Teach Small Models to Reason by Curriculum Distillation

Wangyi Jiang[1,2], **Yaojie Lu**[1], **Hongyu Lin**[1], **Xianpei Han**[1*], **Le Sun**[1*]
[1]Chinese Information Processing Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{jiangwangyi2020,luyaojie,hongyu,xianpei,sunle}@iscas.ac.cn

Abstract

Large Reasoning Models (LRMs) show strong System-2-style reasoning, but at the cost of significant computational overhead. In contrast, efficient System-1-style Large Language Models (LLMs) often struggle on complex tasks. We identify a critical asymmetry between these two paradigms: LRMs can implicitly self-distill their own reasoning, solving hard problems with near System-1-style efficiency while retaining superior performance. LLMs, however, lack such deep internal modes and collapse when forced to rely on their own reasoning rather than imitating external traces. This asymmetry explains why direct distillation from strong LRMs to weaker LLMs often fails: student models struggle to learn from LRMs' overly complex explicit reasoning and gain little from their overly compact implicit solutions. To address this, we introduce a two-stage curriculum distillation framework, which first builds a robust internal problem-solving student model and then teaches the student model to externalize this latent knowledge as explicit reasoning. On challenging mathematical benchmarks, our method significantly outperforms single-stage baselines, creating compact models with strong reasoning ability.

1 Introduction

Recent advances in language modeling have led to the emergence of two distinct classes of models. While conventional Large Language Models (LLMs), such as `DeepSeek-V3` (DeepSeek-AI et al., 2025b), `Llama3` (Grattafiori et al., 2024), and `Qwen2.5` (Qwen et al., 2025), have achieved groundbreaking progress, they often fall short in tasks requiring complex, multi-step reasoning like mathematics, multi-hop question answering, and program verification. This limitation has spurred the development of an emergent class of Large Reasoning Models (LRMs) (OpenAI et al., 2024; Anthropic, 2025; Comanici et al., 2025). The distinction between these two archetypes is often analogized to the dual-process theory of cognition: LLMs are optimized for fast, intuitive responses akin to System-1-style thinking, whereas LRMs are engineered to externalize a slow and deliberate thought process that mirrors System-2-style thinking. `DeepSeek-R1` (DeepSeek-AI et al., 2025a), training with innovative reinforcement learning from verifiable rewards, has become the flagship among LRMs due to its open-source accessibility and superior performance.

While LRMs set new benchmarks on challenging tasks like mathematical problem solving (Seed et al., 2025; MiniMax et al., 2025; Bercovich et al., 2025), their deliberative process introduces significant computational overhead. The verbose reasoning traces can lead to excessive inference times, a phenomenon known as *overthinking* (Chen et al., 2025), making them impractical for efficiency-critical applications. Conversely, the System-1-style speed of LLMs, though broadly useful, is often insufficient for complex tasks where step-by-step reasoning is essential for accuracy. This creates a fundamental trade-off: the analytical power of System-2-style at the cost of efficiency, or the speed of System-1-style at the risk of failure on hard problems.

This paper investigates whether this trade-off is fundamental, uncovering a critical asymmetry between the two paradigms. Our empirical analysis reveals that LRMs possess a deep and flexible internal mode of the reasoning process, one that can be modulated to balance performance and efficiency. By moderately pruning its thought process, an LRM can achieve a more efficient, linear style of reasoning. More strikingly, when compelled to bypass explicit thought generation entirely, it can implicitly self-distill its elaborate thinking, solving difficult problems with System-1-style efficiency while still substantially outperforming its

*Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7424–7434
November 4-9, 2025

LLM counterpart. Standard LLMs, in contrast, lack this modulatable internal mode. Our context titration experiments show that as problem difficulty increases, their intrinsic problem-solving capabilities collapse, making them reliant on extensive external reasoning scaffolds. They function not as autonomous reasoners, but as proficient completion engines, adept at following a provided logical path but unable to forge one on their own. This fundamental asymmetry explains why direct distillation often fails: the LRM's explicit reasoning trace is too complex for a student model to imitate, while its implicitly generated, dense solution provides an insufficient learning signal.

To bridge this gap, we introduce a two-stage curriculum distillation framework designed to transfer an LRM's sophisticated, dual-process capabilities to a smaller student LLM. Our approach is inspired by principles of effective learning, following a logical progression from intuition to articulation. The first stage instills a foundational internal mode by training the student model on a curriculum mixing standard LLM reasoning for simple problems with the LRM's dense, implicit solutions for the most challenging ones. Once this latent problem-solving apparatus is established, the second stage teaches the student model to externalize this understanding, training it on the LRM's pruned, explicit reasoning traces for those same hard problems. This method enables the student to first internalize a robust solution strategy and then learn to articulate it effectively.

Our contributions are threefold. First, we empirically characterize the distinct reasoning mechanisms of LRMs and LLMs, identifying the LRM's novel capacity for implicit self-distillation. Second, we diagnose the failure modes of single-stage distillation, attributing them to the fundamental mismatch in complexity and density between the teacher's reasoning and the student's learning capacity. Third, we propose and validate our curriculum distillation framework, demonstrating that it significantly outperforms single-stage baselines on challenging mathematical benchmarks. Our work yields compact models that inherit the sophisticated reasoning of leading LRMs, effectively bridging the divide between System-1-style and System-2-style models.

2 Related Work

2.1 Large Reasoning Models

Recent Large Reasoning Models (LRMs) are distinguished by their reliance on sophisticated reasoning techniques that generate extended, structured thought processes (OpenAI et al., 2024; DeepSeek-AI et al., 2025a; Team, 2025). At test-time, these capabilities are typically realized through two distinct computational approaches.

Sequential reasoning. This approach involves generating a comprehensive chain-of-thought that incorporates reflection and verification within a single forward pass. Instilling these capabilities often requires intensive training paradigms, such as reinforcement learning from process-based feedback or iterative self-improvement (Zelikman et al., 2022; Lambert et al., 2025). `DeepSeek-R1` leverages an innovative technique, reinforcement learning from verifiable rewards, to unlock the reasoning capabilities of models without any supervised data.

Parallel reasoning. This strategy involves generating and subsequently aggregating multiple solution candidates, including techniques such as Best-of-N sampling and search-guided methods like Monte Carlo Tree Search (MCTS) (Snell et al., 2024; Brown et al., 2024). In these methods, numerous reasoning paths are explored, evaluated, and then consolidated using search algorithms or external verification mechanisms.

2.2 Reasoning Efficiency

A widely observed issue with LRMs is the *overthinking* problem (Chen et al., 2025), where models engage in unnecessarily complex or redundant reasoning, leading to significant computational overhead and increased inference latency. To address this, recent research has focused on improving *reasoning efficiency*, aiming to teach LLMs to generate more concise reasoning without compromising performance.

Adaptive Reasoning. This approach enables the model to dynamically adjust its reasoning depth and length according to the complexity of the problem. For instance, Jiang et al. (2025) introduce *Hybrid Group Policy Optimization* to train a model that can adaptively select between long-chain and short-chain reasoning paths. Drawing inspiration from cognitive science, Cheng et al. (2025) incorporate special fast and slow thinking tokens, allowing

the model to dynamically switch between rapid, intuitive responses and more deliberate, step-by-step analysis.

Early Exit. This strategy involves terminating the reasoning process once a confident answer is reached. The core challenge is determining the optimal moment to stop. Qiao et al. (2025) propose a confidence injection technique that inserts high-confidence phrases into the model's reasoning. By monitoring the confidence level, generation can be halted once it surpasses a predefined threshold. A more direct method, *elastic-reasoning* (Xu et al., 2025), imposes a strict token budget on the reasoning process and forcibly concludes it when the allocation is exhausted.

3 Probing the Interchangeability of Reasoning Paradigms

In this section, we examine the behavior of LLMs and LRMs under non-default reasoning paradigms.

3.1 Evaluating LRM Behavior Across Reasoning Paradigms

We construct an experimental framework to evaluate the performance of LRMs under several distinct paradigms designed to elicit different reasoning modes.

3.1.1 Reasoning Paradigms

Think. This is the standard operational mode for LRMs, wherein the model generates a sequence of intermediate reasoning steps before concluding with a final solution. This paradigm is intended to mimic a deliberative, System-2-style cognitive process.

NoRethink. The widely observed *overthinking* problem (Chen et al., 2025) refers to the tendency of LRMs to engage in redundant self-reflection and verification. To counteract this, the *NoRethink* paradigm enforces a more linear reasoning process, discouraging loops or branches. This is implemented during inference by prohibiting the generation of specific keywords known to trigger self-correction, a technique similar to that of Wang et al. (2025). By forbidding the following words, we compel the model to be a more efficient and concise thinker:

```
["alternatively", "another", "but",
↪  "hmm", "verify", "wait"]
```

NoThink. In this experimental condition, we compel the LRMs to bypass explicit, step-by-step reasoning and produce an immediate solution. This is achieved by prepending the prompt with a non-informative thought template, adapted from Ma et al. (2025), thereby simulating an intuitive, System-1-style response:

```
<think> Okay, I think I have finished
↪  thinking. </think>
```

Instruct. This condition serves as a crucial baseline, representing the performance of the underlying LLM from which the LRM is derived. In this paradigm, the base LLM, which has not been specialized for advanced reasoning, is prompted to generate a standard chain-of-thought solution.

3.1.2 Experiments

Experimental Setup. We evaluate the mathematical problem-solving capabilities of the `DeepSeek-R1-Distill-Qwen` model family at three scales: `7B`, `14B`, and `32B`. These models are distilled from `DeepSeek-R1` (DeepSeek-AI et al., 2025a) and are based on `Qwen2.5` and `Qwen2.5-Math` (Qwen et al., 2025; Yang et al., 2024). For our *Instruct* baseline, we use the corresponding instruction-tuned versions of these base models: `Qwen2.5-14B/32B-Instruct` and `Qwen2.5-Math-7B-Instruct`. The evaluation is conducted on the `MATH` dataset (Hendrycks et al., 2021), which consists of 7,500 problems categorized into five difficulty levels. For each problem, we generate four solutions using a temperature of 0.6 and a top-p of 0.95, reporting both Pass@1 and total token usage.

Main Results. Our analysis, illustrated in Figure 1, reveals several consistent and significant trends across all evaluated model scales. These findings underscore a clear trade-off between reasoning paradigms, performance, and computational cost. We summarize the key conclusions below.

Explicit Reasoning Yields Peak Performance at High Cost. The *Think* paradigm consistently achieves the highest Pass@1 accuracy across all difficulty levels. This result validates the effectiveness of generating explicit, intermediate reasoning steps. This superior performance, however, entails the highest token consumption, which scales sharply with problem difficulty, demonstrating the significant computational overhead of deliberative reasoning.

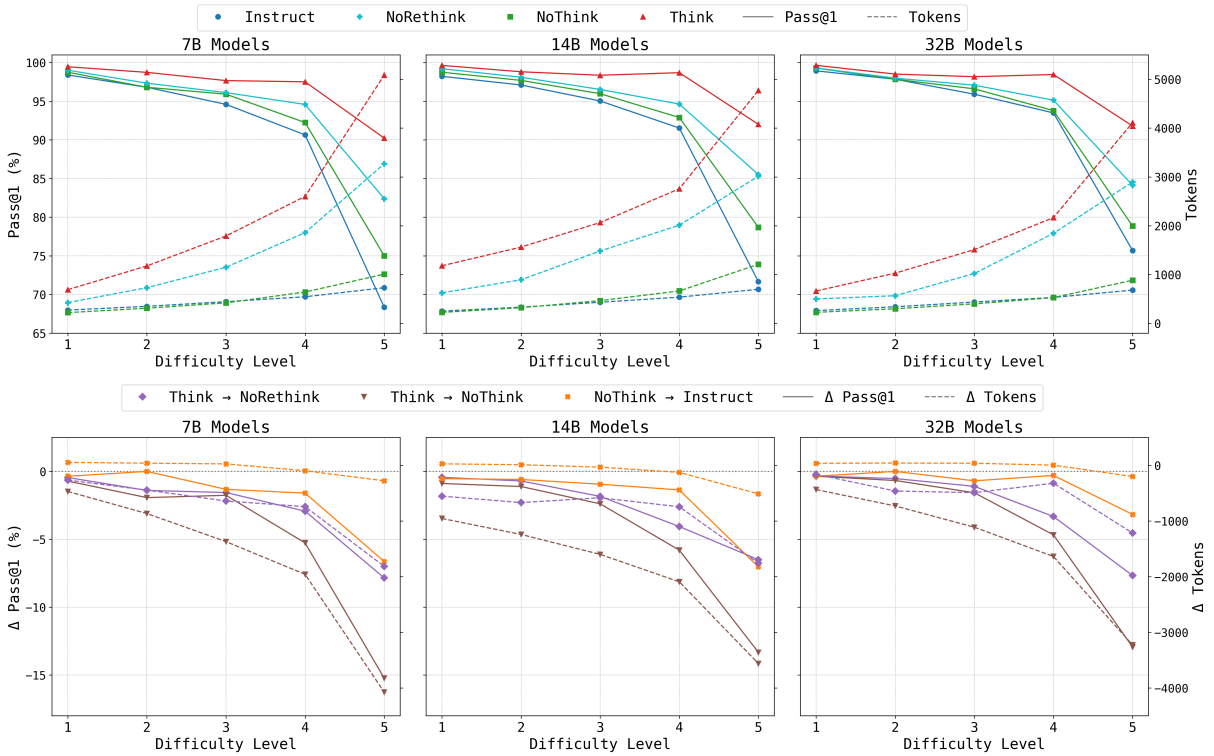

Figure 1: Performance comparison across reasoning paradigms and model scales. The top row presents Pass@1 scores and token consumption for the *Think*, *NoRethink*, *NoThink*, and *Instruct* conditions across five difficulty levels. The bottom row provides a differential analysis, illustrating the performance gaps and corresponding token consumption differences between paradigms. Results are shown for 7B, 14B, and 32B models evaluated on mathematical reasoning tasks of increasing complexity.

Internalized Reasoning Surpasses Instruction Following on Complex Problems. At the other end of the spectrum, the *Instruct* and *NoThink* conditions are the most token-efficient. While their performance is nearly identical on low-difficulty problems, a significant divergence emerges as task complexity increases. The *NoThink* paradigm progressively outperforms the *Instruct* baseline on harder problems, establishing a performance gap of up to 7 percentage points. Notably, this substantial accuracy gain is achieved with only a marginal increase in token usage, suggesting that the LRM's internalized reasoning capability is more effective than the base LLM's standard chain-of-thought process.

Restricted Reasoning Offers a Balanced Trade-off. The *NoRethink* paradigm consistently occupies an intermediate position in both performance and efficiency. By precluding self-correction, it strikes a more favorable balance between the high accuracy of *Think* and the token efficiency of *NoThink*. This demonstrates that pruning the reasoning process can effectively mitigate computational costs while retaining a majority of the performance benefits of a full thought process.

In essence, our findings reveal a clear trade-off between reasoning depth and efficiency. While *Think* achieves peak accuracy through costly deliberation, *NoRethink* proves that pruning this process yields a balanced outcome. Most importantly, *NoThink* confirms the LRM's fundamental value by showcasing its superior and efficient internalized reasoning capability.

3.2 Quantifying LLM Reliance on External Reasoning Scaffolds

In stark contrast to an LRM's ability to internalize reasoning, a standard LLM's capacity for System-2-style thinking appears heavily dependent on external guidance. To probe the extent of this dependency, we design a *context titration experiment* to quantify an LLM's reliance on explicit thinking processes.

Experimental Setup. Our experiment begins by sampling problems from the MATH dataset, stratified into two groups: *Low Difficulty (Level 1,2,3)* and *High Difficulty (Level 4,5)*. The source of the external reasoning scaffold for each problem is a complete, successful solution generated by an

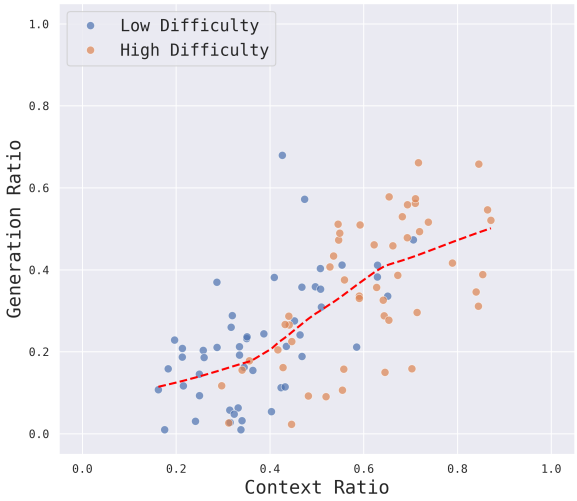

Figure 2: A scatter plot illustrating the relationship between the fraction of context provided to an LLM (*Context Ratio*) and the relative length of its generated solution (*Generation Ratio*).

LRM under the *Think* paradigm. We first segment this LRM trace into a sequence of discrete logical units, where each unit represents a self-contained conceptual or computational step (e.g., defining a subgoal, deriving an intermediate result). The process is iterative: we provide the LLM (*Instruct*) with the first unit as a contextual prefix and task it with solving the problem. If it fails across four generated samples, we expand the context to include the next unit and repeat the process. This continues until a correct solution is obtained, thereby identifying the minimal reasoning scaffold the LLM required to succeed.

To quantify the LLM's behavior, we introduce two metrics: 1) the *Context Ratio*, representing the fraction of the LRM's total reasoning trace provided to the LLM, and 2) the *Generation Ratio*, defined as the token length of the LLM's generated solution divided by the token length of the LRM's remaining ground-truth reasoning and solution.

Main Results. Our findings, depicted in Figure 2, reveal a strong positive correlation between the length of provided reasoning context and the LLM's success, a dependency that is significantly intensified by problem difficulty. We summarize our key conclusions as follows.

LLM's Reliance on Scaffolding Increases with Difficulty. For low-difficulty problems, the LLM exhibits a degree of autonomous reasoning, requiring a median *Context Ratio* of 33% to arrive at a correct solution. However, this capacity falters on high-difficulty problems, where the necessary context dramatically increases to a median of 61%. This suggests the LLM's intrinsic problem-solving ability degrades as complexity rises.

Shift from Problem-Solver to Completion Engine. The steep increase in required context for harder problems indicates a fundamental shift in the LLM's function. Rather than reasoning autonomously, it increasingly acts as a proficient completion engine, merely executing the final, well-defined steps of a plan laid out for it. It succeeds not by reasoning, but by completing a nearly-solved problem.

Internalized vs. Externalized Reasoning. This observed behavior contrasts sharply with the LRM, whose strength lies in its internalized cognitive model. The LRM excels in complex scenarios precisely because it can leverage this internal apparatus, a faculty so robust that it enables correct solutions even when explicit step-by-step generation is suppressed (*NoThink*). While the base LLM needs an external script, the LRM relies on an internal one.

This distinction underscores our central hypothesis: the crucial transformation from an LLM to an LRM is not about learning to produce longer reasoning chains, but about cultivating a compact and flexible internal reasoning capability. This cognitive apparatus can be explicitly unfolded into a verifiable trace (*NoRehink*) or executed implicitly for greater efficiency (*NoThink*).

4 Distill Reasoning Capabilities into Small LLMs

This section investigates the distillation of advanced reasoning abilities from LRMs into smaller LLM counterparts. A fundamental challenge in this process is the inherent difficulty smaller models have in replicating the long and complex reasoning chains produced by larger teacher models. Recent findings by Li et al. (2025) corroborate this, demonstrating that smaller models benefit more from concise reasoning traces and supervision from moderately-sized teachers.

4.1 Single-Stage Reasoning Distillation

A common approach for transferring reasoning skills is single-stage distillation, where a smaller student model is fine-tuned directly on the reasoning traces generated by a more capable teacher model. We design a series of experiments

	AIME 24	AIME 25	AMC 23	MATH 500	Minerva	Olympiad	Average
				Qwen2.5-3B			
Instruct	$4.79_{\pm 2.69}$	$2.50_{\pm 1.20}$	$38.28_{\pm 5.65}$	$62.98_{\pm 1.78}$	$24.63_{\pm 2.19}$	$26.72_{\pm 1.36}$	26.65
NoThink	$5.31_{\pm 2.65}$	$1.15_{\pm 0.53}$	$29.53_{\pm 5.23}$	$59.95_{\pm 1.80}$	$20.08_{\pm 1.99}$	$24.31_{\pm 1.29}$	23.39
NoRethink	$3.42_{\pm 1.44}$	$3.02_{\pm 0.97}$	$26.80_{\pm 4.84}$	$57.73_{\pm 1.77}$	$18.43_{\pm 1.48}$	$22.24_{\pm 1.22}$	21.94
Think	$2.81_{\pm 1.29}$	$3.75_{\pm 1.90}$	$21.80_{\pm 3.89}$	$56.70_{\pm 1.75}$	$14.38_{\pm 1.55}$	$21.16_{\pm 1.19}$	20.08
Think: Solution	$2.81_{\pm 1.30}$	$1.67_{\pm 0.89}$	$25.94_{\pm 4.77}$	$57.73_{\pm 1.82}$	$18.43_{\pm 1.82}$	$21.98_{\pm 1.24}$	21.43
				Qwen2.5-3B-Instruct			
Instruct	$4.06_{\pm 1.79}$	$1.77_{\pm 0.86}$	$35.62_{\pm 5.31}$	$62.82_{\pm 1.79}$	$25.00_{\pm 2.23}$	$27.31_{\pm 1.39}$	26.10
NoThink	$5.10_{\pm 2.48}$	$2.29_{\pm 0.92}$	$28.67_{\pm 4.73}$	$61.20_{\pm 1.78}$	$22.52_{\pm 2.14}$	$25.54_{\pm 1.33}$	24.22
NoRethink	$3.33_{\pm 1.55}$	$3.23_{\pm 1.49}$	$25.94_{\pm 4.67}$	$58.43_{\pm 1.75}$	$19.21_{\pm 1.88}$	$22.24_{\pm 1.18}$	22.06
Think	$3.02_{\pm 1.73}$	$4.48_{\pm 2.29}$	$24.22_{\pm 3.99}$	$56.95_{\pm 1.76}$	$15.12_{\pm 1.58}$	$21.15_{\pm 1.19}$	20.82
Think: Solution	$3.02_{\pm 1.41}$	$2.40_{\pm 1.29}$	$28.52_{\pm 5.19}$	$57.37_{\pm 1.82}$	$20.31_{\pm 2.00}$	$21.56_{\pm 2.15}$	22.20

Table 1: Performance of `Qwen2.5-3B` and `Qwen2.5-3B-Instruct` student models after single-stage reasoning distillation. Models are fine-tuned on datasets generated by a 32B LLM teacher (*Instruct*) and a 32B LRM teacher under various conditions (*NoThink*, *NoRethink*, and *Think*). We report Pass@1 scores ($\pm$ means standard deviation) across six mathematical reasoning benchmarks.

to evaluate the effectiveness of this paradigm. We distill reasoning capabilities from two teacher models, an LLM `Qwen2.5-32B-Instruct`, and an LRM `DeepSeek-R1-Distill-Qwen-32B`, into smaller student models, `Qwen2.5-3B` and `Qwen2.5-3B-Instruct`. Crucially, both teacher models originate from the same foundational model, `Qwen2.5-32B`, ensuring they share a common basis of world knowledge. However, their distinct post-training strategies result in different reasoning styles. This controlled setup allows us to isolate and evaluate the impact of the teacher's reasoning approach on the distillation process.

Reasoning Traces Extraction. We derive our distillation data from the 7,500 problems in the `MATH` training set (Hendrycks et al., 2021). To generate solutions for each problem, we prompted two teacher models and used rejection sampling to ensure correctness, verifying each output with the Math-Verify toolkit (Face, 2025). Our generation process was configured to maximize the yield of correct solutions, using a temperature of 0.6, a top-p of 0.95, a maximum sequence length of 16,384 tokens, and up to four sampling attempts per problem. We prompted the `Qwen2.5-32B-Instruct` teacher in its standard *Instruct* mode and the `DeepSeek-R1-Distill-Qwen-32B` teacher using the *Think*, *NoRethink*, and *NoThink* modes. The final dataset comprises the 6,445 problems for which all prompting modes yielded a correct solution. For a more comprehensive comparison, this dataset also includes the final condensed solutions from the LRM's *Think* mode.

Training and Evaluations. We instruction-tuned both *3B* student models for two epochs using the LlamaFactory framework (Zheng et al., 2024), a cosine learning rate scheduler, and a maximum learning rate of 1×10^{-5}. To evaluate performance, we used a suite of mathematical reasoning benchmarks: `AIME 2024` (Art of Problem Solving, 2024), `AIME 2025` (Art of Problem Solving, 2025), `AMC 2023` (Art of Problem Solving, 2023), `MATH 500` (Hendrycks et al., 2021), `Minerva` (Lewkowycz et al., 2022), and `Olympiad Bench` (He et al., 2024). Our evaluation protocol involved generating multiple samples per problem with a temperature of 0.6, a top-p of 0.95, and a maximum sequence length of 16,384 tokens. We generated 32 samples for each problem in the `AIME` and `AMC` benchmarks and 8 samples for all other benchmarks. Performance is reported using the Pass@1 metric across all datasets.

Experimental Results. The experimental results, detailed in Table 1, demonstrate a fundamental trade-off between teaching for general proficiency and specialized skill. This tension highlights the

limitations of a direct imitation approach and motivates the need for a more structured distillation strategy. Our key conclusions are as follows.

A Foundational Trade-off Between Teachers. We observe a distinct trade-off between teacher models. The LLM teacher (*Instruct*) consistently produces the best-performing student models on average, indicating that its straightforward chain-of-thought format provides a highly effective and learnable foundation for general reasoning. In contrast, distillation from the LRM's complex, explicit reasoning (*Think*) proves to be the least effective method, confirming that small models cannot easily replicate long, intricate logical chains. This establishes a clear dilemma: the teacher with the most comprehensible reasoning style lacks the specialized knowledge for the hardest problems, while the expert teacher's explicit thoughts are too complex to be learned effectively through simple imitation.

Implicit Heuristics are More Transferable than Explicit Traces. A promising finding is the relative success of the *NoThink* paradigm. While still lagging the *Instruct* baseline on average, it outperforms all other LRM-based methods and shows a competitive advantage on the most difficult problems (`AIME 24` and `AIME 25`). This critical result suggests that the student model can successfully internalize the LRM's advanced problem-solving heuristics when presented in a concise, implicit format. The core value being transferred is not the step-by-step articulation of the reasoning, but the underlying heuristic leap required to solve the problem.

The results indicate that single-stage distillation is a suboptimal method for transferring advanced reasoning. Effective knowledge transfer necessitates two elements not found in any single teacher model: first, a foundational capacity for explicit, procedural reasoning on general tasks, and second, exposure to the sophisticated, implicit heuristics of an expert for challenging tasks. A naive combination data from these distinct paradigms is prone to failure because the student may fail to reconcile the divergent reasoning processes. Therefore, a curricular learning approach is warranted, one that first establishes a robust reasoning foundation before systematically introducing more advanced, implicit heuristics.

4.2 Curriculum Reasoning Distillation

To resolve the challenges identified in the single stage approach, we propose a curriculum based distillation framework. This methodology draws inspiration from the educational principle of the *Zone of Proximal Development*, which suggests that learning is most effective when complex concepts are introduced after foundational knowledge has been established. Our framework implements this concept through a two stage process designed to develop a robust internal reasoning model within the student, rather than promoting the superficial imitation of reasoning traces.

Stage 1: Foundational Distillation of Implicit Heuristics. The initial stage of the curriculum focuses on establishing a strong foundation of implicit problem solving heuristics. We construct a composite dataset that strategically combines data from both teacher models. For problems of low to moderate difficulty (Levels 1 through 4), we use the effective and learnable chain of thought data from the LLM teacher (*Instruct*). For the most challenging problems (Level 5, which constitute approximately 24% of the data), we substitute this with the LRM's concise *NoThink* solutions. This selective data composition introduces the LRM's advanced, dense reasoning patterns only on tasks where they are most critical. The design compels the student model to internalize sophisticated heuristics without being overwhelmed by long, explicit reasoning chains.

Stage 2: Unfolding Implicit Knowledge into Explicit Reasoning. Building upon the foundation established in the first stage, the second stage teaches the student model to articulate its acquired implicit knowledge as an explicit and verifiable chain of thought. The learning objective shifts from internalizing heuristics to externalizing them. To achieve this, we fine tune the Stage 1 model on a revised data mixture. For less complex problems, we reinforce efficient problem solving using the LRM's *NoThink* outputs. For the most complex Level 5 problems, we introduce the LRM's externalized reasoning traces from the *NoRethink* paradigm. Since the student has already been primed with the appropriate implicit heuristics for these problems, it is not learning the complex reasoning without prior conceptual grounding. Instead, it learns to verbalize a thought process founded on logic it has already begun to master. This progressive transition from implicit intuition to explicit articulation aims to foster a genuine and flexible reasoning capability that emulates the LRM's own functionality.

	AIME 24	AIME 25	AMC 23	MATH 500	Minerva	Olympiad	Average
			Qwen2.5-3B				
Instruct	$4.79_{\pm 2.69}$	$2.50_{\pm 1.20}$	$38.28_{\pm 5.65}$	$62.98_{\pm 1.78}$	$24.63_{\pm 2.19}$	$26.72_{\pm 1.36}$	26.65
NoThink	$5.31_{\pm 2.65}$	$1.15_{\pm 0.53}$	$29.53_{\pm 5.23}$	$59.95_{\pm 1.80}$	$20.08_{\pm 1.99}$	$24.31_{\pm 1.29}$	23.39
Mix-Long	$4.06_{\pm 2.60}$	$1.87_{\pm 0.82}$	$34.14_{\pm 5.40}$	$61.52_{\pm 1.74}$	$23.62_{\pm 2.08}$	$26.11_{\pm 1.34}$	25.22
Stage 1: Instruct + NoThink							
+ Stage2: NoThink Only	$6.15_{\pm 3.05}$	$2.91_{\pm 1.45}$	$38.81_{\pm 5.51}$	$63.43_{\pm 1.75}$	$26.13_{\pm 2.20}$	$28.67_{\pm 1.45}$	27.68
+ Stage2: NoThink + Think	$6.43_{\pm 3.01}$	$3.05_{\pm 1.51}$	$38.23_{\pm 5.25}$	$62.81_{\pm 1.75}$	$25.82_{\pm 2.32}$	$28.99_{\pm 1.47}$	27.56
+ Stage2: NoThink + NoRethink	$6.98_{\pm 3.12}$	$3.44_{\pm 1.65}$	$39.43_{\pm 5.48}$	$64.02_{\pm 1.74}$	$26.88_{\pm 2.31}$	$29.51_{\pm 1.41}$	28.38
			Qwen2.5-3B-Instruct				
Instruct	$4.06_{\pm 1.79}$	$1.77_{\pm 0.86}$	$35.62_{\pm 5.31}$	$62.82_{\pm 1.79}$	$25.00_{\pm 2.23}$	$27.31_{\pm 1.39}$	26.10
NoThink	$5.10_{\pm 2.48}$	$2.29_{\pm 0.92}$	$28.67_{\pm 4.73}$	$61.20_{\pm 1.78}$	$22.52_{\pm 2.14}$	$25.54_{\pm 1.33}$	24.22
Mix-Long	$3.75_{\pm 1.77}$	$1.98_{\pm 1.11}$	$35.00_{\pm 5.62}$	$62.38_{\pm 1.77}$	$25.60_{\pm 2.20}$	$26.41_{\pm 1.35}$	25.85
Stage 1: Instruct + NoThink							
+ Stage2: NoThink Only	$5.91_{\pm 2.31}$	$2.66_{\pm 1.20}$	$36.54_{\pm 4.72}$	$63.95_{\pm 1.79}$	$26.15_{\pm 2.28}$	$28.95_{\pm 1.46}$	27.36
+ Stage2: NoThink + Think	$6.02_{\pm 2.80}$	$2.81_{\pm 1.35}$	$36.19_{\pm 5.15}$	$63.40_{\pm 1.76}$	$25.74_{\pm 2.10}$	$29.21_{\pm 1.49}$	27.23
+ Stage2: NoThink + NoRethink	$6.44_{\pm 2.19}$	$3.01_{\pm 1.42}$	$37.19_{\pm 5.11}$	$64.21_{\pm 1.76}$	$27.05_{\pm 2.33}$	$29.86_{\pm 1.32}$	27.96

Table 2: Performance comparison of the two-stage curriculum distillation framework against single-stage baselines. We report Pass@1 scores ($\pm$ means standard deviation) for the `Qwen2.5-3B` and `Qwen2.5-3B-Instruct` student models across six mathematical reasoning benchmarks. The single-stage baselines represent the best-performing methods from Table 1.

Experimental Setup. We follow the training and evaluation protocol established in our single-stage experiments to facilitate a direct comparison. Our two-stage training procedure is configured as follows. For Stage 1, we use solutions from the `Qwen2.5-32B-Instruct` teacher (*Instruct*) for `MATH` problems of difficulty levels 1-4. For the most difficult problems (level 5), we incorporate solutions from the `DeepSeek-R1-Distill-Qwen-32B` teacher's *NoThink* paradigm. Subsequently, in Stage 2, the training data comprises the LRM teacher's *NoThink* outputs for levels 1-4 and its explicit reasoning traces (*NoRethink*) for level 5. Our baseline is the *Mix-Long* method (Li et al., 2025), which randomly combines *Instruct* outputs with *Think* outputs in a 4:1 ratio. We conduct evaluations on six mathematical reasoning benchmarks, employing uniform sampling parameters and reporting the Pass@1 score.

Experimental Results. The experimental results, presented in Table 2, validate the effectiveness of our two-stage curriculum distillation framework. This structured approach consistently yields student models that outperform all single-stage baselines, confirming that a progressive learning strategy is more effective than direct imitation for transferring complex reasoning skills. We summarize our key findings below.

Curriculum Distillation Substantially Outperforms Baselines. Our curriculum-trained models achieve a clear and significant performance improvement over the strongest single-stage methods. For instance, our optimal `Qwen2-3B` model reaches an average Pass@1 score of 28.38 across all benchmarks. This is a substantial gain compared to the 26.65 score from the best single-stage baseline, which was trained on the LLM teacher's *Instruct* data. This outcome demonstrates the value of a carefully scaffolded educational progression.

Pruned Reasoning is the Optimal Scaffold for Articulation. The primary advantage of our design is most evident in the Stage 2 results. The strategy of using streamlined *NoRethink* traces to teach the articulation of complex reasoning yields the best performance. This approach surpasses the alternatives of either reinforcing implicit knowledge with more *NoThink* data or introducing the unabridged reasoning from the *Think* paradigm. After Stage 1, the student model has already acquired the implicit problem-solving heuristics for hard problems. At this point, the verbose and often redundant *Think* traces are suboptimal, as they overwhelm the model instead of refining its skills.

The Curriculum Successfully Bridges Intuition and Expression. The success of the *NoRethink* traces in Stage 2 confirms our core hypothesis. The streamlined traces act as a ideal conceptual scaffold. They provide a clean, logical template that

teaches the student how to effectively structure and externalize the implicit knowledge it has already acquired. By first fostering the internalization of heuristics and then teaching the model to articulate that knowledge using concise examples, we cultivate a more robust reasoning capability that is challenging to achieve with single-stage methods.

5 Conclusion

In this work, we demonstrate a fundamental asymmetry between LRMs and conventional LLMs, which originates from the LRM's intrinsic capacity for implicit self-distillation. We show that this disparity causes direct distillation to fail by creating an intractable learning problem: the LRM's explicit reasoning traces are too complex, while its implicit solutions are too information-dense for a student model to effectively learn from. To address this challenge, we propose a novel curriculum distillation framework. Our two-stage method first trains a student on an LRM's implicit solutions to build a foundational reasoning capability, then fine-tunes it on explicit reasoning traces to develop explicit reasoning skills. Empirical evaluation confirms that our method significantly outperforms single-stage baselines, yielding compact models that successfully inherit the dual-process reasoning of advanced LRMs and mitigate the trade-off between analytical depth and inference efficiency.

Limitations

Our study has several limitations. First, our investigation is confined to mathematical reasoning, and the generalizability of our curriculum distillation framework to other domains, such as code generation, instruction following, or multi-modal reasoning, remains unexplored. Second, we focus on reasoning paradigms and distillation strategies, without systematically investigating how variations in data diversity or annotation style might affect the transfer of reasoning skills. Third, our experiments are limited to English-language benchmarks; the effectiveness of our approach in multilingual or culturally diverse settings is yet to be examined. Finally, while our framework improves the efficiency of smaller models, it presupposes access to powerful LRMs as teachers, which may not be feasible in resource-constrained environments.

Acknowledgments

We sincerely than the reviewers for their insightful comments and valuable suggestions. This work was supported by National Key R&D Program of China (2024YFC3308000), Beijing Natural Science Foundation (L243006), the Natural Science Foundation of China (No. 62306303, 62476265), the Basic Research Program of ISCAS (Grant No. ISCAS-ZD-202402, ISCAS-JCZD-202303).

References

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4.

Art of Problem Solving. 2023. 2023 amc 12a.

Art of Problem Solving. 2024. 2024 aime i.

Art of Problem Solving. 2025. 2025 aime i.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 117 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do not think that much for 2+3=? on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.

Xiaoxue Cheng, Junyi Li, Zhenduo Zhang, Xinyu Tang, Wayne Xin Zhao, Xinyu Kong, and Zhiqiang Zhang. 2025. Incentivizing dual process thinking for efficient large language model reasoning. *Preprint*, arXiv:2505.16315.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,

Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Hugging Face. 2025. Math-Verify.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025. Think only when you need with large hybrid-reasoning models. *Preprint*, arXiv:2505.14631.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *Preprint*, arXiv:2504.09858.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, and 109 others. 2025. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *Preprint*, arXiv:2506.13585.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Ziqing Qiao, Yongheng Deng, Jiali Zeng, Dong Wang, Lai Wei, Fandong Meng, Jie Zhou, Ju Ren, and Yaoxue Zhang. 2025. Concise: Confidence-guided compression in step-by-step efficient reasoning. *Preprint*, arXiv:2505.04881.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

ByteDance Seed, :, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiying Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, and 255 others. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *Preprint*, arXiv:2504.13914.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. 2025. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. *Preprint*, arXiv:2506.08343.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. 2025. Scalable chain of thoughts via elastic reasoning. *Preprint*, arXiv:2505.05315.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.