (10% warmup)
Training configuration	Epochs: 3 Batch size / device: 4 Gradient accumulation: 4 Max gradient norm: 1.0

C.3 Parameter Setting

The detailed hyperparameter configurations for fine-tuning experiments are summarized in Table 6.

D Additional Results

In this section, we further evaluate the two group based curriculum variants, **Group Forward Curriculum (GFC)** and **Group Reverse Curriculum (GRC)**, across all five metric dimensions introduced in Section 2.1: *Problem Difficulty*, *Model Surprisal*, *Confidence Margin*, *Predictive Uncertainty*, and *Decision Variability*. For each metric, we construct both GFC and GRC schedules and assess their impact on model reasoning performance under identical training settings. The following tables summarize the quantitative results for each metric dimension, providing a detailed comparison between the forward and reverse group based curricula.

D.1 Problem Difficulty

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	RS↑	56.05	76.56	55.86	33.98	24.22
	RS↓	52.73	75.98	53.32	33.20	22.27
	SC↑	56.64	76.95	55.27	30.08	26.76
	SC↓	53.52	73.05	52.15	36.33	22.27
	CD↑	52.34	77.15	53.52	35.74	23.44
	CD↓	52.73	75.98	50.98	35.94	22.46
	ACC↑	52.34	77.15	55.47	37.50	24.61
	ACC↓	56.04	77.93	57.81	35.74	22.66
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	RS↑	54.49	71.68	60.16	35.94	16.41
	RS↓	59.57	72.46	63.28	36.52	16.99
	SC↑	53.71	72.27	60.74	36.72	16.99
	SC↓	53.71	74.80	61.91	37.89	18.16
	CD↑	54.49	70.90	57.23	39.06	16.21
	CD↓	53.91	72.66	62.30	38.48	18.55
	ACC↑	55.47	74.02	62.50	37.70	15.04
	ACC↓	51.95	73.24	60.74	38.28	18.16
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	RS↑	60.35	75.00	64.26	35.35	19.73
	RS↓	61.52	77.54	69.92	38.67	20.70
	SC↑	60.16	72.07	67.97	35.74	19.53
	SC↓	51.95	70.51	58.40	36.52	22.85
	CD↑	59.18	78.32	66.60	38.28	22.27
	CD↓	56.84	78.52	64.26	38.87	21.29
	ACC↑	54.10	74.80	61.52	35.35	18.95
	ACC↓	59.98	77.93	69.73	41.80	20.90

Table 7: Accuracy (%) on five mathematical reasoning benchmarks for curricula constructed from *problem difficulty metrics (RS, SC, CD, ACC)*. The **baseline** uses randomly shuffled data. ↑ and ↓ indicate **group-level ascending (GFC)** and **group-level descending (GRC)** curricula, respectively. See Sec. 2.1 for metric definitions.

We consider four difficulty metrics for problems: *Reasoning Steps (RS)*, *Symbol Complexity (SC)*, *Comprehension Difficulty (CD)*, and *Acc@K (ACC)*, as described in Section 2.1.1. The results of Forward CL (FCL) and Reverse CL (RCL) based on these metrics are shown in Table 2, while their group-based counterparts, Group Forward Curriculum (GFC) and Group Reverse Curriculum (GRC), are reported in Table 7.

D.2 Model Surprisal

Model surprisal is evaluated through *Token-Level Perplexity (TLP)* and *Sequence-Level Perplexity (SLP)*, with formal definitions provided in Section 2.1.2. FCL and RCL results appear in Table 2, with their group-based versions (GFC and GRC) in Table 8.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	SLP↑	56.64	76.95	53.12	36.91	23.44
	SLP↓	51.17	74.80	52.73	38.48	21.48
	TLP↑	54.30	75.00	58.98	32.62	21.68
	TLP↓	50.20	76.37	55.47	31.25	22.07
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	SLP↑	53.71	64.06	52.54	38.87	15.23
	SLP↓	54.49	75.58	63.67	38.28	15.82
	TLP↑	54.30	74.80	58.20	39.06	16.21
	TLP↓	54.49	75.78	63.67	33.59	15.43
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	SLP↑	62.50	75.78	68.16	37.89	20.31
	SLP↓	55.47	75.78	60.74	35.94	19.73
	TLP↑	56.64	74.41	59.18	37.50	19.34
	TLP↓	57.42	75.59	60.94	38.28	20.90

Table 8: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed from **model surprisal metrics** (SLP, TLP). **Baseline**, ↑, and ↓ follow the same notation as in Table 7.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	LG↑	52.93	73.05	54.40	34.48	21.68
	LG↓	56.05	75.20	53.32	34.96	22.26
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	LG↑	53.52	74.61	62.11	41.60	18.36
	LG↓	53.91	73.63	60.16	41.02	18.95
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	LG↑	61.52	75.98	64.84	39.45	21.09
	LG↓	58.40	73.05	66.41	38.67	21.47

Table 9: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed from **confidence margin** (LG). **Baseline**, ↑, and ↓ follow the same notation as in Table 7.

D.3 Confidence Margin

Confidence margin is measured using the *Logit Gap (LG)* metric, formally defined in Section 2.1.2.

Results for FCL and RCL are presented in Table 3, and the corresponding group-based variants, GFC and GRC, are shown in Table 9.

D.4 Predictive Uncertainty

Predictive uncertainty is measured using the *Token-Level Entropy (TLE)* and *Sequential-Level Entropy (SLE)*, formally defined in Section 2.1.2. Results for FCL and RCL are presented in Table 4, and the corresponding group-based variants, GFC and GRC, are shown in Table 10.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$SLE_{\uparrow}$	54.49	76.76	57.81	33.01	26.37
	$SLE_{\downarrow}$	55.27	76.37	53.21	26.56	25.98
	$TLE_{\uparrow}$	52.54	74.02	52.34	37.30	22.07
	$TLE_{\downarrow}$	52.73	74.80	51.95	37.89	22.85
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$SLE_{\uparrow}$	53.91	68.55	61.91	38.67	17.77
	$SLE_{\downarrow}$	54.30	75.39	64.65	39.65	13.67
	$TLE_{\uparrow}$	51.56	71.48	56.84	39.45	15.23
	$TLE_{\downarrow}$	49.61	72.85	59.38	39.26	14.45
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$SLE_{\uparrow}$	58.20	78.32	66.02	38.09	20.12
	$SLE_{\downarrow}$	53.71	71.68	61.13	38.67	19.34
	$TLE_{\uparrow}$	55.66	77.54	63.48	39.06	19.53
	$TLE_{\downarrow}$	54.69	74.22	62.50	39.45	22.66

Table 10: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed from **predictive uncertainty metrics** (SLE, TLE). **Baseline**, ↑, and ↓ follow the same notation as in Table 7.

D.5 Decision Variability

Decision variability is evaluated using the $Acc@K$ metric (VACC), as defined in Section 2.1.2. Results for FCL and RCL are reported in Table 5, while their group-based variants, GFC and GRC, are presented in Table 11.

D.6 Result Analysis of FCL/RCL and GFC/GRC

We compare fine-grained curricula (FCL/RCL) with their group-based variants (GFC/GRC). Overall, GFC and GRC yield more stable and consistent reasoning performance across benchmarks and model scales.

(a) Stability and direction. GFC/GRC produce smoother results and fewer fluctuations between curriculum directions. Grouping samples before scheduling reduces local ranking noise, improving directional consistency—e.g., GFC benefits the *RS* metric on *Gemma-4B*, while GRC enhances *ACC* on *Llama-8B*.

Model	Strategy	MMQA	ASDiv	GSM8K	MBench	MATH
Gemma-4B	Baseline	51.17	73.44	53.52	35.35	21.68
	$VACC_{\uparrow}$	48.44	75.39	49.41	33.20	23.63
	$VACC_{\downarrow}$	56.25	71.88	57.03	25.78	18.75
Mistral-7B	Baseline	54.30	68.55	60.35	36.72	18.75
	$VACC_{\uparrow}$	54.49	68.55	60.35	40.43	15.82
	$VACC_{\downarrow}$	52.93	73.44	61.52	43.16	16.99
Llama-8B	Baseline	59.57	74.02	62.50	38.09	18.55
	$VACC_{\uparrow}$	49.61	76.37	58.01	32.03	49.41
	$VACC_{\downarrow}$	58.20	78.32	65.23	40.43	58.20

Table 11: Accuracy (%) on five mathematical reasoning benchmarks under curricula constructed from **decision variability** (VACC). **Baseline**, ↑, and ↓ follow the same notation as in Table 7.

(b) Robustness of effects. While FCL/RCL occasionally reach higher peaks, their performance varies more with direction choice. GFC/GRC achieve more balanced gains, suggesting that grouped scheduling stabilizes optimization and mitigates overfitting to specific ranking orders.

(c) Model-scale sensitivity. Smaller models (*Gemma-4B*, *Mistral-7B*) show larger differences between FCL/RCL and GFC/GRC, whereas larger ones (*Llama-8B*) are less sensitive to ranking granularity. This indicates that coarser grouping becomes increasingly effective with model size.

Summary. Group-based curricula improve the stability, robustness, and interpretability of curriculum learning by suppressing noisy ranking signals and yielding more consistent training behavior.