LEARNING MORE BY TRAINING LESS: UNCERTAINTY-GUIDED PROGRESSIVE CONTRACTION FOR EFFICIENT REASONING POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models have achieved remarkable advances in complex reasoning. However, scaling reasoning post-training with ever-larger corpora can substantially increase training cost without proportional performance gains. In this study, we therefore ask: how far can reasoning supervision be compressed without sacrificing capability? To this end, we propose an uncertainty-guided progressive contraction framework that concentrates supervision and optimization on the reasoning trajectories most valuable for learning. Specifically, rather than treating a teacher-generated chain-of-thought (CoT) as a single realized path, we construct *uncertainty-profiled reasoning supervision* by building Monte Carlo reasoning trees around intermediate reasoning states to expose process-level uncertainty. This uncertainty first contracts the original corpus into a compact teacher-informed pool. A second, learner-aware contraction then performs online routing during SFT, concentrating optimization according to the student's evolving learning state. Together, the two stages progressively contract both supervision and optimization, substantially reducing the student training budget. Experiments across diverse reasoning benchmarks and LLM families show that our method uses as little as 0.22% of the optimization budget of full-data SFT, over 450$\times$ less training exposure, while maintaining or improving reasoning performance. Notably, our framework generalizes well across heterogeneous student families, black-box teachers, and weak-to-strong transfer. This indicates that it is not only a way to accelerate training but can also enhance reasoning capability.

1 INTRODUCTION

Large language models (LLMs) have made remarkable progress in complex reasoning, driven in part by increasingly capable pretrained models and post-training on ever-larger reasoning corpora (Wei et al., 2023; Guo et al., 2025). A prevailing line of work is to scale supervision with more reasoning trajectories to train on (Zelikman et al., 2022b; GLM-5-Team et al., 2026). However, this scaling incurs substantial training cost without proportional performance gains, motivating a shift from simply increasing data volume toward identifying reasoning trajectories with greater learning value He et al. (2026); Hou et al. (2026). Existing efforts broadly explore two complementary directions: constructing more effective reasoning supervision through stronger teacher generation and trajectory augmentation (Guo et al., 2025; Li et al., 2023), and improving the utilization of fixed corpora through filtering or routing based on correctness (Slinko et al., 2026; Guha et al., 2025), difficulty (Tian et al., 2025; Huang et al., 2026), or quality (Li et al., 2026; Just et al., 2026). While achieving promising results, neither direction is sufficient for extreme optimization efficiency. As a result, achieving such efficiency requires coupling supervision construction with downstream allocation: first exposing where learning value resides, and then progressively concentrating optimization on it.

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

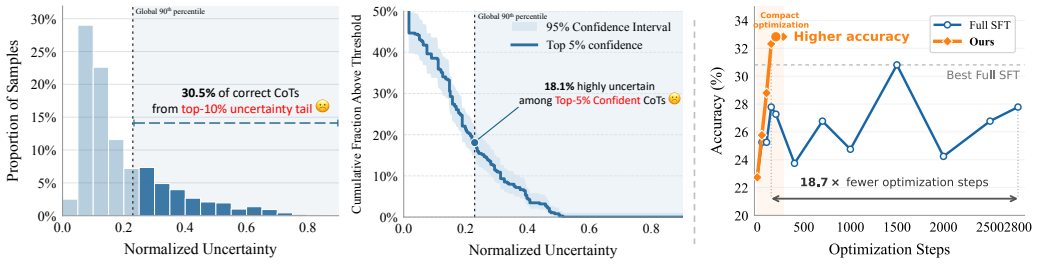


Figure 1: **Correct and confident CoTs can still conceal fragile reasoning.** *Left*: Even among the top-5% teacher-confident CoTs, **18.1%** fall within the global top-10% uncertainty tail. *Mid*: Likewise, **30.5%** of correct CoTs fall within the same uncertainty tail. *Right*: Motivated by this, we propose an uncertainty-guided progressive contraction framework. By progressively narrowing supervision from uncertainty-informative teacher reasoning to learner-relevant trajectories, we achieve competitive reasoning performance with 18.7× fewer post-training steps.

To expose learning-critical information, we exploit the internal structure of reasoning. In fact, reasoning is inherently non-uniform: a solution often hinges on a small number of intermediate states where plausible continuations diverge toward different outcomes, which we term *reasoning forks*. Such forks naturally manifest as process-level uncertainty, as alternative continuations from the same reasoning state may lead to both successful and unsuccessful solutions. As shown in Fig. 1, even highly confident or correct Chain of Thoughts (CoTs) can conceal substantial process-level uncertainty, suggesting that terminal confidence and correctness may mask fragile intermediate reasoning (Zelikman et al., 2022a; Zhao, 2026; Mao et al., 2026; Ye et al., 2025b). Consistent with prior work showing that uncertainty reveals local failures and reasoning boundary (Kadavath et al., 2022; Zhang et al., 2026c; Mei et al., 2025), we therefore view process uncertainty as a behavioral signature of the reasoning capability boundary.

Motivated by this, we propose an uncertainty-guided progressive contraction framework for efficient reasoning post-training. We first construct uncertainty-profiled reasoning supervision trajectories by building Monte Carlo Reasoning Trees (MCRTs) around intermediate states and estimating correct-answer reachability and branch-outcome uncertainty from sampled continuations. Aggregated process uncertainty contracts the original corpus into a compact teacher-informed candidate pool along with uncertainty profiles. During student SFT, however, the uncertainty-profiled CoTs remain learner-agnostic and do not reflect the student's evolving training state. Thus, we introduce a second learner-aware contraction, which narrows the candidate set by routing optimization according to the student's evolving learning state. Together, the two stages progressively concentrate supervision from teacher-informative to learner-relevant trajectories, reducing both data usage and, more importantly, the optimization required to learn from it. As a result, our method uses as little as 0.22% of the optimization budget of full-data SFT, **over 450× less** training exposure, while maintaining or improving reasoning performance.

Experimental results across diverse reasoning benchmarks and LLM architectures demonstrate that our method substantially accelerates SFT training while maintaining or improving reasoning performance. Our framework presents strong cross-student transferability: reasoning supervision distilled from the same Qwen3-8B (Yang et al., 2025a) teacher generalizes effectively across heterogeneous student families, including Qwen3.5-4B/9B (Qwen Team, 2026), Gemma-4 E4B (Team et al., 2026), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct (Jiang et al., 2023). Importantly, the gains extend beyond efficiency: our 9B student achieves competitive reasoning performance with substantially larger reasoning models ranging from 30B to 671B parameters. Notably, because uncertainty is characterized purely through observable reasoning behavior, the same framework applies to **API-only black-box teachers** (e.g., DeepSeek-V4-Pro (DeepSeek-AI et al., 2026) and Kimi-K2.6 (Moonshot AI, 2026)) and teachers substantially **smaller than their students**. For instance, a Qwen3-1.7B teacher can improve a much stronger Qwen3.5-14B student, demonstrating effective reasoning transfer.

Our main contributions are as follows: **(1)** We introduce an uncertainty-guided progressive contraction framework for data-efficient reasoning post-training, jointly integrating supervision construction and downstream training allocation to progressively concentrate learning on high-value reasoning trajectories. **(2)** We introduce uncertainty-profiled reasoning supervision, modeling MCRTs to expose process-level reasoning forks from observable model behavior without requiring model-internal access. **(3)** We develop learner-aware routing that further contracts the active training set during SFT according to the student's evolving learning state, enabling enhanced training acceleration. **(4)** Across diverse benchmarks, model families, and teacher regimes, our method reduces reasoning optimization to as little as $< 0.74\%$, while maintaining or improving reasoning performance, establishing a practical pathway toward data-efficient reasoning post-training for future research.

2 The Proposed Method

2.1 Problem Formulation and Overview

We consider reasoning distillation from a teacher model T to a student model S_θ over a training set $\mathcal{D} = (x_i, a_i)_{i=1}^N$. For each problem x_i, the teacher generates a reasoning trajectory $\tau_i = (si, 1, \ldots, s_{i,m_i}, \hat{a}_i)$. Rather than treating all verified CoTs as equally valuable supervision, we observe that their learning value depends on two factors: whether the trajectory exposes informative structures in the teacher's reasoning space and whether such information remains unresolved by the current learner.

Based on this observation, we introduce *progressive supervision contraction*, which progressively narrows supervision through two stages: (1) teacher-side uncertainty contraction and (2) learner-aware optimization contraction. *In the first stage*, we treat each CoT as a backbone trajectory and construct an MCRT around its intermediate states to characterize process-level uncertainty. Each trajectory is assigned a teacher-side uncertainty score q_i^T, and we retain an uncertainty-rich candidate pool $\mathcal{P}_\rho$ with ratio ρ: $\mathcal{P}_\rho = \mathrm{Top}_\rho(\mathcal{C}; q_i^T)$. Since this stage relies only on observable teacher behavior, $\mathcal{P}_\rho$ is student-agnostic and can be reused across different learners without repeating uncertainty profiling. *In the second stage*, given the uncertainty-informed pool, we further allocate optimization according to the current learner state. Specifically, we route training toward the M trajectories with the largest normalized response loss: $\mathcal{H}_M = \arg\max_{\mathcal{H}\subseteq\mathcal{P}_\rho, |\mathcal{H}|=M} \sum_{i\in\mathcal{H}} \ell_i$, where ℓ_i is obtained from standard forward passes without additional optimization or auxiliary computation. Subsequent training is performed exclusively on $\mathcal{H}_M$. Together, the two stages reduce both the amount of data exposure and the optimization required for learning. In practice, this reduces the student-side training budget to less than 1% of vanilla full-set SFT while maintaining downstream reasoning performance.

2.2 Uncertainty-Guided Reasoning and Profiling

Uncertainty-Revealing Reasoning Elicitation. For each problem x_i, we prompt the teacher to generate a structured reasoning trajectory $\tau_i = (s_{i,1}, \ldots, s_{i,m_i}, \hat{a}_i)$, where each $s_{i,t}$ corresponds to a semantically coherent reasoning step. The trajectory induces prefix states $h_{i,t} = (x_i, s_{i,1:t}), t = 1, \ldots, m_i$, which serve as anchors for subsequent behavioral probing. We treat τ_i as a backbone trajectory through the teacher's reasoning space, and characterize the local reasoning stability around each $h_{i,t}$ through sampled continuations.

Monte Carlo Reasoning-Tree Profiling. Given a backbone trajectory τ_i, we characterize the local stability of each intermediate state $h_{i,t}$ by sampling K independent prefix-conditioned continuations, $R_{i,t}^{(k)} \sim \pi_{\mathcal{T}}(\cdot \mid h_{i,t})$. Let $\hat{a}_{i,t}^{(k)}$ denote the resulting final answer and $Z_{i,t}^{(k)} = \mathbb{I}[\hat{a}_{i,t}^{(k)} = a_i]$ its correctness. We model $Z_{i,t}^{(k)} \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p_{i,t})$, where $p_{i,t}$ denotes the correct-answer reachability from $h_{i,t}$, and estimate it by $\hat{p}_{i,t} = K^{-1} \sum_{k=1}^K Z_{i,t}^{(k)}$. Finite-sample concentration is provided in Appendix B.

We then quantify local process uncertainty through the normalized Bernoulli variance

$$U_{i,t} = 4p_{i,t}(1 - p_{i,t}) \in [0, 1], \tag{1}$$

with the plug-in estimator

$$\widehat{U}_{i,t} = 4\hat{p}_{i,t}(1 - \hat{p}_{i,t}). \tag{2}$$

This quantity directly measures branch-outcome instability. For two independent continuations rooted at the same state,

$$U_{i,t} = 2\Pr\left(Z_{i,t}^{(1)} \neq Z_{i,t}^{(2)} \mid h_{i,t}\right). \tag{3}$$

Hence, high $U_{i,t}$ identifies states whose plausible continuations diverge in eventual correctness, rather than states that are simply difficult or incorrect.

Notably, the entire profiling procedure is behavior-based: it operates only on textual prefixes, sampled continuations, and verifiable outcomes, without requiring logits, hidden states, gradients, or model parameters. This makes the same uncertainty estimator directly applicable to heterogeneous and API-only black-box teachers, provided that prefix-conditioned sampling and outcome verification are available.

Finally, we aggregate state-wise uncertainty into the reasoning uncertainty profile $\mathcal{Z}_i = \left\{(\hat{p}_{i,t}, \widehat{U}_{i,t})\right\}_{t=1}^{m_i}$ using the mean process uncertainty

$$\bar{U}_i = \frac{1}{m_i} \sum_{t=1}^{m_i} \widehat{U}_{i,t}, \tag{4}$$

which normalizes for trajectory length. A larger $\bar{U}_i$ indicates that the realized trajectory more frequently passes through states whose alternative continuations disagree in eventual correctness.

Uncertainty-Guided Supervision Contraction. We use the trajectory-level uncertainty $\bar{U}_i$ as the teacher-side informativeness score, i.e., $q_i^T = \bar{U}_i$. Given a candidate-pool ratio ρ, we perform the first supervision contraction by retaining the verified trajectories with the largest process uncertainty:

$$\mathcal{P}_\rho = \arg \max_{\substack{\mathcal{S} \subseteq \mathcal{C}, \\ |\mathcal{S}| = \lfloor \rho |\mathcal{C}| \rfloor}} \sum_{i \in \mathcal{S}} \bar{U}_i. \tag{5}$$

The resulting pool $\mathcal{P}_\rho$ concentrates supervision on verified trajectories that more frequently traverse states with unstable continuation outcomes, forming a compact teacher-informed candidate space for subsequent learner-aware optimization.

2.3 Learner-Aware Data Routing

The teacher-informed pool $\mathcal{P}_\rho$ reduces the amount of supervision presented to the student before training. We next perform the second contraction during SFT, routing optimization toward a small subset of trajectories according to the learner's evolving training state. Concretely, we first perform a single probing epoch over $\mathcal{P}_\rho$. For each trajectory $i \in \mathcal{P}_\rho$, let $\mathcal{R}_i$ denote the supervised response-token positions. We derive the learner-side routing score q_i^S using the length-normalized response loss

$$q_i^S = \ell_i = -\frac{1}{|\mathcal{R}_i|} \sum_{j \in \mathcal{R}_i} \log p_{\theta^{(w)}}\left(\tau_{i,j} \mid x_i, \tau_{i,<j}\right). \tag{6}$$

A larger ℓ_i indicates that the corresponding teacher trajectory remains less well captured by the current student and therefore receives higher routing priority.

Table 1: Performance across competition mathematics, general mathematics, and science benchmarks using Qwen3.5-4/9B. Training budget denotes cumulative student-side sample exposure relative to full-set SFT (100%). Overhead reports wall-clock GPU hours on a 2-H200-GPU server.

	Efficiency		Competition Math						General Math		Science
Method	**Budget**	**Overhead**	**AIME24**	**AIME25**	**AIME26**	**AMC**	**HMMT25**	**HMMT26**	**MATH-500**	**GSM8K**	**GPQA-D**
Qwen3.5-4B Student											
Full SFT	100%	8.47	73.3	56.7	66.7	81.4	40.0	39.4	97.8	92.3	75.8
Ours ($\rho = 30\%$)	1.91%	0.07	86.7	83.3	86.7	96.3	76.7	69.7	98.4	96.4	77.8
Ours ($\rho = 10\%$)	0.74%	0.03	86.7	83.3	83.3	94.8	76.7	78.8	98.0	96.2	77.3
Ours ($\rho = 5\%$)	0.45%	0.02	86.7	73.3	80.0	93.3	70.0	69.7	97.8	96.3	77.3
Ours ($\rho = 1\%$)	0.22%	0.01	90.0	83.3	86.7	94.0	76.7	69.7	98.4	96.3	77.8
Qwen3.5-9B Student											
Full SFT	100%	9.23	90.0	86.7	90.0	93.3	73.3	63.6	97.8	96.7	78.8
Ours ($\rho = 30\%$)	1.91%	0.07	93.3	93.3	90.0	97.0	86.7	81.8	98.6	97.4	81.8
Ours ($\rho = 10\%$)	0.74%	0.03	93.3	90.0	90.0	97.8	90.0	81.8	98.2	96.6	81.8
Ours ($\rho = 5\%$)	0.45%	0.02	93.3	90.0	90.0	97.8	83.3	78.8	98.6	97.7	78.8
Ours ($\rho = 1\%$)	0.22%	0.01	90.0	90.0	86.7	97.0	86.7	78.8	98.4	97.4	81.3

In this way, we contract the active training set by retaining the M trajectories with the largest learner-side routing scores:

$$\mathcal{H}_M = \arg \max_{\mathcal{H} \subseteq \mathcal{P}_\rho, |\mathcal{H}|=M} \sum_{i \in \mathcal{H}} \ell_i. \tag{7}$$

Thus, after the probing epoch, subsequent SFT optimization is performed exclusively over $\mathcal{H}_M$. The routing scores are updated during training, allowing the active subset to adapt to the evolving learner state. We provide a detailed algorithm pipeline in Appendix C.

Efficiency Analysis. Together, the two-stage contraction directly translates into a substantially smaller student optimization budget. For instance, standard three-epoch SFT over N trajectories processes $3N$ instances in total. In contrast, our method trains on the largely contracted teacher-side pool only once and restricts the subsequent two epochs to the learner-routed active set $3N \longrightarrow |\mathcal{P}_\rho| + 2|\mathcal{H}_M| = \lfloor \rho N_c \rfloor + 2M$, where $M \ll N$. Accordingly, the relative optimization budget is

$$\mathcal{B} = \frac{\lfloor \rho N_c \rfloor + 2M}{3N}, \tag{8}$$

where $N_c \leq N$ denotes the number of verified teacher trajectories. This corresponds to training on only a small fraction of the trajectory exposures required by standard full-set SFT. In practice, this reduces student-side training exposure to below **0.74%** of vanilla full-set SFT, i.e., over **135× fewer** sample-level exposures, without sacrificing downstream reasoning performance. We discuss more details in Appendix D.

3 Experiment

3.1 Experimental Setup

Models and Training Data. We evaluate our framework across diverse student scales, architectures, and teacher regimes. Unless specified, we use Qwen3-8B Yang et al. (2025a) as the teacher, and Qwen3-14B and Qwen3.5-4B/9B Qwen Team (2026) as student models. For cross-family generalization, we further evaluate Gemma-4-E4B-IT Team et al. (2026), Llama-3.2-3B-Instruct Grattafiori et al. (2024), and Mistral-7B-Instruct Jiang et al. (2023). Moreover, we consider API-only closed-source teachers, including DeepSeek-V4-Pro DeepSeek-AI et al. (2026) and Kimi-K2.6 Moonshot AI (2026), as well as a weaker Qwen3-1.7B

teacher for weak-to-strong transfer. We train on DAPO-17K Yu et al. (2025), while closed-source teacher experiments use the first 1000 samples from Nemotron-Math Du et al. (2025).

Training and Evaluation. We train models with LoRA-based SFT Hu et al. (2021). More training details, including optimization settings and configurations, are provided in Appendix E. We evaluate reasoning ability across competition mathematics, general mathematical reasoning, and scientific reasoning, including AIME24/25/26 Zhang & Math-AI (2025); Dekoninck et al. (2026), AMC Mathematical Association of America (2024), HMMT25 HMMT (2025), HMMT26 HMMT (2026), Math-500 Hendrycks et al. (2021), GSM8K Cobbe et al. (2021), and GPQA-Diamond (GPQA-D) Rein et al. (2023).

3.2 Main Results

Uncertainty-Guided Data-Efficient Training. As shown in Table 1, we evaluate reasoning distillation across competition mathematics, general mathematics, and scientific reasoning, using Qwen3-8B as a common teacher and Qwen3.5-4B/9B as students on DAPO-17k . It can be seen that progressive contraction yields substantial reductions in both training budget and practical training costs while preserving or improving reasoning performance. Across Qwen3.5-4B and 9B students, our method uses only **0.22–1.91%** of the full-SFT optimization budget, reducing SFT time from 8.47/9.23 hours to only **0.01–0.07 hours**. Under $\rho = 1\%$, our method brings over **800**$\times$ wall-clock acceleration, yet the resulting models remain competitive with or outperform full-set SFT across most benchmarks. Importantly, this efficiency persists as the student scales from 4B to 9B, highlighting its generality. Moreover, performance remains remarkably stable under increasingly aggressive contraction, demonstrating that our progressive contraction effectively concentrates optimization on high-value supervision.

Cross-Family Student Generalization. We next examine whether the optimization efficiency generalizes across distinct student model families. Using the same Qwen3-8B teacher, we evaluate three heterogeneous students: Gemma-4-E4B-IT, LLaMA-3.2-3B-Instruct, and Mistral-7B-Instruct, covering different perspectives of reasoning capability. As shown in Fig. 2, our method maintains or improves overall reasoning performance across all three student families while using only approximately **0.45%** of the full-SFT student-side optimization budget, corresponding to over **200**$\times$ **training acceleration**. The consistent effectiveness under such aggressive contraction demonstrates that the efficiency gains are not specific to a particular student architecture. Importantly, once the teacher-side uncertainty profile is constructed from the same Qwen3-8B teacher, it can be shared across students, highlighting strong cross-family transfer and effectiveness.

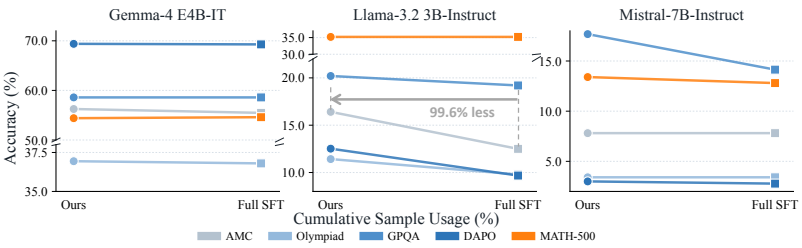


Figure 2: Cross-family student generalization under a shared Qwen3-8B teacher. We evaluate Gemma, LLaMA, and Mistral students under the same teacher-side uncertainty profiling. The cumulative training budget is reduced by over 99.6% compared to full-set SFT.

Weak-to-Strong Reasoning Transfer. Conventional distillation typically assumes a more capable teacher than the target student. In contrast, we use Qwen3-1.7B as a weak teacher for the stronger Qwen3.5-4B and Qwen3-14B students. As shown in Table 2, under $\rho = 5\%$, the 1.7B teacher still provides effective uncertainty-guided supervision, yielding gains of up to 14.4% while requiring merely ≤ 0.1 GPU hours of student training. These results provide strong evidence for *weak-to-strong reasoning transfer*, suggesting that supervision value need not be determined solely by teacher strength: uncertainty-guided profiling can

Table 2: Weak-to-strong reasoning transfer with a 1.7B teacher and Qwen3.5-4B and Qwen3-14B students with $\rho = 5\%$ on DAPO-17K. Overhead is the wall-clock time (h) on a 2-H200-GPU server.

Model	AIME24	AIME25	GPQA-D	MATH-500	GSM8K	Overhead
Qwen3-1.7B (*Weak teacher*)	48.3	36.8	40.1	83.4	90.0	-
Qwen3.5-4B	73.8	65.6	76.2	98.0	96.1	-
+ Ours (5%)	**83.3**$^{+9.5}$	**80.0**$^{+14.4}$	**76.3**$^{+0.1}$	**98.6**$^{+0.6}$	**96.4**$^{+0.2}$	+0.02h
Qwen3-14B	76.0	67.3	61.5	96.0	89.8	-
+ Ours (5%)	**86.7**$^{+10.7}$	**70.0**$^{+2.7}$	**62.2**$^{+0.7}$	95.6$^{-0.4}$	**95.5**$^{+5.7}$	+0.10h

Table 3: Comparison with larger reasoning models across various reasoning benchmarks. **Bold** and underlined values denote the best and second-best results, respectively.

Model	Params	AIME24	AIME25	AIME26	HMMT25	HMMT26	MATH-500	GPQA-D
Larger Reasoning Models								
Gemma-4-it	12B	–	72.9	77.5	63.3	72.7	–	78.8
Ministral-3-Reasoning-2512	14B	89.8	82.9	85.0	67.1	42.4	–	71.2
Phi-4-Reasoning-Plus	14B	81.3	68.4	73.6	50.3	51.5	83.0	68.9
Nemotron-3-Nano	30B	–	89.1	89.9	84.6	–	–	74.1
GLM-4.5-Air	106B	89.4	83.3	–	69.2	–	98.1	–
Qwen3-235B-A22B-Thinking	235B	85.7	92.3	–	83.9	–	98.0	81.1
LongCat-Flash-Thinking	560B	**93.3**	90.6	–	83.7	–	**99.2**	81.5
DeepSeek-R1-0528	671B	91.4	87.5	–	79.4	–	98.3	81.0
GPT-5 Nano (high)	N/A	–	85.2	76.7	75.6	69.7	–	71.2
Ours-9B	**9B**	**93.3**	**93.3**	**90.0**	**90.0**	**81.8**	98.6	**81.8**

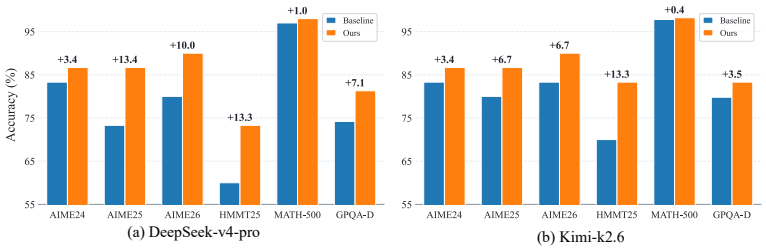
expose locally informative reasoning trajectories, while learner-aware routing concentrates training on those that remain useful to the stronger student.

Black-Box Distillation with Closed-Source Teachers. A practical advantage of our framework is that teacher-side uncertainty profiling is entirely behavior-based. Thus, our method can be directly applied to API-only closed-source teacher models without access to their internal parameters or gradients. Specifically, we use DeepSeek-V4-Pro and Kimi-K2.6 as closed-source teachers, paired with Qwen3.5-4B and Qwen3.5-9B, respectively. We move beyond DAPO-17k and use the first 1000 samples from Nemotron-Math Du et al. (2025) to keep API querying practical. As shown in Fig. 3, uncertainty-guided progressive contraction consistently improves over vanilla distillation across both closed-source teachers and diverse reasoning benchmarks, with gains of up to **13.4%**. Importantly, no modification to the uncertainty estimator is required when moving from open-weight to API-only teachers: process uncertainty is recovered solely from observable continuation behavior, highlighting its practicality.

Comparison with Substantially Larger Reasoning Models. As shown in Table 3, our compact 9B student remains highly competitive with reasoning models that are one to two orders of magnitude larger. Notably, this suggests that the supervision identified by progressive contraction not only improves optimization efficiency, but also translates into substantial capability gains for the resulting student. Thus, our framework can produce a substantially stronger reasoner without relying on increased model scale.

3.3 Mechanism Analysis and Ablations

Teacher-Side Uncertainty Identifies High-Value Reasoning Trajectories. To evaluate whether uncertainty-guided reasoning provides more effective supervision than standard distillation, we prompt the same teacher to generate standard CoTs and train the students using the same optimization budget. As shown



(a) DeepSeek-v4-pro (b) Kimi-k2.6

Figure 3: Generalization to closed-source teachers using (a) DeepSeek-V4-Pro with Qwen3.5-4B, and (b) Kimi-K2.6 with Qwen3.5-9B.

Table 4: Uncertainty-based selection versus random selection under matched budgets using Qwen3.5-4B. We report average results across 7 benchmarks.

Pool Ratio	**1%**	**5%**	**10%**	**30%**
Random	78.1	76.2	75.6	76.5
High-U	**79.7**	**80.6**	**80.1**	**78.7**
Δ	**+1.6**	**+4.4**	**+4.5**	**+2.2**

Table 5: Matched-budget comparison with vanilla distillation under a 1% training budget.

Student	**Method**	**AIME24**	**AIME25**	**AIME26**	**AMC**	**GPQA-D**	**HMMT25**	**HMMT26**	**MATH500**	**GSM8K**
Qwen3.5-4B	Vanilla Distill.	86.7	73.3	76.7	93.3	**77.8**	73.3	66.7	97.0	96.2
	Ours	**90.0**	**83.3**	**86.7**	**94.0**	**77.8**	**76.7**	**69.7**	**98.4**	**96.4**
Qwen3.5-9B	Vanilla Distill.	86.7	86.7	83.3	95.5	80.3	80.0	63.6	97.8	97.0
	Ours	**90.0**	**90.0**	**86.7**	**97.0**	**81.3**	**86.7**	**78.8**	**98.4**	**97.4**

Table 6: Effect of teacher-side uncertainty profiling and learner-aware routing on Qwen3.5-4B with $\rho = 10\%$.

Components		Benchmark Accuracy (%)					
Teacher Unc.	Learner Routing	AIME24	AIME25	AIME26	GPQA	HMMT25	HMMT26
×	×	83.3	73.3	76.7	75.8	63.3	54.5
✓	×	80.0	76.7	76.7	76.3	73.3	57.6
×	✓	83.3	76.7	80.0	76.3	73.3	63.6
✓	✓	**86.7**	**83.3**	**83.3**	**77.3**	**76.7**	**78.7**

in Table 5, our method consistently outperforms vanilla distillation on both Qwen3.5-4B and 9B, improving the average performance by 3.6% and 4.0%, respectively. Because the student optimization budget is strictly matched, the gains reflect more effective allocation of the optimization budget.

To isolate teacher-side uncertainty, we disable learner-aware routing and compare uncertainty-based selection with matched-size random selection among the same verified-correct trajectories. As shown in Table 4, uncertainty-based selection consistently improves average performance by 1.6%-4.5%. This shows that final-answer correctness alone does not fully characterize supervision value; process uncertainty further distinguishes more informative reasoning trajectories among successful CoTs.

Efficiency Comes from Better Optimization Budget Allocation. To better understand the efficiency of our method, as shown in Figure 1, we track the optimization trajectories and compare them with the full SFT. It shows that our method surpasses the full SFT within only 150 steps, whereas full SFT requires about 1500 steps to reach its optimum, **corresponding to** $10\times$ **fewer optimization steps**. These results show that progressive contraction substantially increases the learning progress by concentrating optimization on high-value supervision, thereby improving reasoning performance with dramatically less student training.

Teacher Uncertainty and Learner Routing Are Complementary. We ablate both stages of progressive contraction. As shown in Table 6, each component improves accuracy, while their combination yields the best performance. Importantly, both stages contribute along complementary axes: the former improves the informativeness of the candidate supervision space, while learner-aware routing concentrates SFT updates on trajectories with greater student-specific learning demand. Together, they provide a new perspective

by concentrating increasingly effective supervision into an increasingly compact optimization set, enabling stronger reasoning performance and substantially higher learning efficiency simultaneously. The substantially larger gain from their combination supports the central design of progressive contraction: first preserve informative reasoning supervision, and then allocate optimization according to training dynamics.

4 Related Work

Uncertainty in LLM Reasoning and Learning. Uncertainty quantification has been studied as a means of characterizing the reliability and knowledge boundaries of LLMs (Shorinwa et al., 2025; Li et al., 2025; Ye et al., 2025a). Existing approaches estimate uncertainty through token probabilities, verbalized confidence, sampling-based disagreement, and semantic entropy (Manakul et al., 2023; Farquhar et al., 2024; Luo et al., 2026). Recent studies further show that stronger or longer reasoning does not necessarily yield better uncertainty awareness, with modern reasoning models often remaining overconfident or increasingly miscalibrated on incorrect predictions (Mei et al., 2025; Yeom et al., 2026). Beyond final-answer confidence, process-level uncertainty has been modeled over intermediate reasoning states through Monte Carlo continuation, confidence probing, and calibration-aware supervision (Wang et al., 2026b; Park et al., 2026; Yan et al., 2026). For instance, RLCM (Wang et al., 2026a) introduces confidence-aware process supervision and improves calibration while preserving reasoning accuracy. These works motivate treating uncertainty as a structured signal of reasoning stability and capability boundaries throughout multi-step reasoning. (Wang et al., 2026b; Park et al., 2026; Yan et al., 2026).

Data-Efficient Distillation and Adaptive Data Routing. Knowledge distillation provides a data-centric mechanism for transferring capabilities from stronger teachers to smaller learners (Hsieh et al., 2023; Shridhar et al., 2023; Zhang et al., 2026d). For reasoning tasks, recent work increasingly distills intermediate rationales or long-form CoTs rather than final answers alone, enabling students to acquire richer reasoning behaviors (Liu et al., 2026; Wu et al., 2026a;b). Beyond increasing the amount of reasoning supervision, growing evidence suggests that its effectiveness depends strongly on trajectory form and learner compatibility (Yang et al., 2026b; Zhang et al., 2026b). Accordingly, reasoning supervision has been adapted through trajectory filtering (Yang et al., 2026b), prefix alignment (Liu et al., 2026), difficulty-aware supervision (Wu et al., 2026a), and student-conditioned trajectory selection (Zhang et al., 2026a; He et al., 2026). In parallel, data selection and valuation methods allocate training computation toward more useful examples (Yang et al., 2025b; Mirzasoleiman et al., 2020), using signals such as quality (Wang et al., 2024; Yang et al., 2025c), diversity (Zheng et al., 2023), gradient or influence information (Yang et al., 2022; Wang et al., 2026c), and model-dependent learning dynamics (Engstrom et al., 2024; Wang et al., 2025). Representative approaches include DEITA (Liu et al., 2024), LESS (Xia et al., 2024), Data Agent (Yang et al., 2026a), and RHO-Loss (Mindermann et al., 2022). These methods can be broadly viewed as data routing: allocating limited optimization toward supervision with greater expected learning value.

5 Conclusion

We propose an uncertainty-guided progressive contraction framework that exposes and concentrates learning value in reasoning supervision, substantially accelerating reasoning post-training without sacrificing capability. Across diverse benchmarks and architectures, our method consistently improves data efficiency with strong generalization. These results highlight that effective reasoning supervision depends not only on scale, but also on how learning value is identified and allocated throughout training. More broadly, we hope this work inspires further research on the shift from scaling reasoning data toward more efficient and accessible reasoning post-training, while providing a practical tool to reduce training burden and enable strong reasoning models under limited computational resources for the community.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improve clarity, and polish the presentation of this manuscript. All AI-assisted material was carefully reviewed and verified by the authors.

ETHICS STATEMENT

All datasets used in this work are publicly available benchmarks, including DAPO-17k, Nemotron-Math, AIME, AMC, HMMT, MATH-500, GSM8K, and GPQA-Diamond. These datasets were obtained from their original public sources and used in accordance with their respective licenses and terms of use. The reasoning trajectories used in this work are generated from publicly available or authorized teacher models, including open-weight and API-based models, following the corresponding model usage policies. No new data involving human subjects or personally identifiable information are used in this work.

REPRODUCIBILITY STATEMENT

We provide detailed implementation details, training configurations, and evaluation protocols in the appendix to facilitate reproducibility. The appendix further includes the complete algorithm description, uncertainty profiling procedure, and additional analyses of the proposed Monte Carlo Reasoning Tree estimation. All datasets and preprocessing pipelines used in our experiments are described in detail. We will release the source code, model configurations, and experimental scripts.

REFERENCES

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang,

W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. 2026. URL https://arxiv.org/abs/2605.00674.

Wei Du, Shubham Toshniwal, Branislav Kisacanin, Sadegh Mahdavi, Ivan Moshkov, George Armstrong, Stephen Ge, Edgar Minasyan, Feng Chen, and Igor Gitman. Nemotron-math: Efficient long-context distillation of mathematical reasoning from multi-mode supervision, 2025. URL https://arxiv.org/abs/2512.15489.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. Dsdm: Model-aware dataset selection with datamodels. *arXiv preprint arXiv:2401.12926*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang

Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, An-

nie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chin-

may Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Yingfa Chen, Chaojun Xiao, Xu Han, and Lijie Wen. Student-in-the-loop chain-of-thought distillation via generation-time selection, 2026. URL https://arxiv.org/abs/2604.02819.

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, 2021. URL https://arxiv.org/abs/2103.03874.

HMMT. Hmmt 2025. https://www.hmmt.org/, 2025. Accessed: 2025.

HMMT. Hmmt 2026. https://www.hmmt.org/, 2026. Accessed: 2026.

Zhinan Hou, Jiaqi Zhang, Xunliang Cai, and Keyou You. What matters in on-policy distillation? a perspective on data efficiency and data selection, 2026. URL https://arxiv.org/abs/2609.05198.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023. URL https://arxiv.org/abs/2305.02301.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jiahao Huang, Fei Cheng, Junfeng Jiang, and Akiko Aizawa. Tailoring the curriculum: Student-centered reasoning distillation via dynamic data-model compatibility, 2026. URL https://arxiv.org/abs/2605.29229.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Hoang Anh Just, Myeongseob Ko, and Ruoxi Jia. The signal is in the steps: Local scoring for reasoning data selection, 2026. URL https://arxiv.org/abs/2510.03988.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.150. URL https://aclanthology.org/2023.acl-long.150/.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5131–5157, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.256. URL https://aclanthology.org/2025.acl-long.256/.

Xiaoyuan Li, Yubo Ma, Chengpeng Li, Fengbin Zhu, Yiyao Yu, Keqin Bao, Wenjie Wang, Fuli Feng, and Dayiheng Liu. Unified data selection for llm reasoning, 2026. URL https://arxiv.org/abs/2605.22389.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 22353–22373, 2024.

Zhenghao Liu, Zhuoyang Wu, Xinze Li, Yukun Yan, Shuo Wang, Zulong Chen, Yu Gu, Ge Yu, and Maosong Sun. Long-chain reasoning distillation via adaptive prefix alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18041–18059, 2026.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator. *Advances in Neural Information Processing Systems*, 38:163514–163543, 2026.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Zhenjiang Mao, Anirudhh Venkat, Artem Bisliouk, Sindhura Kumbakonam Subramanian, Akshat Kothiyal, Saithej Singhu, and Ivan Ruchkin. Confidence over time: Confidence calibration with temporal logic for large language model reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9952–9976, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.484. URL https://aclanthology.org/2026.findings-acl.484/.

Mathematical Association of America. AMC 10/12: American Mathematics Competitions 2022–2024. https://www.maa.org/math-competitions/amc-1012, 2024. Accessed: 2024.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know?, 2025. URL https://arxiv.org/abs/2506.18183.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Moonshot AI. Kimi K2.6: Advancing open-source coding. Technical blog, 2026. URL https://www.kimi.ai/blog/kimi-k2-6. Accessed: 2026-09-26.

Young-Jin Park, Kristjan Greenewald, Kaveh Alimohammadi, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models. *Advances in neural information processing systems*, 38:38852–38895, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models, 2023. URL https://arxiv.org/abs/2212.00193.

Igor Slinko, Ilia Zavidnyi, Egor Bogomolov, and Yaroslav Zharov. Step rejection fine-tuning: A practical distillation recipe, 2026. URL https://arxiv.org/abs/2605.10674.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi

Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Yunjie Ji, Han Zhao, and Xiangang Li. Deepdistill: Enhancing llm reasoning capabilities via large-scale difficulty-graded data training, 2025. URL https://arxiv.org/abs/2504.17565.

Jiachen T Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. Greats: Online selection of high-quality data for llm training in every iteration. *Advances in Neural Information Processing Systems*, 37: 131197–131223, 2024.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Nice data selection for instruction tuning in llms with non-differentiable evaluation metric. In *Forty-second International Conference on Machine Learning*, 2025.

Liaoyaqi Wang, Chunsheng Zuo, William Jurayj, Benjamin Van Durme, and Anqi Liu. Process supervision of confidence margin for calibrated llm reasoning, 2026a. URL https://arxiv.org/abs/2604.23333.

Liaoyaqi Wang, Chunsheng Zuo, William Jurayj, Benjamin Van Durme, and Anqi Liu. Process supervision of confidence margin for calibrated llm reasoning. *arXiv preprint arXiv:2604.23333*, 2026b.

Shaobo Wang, Xuan Ouyang, Tianyi Xu, Yuzheng Hu, Jialin Liu, Guo Chen, Tianyu Zhang, Junhao Zheng, Kexin Yang, Xingzhang Ren, et al. Opus: towards efficient and principled data selection in large language model pre-training in every iteration. *arXiv preprint arXiv:2602.05400*, 2026c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Yifan Wu, Jingze Shi, Bingheng Wu, Jiayi Zhang, Xiaotian Lin, Yizhang Zhu, Zhaoyang Yu, Bang Liu, Chenglin Wu, Nan Tang, et al. Concise math reasoning via difficulty-aware distillation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43401–43427, 2026a.

Zhuoyang Wu, Xinze Li, Zhenghao Liu, Yukun Yan, Zhiyuan Liu, Minghe Yu, Cheng Yang, Yu Gu, Ge Yu, and Maosong Sun. Enhancing long-chain reasoning distillation through error-aware self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1465–1482, 2026b.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning, 2024. URL https://arxiv.org/abs/2402.04333.

Hang Yan, Fangzhi Xu, Rongman Xu, Yifei Li, Jian Zhang, Haoran Luo, Xiaobao Wu, Luu Anh Tuan, Haiteng Zhao, Qika Lin, et al. Mur: Momentum uncertainty guided reasoning for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23103, 2026.

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

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*, 2022.

Suorong Yang, Peijia Li, Furao Shen, and Jian Zhao. Reinforcement learning-guided data selection via redundancy assessment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1004–1015. IEEE, 2025b.

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*, volume 2025, pp. 65723–65741, 2025c.

Suorong Yang, Fangjian Su, Hai Gan, Ziqi Ye, Jie Li, Baile Xu, Furao Shen, and Soujanya Poria. Data agent: Learning to select data via end-to-end dynamic optimization. *arXiv preprint arXiv:2603.07433*, 2026a.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, et al. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42123–42150, 2026b.

Peng Ye, Chenyu Huang, Mingzhu Shen, Tao Chen, Yongqi Huang, and Wanli Ouyang. Dynamic model merging with mixture of weights. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a. doi: 10.1109/TCSVT.2025.3544680.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025b. URL https://arxiv.org/abs/2502.03387.

Jewon Yeom, Jaewon Sok, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Epicar: Knowing what you don't know matters for better reasoning in llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22414–22443, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022a. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022b. URL https://arxiv.org/abs/2203.14465.

Lechen Zhang, Yunxiang Zhang, Wei Hu, and Lu Wang. Skill-aware data selection and fine-tuning for data-efficient reasoning distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 595–604, 2026a.

Lechen Zhang, Yunxiang Zhang, Wei Hu, and Lu Wang. Skill-aware data selection and fine-tuning for data-efficient reasoning distillation, 2026b. URL https://arxiv.org/abs/2601.10109.

Qingjie Zhang, Yujia Fu, Yang Wang, Liu Yan, Tao Wei, Ke Xu, Minlie Huang, and Han Qiu. Stop before you fail: Operational capability boundaries for mitigating unproductive reasoning in large reasoning models, 2026c. URL https://arxiv.org/abs/2509.24711.

Ruichen Zhang, Rana Muhammad Shahroz Khan, Zhen Tan, Dawei Li, Song Wang, and Tianlong Chen. The quest for efficient reasoning: A data-centric benchmark to cot distillation, 2026d. URL https://arxiv.org/abs/2505.18759.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Xinghao Zhao. Entropy trajectory shape predicts llm reasoning reliability: A diagnostic study of uncertainty dynamics in chain-of-thought, 2026. URL https://arxiv.org/abs/2603.18940.

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.

A UNCERTAINTY-AWARE REASONING ELICITATION

Elicitation design. Stage 1 begins with an uncertainty-aware elicitation prompt that structures the teacher's reasoning into explicit intermediate steps while encouraging lightweight self-reported uncertainty cues during generation. The purpose of these annotations is to make potentially uncertain reasoning states explicit in the generated trace and provide well-defined anchors for the subsequent reasoning-tree probing described in Sec. 2.2. Importantly, the annotations themselves are not used as the final uncertainty score for data routing. Our process uncertainty is instead grounded behaviorally through Monte Carlo continuations from each intermediate reasoning state.

The elicitation procedure operates entirely through natural-language interaction with the teacher and therefore requires no access to logits, hidden states, gradients, or architecture-specific internals. We use the following prompt template for teacher generation.

Teacher Prompt

```
You are a reasoning teacher generating supervision for a smaller student
    model.

Solve the following problem step by step, while preserving semantically
meaningful intermediate reasoning states.

Requirements:
1. Use numbered reasoning steps.
2. Keep each step concise, self-contained, and logically necessary.
3. Preserve important intermediate deductions, alternative considerations,
   self-corrections, and non-trivial calculations when they are relevant to
   reaching the solution.
4. Avoid collapsing multiple distinct reasoning transitions into a single
    step.
5. Do not introduce unnecessary explanations or redundant steps.
6. End with the final answer in the format [[answer]].

Question:
{question}
```

Monte Carlo Rollout Prompt

```
\textbf{System:}

You are a reasoning teacher generating supervision for a smaller student
    model.

\textbf{User:}

Continue the reasoning from the given prefix and finish the solution.
The very last line must be exactly [[answer]] <number>, where <number>
is the final numeric answer.

Question:
{question}

Reasoning prefix:
```

```
{reasoning\_prefix}
```

Output Representation. The teacher returns a structured reasoning trajectory

$$\tau_i^{\text{ann}} = (s_{i,1}, \ldots, s_{i,m_i}, \hat{a}_i),$$

where uncertainty-bearing reasoning steps are explicitly annotated with the auxiliary marker `[uncertain]`. A representative output is shown below.

Representative Teacher Output

```
Step 1: ...

Step 2 [uncertain]: ...

Step 3: ...

Therefore, the final answer is [[...]].
```

For every generated example, we retain two versions of the trajectory. The annotated version, τ_i^{ann}, preserves the teacher's self-reported uncertainty cues for analysis, while the clean version, τ_i, removes the auxiliary markers and serves as the actual distillation target:

$$\tau_i = \text{Clean}\left(\tau_i^{\text{ann}}\right).$$

We additionally extract the terminal prediction $\hat{a}_i$ and record

$$c_i = \mathbb{I}[\hat{a}_i = a_i],$$

which associates the generated reasoning path with its observed correctness outcome. No trajectory is discarded during this elicitation step. The structured reasoning states, and the associated correctness signal are subsequently used to construct the prefix-rooted Monte Carlo reasoning trees in Stage 1.

From elicited uncertainty to behavioral uncertainty. The `[uncertain]` annotations provide a semantic view of uncertainty expressed by the teacher itself, whereas the rollout-based uncertainty used by TALE characterizes the behavior of the underlying reasoning process. Specifically, each intermediate state $h_{i,t}$ is expanded through multiple independent continuations, producing the step-wise uncertainty profile

$$\mathcal{Z}_i = \left\{(\hat{p}_{i,t}, \widehat{U}_{i,t})\right\}_{t=1}^{m_i}.$$

Thus, the elicitation prompt exposes uncertainty at generation time, while Monte Carlo reasoning-tree probing converts the generated trajectory into behaviorally grounded, uncertainty-profiled supervision.

B Finite-Sample Analysis of MCRT Uncertainty Profiling

This section provides finite-sample guarantees for the Monte Carlo quantities used in our reasoning uncertainty profiling. For clarity, we consider a fixed intermediate state $h_{i,t}$ and omit the subscripts i, t when no ambiguity arises. Recall that, under the fixed stochastic decoding policy used for MCRT rollouts,

$$Z^{(k)} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \qquad k = 1, \ldots, K, \tag{9}$$

where p denotes the correct-answer reachability from the current reasoning state, and

$$\hat{p} = \frac{1}{K} \sum_{k=1}^{K} Z^{(k)} \tag{10}$$

is its Monte Carlo estimate.

Concentration of Correct-Answer Reachability. Since $Z^{(k)} \in [0,1]$ are conditionally independent given the reasoning state, Hoeffding's inequality gives, for any $\epsilon > 0$,

$$\Pr\left(|\hat{p} - p| \geq \epsilon \mid h_{i,t}\right) \leq 2\exp(-2K\epsilon^2). \tag{11}$$

Equivalently, to ensure $|\hat{p} - p| < \epsilon$ with probability at least $1 - \delta$, it is sufficient to use

$$K \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}. \tag{12}$$

Thus, the probability of a fixed estimation error decays exponentially with the rollout budget K, establishing $\hat{p}$ as a consistent Monte Carlo estimate of state-wise correct-answer reachability.

Concentration of Process Uncertainty. Our process uncertainty is defined as

$$U = 4p(1-p), \qquad \widehat{U} = 4\hat{p}(1-\hat{p}). \tag{13}$$

Let $f(p) = 4p(1-p)$. Since

$$|f'(p)| = |4 - 8p| \leq 4, \qquad p \in [0,1], \tag{14}$$

f is 4-Lipschitz over $[0,1]$. Therefore,

$$|\widehat{U} - U| \leq 4|\hat{p} - p|. \tag{15}$$

Combining this relation with Eq. equation 11 yields

$$\Pr\left(|\widehat{U} - U| \geq \epsilon \mid h_{i,t}\right) \leq 2\exp\left(-\frac{K\epsilon^2}{8}\right). \tag{16}$$

Hence, the process uncertainty estimate inherits the consistency of the underlying Monte Carlo reachability estimator. As K increases, the empirical branch-instability score $\widehat{U}$ converges to its population counterpart U.

Finite-K Bias. For completeness, the plug-in estimator $\widehat{U}$ has a small finite-sample bias. Since $\mathrm{Var}(\hat{p}) = p(1-p)/K$, we have

$$\begin{aligned} \mathbb{E}[\widehat{U} \mid h_{i,t}] &= 4\mathbb{E}[\hat{p}(1-\hat{p})] \\ &= 4\left(1 - \frac{1}{K}\right) p(1-p) \\ &= \left(1 - \frac{1}{K}\right) U. \end{aligned} \tag{17}$$

Thus, $\widehat{U}$ slightly underestimates the population uncertainty for finite K. An unbiased version can be obtained as

$$\widehat{U}^{\,\mathrm{unb}} = \frac{K}{K-1}\widehat{U}, \qquad K > 1. \tag{18}$$

Importantly, our experiments use the same rollout budget K for every reasoning state. The correction factor $K/(K-1)$ is therefore constant across samples and does not alter the ranking used for uncertainty-guided candidate selection. We consequently use the simpler plug-in form $\widehat{U}$ throughout the main text.

C ALGORITHM PIPELINE

Algorithm 1 Uncertainty-Profiled Supervision with Learner-Aware Routing

Require: Teacher model T, student model S_θ, training set $\mathcal{D} = \{(x_i, a_i)\}_{i=1}^{N}$, rollout number K, candidate-pool ratio ρ, epoch E, data routing size M

Offline Stage I: Teacher-Side Uncertainty Profiling and Pool Construction
for each $(x_i, a_i) \in \mathcal{D}$ **do**
  Uncertainty-revealing elicitation:
  Generate a structured reasoning trajectory $\tau_i = (s_{i,1}, \ldots, s_{i,m_i}, \hat{a}_i)$ from T using targeted prompt to expose uncertainty-bearing intermediate steps
  Record terminal correctness $c_i \leftarrow \mathbb{I}[\hat{a}_i = a_i]$
  Step-level Monte Carlo trust estimation:
  for $t = 1, \ldots, m_i$ **do**
    Form reasoning state $h_{i,t} \leftarrow (x_i, s_{i,1:t})$
    Sample K independent continuations $\{\hat{a}_{i,t}^{(k)}\}_{k=1}^{K}$ from T conditioned on $h_{i,t}$
    $Z_{i,t}^{(k)} \leftarrow \mathbb{I}[\hat{a}_{i,t}^{(k)} = a_i]$
    Estimate state trust $\hat{p}_{i,t} \leftarrow \dfrac{1}{K}\sum_{k=1}^{K} Z_{i,t}^{(k)}$
    Estimate branch-outcome uncertainty $\widehat{U}_{i,t} \leftarrow 4\hat{p}_{i,t}(1 - \hat{p}_{i,t})$
  end for
  Trajectory-level profiling: $\mathcal{Z}_i \leftarrow \{(\hat{p}_{i,t}, \widehat{U}_{i,t})\}_{t=1}^{m_i}$
  Aggregate trajectory uncertainty $\bar{U}_i \leftarrow \dfrac{1}{m_i}\sum_{t=1}^{m_i} \widehat{U}_{i,t}$
end for
Teacher-side candidate pooling: $\mathcal{C} \leftarrow \{i : c_i = 1\}$
$\mathcal{P}_\rho \leftarrow \arg\max_{\mathcal{S}\subseteq\mathcal{C},\, |\mathcal{S}|=\lfloor\rho|\mathcal{C}|\rfloor} \sum_{i\in\mathcal{S}} \bar{U}_i$
Online Stage II: Learner-Aware Online Supervision Routing
Learner probing: Perform one standard optimization over $\mathcal{P}_\rho$
for each $i \in \mathcal{P}_\rho$ **do**
  Record normalized response loss $\ell_i \leftarrow -\dfrac{1}{|\tau_i|}\sum_{j=1}^{|\tau_i|} \log p_\theta(\tau_{i,j} \mid x_i, \tau_{i,<j})$
end for
Learner-aware routing: $\mathcal{H}_M \leftarrow \arg\max_{\mathcal{H}\subseteq\mathcal{P}_\rho,\, |\mathcal{H}|=M} \sum_{i\in\mathcal{H}} \ell_i$
Budget-contracted training:
for $e = 2, \ldots, E$ **do**
  Fine-tune S_θ for one epoch using only $\mathcal{H}_M$
end for
  return Fine-tuned student model S_θ

Table 7: Practical training efficiency under different uncertainty-guided pool ratios. *Pool Retention* denotes the fraction of Stage-1 trajectories retained in $\mathcal{P}_\rho$. *Actual Budget* measures the cumulative number of processed samples relative to standard three-epoch SFT, i.e., $(|\mathcal{P}_\rho| + 2M)/(3N)$ with $M = 40$.

Dataset	ρ	$\lvert\mathcal{P}_\rho\rvert$	**Pool Retention** ↓	**Actual Budget** ↓
DAPO-17K	1%	30	0.19%	**0.19%**
	5%	149	0.96%	**0.49%**
	10%	298	1.92%	**0.81%**
	30%	895	5.76%	**2.09%**
GSM8K	1%	65	0.87%	**0.65%**
	5%	323	4.32%	**1.80%**
	10%	646	8.64%	**3.24%**
	30%	1,938	25.93%	**9.00%**

D More Analyses on the Data-Efficient Training

In this work, we introduce a progressive data contraction to enhance model capability boundaries while reducing training overhead. As shown in Table 7, teacher-side uncertainty profiling first retains only verified correct CoTs and then selects the top-ρ trajectories according to their aggregated reasoning uncertainty. Consequently, the practical candidate pool can be substantially smaller than the nominal ratio ρ of the original dataset. For example, with $\rho = 10\%$, only 298 DAPO trajectories and 646 GSM8K trajectories remain.

Learner-aware routing further reduces the training footprint. The candidate pool is used only in the first student-training epoch, while subsequent epochs operate on a much smaller active set selected according to student-side difficulty, with as few as $M = 40$ trajectories in our default setting. Therefore, routing reduces not only the amount of processed supervision but also the actual number of optimization steps. Overall, the successive contraction from correctness filtering to uncertainty-guided pooling and learner-aware routing yields a training budget far below that implied by the nominal pool ratio. Importantly, as shown in Sec. 3, this substantial reduction in training cost does not compromise downstream reasoning performance and can even lead to consistent improvements over substantially more expensive SFT baselines.

E Implementation Details

Training objective. We fine-tune all student models using LoRA-based supervised fine-tuning (SFT) (Hu et al., 2021). Each example is formatted as a two-turn conversation, with the original problem as the user message and the teacher-generated reasoning trajectory as the assistant response. We use the official chat template associated with each student model. The causal language-modeling loss is computed only over assistant-response tokens; user prompt tokens and padding positions are masked. To ensure that the learner-side difficulty score is comparable across examples of different lengths, the loss of each example is normalized by its number of supervised response tokens.

Optimization and parameter-efficient fine-tuning. We insert LoRA adapters into the query and value projections of each Transformer block while keeping the backbone parameters frozen. Unless otherwise specified, models are trained for three epochs with AdamW in bfloat16 precision. We use dynamic within-batch padding and gradient checkpointing to reduce memory consumption. Examples exceeding the maximum sequence length are excluded rather than silently truncated, since right truncation may remove the final answer. The complete training configuration is summarized in Table 8.

Table 8: Training configuration used in the main experiments. The batch size denotes the global batch size across all GPUs.

Category	Hyperparameter	Setting
LoRA	Target modules	Query and value projections (`q_proj`, `v_proj`)
	Rank r	8
	Scaling factor α	32
	Dropout	0.0
	Trainable parameters	LoRA adapters only
Optimization	Optimizer	AdamW
	Peak learning rate	5×10^{-5}
	Weight decay	0.01
	Learning-rate schedule	Linear decay to 0.1 of the initial rate
	Gradient clipping	1.0
	Training epochs	3
	Global batch size	8
	Gradient accumulation	1
	Numerical precision	bfloat16
Data processing	Training format	Model-specific chat template
	Supervised tokens	Assistant response only
	Per-example normalization	Mean over supervised response tokens
	Maximum sequence length	8,192 tokens
	Padding	Dynamic, left padding within each mini-batch

Learner-aware training schedule. For TALE, the student first performs one uniform pass over the candidate pool. During this epoch, we record the first-observation response loss $\ell_i^{(0)}$ for every example. We subsequently select the M examples with the largest losses and keep this subset fixed for the remainder of training. During each of the remaining two epochs, examples are sampled with replacement according to

$$q_i = \frac{\exp\left(\ell_i^{(0)}/\tau\right)}{\sum_{j \in \mathcal{H}_M} \exp\left(\ell_j^{(0)}/\tau\right)}, \qquad i \in \mathcal{H}_M, \tag{19}$$

where $\mathcal{H}_M$ is the fixed high-loss subset and τ is the sampling temperature. The resulting probabilities determine only the frequency with which examples are sampled; we do not apply additional importance weights to the language-modeling loss. Within each controlled comparison, all methods use the same student initialization, data preprocessing, LoRA configuration, optimizer, batch size, sequence-length limit, and training precision. They differ only in the data-selection or sampling strategy under study.

F Additional Analysis of Teacher-Side Uncertainty

To better understand what information is captured by teacher-side process uncertainty, we compare our criterion with several alternative statistics derived from the same reasoning process. We keep the teacher trajectories, candidate-pool size, and downstream training protocol identical, varying only the criterion used to rank final-answer-correct trajectories. In addition to random selection, we consider criteria based on boundary uncertainty, average rollout success, average trust, and the largest positive transition along the reasoning

Table 9: Comparison of trajectory-level selection criteria in the 7B→3B GSM8K setting. All methods select 30% of final-answer-correct teacher CoTs under the same training protocol; only the selection criterion is varied.

Selection Criterion	**Acc.**
Random Correct	81.58
Highest Boundary	80.21
Medium Avg. Success	79.38
Highest Avg. Trust	78.47
Max Positive Transition	80.14
Highest Avg. Uncertainty	**82.11**

trajectory. These alternatives capture different aspects of the teacher's reasoning dynamics, ranging from overall confidence to localized changes in reasoning reliability.

As shown in Table 9, selecting trajectories with the highest average process uncertainty achieves the best downstream accuracy of 82.11%, outperforming random correct selection and all alternative process statistics. Notably, neither selecting trajectories close to the decision boundary nor prioritizing individual reliability transitions provides the same benefit. This suggests that the useful signal is not captured by a single uncertain state or local transition; rather, aggregating uncertainty over the reasoning process provides a more informative characterization of trajectory-level supervision value.

G More Analyses on Learner Routing Size

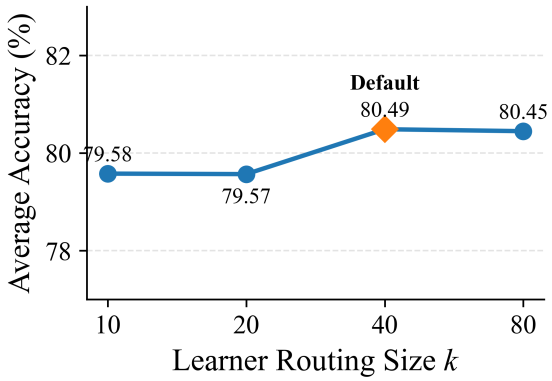


Figure 4: **Sensitivity to the learner-routing size k.** With the uncertainty pool fixed at $\rho = 30\%$, performance remains stable across an $8\times$ range of active-set sizes. The seven-benchmark average varies by only 1.12 points across $k \in \{10, 20, 40, 80\}$, with $k = 40$ used as our default setting.

In our learner-aware routing, we concentrate optimization on an extremely compact subset with a size of k. In this section, we further examine the stability of k by varying $k \in \{10, 20, 40, 80\}$, where $k = 40$ is the default setting used in all our main experiments. As shown in Figure 4, we report the average accuracy over seven reasoning benchmarks: AIME24, AIME25, AIME26, AMC, HMMT25, HMMT26, and GPQA-D using Qwen3.5-4B on DAPO17k. It can be observed that the performance exhibits a broad plateau across

substantially different routing sizes. The size of data routing has limited influence on model performance but more on training efficiency. Instead, the broad stability across k supports the progressive supervision-contraction principle underlying our method: once teacher-side uncertainty has exposed an informative reasoning space, effective optimization can be sustained over a wide range of compact learner-specific active sets. The precise number of retained trajectories is therefore secondary to concentrating student updates on supervision that remains unresolved by the learner.

H Analysis of Candidate-Pool Distribution

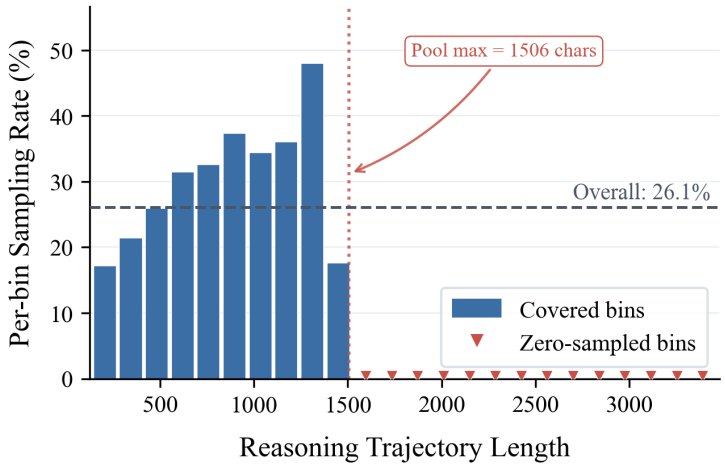


Figure 5: Distribution of candidate-pool selection on DAPO-17k using Qwen3.5-8B as the teacher model.

To better understand what types of reasoning trajectories are retained by the teacher-side uncertainty criterion, we analyze the candidate-pool composition with respect to reasoning trajectory length. Figure 5 reports the per-bin sampling rate on DAPO-17k, where trajectories are grouped according to their reasoning length and Qwen3.5-8B is used as the teacher model. The dashed horizontal line denotes the overall sampling rate of 26.1%.

Interestingly, the selection probability does not increase monotonically with trajectory length. Instead, the candidate pool is enriched in intermediate-length trajectories, with the sampling rate increasing over a moderate length range and then dropping sharply for longer trajectories. In particular, the longest trajectory retained in the candidate pool contains approximately 1506 characters, despite substantially longer trajectories being present in the original dataset.

This distribution suggests that the uncertainty criterion captures a different property from trajectory length alone. If uncertainty merely acted as a proxy for long reasoning chains, one would expect longer trajectories to be preferentially retained. The observed non-monotonic pattern instead indicates that high-value uncertainty tends to emerge within a particular region of the reasoning-length distribution rather than simply at its longest tail. Together with the length-controlled results in Table 10, this analysis further supports that the effectiveness of uncertainty-based selection cannot be explained solely by a preference for longer reasoning trajectories.

I Is Process Uncertainty Simply a Proxy for Reasoning Length?

Table 10: Controlling for trajectory length in the 7B→3B GSM8K setting. We select the same-sized 30% pool using sequence length, token count, or process uncertainty, while keeping the downstream training protocol fixed.

Selection Criterion	Acc.	Avg. Tokens	Avg. Length
Max Length	80.97	265.54	790.79
Max Tokens	78.62	269.91	778.30
Uncertainty	**82.11**	**191.15**	**550.54**

The length-distribution analysis in Sec H shows that the candidate pool exhibits a non-monotonic preference over reasoning length, rather than systematically favoring longer CoTs. We further conduct a controlled experiment to examine whether the benefit of uncertainty-based selection can be explained by trajectory length alone. Specifically, we use Qwen2.5-7B as the teacher to distill GSM8K and train Qwen2.5-3B as the student model. Here, the candidate pool is constructed using three criteria: response length, token count, and our process uncertainty. All candidate pools are subsequently trained under the same protocol.

As shown in Table 10, despite containing substantially shorter trajectories on average, uncertainty selection achieves the best accuracy. Notably, the uncertainty-selected pool contains approximately 28–29% fewer tokens on average than either length-based alternative. Together with the non-monotonic length distribution, these results provide direct evidence that **the effectiveness of process uncertainty cannot be attributed to selecting longer or more verbose reasoning trajectories**. Instead, uncertainty identifies informative reasoning processes beyond superficial sequence length.

J Limitations and Future Work

Our current framework is primarily developed for verifiable reasoning settings, where sampled continuations can be reliably evaluated against known answers. Extending uncertainty profiling to open-ended or weakly verifiable tasks remains an important direction, potentially requiring reliable learned or model-based verification.

Second, Monte Carlo reasoning-tree construction introduces additional teacher-side inference cost. Although this profiling is performed once and can be reused across students, reducing the required number of rollouts or developing more efficient uncertainty estimators would further improve practical efficiency.

Finally, our current framework performs contraction at the trajectory level. Extending progressive contraction to finer-grained reasoning units, such as steps or local reasoning branches, may further improve optimization efficiency and provide more precise control over where learning is allocated. We view this as a promising direction toward more fine-grained and general reasoning post-training.