

Small Models Struggle to Learn from Strong Reasoners

Yuetai Li[♣] Xiang Yue[◇] Zhangchen Xu[♣] Fengqing Jiang[♣] Luyao Niu[♣]

Bill Yuchen Lin[♣] Bhaskar Ramasubramanian[♠] Radha Poovendran[♣]

♣University of Washington ◇Carnegie Mellon University ♠Western Washington University
{yuetaili, zxu9, fqjiang, luyaoni, byuchen1, rp3}@uw.edu,
xyue2@andrew.cmu.edu, ramasub@wwu.edu

Abstract

Large language models (LLMs) excel in complex reasoning tasks, and distilling their reasoning capabilities into smaller models has shown promise. However, we uncover an interesting phenomenon, which we term the *Small Model Learnability Gap*: small models ($\leq 3B$ parameters) do not consistently benefit from long chain-of-thought (CoT) reasoning or distillation from larger models. Instead, they perform better when fine-tuned on shorter, simpler reasoning chains that better align with their intrinsic learning capacity. To address this, we propose Mix Distillation, a simple yet effective strategy that balances reasoning complexity by combining long and short CoT examples or reasoning from both larger and smaller models. Our experiments demonstrate that Mix Distillation significantly improves small model reasoning performance compared to training on either data alone. These findings highlight the limitations of direct strong model distillation and underscore the importance of adapting reasoning complexity for effective reasoning capability transfer.

1 Introduction

Large language models (LLMs) (Anthropic, 2023; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a) have demonstrated remarkable performance in complex reasoning tasks, enabling

advancements in mathematical problem-solving, logical inference, and structured decision-making (Cobbe et al., 2021; Shao et al., 2024; Yang et al., 2024). A key advancement in improving LLM complex reasoning capability is the chain-of-thought (CoT) prompting. This technique decomposes complex problems into intermediate reasoning steps, enhancing both performance and interpretability. (Wei et al., 2023).

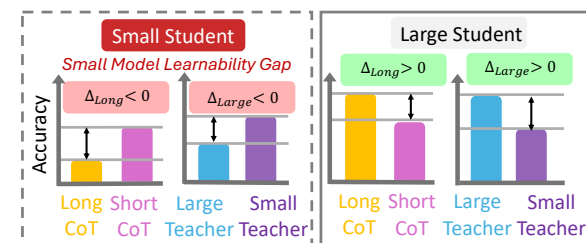


图 1: Small student models ($\leq 3B$ parameters) do not consistently benefit from long CoT reasoning or distillation from large teacher models. Instead, they perform better when fine-tuned on shorter CoT reasoning or distilled from smaller teachers, which better matches their intrinsic learning capacity. We term this phenomenon the *Small Model Learnability Gap*.

However, the high computational cost of LLMs hinders their deployment on resource-constrained devices, motivating the development of smaller models that offer similar capabilities at reduced cost. A widely adopted strategy to achieve this is distillation (Agarwal et al., 2024; Hinton et

Small Models Struggle to Learn from Strong Reasoners

Yuetai Li[♣] Xiang Yue[◇] Zhangchen Xu[♣] Fengqing Jiang[♣] Luyao Niu[♣]

Bill Yuchen Lin[♣] Bhaskar Ramasubramanian[♠] Radha Poovendran[♣]

♣University of Washington ◇Carnegie Mellon University ♠Western Washington University
{yuetaili, zxu9, fqjiang, luyaoni, byuchen1, rp3}@uw.edu,
xyue2@andrew.cmu.edu, ramasub@wwu.edu

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址 https://github.com/binary-husky/gpt_academic/。项目在线体验地址 <https://auth.gpt-academic.top/>。当前大语言模型: Qwen2.5-72B-Instruct, 当前语言模型温度设定: 0.3。为了防止大语言模型的意外谬误产生扩散影响, 禁止移除或修改此警告。

大型语言模型 (LLMs) 在复杂的推理任务中表现出色, 将它们的推理能力提炼到较小的模型中也显示出希望。然而, 我们发现了一个有趣的现象, 我们称之为小模型学习能力差距: 小型模型 ($\leq 3B$ 参数) 从长链思考 (CoT) 推理或从大型模型中提炼并不总能受益。相反, 当它们在较短、较简单的推理链上进行微调时, 表现更好, 这些推理链更符合它们的内在学习能力。为了解决这一问题, 我们提出了混合提炼 (Mix Distillation), 这是一种简单而有效的策略, 通过结合长链和短链CoT示例或从大型和小型模型中推理, 来平衡推理复杂性。我们的实验表明, 与单独使用任何一种数据进行训练相比, 混合提炼显著提高了小型模型的推理性能。这些发现突显了直接强模型提炼的局限性, 并强调了适应推理复杂性以有效转移推理能力的重要性。

1 Introduction

大型语言模型 (LLMs) (Anthropic, 2023; Brown et al., 2020; OpenAI, 2023; Touvron et al.,

2023a) 在复杂的推理任务中展示了卓越的性能, 推动了数学问题解决、逻辑推理和结构化决策的进展 (Cobbe et al., 2021; Shao et al., 2024; Yang et al., 2024)。在提高 LLM 复杂推理能力的关键进展之一是链式思维 (CoT) 提示。这种技术将复杂问题分解为中间推理步骤, 提高了性能和可解释性 (Wei et al., 2023)。

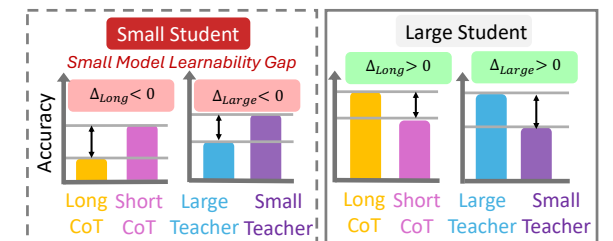


图 1: 小模型 ($\leq 3B$ 参数) 并不能始终从长链路推理 (CoT reasoning) 或从大模型蒸馏中获益。相反, 当它们在较短的链路推理上进行微调或从较小的教师模型中蒸馏时, 表现会更好, 这更符合它们的内在学习能力。我们将这种现象称为小模型学习能力差距。

然而, 大型语言模型 (LLMs) 的高计算成本阻碍了它们在资源受限设备上的部署, 这促使了开发更小的模型, 这些模型以较低的成本提供类似的功能。实现这一目标的一种广泛采用的策略是蒸馏 (Agarwal et al., 2024; Hinton et al., 2015; Kim et al., 2024a), 其中由强大教师模型生成的CoT序列用于微调较弱的学生模型。自然地, 人们可能会期望从更强模型蒸馏出的CoT序列会一致地提高小型模型的复杂推

al., 2015; Kim et al., 2024a), where CoT sequences generated by a strong teacher model are used to fine-tune a weaker student model. Naturally, one might expect that distilling CoT sequences from stronger models would consistently improve small models’ complex reasoning capabilities (Agarwal et al., 2024; DeepSeek-AI et al., 2024; Min et al., 2024; Tunstall et al., 2023).

However, we reveal an interesting phenomenon, which we term the *Small Model Learnability Gap* (Fig. 1): small models do not consistently benefit from the complex reasoning sequences provided by strong teachers, such as long CoT reasoning or distillation from large models. In our experiments, we observe that when small models are exposed to long and intricate reasoning traces, they struggle to internalize the multi-step logic due to their constrained ability. Instead, small models perform better when fine-tuned on *shorter, simpler reasoning chains* that align more closely with their intrinsic learning capacity. This suggests that small models struggle to process overly elaborate reasoning traces or adapt to the distribution shifts introduced by stronger teachers, ultimately limiting their ability to generalize effectively.

To address the challenge described above, we propose *Mix Distillation*, a simple yet effective approach that balances reasoning complexity by blending different types of reasoning traces. Specifically, our method comprises two configurations: (1) *Mix-Long* – A combination of long and short CoT examples, ensuring that small models are exposed to both detailed and concise reasoning steps. (2) *Mix-Large* – A mixture of responses from both larger and smaller models, allowing small models to learn from reasoning chains that are better suited to their capacity.

Our experiments demonstrate that *Mix Distillation* consistently improves small model reasoning performance compared to standard distillation.

For instance, Qwen2.5-3B-Instruct improves by more than 8 points on MATH and AMC using Mix-Long, compared to direct training on long CoT data. Qwen2.5-3B-Instruct gains more than 7 points on MATH, AIME and AMC using Mix-Large compared with training on large teacher CoT data.

These findings highlight a fundamental limitation of direct strong model distillation and emphasize the importance of *adapting reasoning complexity* for effective knowledge transfer. By carefully designing distillation strategies, we provide new insights into overcoming the constraints of small model learning, making them more effective at reasoning-intensive tasks.

2 Preliminaries

2.1 Notation

Let $x = (x_1, x_2, \dots, x_n)$ represent an input sequence (e.g., a prompt), and $y = (y_1, y_2, \dots, y_m)$ be the corresponding output sequence. We consider a LLM parameterized by θ , which predicts the next token following a conditional distribution $\pi_\theta(y_t | x, y_{1:t-1})$. We denote by $\text{CoT}(y) \subseteq y$ the subset of tokens in the generated output that encodes a *chain-of-thought*, often serving as a reasoning trace or explanatory sequence.

Throughout this work, we use the term **short CoT**, to describe concise reasoning paths to arrive at solutions (Min et al., 2024; Yeo et al., 2025) and **long CoT** to describe an extended reasoning sequence that is not only longer but also demonstrates more complex reflective thoughts (Qwen, 2024b; Yeo et al., 2025). Additionally, we use the term **large teacher CoT** to refer to the reasoning trace generated by a larger teacher model, and the term **small teacher CoT** for the reasoning steps produced by a smaller teacher model. Please see Appendix D for more examples.

能力 (Agarwal et al., 2024; DeepSeek-AI et al., 2024; Min et al., 2024; Tunstall et al., 2023)。

然而，我们揭示了一个有趣的现象，我们称之为 **小型模型学习差距** (图 1)：小型模型并不总是从强大教师提供的复杂推理序列中受益，例如长的CoT推理或从大型模型蒸馏。在我们的实验中，我们观察到，当小型模型接触到长而复杂的推理轨迹时，由于其能力的限制，它们难以内化多步骤逻辑。相反，小型模型在微调时使用 **较短、较简单的推理链** 时表现更好，这些推理链更符合它们的内在学习能力。这表明小型模型难以处理过于复杂的推理轨迹或适应由更强教师引入的分布变化，最终限制了它们的有效泛化能力。

为了解决上述挑战，我们提出了一种简单而有效的 **混合蒸馏** 方法，该方法通过混合不同类型的推理轨迹来平衡推理复杂性。具体而言，我们的方法包括两种配置：(1) **混合长 – 长短CoT示例的组合**，确保小型模型同时接触到详细和简洁的推理步骤。(2) **混合大 – 较大和较小模型响应的混合**，使小型模型能够从更适合其能力的推理链中学习。

我们的实验表明，与标准蒸馏相比，混合蒸馏一致地提高了小型模型的推理性能。例如，Qwen2.5-3B-Instruct 在MATH和AMC上使用混合长时比直接在长CoT数据上训练提高了超过8分。Qwen2.5-3B-Instruct 在MATH、AIME和AMC上使用混合大时比在大型教师CoT数据上训练提高了超过7分。

这些发现突显了直接强模型蒸馏的根本局限性，并强调了 **适应推理复杂性** 对于有效知识转移的重要性。通过精心设计蒸馏策略，我们为克服小型模型学习的限制提供了新的见解，使它们在推理密集型任务中更加有效。

2 Preliminaries

2.1 Notation

令 $x = (x_1, x_2, \dots, x_n)$ 表示一个输入序列 (例如，一个提示)， $y = (y_1, y_2, \dots, y_m)$

为相应的输出序列。我们考虑一个由 θ 参数化的大型语言模型 (LLM)，该模型根据条件分布 $\pi_\theta(y_t | x, y_{1:t-1})$ 预测下一个标记。我们用 $\text{CoT}(y) \subseteq y$ 表示生成输出中编码 **思维链** 的标记子集，通常作为推理轨迹或解释序列。

在本文中，我们使用术语 **短 CoT** 来描述简洁的推理路径，以达到解决方案 (Min et al., 2024; Yeo et al., 2025)，并使用术语 **长 CoT** 来描述不仅更长而且展示更复杂反思思维的扩展推理序列 (Qwen, 2024b; Yeo et al., 2025)。此外，我们使用术语 **大教师 CoT** 来指代由较大教师模型生成的推理轨迹，以及术语 **小教师 CoT** 来指代较小教师模型生成的推理步骤。更多示例请参见附录 D。

2.2 Supervised Fine-Tuning (SFT)

监督微调 (SFT) 被广泛采用以增强大型语言模型 (LLMs) 在数据集 $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$ 上的推理能力，其中 y^i 可以是短的CoT、长的CoT、强模型的CoT或弱模型的CoT序列。SFT过程通过最小化指令数据集 \mathcal{D} 上的负对数似然损失来更新语言模型的参数 θ 。

3 Small Model Learnability Gap

在本节中，我们使用不同的CoT数据对student模型进行微调。然后，我们通过微调模型的性能揭示了小模型的学习能力差距。

3.1 Experiment Setup

数据集。 我们使用了MATH (Hendrycks et al., 2021) 的7,500个提示集。该数据集涵盖了高级微积分、几何和线性代数等七个数学主题。

学生模型。 我们的研究考虑了来自Qwen (Qwen, 2024a) 和 Llama (Meta, 2024a,b) 模型家族的十个学生模型，这些模型的规模各不相同。这些模型包括 Qwen2.5-0.5B、Qwen2.5-1.5B、Qwen2.5-3B、Qwen2.5-7B、Qwen2.5-14B 和 Qwen2.5-32B 的指令版本，以及 Llama3.2-1B、Llama3.2-3B、Llama3.1-8B

2.2 Supervised Fine-Tuning (SFT)

Supervised fine-tuning (SFT) is widely adopted to enhance reasoning capabilities of LLMs on a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$, where y^i can be short CoT, long CoT, strong model CoT or weak model CoT sequences. The SFT process updates the parameters θ of a language model by minimizing the negative log-likelihood loss over the instruction dataset \mathcal{D} .

3 Small Model Learnability Gap

In this section, we fine-tune student models using different CoT data. We then reveal the small model learnability gap given the performance of fine-tuned models.

3.1 Experiment Setup

Datasets. We use the 7,500 prompt set of MATH (Hendrycks et al., 2021). This dataset encompasses seven math topics such as advanced calculus, geometry, and linear algebra.

Student models. Our study considers ten student models from the Qwen (Qwen, 2024a) and Llama (Meta, 2024a,b) model families of varying sizes. These models include the Instruct version of Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, and Qwen2.5-32B, and the Instruct version of Llama3.2-1B, Llama3.2-3B, Llama3.1-8B, and Llama3.3-70B. A comprehensive overview of the student models is presented in Table 4 of Appendix A.

Teacher models. To compare long CoT with short CoT, we use QwQ-32B-Preview (Qwen, 2024b) to generate long CoT sequences and Qwen2.5-32B-Instruct as the response generator for short CoT. Within each model family, we designate the larger scale model as the large teacher and the smaller scale model as the small teacher. This includes Qwen2.5-72B-Instruct

vs Qwen2.5-3B-Instruct, Llama3.1-70B-Instruct vs Llama3.1-8B-Instruct, and Gemma2-27B-it vs Gemma2-9B-it.

Evaluation Benchmarks. We evaluate the reasoning capability of fine-tuned student models on a set of commonly used benchmarks, including MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), AMC 2023, AIME 2024, and the English math subset of OlympiadBench (He et al., 2024). These benchmarks span a wide range of challenge levels, from elementary mathematics to advanced competition problems. We define the student model performance as the average score on five benchmarks. Unless otherwise specified, all fine-tuned models are evaluated in a zero-shot setting using greedy decoding. We set the maximum generation tokens as 16k. Please see Appendix A for detailed experimental setup.

We define the following performance scores:

- P_{Long} : Performance score of a student model fine-tuned on long CoT data.
- P_{Short} : Performance score of a student model fine-tuned on short CoT data.
- P_{Large} : Performance score of a student model fine-tuned on CoT from a larger teacher.
- P_{Small} : Performance score of a student model fine-tuned on CoT from a smaller teacher.

Training Setup. Teacher models generate responses by rejection sampling (Dong et al., 2023; Gulcehre et al., 2023; Tong et al., 2024; Yuan et al., 2023; Yue et al., 2023; Zelikman et al., 2022). By default, teacher models employ greedy decoding. By combining the math problem instructions with corresponding solutions generated by teacher models, we construct problem-solution pairs to fine-tune student models. We train the models using the

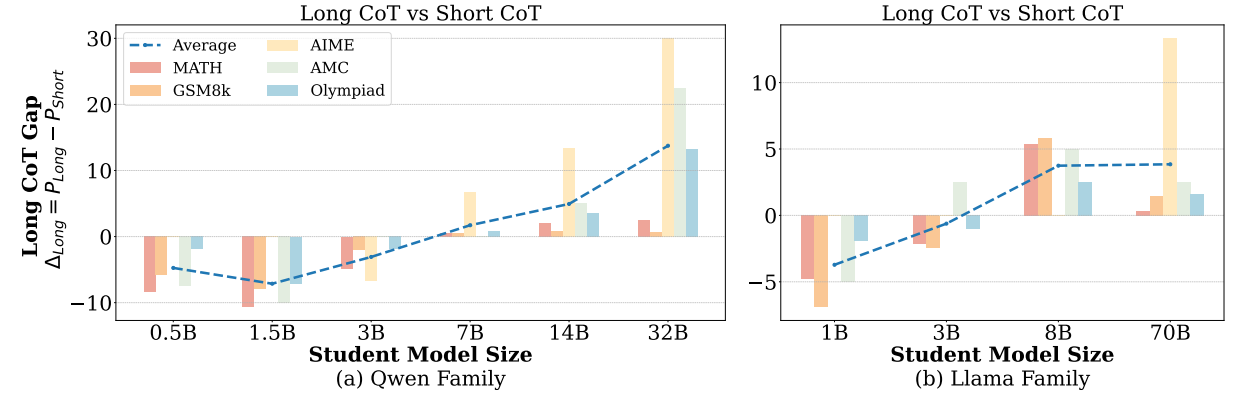


图 2: 长链思考差距 ($\Delta_{Long} = P_{Long} - P_{Short}$) 在不同模型大小的学生模型中, 对于 (a) Qwen 系列 (b) Llama 系列。对于教师模型, 选择 QwQ-32B-Preview 生成长链思考响应, 而选择 Qwen2.5-32B-Instruct 生成短链思考响应。负 (正) Δ_{Long} 表示长链思考比短链思考更差 (更好)。我们的结果表明, 对于较小的学生模型, 短链思考更好 (由 $\Delta_{Long} < 0$ 表示), 而对于较大的学生模型, 长链思考更好 (由 $\Delta_{Long} > 0$ 表示)。

和 Llama3.3-70B 的指令版本。学生模型的详细概述见附录 A 中的表 4。

教师模型. 为了比较长链思维与短链思维, 我们使用 QwQ-32B-Preview (Qwen, 2024b) 生成长链思维序列, 并使用 Qwen2.5-32B-Instruct 作为短链思维的响应生成器。在每个模型家族中, 我们将较大规模的模型指定为大教师, 较小规模的模型指定为小教师。这包括 Qwen2.5-72B-Instruct 与 Qwen2.5-3B-Instruct、Llama3.1-70B-Instruct 与 Llama3.1-8B-Instruct 以及 Gemma2-27B-it 与 Gemma2-9B-it。

评估基准. 我们在一组常用的基准上评估微调学生模型的推理能力, 包括 MATH (Hendrycks et al., 2021)、GSM8K (Cobbe et al., 2021)、AMC 2023、AIME 2024 以及 OlympiadBench (He et al., 2024) 的英语数学子集。这些基准涵盖了从基础数学到高级竞赛题目的广泛难度范围。我们将学生模型的性能定义为在五个基准上的平均得分。除非另有说明, 所有微调模型均在零样本设置下使用贪婪解码进行评估。我们将最大生成标记设置为 16k。详细实验设置请参见附录 A。

我们定义以下性能得分:

- P_{Long} : 在长CoT数据上微调的学生模型的性能得分。
- P_{Short} : 在短CoT数据上微调的学生模型的性能得分。
- P_{Large} : 在较大教师模型上微调的学生模型的性能得分。
- P_{Small} : 较小教师模型微调后学生模型的性能得分。

训练设置. 教师模型通过拒绝采样生成响应 (Dong et al., 2023; Gulcehre et al., 2023; Tong et al., 2024; Yuan et al., 2023; Yue et al., 2023; Zelikman et al., 2022) 默认情况下, 教师模型采用贪婪解码。通过将数学问题指令与教师模型生成的相应解决方案相结合, 我们构建了问题-解决方案对以微调学生模型。我们使用LLaMA-Factory框架 (Zheng et al., 2024) 训练模型。对于规模小于14B的学生模型, 我们使用全参数SFT, 并实施最大学习率为 10^{-5} 的余弦学习率调度, 以微调学生模型两个周期 (Touvron et al., 2023b)。对于大于14B的学生模型, 我们采用LoRA微调, 学习率为 10^{-4} , 微

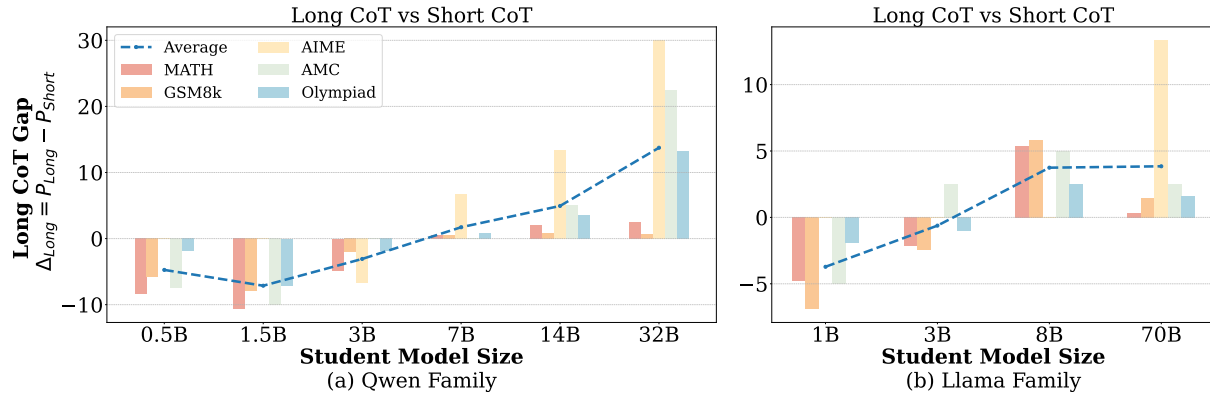


图 2: Long CoT Gap ($\Delta_{Long} = P_{Long} - P_{Short}$) of student models with different model sizes for (a) Qwen family (b) Llama family. For teacher models, QwQ-preview-32B is chosen to generate long CoT responses, while Qwen2.5-32B-Instruct is chosen to generate short CoT responses. Negative (Positive) Δ_{Long} indicates that long CoT is worse (better) than short CoT. Our results demonstrate that short CoT is better for smaller student models (indicated by $\Delta_{Long} < 0$), while long CoT is better for larger student models (indicated by $\Delta_{Long} > 0$).

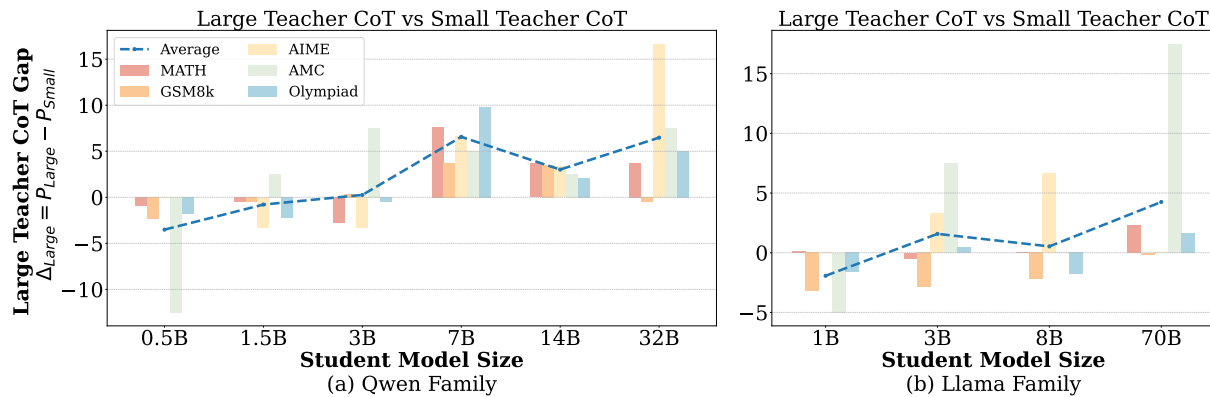


图 3: Large model CoT Gap ($\Delta_{Large} = P_{Large} - P_{Small}$) of student models with different model sizes for (a) Qwen family (b) Llama family. For teacher models, Qwen2.5-72B-Instruct is chosen as the large teacher to generate responses, while Qwen2.5-3B-Instruct is chosen as the small teacher to generate responses. Negative (positive) Δ_{Large} indicates that large teacher CoT is worse (better) than small teacher CoT. Our results demonstrate that small teacher CoT is better for smaller student models (indicated by $\Delta_{Large} < 0$), while large model CoT is better for larger student models (indicated by $\Delta_{Large} > 0$).

LLaMA-Factory framework (Zheng et al., 2024).

For student models of scale less than 14B, we use full-parameter SFT and implement a cosine learning rate schedule with a maximum learning rate of 10^{-5} to fine-tune student models for two epochs (Touvron et al., 2023b). For student models larger than 14B, we adopt LoRA fine-tuning with a learning rate of 10^{-4} for two epochs. Detailed hyperparameters and information about the experimental platform are provided in Appendix A.

3.2 Long CoT Gap

This section evaluates the reasoning capabilities of student models fine-tuned over long CoT data and short CoT data. We quantify the performance difference between long and short CoT data using *long CoT gap* Δ_{Long} , defined as:

$$\Delta_{Long} = P_{Long} - P_{Short}.$$

Figure 2 provides a comprehensive overview of the long CoT gap Δ_{Long} across different student

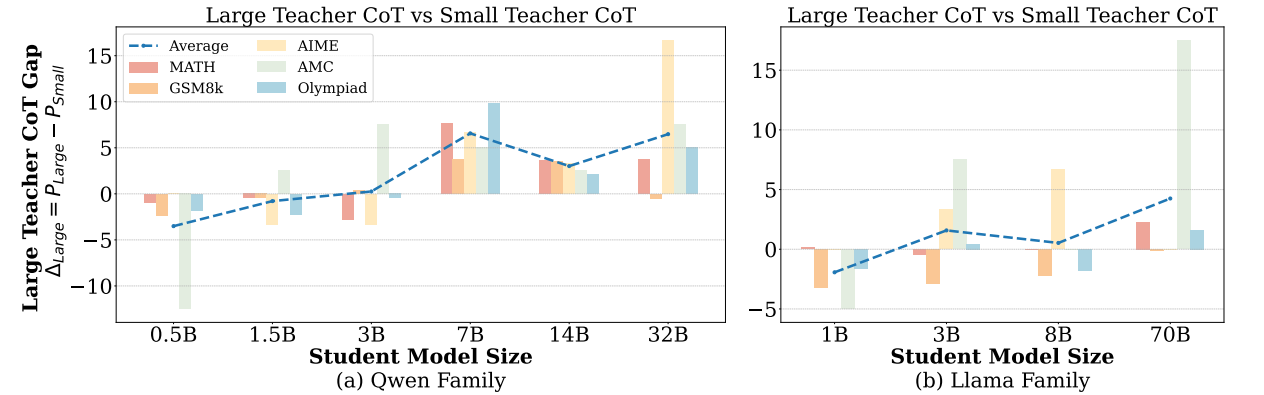


图 3: 大型模型CoT差距 ($\Delta_{Large} = P_{Large} - P_{Small}$) 在不同模型大小的学生模型中, (a) Qwen系列 (b) Llama系列。对于教师模型, 选择Qwen2.5-72B-Instruct作为大型教师生成响应, 而Qwen2.5-3B-Instruct作为小型教师生成响应。负 (正) Δ_{Large} 表示大型教师CoT比小型教师CoT差 (好)。我们的结果表明, 对于较小的学生模型, 小型教师CoT更好 (由 $\Delta_{Large} < 0$ 表示), 而对于较大的学生模型, 大型模型CoT更好 (由 $\Delta_{Large} > 0$ 表示)。

调两个周期。详细的超参数和实验平台信息见附录 A。

3.2 Long CoT Gap

本节评估了在长链思考数据和短链思考数据上微调的学生模型的推理能力。我们使用长链思考差距 Δ_{Long} 来量化长链思考数据和短链思考数据之间的性能差异, 定义为:

$$\Delta_{Long} = P_{Long} - P_{Short}.$$

图 2 提供了不同学生模型在长 CoT 间隙 Δ_{Long} 方面的全面概述。MATH、GSM8K、AIME、AMC 和 OlympiadBench 的详细基准分数详见附录 B 中的表 7。我们报告以下关键结论。

要点 1: 长 CoT 间隔

小型学生模型往往从短 CoT 中受益更多, 而大型学生模型则从长 CoT 中获得更大的优势。

我们观察到, 对于较大的模型, 长链思维 (CoT) 更加有效, 一致地提高了大多数数学基准的性能。例如, 学生模型 Qwen2.5-32B-Instruct 在所有数学指标上平均提高了约 15 分。

然而, 长链思维数据对较小的模型并不有效, 与短链思维相比, 改进幅度显著较小。在 MATH 和 AMC 基准上, 学生模型 Qwen2.5-1.5B-Instruct 在使用长链思维数据微调时, 性能降低了超过 10 分。这表明较小的模型可能无法有效地学习和利用长链思维范式。请参见第 3.4 节以获取更多归因分析。

Student Model	P_{Long}	P_{Short}	Δ_{Long}	Better?
Qwen2.5-0.5B	14.8	19.5	-4.7	Short
Qwen2.5-1.5B	27.0	34.2	-7.1	Short
Qwen2.5-3B	40.3	43.4	-3.1	Short
Qwen2.5-7B	48.9	47.2	1.7	Long
Qwen2.5-14B	59.2	54.3	4.9	Long
Qwen2.5-32B	73.0	59.3	13.7	Long
Llama-3.2-1B	15.8	19.5	-3.7	Short
Llama-3.2-3B	32.5	33.1	-0.6	Short
Llama-3.1-8B	35.2	31.5	3.7	Long
Llama-3.3-70B	58.2	54.3	3.8	Long

表 1: 比较长链思维 (P_{Long}) 和短链思维 (P_{Short}) 微调的平均性能。我们发现, 小型学生模型可能难以从长链思维数据中学习。

3.3 Large Teacher CoT Gap

我们研究小型模型如何从大型教师模型和小型教师模型中学习。我们定义一个大型教

models. The detailed benchmark scores on MATH, GSM8K, AIME, AMC, and OlympiadBench are deferred to Table 7 in Appendix B. We report the following key takeaways.

Takeaway 1: Long CoT Gap

Small student models tend to benefit more from short CoT, while large student models gain greater advantages from long CoT.

We observe that long CoT is more effective for larger models, consistently leading to improved performance across most math benchmarks. For example, the student model Qwen2.5-32B-Instruct improves about 15 points across all math metrics on average.

However, long CoT data is not effective for smaller models, yielding significantly less improvement compared to short CoT. On the MATH and AMC benchmarks, student model Qwen2.5-1.5B-Instruct performs over 10 points lower when fine-tuned with long CoT data. This shows that smaller models may not be able to effectively learn and utilize the long CoT paradigm. Please see more attribution analysis in Section 3.4.

Student Model	P_{Long}	P_{Short}	Δ_{Long}	Better?
Qwen2.5-0.5B	14.8	19.5	-4.7	Short
Qwen2.5-1.5B	27.0	34.2	-7.1	Short
Qwen2.5-3B	40.3	43.4	-3.1	Short
Qwen2.5-7B	48.9	47.2	1.7	Long
Qwen2.5-14B	59.2	54.3	4.9	Long
Qwen2.5-32B	73.0	59.3	13.7	Long
Llama-3.2-1B	15.8	19.5	-3.7	Short
Llama-3.2-3B	32.5	33.1	-0.6	Short
Llama-3.1-8B	35.2	31.5	3.7	Long
Llama-3.3-70B	58.2	54.3	3.8	Long

表 1: Comparison of the average performance between fine-tuning with long CoT (P_{Long}) and short CoT (P_{Short}). We find that small student models may struggle to learn from long CoT data.

Student Model	P_{Large}	P_{Small}	Δ_{Large}	Better?
Qwen2.5-0.5B	16.9	20.4	-3.5	Weak
Qwen2.5-1.5B	32.2	33.0	-0.8	Weak
Qwen2.5-3B	39.7	39.4	0.3	Strong
Qwen2.5-7B	48.9	42.3	6.6	Strong
Qwen2.5-14B	52.9	49.9	3.0	Strong
Qwen2.5-32B	59.5	53.0	6.5	Strong
Llama-3.2-1B	16.5	18.5	-1.9	Weak
Llama-3.2-3B	32.8	31.2	1.6	Strong
Llama-3.2-8B	25.6	25.1	0.5	Strong
Llama-3.2-70B	57.6	53.3	4.3	Strong

表 2: Comparison of average performance between fine-tuning with large teacher CoT (P_{Long}) and small teacher CoT (P_{Small}). We find that small student models may struggle to learn from large teacher CoT data.

3.3 Large Teacher CoT Gap

We investigate how effective small models may learn from large teacher and small teachers. We define a *large teacher CoT gap* as:

$$\Delta_{Large} = P_{Large} - P_{Small}.$$

Figure 3 provides a comprehensive comparison of the Δ_{Large} incurred by all student models. The detailed benchmark scores of MATH, GSM8K, AIME, AMC and OlympiadBench are deferred to Table 8 in Appendix B. More experimental results of different teacher models, including Llama3.1-70B vs Llama3.1-8B and Gemma2-27B vs Gemma2-9B are in Table 9 of Appendix B.

We observe that larger student models learn effectively from large teacher CoT. For example, Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct student models improve over 5 points on average, with Qwen2.5-32B-Instruct achieving more than a 15 point increase on the AIMC benchmark. However, smaller models do not learn effectively from large teacher models such as Qwen2.5-72B-Instruct. Instead, small teacher models such as Qwen2.5-3B-Instruct may serve as better teacher models for small stu-

Student Model	P_{Large}	P_{Small}	Δ_{Large}	Better?
Qwen2.5-0.5B	16.9	20.4	-3.5	Weak
Qwen2.5-1.5B	32.2	33.0	-0.8	Weak
Qwen2.5-3B	39.7	39.4	0.3	Strong
Qwen2.5-7B	48.9	42.3	6.6	Strong
Qwen2.5-14B	52.9	49.9	3.0	Strong
Qwen2.5-32B	59.5	53.0	6.5	Strong
Llama-3.2-1B	16.5	18.5	-1.9	Weak
Llama-3.2-3B	32.8	31.2	1.6	Strong
Llama-3.2-8B	25.6	25.1	0.5	Strong
Llama-3.2-70B	57.6	53.3	4.3	Strong

表 2: 比较大型教师 CoT (P_{Long}) 和小型教师 CoT (P_{Small}) 微调的平均性能。我们发现小型学生模型可能难以从大型教师 CoT 数据中学习。

师 CoT 差距为:

$$\Delta_{Large} = P_{Large} - P_{Small}.$$

图 3 提供了所有学生模型所承受的 Δ_{Large} 的全面比较。MATH、GSM8K、AIME、AMC 和 OlympiadBench 的详细基准分数被推迟到附录 B 的表 8 中。不同教师模型的更多实验结果, 包括 Llama3.1-70B 与 Llama3.1-8B 以及 Gemma2-27B 与 Gemma2-9B 的比较, 见附录 B 的表 9。

我们观察到, 较大的学生模型能够有效地从大型教师 CoT 中学习。例如, Qwen2.5-7B-Instruct 和 Qwen2.5-32B-Instruct 学生模型的平均分数提高了超过 5 分, 其中 Qwen2.5-32B-Instruct 在 AIMC 基准测试中的分数提高了超过 15 分。然而, 较小的模型无法从大型教师模型 (如 Qwen2.5-72B-Instruct) 中有效学习。相反, 较小的教师模型 (如 Qwen2.5-3B-Instruct) 可能更适合较小的学生模型。例如, Qwen2.5-0.5B-Instruct 在 AMC 基准测试中的表现下降了超过 10 分。

需要注意的是, 先前的研究 (Kim et al., 2024b) 也表明, 更强的模型不一定是更好的教师, 强调了响应生成器和教师端因素。我们的工作不同之处在于, 我们将这种现象主要归因于学生模型的大小。

要点 2: 大教师 CoT 差距

小的学生模型倾向于从较小的教师模型中学习得更好, 而大型学生模型则从大型教师模型中受益更多。

3.4 Analysis of Small Model Learnability Gap

领域知识影响可学习性差距。我们观察到, 尽管数学专家模型的模型规模较小, 但在图 4 中, 与通用模型相比, 它们在长 CoT 和大型教师 CoT 数据上的可学习性差距更小。具体来说, 我们比较了学生模型 Qwen2.5-Math-1.5B-Instruct 和 Qwen2.5-1.5B-Instruct 之间的可学习性差距。我们的研究发现, 小规模数学专家模型的长 CoT 差距显著小于通用小模型。此外, 当使用大型教师 CoT 进行微调时, Qwen2.5-Math-1.5B 的性能提升超过了 Qwen2.5-1.5B, 这表明数学专家模型从大型教师 CoT 中受益更大。我们推测, 导致小模型可学习性差距的一个关键因素是小学生模型在领域内的知识有限。我们总结了这一观察结果, 如下所述。

要点 3: 领域知识的影响

小模型的有限领域知识可能阻碍它们从强大的推理教师那里学习。

基础模型表现出更显著的学习能力差距。我们观察到, 在图 5 中, 基础模型通常比指令模型表现出更显著的学习能力差距。这表明, 对于小型基础模型来说, 从长的 CoT 数据或大型教师 CoT 中有效学习更具挑战性。

要点 4: 基础模型 vs 指令模型

小型基础模型比指令模型经历更显著的学习能力差距。

说话风格的转变。我们采用 (Lin et al., 2023) 的方法, 评估在长 CoT 和大型教师 CoT 数

Student Model	Distillation Method	MATH	AMC	GSM8k	Olympiad Bench	AIME	Average
Qwen2.5-3B	Long CoT	56.2	37.5	80.0	24.4	<u>3.3</u>	40.3
	Short CoT	61.0	37.5	82.0	26.4	10.0	43.4
	Strong Model CoT	57.5	35.0	80.0	25.9	0.0	39.7
	Weak Model CoT	60.3	27.5	79.5	26.4	<u>3.3</u>	39.4
	Deepseek-R1-32B (Long CoT)	50.7	20.0	81.2	15.7	0.0	33.5
	<i>Ours</i>						
	Mix-Long	<u>64.7</u>	45.0	<u>81.4</u>	<u>28.6</u>	10.0	45.9
	Mix-Large	65.8	<u>42.5</u>	81.7	29.0	10.0	<u>45.8</u>
Llama3.2-3B	Long CoT	48.7	17.5	75.1	<u>17.6</u>	<u>3.3</u>	32.5
	Short CoT	50.9	15.0	77.5	18.7	<u>3.3</u>	33.1
	Strong Model CoT	47.4	25.0	71.2	16.9	<u>3.3</u>	32.8
	Weak Model CoT	47.9	17.5	74.1	16.4	<u>3.3</u>	31.2
	Deepseek-R1-32B (Long CoT)	48.5	17.5	<u>77.7</u>	16.1	6.7	33.3
	<i>Ours</i>						
	Mix-Long	53.0	<u>22.5</u>	79.4	17.2	<u>3.3</u>	35.1
	Mix-Large	<u>51.8</u>	25.0	76.3	17.2	<u>3.3</u>	<u>34.7</u>

表 3: **Mix Distillation** outperforms the baseline models across most metrics. We use Llama3.2-3B-Instruct and Qwen2.5-3B-Instruct as the student model and 7.5k samples in MATH dataset as the training set. We distill different teacher models to generate responses as the baseline. Our proposed Mix-Long combines long CoT data and normal CoT data in a 1:4 ratio, while Mix-Large combines strong model response and weak model response with the same proportion. Experimental results demonstrate that both Mix-Long and Mix-Large surpass baselines in most evaluation metrics. The highest score is bolded, and the second highest score is underlined.

dent models. For instance, the performance of Qwen2.5-0.5B-Instruct degrades by more than 10 points on the AMC benchmark.

Note that prior studies (Kim et al., 2024b) also demonstrated that stronger models are not necessarily stronger teachers, emphasizing response generator and teacher-side factors. Our work differs in that we attribute this phenomenon primarily to the size of the student model.

Takeaway 2: Large Teacher CoT Gap

Small student models tend to learn better from small teachers, while large student models benefit more from large teachers.

3.4 Analysis of Small Model Learnability Gap

Domain knowledge affects learnability gap.

We observe that math expert models, in spite of small model size, exhibit a smaller learnability gap for both long CoT and large teacher CoT data compared to general models in Figure 4. Specifically, we compare the learnability gaps between the student models Qwen2.5-Math-1.5B-Instruct and Qwen2.5-1.5B-Instruct. Our findings show that the long CoT gap of the small math expert model is significantly smaller than that of general small models. Furthermore, the performance improvement of Qwen2.5-Math-1.5B when finetuned with large teacher CoT exceeds that of Qwen2.5-1.5B, suggesting that math expert models benefit more substantially from large teacher CoT. We conjecture that a key factor leading to the small model learn-

Student Model	Distillation Method	MATH	AMC	GSM8k	Olympiad Bench	AIME	Average
Qwen2.5-3B	Long CoT	56.2	37.5	80.0	24.4	<u>3.3</u>	40.3
	Short CoT	61.0	37.5	82.0	26.4	10.0	43.4
	Strong Model CoT	57.5	35.0	80.0	25.9	0.0	39.7
	Weak Model CoT	60.3	27.5	79.5	26.4	<u>3.3</u>	39.4
	Deepseek-R1-32B (Long CoT)	50.7	20.0	81.2	15.7	0.0	33.5
	<i>Ours</i>						
	Mix-Long	<u>64.7</u>	45.0	<u>81.4</u>	<u>28.6</u>	10.0	45.9
	Mix-Large	65.8	<u>42.5</u>	81.7	29.0	10.0	<u>45.8</u>
Llama3.2-3B	Long CoT	48.7	17.5	75.1	<u>17.6</u>	<u>3.3</u>	32.5
	Short CoT	50.9	15.0	77.5	18.7	<u>3.3</u>	33.1
	Strong Model CoT	47.4	25.0	71.2	16.9	<u>3.3</u>	32.8
	Weak Model CoT	47.9	17.5	74.1	16.4	<u>3.3</u>	31.2
	Deepseek-R1-32B (Long CoT)	48.5	17.5	<u>77.7</u>	16.1	6.7	33.3
	<i>Ours</i>						
	Mix-Long	53.0	<u>22.5</u>	79.4	17.2	<u>3.3</u>	35.1
	Mix-Large	<u>51.8</u>	25.0	76.3	17.2	<u>3.3</u>	<u>34.7</u>

表 3: **混合蒸馏** 在大多数指标上优于基线模型。我们使用 Llama3.2-3B-Instruct 和 Qwen2.5-3B-Instruct 作为学生模型，并使用 MATH 数据集集中的 7.5k 样本作为训练集。我们蒸馏不同的教师模型以生成响应作为基线。我们提出的 Mix-Long 以 1:4 的比例结合了长 CoT 数据和正常 CoT 数据，而 Mix-Large 以相同的比例结合了强模型响应和弱模型响应。实验结果表明，Mix-Long 和 Mix-Large 在大多数评估指标上都超过了基线模型。最高分被加粗，第二高分被 下划线 标记。

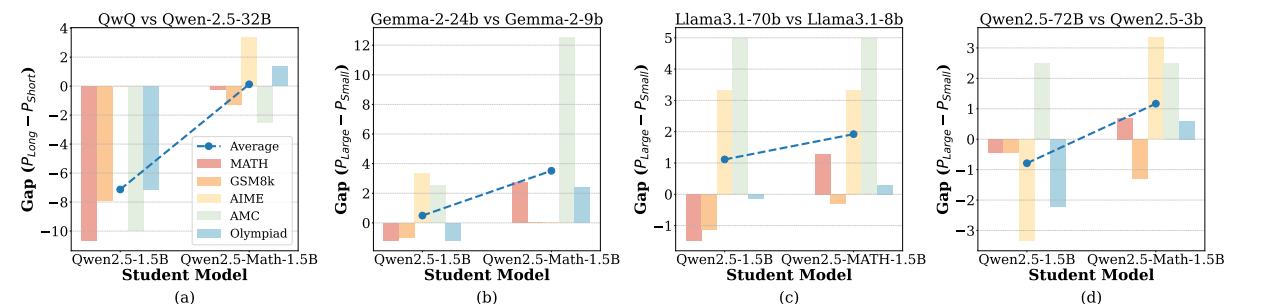


图 4: 数学专家模型通常比通用模型具有更不显著的可学习性差距。正向的差距意味着较长的CoT或较大的教师CoT更好，而负向则意味着更差。这表明数学专家模型能够更容易地从较长的CoT数据或较大的教师CoT中学习。

据上微调前后每个 token 的排名变化。这使我们能够比较微调过程引起的 token 分布变化。然后，我们将排名变化最大的 token 标注为最变化的 token。我们的分析表明，这些 token 主要与表达性和风格元素相关，例如“wait”、“But”和“Let”。更多详情请参见附录 C。

要点 5: 说话风格的变化

长时间的CoT和大规模教师的CoT主要改变了学生模型与说话风格相关的令牌分布。

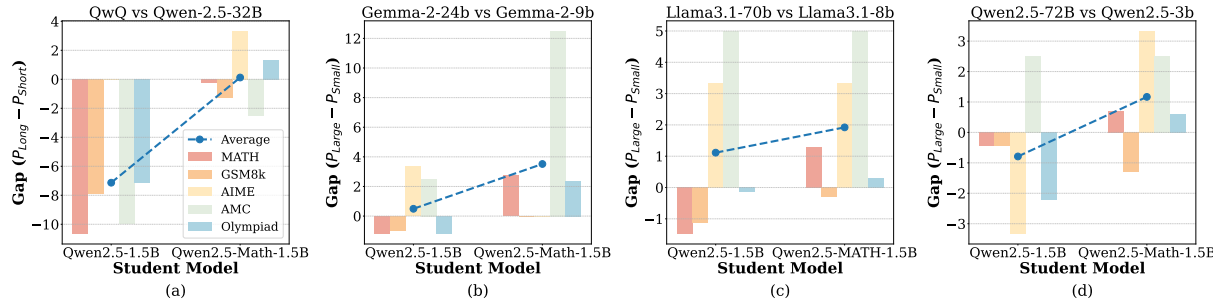


图 4: Math expert models usually have a less significant Learnability Gap than the general models. A positive Gap means long CoT or large teacher CoT is better while negative means worse. This indicates that the math expert model could more easily learn from long CoT data or large teacher CoT.

ability gap is the *limited in-domain knowledge of small student models*. We summarize this observation in the following takeaway.

Takeaway 3: Effect of Domain Knowledge

Limited domain knowledge of small models may hinder their learning from strong reasoning teachers.

Base models exhibit a more significant learnability gap. We observe that base models generally exhibit a more significant learnability gap than Instruct models in Figure 5. This suggests that it is more challenging for small base models to effectively learn from long CoT data or large teacher CoT.

Takeaway 4: Base vs Instruct

Small base models experience more significant learnability gap than Instruct models.

Speaking styles shift. We adopt the method from (Lin et al., 2023) to evaluate the rank shift of each token before and after fine-tuning on long CoT and Large teacher CoT data. This allows us to compare the token distribution shifts induced by the fine-tuning process. We then annotate the tokens that exhibit the largest rank shifts as the most shifted tokens. Our analysis reveals that these to-

kens are predominantly associated with expressive and stylistic elements, such as “wait”, “But”, and “Let”. Please see Appendix C for more details.

Takeaway 5: Speaking Styles Shift

Long CoT and large teacher CoT primarily shift the student model’s distribution of tokens associated with speaking styles.

4 Mix Distillation: Bridge Small Model Learnability Gap

This section presents our Mix Distillation approach to bridge the small model learnability gap.

4.1 Mix Distillation

We propose *Mix Distillation* to address the learnability gap observed in small models. This approach blends easier-to-learn data with more challenging data for small models, thereby leveraging the strengths of both.

Our insight is that small models tend to perform better on data that closely matches their inherent distribution (such as short CoT or small teacher CoT), while they struggle with data that exhibits greater distribution shifts. The token distribution of the mixed long CoT and large teacher CoT data may become closer to that of small models’ inherent distribution, thereby enabling them to learn

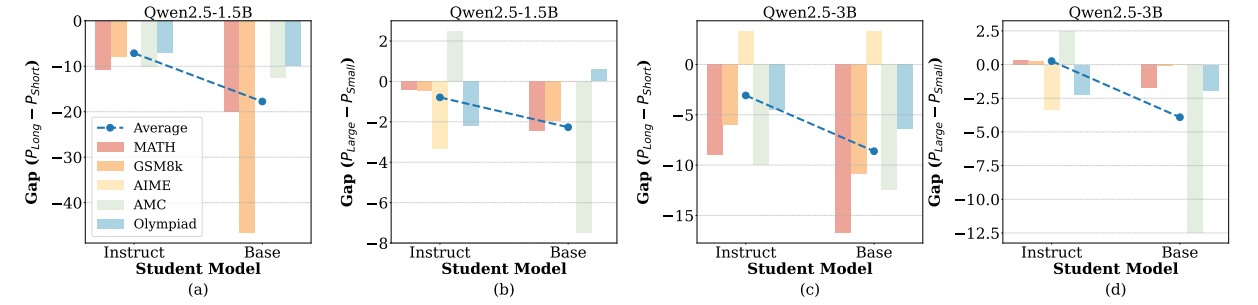


图 5: 基础模型通常表现出比指令模型更显著的学习能力差距。正向差距表明长链思考数据或大型教师链思考数据能提升性能，而负向差距则表明它们有相反的效果。这表明，对于小型基础模型来说，从长链思考数据或大型教师链思考数据中有效学习更具挑战性。

4 Mix Distillation: Bridge Small Model Learnability Gap

本节介绍了我们的混合蒸馏方法，以弥合小型模型学习能力的差距。

4.1 Mix Distillation

我们提出混合蒸馏（Mix Distillation）来解决小模型中观察到的学习能力差距。这种方法将更容易学习的数据与更具挑战性的数据混合，从而发挥两者的优势。

我们的见解是，小模型在与其固有分布更接近的数据上表现更好（例如，较短的CoT或较小的教师CoT），而在表现出更大分布偏移的数据上则表现不佳。混合的长CoT和大教师CoT数据的令牌分布可能会更接近小模型的固有分布，从而使其能够更有效地从具有挑战性的数据集中学习。

我们提出了Mix-Long，它将长CoT和短CoT数据以长CoT的权重 α 和短CoT的权重 $1 - \alpha$ 进行结合。同样，我们提出了Mix-Large，它将大教师CoT与权重 α 和小教师CoT与权重 $1 - \alpha$ 进行结合。

4.2 Experiment Results

我们使用Qwen2.5-3B-Instruct作为学生模型，并使用MATH (7.5k)作为训练集。我们蒸馏不同的教师模型以生成响应作为基线。它们包括QwQ-32B（长CoT），Qwen2.5-32B（短CoT），Qwen2.5-72B（大教

师CoT），Qwen2.5-3B（小教师CoT）。我们添加Deepseek-R1-32B (DeepSeek-AI, 2025)作为教师模型，以生成另一组长CoT数据作为基线。我们在Mix-Long和Mix-Large的两种配置中设置 $\alpha = 0.2$ 。

实验结果表明，Mix-Long和Mix-Large在大多数评估指标上都超过了基线。我们展示了通过混合蒸馏，小的学生模型在单一数据集上训练时可以实现性能的提升。例如，Qwen2.5-3B-Instruct在MATH和AMC上使用Mix-Long比直接在长CoT数据上训练提高了超过8分。与在大教师CoT数据上训练相比，Qwen2.5-3B-Instruct通过Mix-Large在MATH、AIME和AMC上也显示出了超过7分的提升。这表明，小的学生模型从通过混合蒸馏生成的数据集中学习更为容易。

要点 6: 混合蒸馏弥合差距

通过混合长链思维数据（或大型教师链思维）和短链思维数据（或小型教师链思维），小型学生模型可以实现比单独使用任一数据训练更好的性能。

图 6 显示了在不同长 CoT 数据或大教师 CoT 混合权重 α 下的平均性能。我们选择 Qwen2.5-3B-Instruct 作为学生模型，并发现权重 α 为 0.2 时，在 Mix-Long 和 Mix-Large 两种情况下，在五个基准测试中均取得了最高的平均性能。

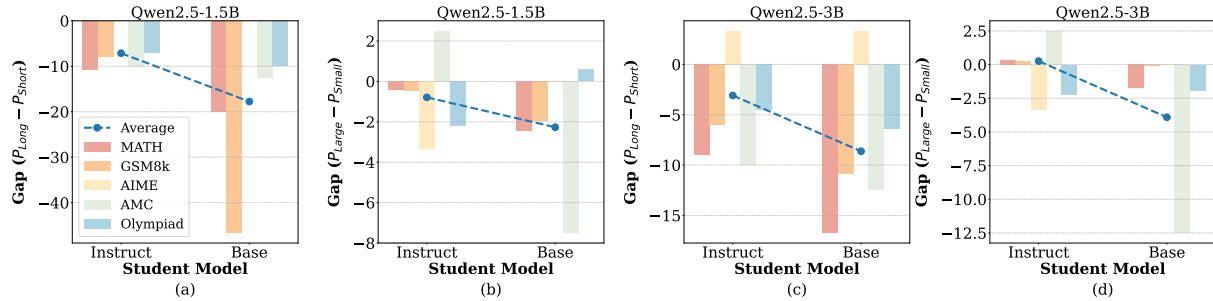


图 5: Base models generally exhibit a more significant learnability gap than Instruct models. A positive gap indicates that long CoT data or large teacher CoT enhance performance, whereas a negative gap suggests they have the opposite effect. This implies that it is more challenging for small base models to effectively learn from long CoT data or large teacher CoT.

more effectively from challenging datasets.

We propose Mix-Long, which combines long CoT and short CoT data with a weight of long CoT α and short CoT $1 - \alpha$. Similarly, we proposed Mix-Large, which combines large teacher CoT with a weight of α and small teacher CoT with a weight of $1 - \alpha$.

4.2 Experiment Results

We use Qwen2.5-3B-Instruct as the student model and MATH (7.5k) as the training set. We distill different teacher models to generate responses as the baseline. They include QwQ-32B (long CoT), Qwen2.5-32B (short CoT), Qwen2.5-72B (large teacher CoT), Qwen2.5-3B (small teacher CoT). We add Deepseek-R1-32B (DeepSeek-AI, 2025) as the teacher model to generate another set of long CoT data as baseline. We set $\alpha = 0.2$ in both configurations of Mix-Long and Mix-Large.

Experimental results demonstrate that both Mix-Long and Mix-Large surpass baselines in most evaluation metrics. We show that the small student model could achieve improved performance by Mix Distillation compared to training on a single dataset. For instance, Qwen2.5-3B-Instruct improves by more than 8 points on MATH and AMC using Mix-Long, compared to direct training on long CoT data. It also shows a more than 7-point gain on MATH,

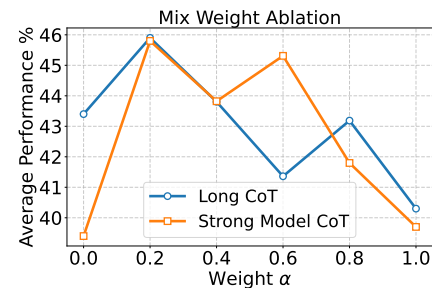


图 6: The average performance varies with the mix weight of long CoT or large teacher CoT data. Qwen2.5-3B-Instruct is chosen as the student model. At a weight of 0.2, mix distillation achieves the highest average performance.

AIME and AMC for Qwen2.5-3B-Instruct by Mix-Large compared with training on large teacher CoT data. This implies that it is easier for small student models to learn from datasets generated by Mix Distillation.

Takeaway 6: Mix Distillation Bridges Gap

By mixing long CoT data (resp. large teacher CoTs) and short CoT data (resp. small teacher CoT), the small student model could achieve better performance compared to training on either data alone.

Figure 6 shows the average performance when taking different mix weight α of long CoT data or large teacher CoT. We choose

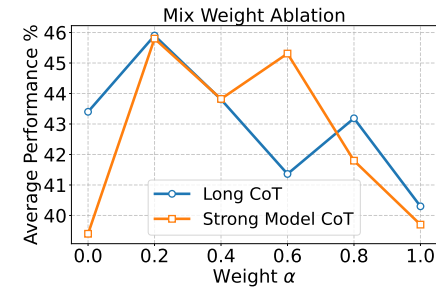


图 6: 平均性能随长CoT或大教师CoT数据的混合权重而变化。选择Qwen2.5-3B-Instruct作为学生模型。在权重为0.2时，混合蒸馏达到最高的平均性能。

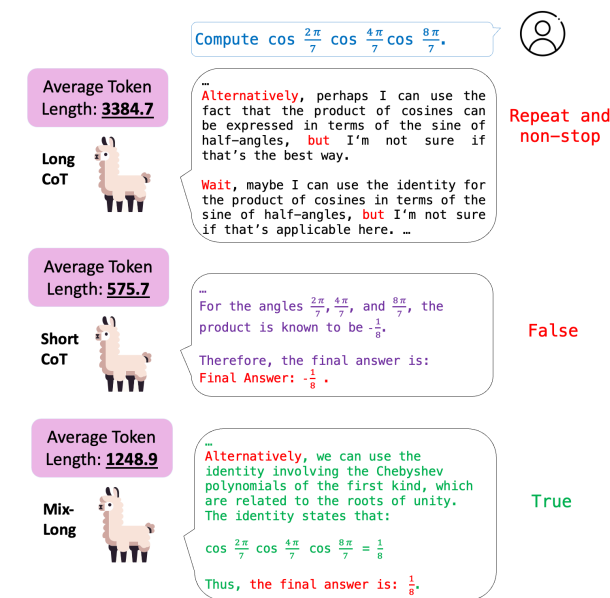


图 7: Mix-Long 案例研究。在长链思考 (CoT) 上微调的模型倾向于过度思考，而在短链思考 (CoT) 上训练的模型则产生了错误的答案。相比之下，Mix-Long 通过引入分支元素（例如，“另选方案”）实现了平衡的推理过程，并得出了正确的答案。

有趣的是，我们发现将长 CoT 和短 CoT 数据混合后，小学生模型的输出结合了长 CoT 的特征，如分叉过程，同时保持了较短的 token 长度，避免了过于复杂的思考。这在图 7 中有所说明。我们观察到，经过长 CoT 数据微调的小学生模型会被重复的思考所淹没，无法停止，而经过短 CoT 数据微调的模型则会产生错误的答案。相比之下，我们提出的 Mix-Long，通过引入分支元素（例如使用“或者”），能够提供正确的答案。此外，由长

CoT、短 CoT 和 Mix-Long 生成的响应的平均 token 长度分别为 3384.7、575.7 和 1248.9。我们建议，将长 CoT 和短 CoT 数据混合是一种实现平衡 CoT 长度的实用方法，从而增强小学生模型的推理能力。

5 Related Work

5.1 Chain-of-Thought

早期关于 CoT 的研究主要集中在短 CoT 上，模型通过简洁的推理路径来达到解决方案 (Lambert et al., 2025; Longpre et al., 2023; Wei et al., 2023; Yu et al., 2024)。最近，研究人员转向了长 CoT 提示，这鼓励生成扩展和详细的推理链 (DeepSeek-AI, 2025; Hou et al., 2025; Kimi Team, 2025; NovaSky, 2025; OpenAI, 2024; Pan et al., 2025; Zeng et al., 2025)。模型系统地探索多条路径（分支）并在特定路径被证明错误时返回到早期的点（回溯）。尽管有几项研究调查了通过蒸馏和强化学习将长 CoT 能力整合到大型语言模型中的方法，但这些努力主要集中在大型模型上。相比之下，我们的工作专门针对训练小型模型所面临的挑战。

5.2 Synthetic Reasoning Data

尽管人工构建的推理数据集已被用于增强大语言模型 (LLM) 的推理能力 (Hendrycks et al., 2021; LI et al., 2024)，但其开发既耗时又费力。最近的进展通过从 LLM 直接生成指令或响应 (Hui et al., 2024; Toshniwal et al., 2024; Xu et al., 2024; Yue et al., 2023; Zhang et al., 2025) 或直接从网络提取数据 (Paster et al., 2023; Yue et al., 2024)，简化了这一过程，生成了更详细和多样的链式思维推理路径。最近的研究调查了各种响应生成器的影响 (Kim et al., 2024b)，表明在指令跟随和推理领域，来自更强教师模型的响应并不一定能产生最有效的学习效果。然而，这些研究尚未认识到学生模型的大小是影响这一现象的关键因素，也未进行如本文所述的更深入的归因和缓解分析。

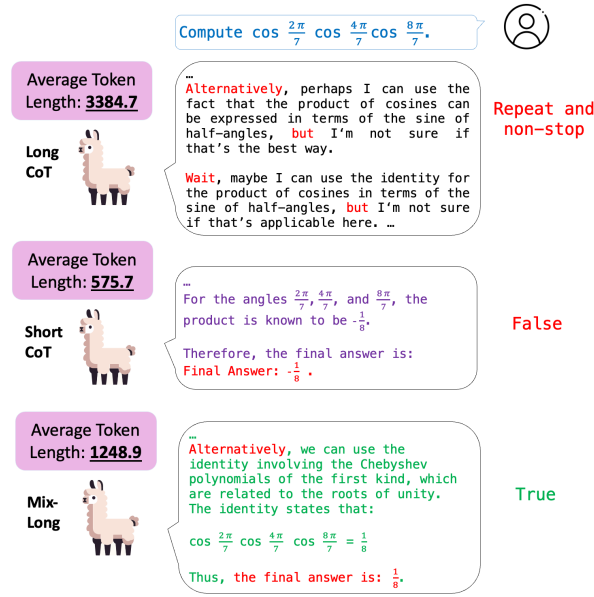


图 7: Case Study of Mix-Long. Models fine-tuned on long CoT tended to overthink, while those trained on short CoT produced incorrect answers. In contrast, Mix-Long, incorporating branching elements (e.g., “Alternatively”), achieved a balanced reasoning process and arrived at the correct answer.

Qwen2.5-3B-Instruct as the student model and find that a weight α of 0.2 achieves the highest average performance across five benchmarks for both Mix-Long and Mix-Large.

Interestingly, we find that after mixing long CoT and short CoT data, the small student model’s output incorporates characteristics of long CoT, such as a branching process, while maintaining a reduced token length and avoiding overly elaborate thinking. This is illustrated in Figure 7. We observed that the small student model fine-tuned on long CoT data becomes overwhelmed by repeated thoughts and fails to stop, whereas the model fine-tuned on short CoT data produces incorrect answers. In contrast, our proposed Mix-Long, which incorporates branching elements (e.g., the use of “Alternatively”), delivers the correct answer. Additionally, the average token lengths of responses generated by long CoT, short CoT, and Mix-Long are 3384.7, 575.7, and 1248.9, respectively. We

suggest that mixing long CoT and short CoT data is a practical approach to achieving a balanced CoT length, thereby enhancing the reasoning capabilities of small student models.

5 Related Work

5.1 Chain-of-Thought

Early research on CoT primarily focused on short CoT, where models produce succinct reasoning paths to reach a solution (Lambert et al., 2025; Longpre et al., 2023; Wei et al., 2023; Yu et al., 2024). Recently, researchers have turned to long CoT prompting, which encourages the generation of extended and detailed reasoning chains (DeepSeek-AI, 2025; Hou et al., 2025; Kimi Team, 2025; NovaSky, 2025; OpenAI, 2024; Pan et al., 2025; Zeng et al., 2025). The model systematically explores multiple paths (branching) and reverts to earlier points if a particular path proves wrong (backtracking). Although several studies have investigated methods such as distillation and reinforcement learning to integrate long CoT capabilities into LLMs, these efforts have predominantly concentrated on large models. In contrast, our work specifically targets the challenges associated with training smaller models.

5.2 Synthetic Reasoning Data

Although human-crafted reasoning datasets have been used to enhance LLM reasoning capabilities (Hendrycks et al., 2021; LI et al., 2024), their development is both time-consuming and labor-intensive. Recent advancements have streamlined this process by generating instructions or responses directly from LLMs (Hui et al., 2024; Toshniwal et al., 2024; Xu et al., 2024; Yue et al., 2023; Zhang et al., 2025) or extracting data directly from web (Paster et al., 2023; Yue et al., 2024), yielding more detailed and diverse chain-of-thought reasoning pathways. Recent study has investigated the impact

6 Conclusion and Future Work

在本文中，我们展示了长链思维（CoT）数据和大型模型响应对于小型学生模型并非始终有益。我们发现，小型模型在使用短链思维和小型模型链思维进行微调时，表现可能更好。我们将这一挑战称为小型模型可学习性差距。其背后的原因可能是小型学生模型在与它们固有分布密切匹配的数据上表现出色，但在显著的分布偏移上表现不佳。为了弥合这一差距，我们引入了混合蒸馏方法，包括Mix-Long，它以一定比例结合了长链思维和短链思维数据，以及Mix-Large，它整合了大型和小型教师的链思维。实验结果表明，Mix-Long和Mix-Large在大多数评估指标上均优于基线，这表明混合蒸馏优于单一数据分布的训练。本文为优化后训练策略以增强小型语言模型的推理能力提供了实用见解。

我们将在未来的工作中探索几个有前景的方向。首先，我们将通过最优地结合多样化的数据源并提出更细粒度的混合算法来改进混合蒸馏，以提升推理能力。其次，我们提议研究如何使强大的推理教师生成更适合调优小型学生模型的数据，从而促进更有效的知识转移。第三，我们将进一步进行关于小型模型可学习性差距的理论和模型可插值性研究。最后，我们将研究哪些SFT方法能产生最佳的初始策略，以最终增强整体模型性能。

Acknowledgment

本工作部分由空军科学研究办公室 (AFOSR) 根据资助协议 FA9550-23-1-0208，海军研究办公室 (ONR) 根据资助协议 N0014-23-1-2386，以及国家自然科学基金会 (NSF) AI 基于代理的网络威胁情报与操作研究所 (ACTION) 根据资助协议 IIS 2229876 支持。

本工作部分由国家自然科学基金会、国土安全部和 IBM 提供的资金支持。本材料中表达的任何观点、发现、结论或建议均属于作者，并不一定反映 NSF 或其联邦机构和行业合作伙伴的观点。

伴的观点。

Limitations

虽然我们的研究为理解数学推理中的小型模型学习能力差距提供了有价值的见解，但我们承认，我们的研究主要集中在这一特定领域，没有评估其他关键技能，如指令遵循、代码生成或多模态理解。我们也没有研究预训练数据组成中的细微变化对小型模型学习能力差距的影响。对不同预训练数据源及其比例如何影响学习结果进行更详细的分析，可以为优化数据选择策略以缓解这一差距提供有价值的见解。

Ethical Statement

本文专注于通过蒸馏技术评估和增强小型语言模型的推理能力。我们在实验中使用的数据集和基准测试是公开可用的。我们不介绍或推荐任何可能造成危害或被滥用的应用。本文不涉及任何伦理问题。

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. [On-policy distillation of language models: Learning from self-generated mistakes](#). *Preprint*, arXiv:2306.13649.
- Anthropic. 2023. [Introducing claude](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- 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. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

of various response generators (Kim et al., 2024b), suggesting that in the domains of instruction following and reasoning, responses from stronger teacher models do not necessarily produce the most effective learning effects for student models. However, these investigations have not recognized student model size as a critical factor influencing this phenomenon, nor have they performed the more attribution and mitigation analyses as in this paper.

6 Conclusion and Future Work

In this paper, we show that long CoT data and large model responses were not uniformly beneficial for small student models. We found that small models may perform better when fine-tuned with short CoT and small model CoT. We termed this challenge as the Small Model Learnability Gap. The reason behind it may be that small student models excel on data that closely match their inherent distribution but struggle with significant distribution shifts. To bridge the gap, we introduced Mix Distillation, including Mix-Long, which combined long CoT and short CoT data in a ratio, and Mix-Large, which integrated large and small teacher CoT. Experimental results showed that both Mix-Long and Mix-Large outperform baselines across most evaluation metrics, which implied mix distillation outperforms training on a single data distribution. This paper provided practical insights for optimizing post-training strategies to enhance small language model reasoning capability.

We will explore several promising directions as future work. First, we will refine mix distillation by optimally combining diverse data sources and proposing more fine-grained mixing algorithms to boost reasoning capabilities. Second, we propose to study how strong reasoning teachers can generate data that is better suited for tuning small student models, thereby facilitating more effective knowledge transfer. Third, we will conduct further

theoretical and model interpolability studies on the small model learnability gap. Lastly, we will investigate which SFT methods yield the best initial policies for subsequent RL procedure, ultimately enhancing overall model performance.

Acknowledgment

This work is partially supported by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-23-1-0208, the Office of Naval Research (ONR) under grant N0014-23-1-2386, and the National Science Foundation (NSF) AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) under grant IIS 2229876.

This work is supported in part by funds provided by the National Science Foundation, Department of Homeland Security, and IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or its federal agency and industry partners.

Limitations

While our study provides valuable insights into the understanding of small model learnability gap in math reasoning, we acknowledge that our research primarily focuses on this specific domain and does not evaluate other crucial skills such as instruction following, code generation, or multimodal understanding. We also did not investigate the impact of fine-grained variations in pre-training data composition on the small model learnability gap. A more detailed analysis of how different pre-training data sources and their proportions affect learning outcomes could offer valuable insights into optimizing data selection strategies for mitigating this gap.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and et al. 2024. *Deepseek-v3 technical report*. *Preprint*, arXiv:2412.19437.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. *Raft: Reward ranked finetuning for generative foundation model alignment*. *Preprint*, arXiv:2304.06767.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. *Reinforced self-training (rest) for language modeling*. *Preprint*, arXiv:2308.08998.

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. *Olympiad-bench: 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*. *Preprint*, arXiv:2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. *Distilling the knowledge in a neural network*. *Preprint*, arXiv:1503.02531.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. *Advancing language model reasoning through reinforcement learning and inference scaling*. *Preprint*, arXiv:2501.11651.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. *Preprint*, arXiv:2106.09685.

Tingfeng Hui, Lulu Zhao, Guanting Dong, Yaqi Zhang, Hua Zhou, and Sen Su. 2024. *Smaller language models are better instruction evolvers*. *Preprint*, arXiv:2412.11231.

Gyeongman Kim, Doohyuk Jang, and Eunho Yang. 2024a. *Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning*. *Preprint*, arXiv:2402.12842.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024b. *Evaluating language models as synthetic data generators*. *Preprint*, arXiv:2412.03679.

Kimi Team. 2025. *Kimi k1.5: Scaling reinforcement learning with llms*. *Preprint*, arXiv:2501.12599.

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 Taffjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. *Tulu 3: Pushing frontiers in open language model post-training*. *Preprint*, arXiv:2411.15124.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. *Numinamath*. [<https://github.com/project-numina/aimo-progress-prize>](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. *The unlocking spell on base llms: Rethinking alignment via in-context learning*. *Preprint*, arXiv:2312.01552.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. *The flan collection: Designing data and methods for effective instruction tuning*. *Preprint*, arXiv:2301.13688.

Meta. 2024a. Llama-3.2-3b. <https://huggingface.co/meta-llama/Llama-3.2-3B>.

Meta. 2024b. Meet llama 3.1. <https://llama.meta.com>.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. *Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems*. *Preprint*, arXiv:2412.09413.

NovaSky. 2025. *Sky-T1: Train your own o1 preview model within \$450*. Accessed: 2025-01-09.

OpenAI. 2023. *Gpt-4 technical report*.

OpenAI. 2024. *Learning to reason with llms*.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. 2025. *Tinyzero*. <https://github.com/Jiayi-Pan/TinyZero>. Accessed: 2025-01-24.

Ethical Statement

This paper focuses on the evaluation and enhancement of reasoning capabilities in small language models through distillation techniques. The dataset and benchmarks used in our experiments are publicly available. We do not introduce or endorse any applications that could cause harm or be misused. This paper does not present any ethical concerns.

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. [On-policy distillation of language models: Learning from self-generated mistakes](#). *Preprint*, arXiv:2306.13649.
- Anthropic. 2023. [Introducing claude](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- 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. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and et al. 2024. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. [Raft: Reward ranked finetuning for generative foundation model alignment](#). *Preprint*, arXiv:2304.06767.
- Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. [Reinforced self-training \(rest\) for language modeling](#). *Preprint*, arXiv:2308.08998.
- 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. [Olympiad-bench: 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](#). *Preprint*, arXiv:2103.03874.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *Preprint*, arXiv:1503.02531.
- Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. [Advancing language model reasoning through reinforcement learning and inference scaling](#). *Preprint*, arXiv:2501.11651.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Tingfeng Hui, Lulu Zhao, Guanting Dong, Yaqi Zhang, Hua Zhou, and Sen Su. 2024. [Smaller language models are better instruction evolvers](#). *Preprint*, arXiv:2412.11231.
- Gyeongman Kim, Doohyuk Jang, and Eunho Yang. 2024a. [Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning](#). *Preprint*, arXiv:2402.12842.
- Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024b. [Evaluating language models as synthetic data generators](#). *Preprint*, arXiv:2412.03679.
- Kimi Team. 2025. [Kimi k1.5: Scaling reinforcement learning with llms](#). *Preprint*, arXiv:2501.12599.
- 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, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. [Tulu 3: Pushing frontiers in open language model post-training](#). *Preprint*, arXiv:2411.15124.
- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. [Openwebmath: An open dataset of high-quality mathematical web text](#). *Preprint*, arXiv:2310.06786.
- Qwen. 2024a. [Qwen2.5: A party of foundation models](#).
- Qwen. 2024b. [Qwq: Reflect deeply on the boundaries of the unknown](#). 2024b.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Keane, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. [Beyond human data: Scaling self-training for problem-solving with language models](#). *Preprint*, arXiv:2312.06585.
- Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. [Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving](#). *Preprint*, arXiv:2407.13690.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisanin, Alexan Ayrapetyan, and Igor Gitman. 2024. [Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data](#). *Preprint*, arXiv:2410.01560.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. [Zephyr: Direct distillation of lm alignment](#). *Preprint*, arXiv:2310.16944.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. [Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing](#). *Preprint*, arXiv:2406.08464.
- 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.
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. [Demystifying long chain-of-thought reasoning in llms](#). *Preprint*, arXiv:2502.03373.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Metamath: Bootstrap your own mathematical questions for large language models](#). *Preprint*, arXiv:2309.12284.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. [Scaling relationship on learning mathematical reasoning with large language models](#). *Preprint*, arXiv:2308.01825.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mammoth: Building math generalist models through hybrid instruction tuning](#). *Preprint*, arXiv:2309.05653.
- Xiang Yue, Toney Zheng, Ge Zhang, and Wenhua Chen. 2024. [Mammoth2: Scaling instructions from the web](#). *Preprint*, arXiv:2405.03548.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). *Preprint*, arXiv:2203.14465.
- Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 2025. [7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient](#). <https://hkust-nlp.notion.site/simpler1-reason>. Notion Blog.
- Dylan Zhang, Qirun Dai, and Hao Peng. 2025. [The best instruction-tuning data are those that fit](#). *Preprint*, arXiv:2502.04194.

- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. [<https://github.com/project-numina/aimo-progress-prize>](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *Preprint*, arXiv:2312.01552.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *Preprint*, arXiv:2301.13688.
- Meta. 2024a. Llama-3.2-3b. <https://huggingface.co/meta-llama/Llama-3.2-3B>.
- Meta. 2024b. Meet llama 3.1. <https://llama.meta.com>.
- Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *Preprint*, arXiv:2412.09413.
- NovaSky. 2025. Sky-T1: Train your own o1 preview model within \$450. Accessed: 2025-01-09.
- OpenAI. 2023. Gpt-4 technical report.
- OpenAI. 2024. Learning to reason with llms.
- Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. 2025. Tinyzero. <https://github.com/Jiayi-Pan/TinyZero>. Accessed: 2025-01-24.
- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text. *Preprint*, arXiv:2310.06786.
- Qwen. 2024a. Qwen2.5: A party of foundation models.
- Qwen. 2024b. Qwq: Reflect deeply on the boundaries of the unknown. 2024b.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Keane, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. Beyond human data: Scaling self-training for problem-solving with language models. *Preprint*, arXiv:2312.06585.
- Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Preprint*, arXiv:2407.13690.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *Preprint*, arXiv:2410.01560.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu,
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 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)*, Bangkok, Thailand. Association for Computational Linguistics.

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.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. [Demystifying long chain-of-thought reasoning in llms](#). *Preprint*, arXiv:2502.03373.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhen-guo Li, Adrian Weller, and Weiyang Liu. 2024. [Metamath: Bootstrap your own mathematical questions for large language models](#). *Preprint*, arXiv:2309.12284.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. [Scaling relationship on learning mathematical reasoning with large language models](#). *Preprint*, arXiv:2308.01825.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mammoth: Building math generalist models through hybrid instruction tuning](#). *Preprint*, arXiv:2309.05653.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhua Chen. 2024. [Mammoth2: Scaling instructions from the web](#). *Preprint*, arXiv:2405.03548.

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

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 2025. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. <https://hkust-nlp.notion.site/simpler1-reason>. Notion Blog.

Dylan Zhang, Qirun Dai, and Hao Peng. 2025. [The best instruction-tuning data are those that fit](#). *Preprint*, arXiv:2502.04194.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 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)*, Bangkok, Thailand. Association for Computational Linguistics.

A Detailed Experimental Setups

Category	Models
Teacher Models	
Long CoT vs ShortCoT	QwQ-32B-Preview vs Qwen2.5-32B-Instruct
Large Teacher vs Small Teacher	
<i>Qwen Family</i>	Qwen2.5-72B-Instruct vs Qwen2.5-3B-Instruct
<i>Llama Family</i>	Llama3.1-70B-Instruct vs Llama3.1-8B-Instruct
<i>Gemma Family</i>	Gemma2-27B-it vs Gemma2-9B-it
Student Models	
<i>Qwen Family</i>	Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct
<i>Llama Family</i>	Llama3.2-1B-Instruct, Llama3.2-3B-Instruct, Llama3.1-8B-Instruct, Llama3.3-70B-Instruct

表 4: Overview of Teacher and Student Models

A.1 Models

表 4 提供了我们论文中使用的学生模型和教师模型的综合概述。

A.2 Training Setup

我们的模型训练使用了LLaMA-Factory (Zheng et al., 2024)，在一台配备四块NVIDIA A100-SXM4-80GB GPU、一个AMD EPYC 7763 64核处理器和512 GB内存的服务器上进行。对于参数量小于14B的学生模型，我们使用全参数微调。当学生模型的参数量大于14B时，我们使用LoRA微调 (Hu et al., 2021)。表5和表6分别列出了全参数微调和LoRA微调的超参数。

教师模型通过拒绝采样生成响应 (Zelikman et al., 2022; Tong et al., 2024; Yue et al., 2023; Singh et al., 2024; Gulcehre et al., 2023; Yuan et al.,

Hyper-parameter	Value
Learning Rate	1×10^{-5}
Number of Epochs	2
Number of Devices	4
Per-device Batch Size	2
Optimizer	Adamw
Learning Rate Scheduler	cosine
Max Sequence Length	16384

表 5: 此表显示了完整参数微调的超参数。

Hyper-parameter	Value
Learning Rate	1×10^{-4}
Number of Epochs	2
Number of Devices	4
Per-device Batch Size	1
Lora Target	full
Learning Rate Scheduler	cosine
Warmup Ratio	0.03
Max Sequence Length	16384

表 6: 此表显示了LoRA微调的超参数。

2023; Dong et al., 2023)。默认情况下，教师模型采用贪婪解码。通过将数学问题指令与教师模型生成的相应解决方案结合起来，我们构建了问题-解决方案对以微调学生模型。我们对不同教师模型生成的解决方案进行成对比较，并筛选出两个模型都正确的問題-解决方案对，以微调学生模型。

A.3 Evaluation Setup

我们评估了微调后的学生模型在一组常用基准测试上的推理能力，包括 MATH (Hendrycks et al., 2021)、GSM8K (Cobbe et al., 2021)、AMC 2023、AIME 2024 以及 Olympiad-Bench (He et al., 2024) 的英语数学子集。

除非另有说明，所有微调模型均在零样本设置下使用贪婪解码进行评估。我们将最大生成令牌数设置为 16k。评估提示如下所示。

在提取了评估模型的最终答案后，我们首先使用精确匹配来确定答案的正确性。如果答

A Detailed Experimental Setups

Category	Models
Teacher Models	
Long CoT vs ShortCoT	QwQ-32B-Preview vs Qwen2.5-32B-Instruct
Large Teacher vs Small Teacher	
<i>Qwen Family</i>	Qwen2.5-72B-Instruct vs Qwen2.5-3B-Instruct
<i>Llama Family</i>	Llama3.1-70B-Instruct vs Llama3.1-8B-Instruct
<i>Gemma Family</i>	Gemma2-27B-it vs Gemma2-9B-it
Student Models	
<i>Qwen Family</i>	Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct
<i>Llama Family</i>	Llama3.2-1B-Instruct, Llama3.2-3B-Instruct, Llama3.1-8B-Instruct, Llama3.3-70B-Instruct

表 4: Overview of Teacher and Student Models

A.1 Models

Table 4 presents a comprehensive overview of student and teacher models used in our paper.

A.2 Training Setup

Our model training is conducted using LLaMA-Factory (Zheng et al., 2024), on a server with four NVIDIA A100-SXM4-80GB GPUs, an AMD EPYC 7763 64-Core Processor, and 512 GB of RAM. We use full parameter fine-tuning on student models less than 14B parameters. When the student model is larger than 14B, we use LoRA fine-tuning (Hu et al., 2021). Table 5 and Table 6 list hyper-parameters for full parameter fine-tuning and LoRA fine-tuning respectively.

Teacher models generate responses by rejec-

Hyper-parameter	Value
Learning Rate	1×10^{-5}
Number of Epochs	2
Number of Devices	4
Per-device Batch Size	2
Optimizer	Adamw
Learning Rate Scheduler	cosine
Max Sequence Length	16384

表 5: This table shows the hyper-parameters for full parameter fine-tuning.

Hyper-parameter	Value
Learning Rate	1×10^{-4}
Number of Epochs	2
Number of Devices	4
Per-device Batch Size	1
Lora Target	full
Learning Rate Scheduler	cosine
Warmup Ratio	0.03
Max Sequence Length	16384

表 6: This table shows the hyper-parameters for LoRA fine-tuning.

tion sampling (Zelikman et al., 2022; Tong et al., 2024; Yue et al., 2023; Singh et al., 2024; Gulcehre et al., 2023; Yuan et al., 2023; Dong et al., 2023). By default, teacher models employ greedy decoding. By combining the math problem instructions with corresponding solutions generated by teacher models, we construct problem-solution pairs to fine-tune student models. We perform pairwise comparisons of solutions generated by different teacher models and filter out problem-solution pairs that are correct for both models to fine-tune student models.

A.3 Evaluation Setup

We evaluate the reasoning capability of fine-tuned student models on a set of commonly used benchmarks, including MATH (Hendrycks et al.,

Prompt

Solve the following math problem and present the final answer in the format: Final Answer: {your answer}
Problem: {problem}
Answer:

案不正确，我们使用Qwen-32B-Instruct作为评判者，将提取的最终答案与 ground truth 进行比较。提示如下。

Prompt

Given a math problem, its correct final answer, and the model’s generated final answer, determine if the model’s answer is correct. Respond with ‘True’ if the it is correct and ‘False’ if it is incorrect.
Problem: {problem}
Correct Final Answer: {ground truth}
Model’s Generated Final Answer: {resp answer}
Your Judgement:

B More Experiments Results

在本节中，我们展示了长CoT间隔和大教师CoT间隔的额外实验结果。

B.1 Long CoT Gap: Additional Results

表 7 显示了不同学生模型在长 CoT 和短 CoT 上微调后每个基准的详细性能分数和差距。QwQ-32B-Preview 被选中生成成长 CoT，而 Qwen-2.5-32B-Instruct 被选中生成短 CoT。我们观察到，小型学生模型从短 CoT 中受益更多，而大型学生模型从长 CoT 中获得更大的优势。

B.2 Large Teacher CoT Gap: Additional Results

表 8 显示了从大型教师模型和小型教师模型蒸馏出的不同学生模型在每个基准测试中的详细性能得分和差距。我们总结了来自 Llama 和 Qwen 家族的 10 个学生模型在不同模型大小下的性能。选择 Qwen-2.5-72B-Instruct 作为大型教师，而 Qwen-2.5-3B-Instruct 作为小型教师。结果如表 8 所示。我们的研究发现，小型学生模型从大型教师蒸馏时的性能可能会比从小型教师蒸馏时下降，而较大的学生模型则从大型教师蒸馏中受益更多。

表 9 显示了不同模型家族的教师模型的更多实验结果，包括 Gemma-27B-it 与 Gemma-9B-it 以及 Llama3.1-72B-Instruct 与 Llama3.1-8B-Instruct 的对比。

C Examples of Speaking Style Shift

我们采用 (Lin et al., 2023) 的方法来评估在长 CoT 和大型教师 CoT 数据上微调后最偏移的 token。图 8 显示了计算过程。这使我们能够比较微调过程引起的 token 分布偏移。我们将表现出最大排名偏移的 token 标注为最偏移的 token。我们选择 Qwen2.5-3B-Instruct 作为学生模型。我们将长 CoT 数据微调后最偏移的 token 的结果放在图 9 和 10 中。大型教师 CoT 数据微调后最偏移的 token 的结果如图 11 所示。我们的分析表明，这些 token 主要与表达性和风格性元素相关，如 “wait”、“But” 和 “Let”。

D Examples of Various CoT Data

本节展示了不同CoT数据的示例，包括长CoT、短CoT、大教师CoT和小教师CoT。参见以下示例。

2021), GSM8K (Cobbe et al., 2021), AMC 2023, AIME 2024, and the English math subset of OlympiadBench (He et al., 2024).

Unless otherwise specified, all fine-tuned models are evaluated in a zero-shot setting using greedy decoding. We set the maximum generation tokens as 16k. The evaluation prompt is shown below.

Prompt

Solve the following math problem and present the final answer in the format: Final Answer: {your answer}

Problem: {problem}

Answer:

After extracting the final answer of the evaluated model, we first employ exact matching to determine the correctness of the answer. If the answer is incorrect, we use Qwen-32B-Instruct as a judge to compare the extracted final answers against that of the ground truth. The prompt is shown below.

Prompt

Given a math problem, its correct final answer, and the model’s generated final answer, determine if the model’s answer is correct. Respond with ‘True’ if the it is correct and ‘False’ if it is incorrect.

Problem: {problem}

Correct Final Answer: {ground truth}

Model’s Generated Final Answer: {resp answer}

Your Judgement:

B More Experiments Results

In this section we present additional experiment results of long CoT gap and large teacher CoT gap.

B.1 Long CoT Gap: Additional Results

Table 7 shows the detailed performance scores and gap of each benchmark for different student models fine-tuned on long CoT and short CoT. QwQ-32B-Preview is chosen to generate long CoT and awhile Qwen-2.5-32B-Instruct is chosen to generate short CoT. We observe that small student models tend to benefit more from short CoT, while large student models gain greater advantages from long CoT.

B.2 Large Teacher CoT Gap: Additional Results

Table 8 shows the detailed performance scores and gap of each benchmark for different student models distilled from large teacher and small teacher. We summarize the performance of 10 student models from the Llama and Qwen families across various model sizes. Qwen-2.5-72B-Instruct is chosen as the large teacher while Qwen-2.5-3B-Instruct is chosen as the small teacher. The results are shown in Table 8. Our findings indicate that small student models may experience degraded performance when distilled from a large teacher compared to a small teacher, whereas larger student models benefit more from distilling a large teacher.

Table 9 shows more experiment results for teacher models in different model families, including Gemma-27B-it vs Gemma-9B-it and Llama3.1-72B-Instruct vs Llama3.1-8B-Instruct.

C Examples of Speaking Style Shift

We adopt the method from (Lin et al., 2023) to evaluate the most shifted tokens after fine-tuning on long CoT and Large teacher CoT data. Figure 8 shows the calculation process. This allows us to compare the token distribution shifts induced by the fine-tuning process. We annotate the tokens that

Model	MATH			GSM8K			AIME			AMC			Olympiad			Average Δ_{Long}
	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	
Llama-3.2-1B	28.6	33.4	-4.78	42.3	49.2	-6.90	0.00	0.00	0.00	2.50	7.50	-5.00	5.48	7.40	-1.92	-3.72
Llama-3.2-3B	48.7	50.9	-2.14	75.1	77.5	-2.42	3.33	3.33	0.00	17.5	15.0	2.50	17.6	18.7	-1.04	-0.619
Llama-3.1-8B	50.0	44.6	5.36	81.4	75.5	5.84	0.00	0.00	0.00	27.5	22.5	5.00	17.3	14.8	2.52	3.74
Llama-3.3-70B	75.3	74.9	0.340	92.7	91.2	1.44	26.7	13.3	13.3	55.0	52.5	2.50	41.3	39.7	1.63	3.85
Qwen2.5-0.5B	23.0	31.5	-8.44	39.5	45.3	-5.84	0.00	0.00	0.00	7.50	15.0	-7.50	4.00	5.93	-1.93	-4.74
Qwen2.5-1.5B	41.6	52.3	-10.7	63.8	71.7	-7.89	0.00	0.00	0.00	17.5	27.5	-10.0	12.3	19.4	-7.11	-7.13
Qwen2.5-3B	56.2	61.0	-4.84	80.0	82.0	-1.98	3.33	10.0	-6.67	37.5	37.5	0.00	24.4	26.4	-1.93	-3.08
Qwen2.5-7B	68.2	67.8	0.460	86.2	85.7	0.560	13.3	6.67	6.67	40.0	40.0	0.00	36.6	35.7	0.889	1.72
Qwen2.5-14B	78.3	76.2	2.04	93.3	92.5	0.760	20.0	6.67	13.3	60.0	55.0	5.00	44.4	40.9	3.56	4.94
Qwen2.5-32B	84.8	82.3	2.44	94.9	94.3	0.610	40.0	10.0	30.0	85.0	62.5	22.5	60.4	47.3	13.2	13.7

表 7: 此表总结了使用长CoT和短CoT数据微调的Llama和Qwen系列模型的性能。它们在MATH、GSM8K、AIME、AMC和OlympiadBench上进行了评估。选择QwQ-32B-Preview生成长CoT，而选择Qwen-2.5-32B-Instruct生成短CoT。我们观察到，小型学生模型从短CoT中受益更多，而大型学生模型从长CoT中获得更大的优势。

Model	MATH			GSM8k			AIME			AMC			Olympiad			Average Δ_{Strong}
	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	
Llama-3.2-1B	29.8	29.6	0.160	44.4	47.5	-3.18	0.00	0.00	0.00	2.50	7.50	-5.00	6.07	7.70	-1.63	-1.93
Llama-3.2-3B	47.4	47.9	-0.500	71.2	74.1	-2.88	3.33	0.00	3.33	25.0	17.5	7.50	16.9	16.4	0.445	1.58
Llama-3.2-8B	37.6	37.6	-0.040	67.0	69.2	-2.20	6.67	0.00	6.67	7.50	7.50	0.00	9.19	11.0	-1.78	0.530
Llama-3.2-70B	74.5	72.2	2.28	92.0	92.2	-0.152	16.7	16.7	0.00	67.5	50.0	17.5	37.3	35.7	1.63	4.25
Qwen2.5-0.5B	30.0	31.0	-0.920	43.1	45.4	-2.35	0.00	0.00	0.00	5.00	17.5	-12.5	6.52	8.30	-1.78	-3.51
Qwen2.5-1.5B	50.3	50.7	-0.440	70.6	71.0	-0.455	0.00	3.33	-3.33	22.5	20.0	2.50	17.8	20.0	-2.22	-0.790
Qwen2.5-3B	57.5	60.3	-2.82	79.9	79.5	0.379	0.00	3.33	-3.33	35.0	27.5	7.50	25.9	26.4	-0.444	0.256
Qwen2.5-7B	71.3	63.6	7.66	87.8	84.1	3.72	6.67	0.00	6.67	40.0	35.0	5.00	38.8	29.0	9.78	6.56
Qwen2.5-14B	76.4	72.8	3.66	93.1	89.6	3.49	6.67	3.33	3.33	47.5	45.0	2.50	41.0	39.0	2.07	3.01
Qwen2.5-32B	80.5	76.8	3.72	92.2	92.7	-0.531	20.0	3.33	16.7	57.5	50.0	7.50	47.4	42.4	5.04	6.48

表 8: 此表总结了在 MATH、GSM8K、AIME、AMC 和 OlympiadBench 上评估的 Llama 和 Qwen 系列模型在使用大型教师 CoT 和小型教师 CoT 微调后的性能。选择 Qwen-2.5-72B-Instruct 作为大型教师，而 Qwen-2.5-3B-Instruct 作为小型教师。我们观察到，与从小型教师蒸馏相比，小型学生模型从大型教师蒸馏时可能会出现性能下降，而较大的学生模型则从大型教师蒸馏中受益更多。

Model	Gemma2-9B vs Gemma2-27B						Llama3.1-8B vs Llama3.1-70B					
	MATH	AMC	Olympiad	AIME	GSM8k	Average	MATH	AMC	Olympiad	AIME	GSM8k	Average
Llama3.2-1B	-1.42	-7.50	0.00	0.00	-0.227	-1.83	-1.42	-5.00	-0.296	3.33	0.152	-0.646
Llama3.2-3B	2.08	-7.50	-0.888	0.00	1.67	-0.928	-0.14	10.0	-0.593	3.33	1.06	2.73
Llama3.1-8B	0.56	0.00	0.078	0.00	-0.516	0.0243	-2.18	7.50	2.67	0.00	-1.29	1.34
Llama3.1-70B	0.02	7.50	-0.741	10.0	0.152	3.39	2.72	17.5	5.48	6.67	0.986	6.67
Qwen2.5-0.5B	-4.56	0.00	0.741	0.00	0.592	-0.645	-1.88	0.00	0.185	0.00	-1.74	-0.688
Qwen2.5-1.5B	-1.20	2.50	-1.19	0.00	-0.986	-0.174	-1.48	5.00	-0.148	3.33	-1.14	1.11
Qwen2.5-3B	0.44	5.00	1.78	0.00	-0.758	1.29	-1.26	5.00	-0.741	-3.33	-1.29	-0.325
Qwen2.5-7B	0.22	5.00	1.04	-3.33	3.94	1.37	3.68	20.0	4.15	3.33	2.81	6.79
Qwen2.5-14B	1.32	2.50	-0.148	0.00	-0.986	0.537	2.18	0.00	0.445	3.33	-0.303	1.13
Qwen2.5-32B	0.10	2.50	1.48	3.44	1.36	1.78	2.72	-2.50	5.63	3.33	0.834	2.00

表 9: 此表展示了从不同教师模型蒸馏出的学生模型的性能，包括 Gemma-27B-it 与 Gemma-9B-it 以及 Llama3.1-72B-Instruct 与 Llama3.1-8B-Instruct。我们观察到，当从大型教师模型蒸馏时，小型学生模型的性能可能会下降，而大型学生模型则能从大型教师模型中获益更多。

Model	MATH			GSM8K			AIME			AMC			Olympiad			Average Δ_{Long}
	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	P_{Long}	P_{Short}	Δ_{Long}	
Llama-3.2-1B	28.6	33.4	-4.78	42.3	49.2	-6.90	0.00	0.00	0.00	2.50	7.50	-5.00	5.48	7.40	-1.92	-3.72
Llama-3.2-3B	48.7	50.9	-2.14	75.1	77.5	-2.42	3.33	3.33	0.00	17.5	15.0	2.50	17.6	18.7	-1.04	-0.619
Llama-3.1-8B	50.0	44.6	5.36	81.4	75.5	5.84	0.00	0.00	0.00	27.5	22.5	5.00	17.3	14.8	2.52	3.74
Llama-3.3-70B	75.3	74.9	0.340	92.7	91.2	1.44	26.7	13.3	13.3	55.0	52.5	2.50	41.3	39.7	1.63	3.85
Qwen2.5-0.5B	23.0	31.5	-8.44	39.5	45.3	-5.84	0.00	0.00	0.00	7.50	15.0	-7.50	4.00	5.93	-1.93	-4.74
Qwen2.5-1.5B	41.6	52.3	-10.7	63.8	71.7	-7.89	0.00	0.00	0.00	17.5	27.5	-10.0	12.3	19.4	-7.11	-7.13
Qwen2.5-3B	56.2	61.0	-4.84	80.0	82.0	-1.98	3.33	10.0	-6.67	37.5	37.5	0.00	24.4	26.4	-1.93	-3.08
Qwen2.5-7B	68.2	67.8	0.460	86.2	85.7	0.560	13.3	6.67	6.67	40.0	40.0	0.00	36.6	35.7	0.889	1.72
Qwen2.5-14B	78.3	76.2	2.04	93.3	92.5	0.760	20.0	6.67	13.3	60.0	55.0	5.00	44.4	40.9	3.56	4.94
Qwen2.5-32B	84.8	82.3	2.44	94.9	94.3	0.610	40.0	10.0	30.0	85.0	62.5	22.5	60.4	47.3	13.2	13.7

表 7: This table summarizes the performance of models in Llama and Qwen families fine-tuned with long CoT and short CoT data. They are evaluated on MATH, GSM8K, AIME, AMC, and OlympiadBench. QwQ-32B-Preview is chosen to generate long CoT and awhile Qwen-2.5-32B-Instruct is chosen to generate short CoT. We observe that small student models tend to benefit more from short CoT, while large student models gain greater advantages from long CoT.

Model	MATH			GSM8k			AIME			AMC			Olympiad			Average Δ_{Strong}
	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	P_{Strong}	P_{Weak}	Δ_{Strong}	
Llama-3.2-1B	29.8	29.6	0.160	44.4	47.5	-3.18	0.00	0.00	0.00	2.50	7.50	-5.00	6.07	7.70	-1.63	-1.93
Llama-3.2-3B	47.4	47.9	-0.500	71.2	74.1	-2.88	3.33	0.00	3.33	25.0	17.5	7.50	16.9	16.4	0.445	1.58
Llama-3.2-8B	37.6	37.6	-0.040	67.0	69.2	-2.20	6.67	0.00	6.67	7.50	7.50	0.00	9.19	11.0	-1.78	0.530
Llama-3.2-70B	74.5	72.2	2.28	92.0	92.2	-0.152	16.7	16.7	0.00	67.5	50.0	17.5	37.3	35.7	1.63	4.25
Qwen2.5-0.5B	30.0	31.0	-0.920	43.1	45.4	-2.35	0.00	0.00	0.00	5.00	17.5	-12.5	6.52	8.30	-1.78	-3.51
Qwen2.5-1.5B	50.3	50.7	-0.440	70.6	71.0	-0.455	0.00	3.33	-3.33	22.5	20.0	2.50	17.8	20.0	-2.22	-0.790
Qwen2.5-3B	57.5	60.3	-2.82	79.9	79.5	0.379	0.00	3.33	-3.33	35.0	27.5	7.50	25.9	26.4	-0.444	0.256
Qwen2.5-7B	71.3	63.6	7.66	87.8	84.1	3.72	6.67	0.00	6.67	40.0	35.0	5.00	38.8	29.0	9.78	6.56
Qwen2.5-14B	76.4	72.8	3.66	93.1	89.6	3.49	6.67	3.33	3.33	47.5	45.0	2.50	41.0	39.0	2.07	3.01
Qwen2.5-32B	80.5	76.8	3.72	92.2	92.7	-0.531	20.0	3.33	16.7	57.5	50.0	7.50	47.4	42.4	5.04	6.48

表 8: This table summarizes the performance of models in Llama and Qwen families fine-tuned with large teacher CoT and small teacher CoT when evaluated on MATH, GSM8K, AIME, AMC, and OlympiadBench. Qwen-2.5-72B-Instruct is chosen as the large teacher while Qwen-2.5-3B-Instruct is chosen as the small teacher. We observe that small student models may experience degraded performance when distilled from a large teacher compared to a small teacher, whereas larger student models benefit more from the distilling a large teacher.

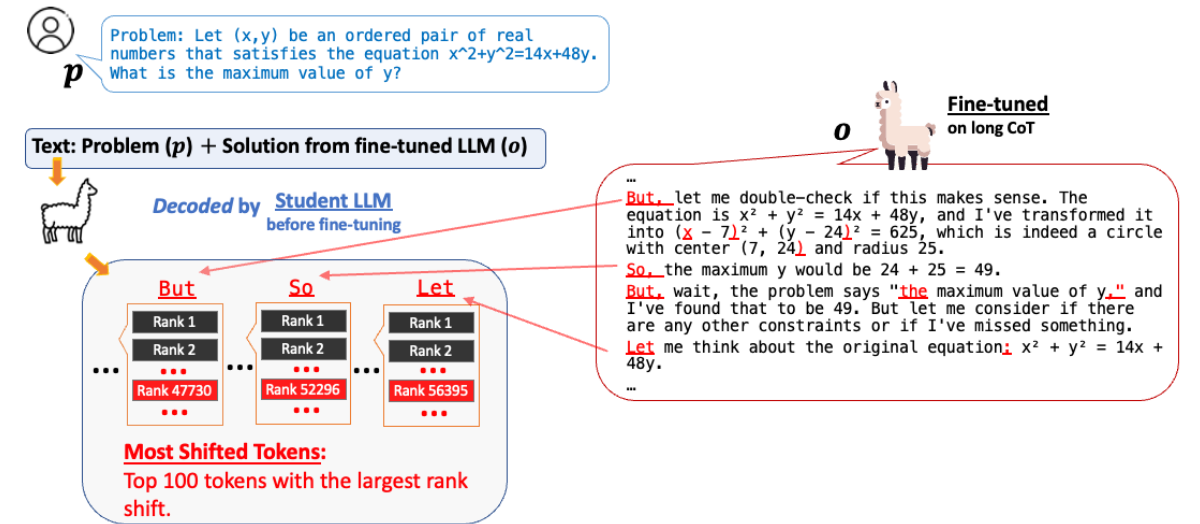
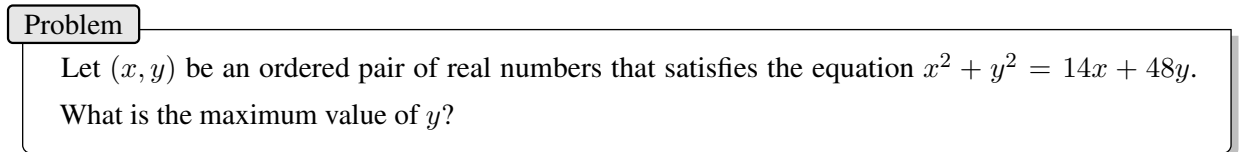


图 8: 计算最偏移的令牌。我们在微调之前解码由学生模型生成的每个令牌，这些令牌是由微调后的LLM生成的。然后我们计算由微调模型生成的每个令牌在学生模型中的排名偏移。我们将表现出最大排名偏移的令牌标注为最偏移的令牌。我们发现这些令牌主要与表达性和风格性元素相关，例如“但是”和“让”。



Model	Gemma2-9B vs Gemma2-27B						Llama3.1-8B vs Llama3.1-70B					
	MATH	AMC	Olympiad	AIME	GSM8k	Average	MATH	AMC	Olympiad	AIME	GSM8k	Average
Llama3.2-1B	-1.42	-7.50	0.00	0.00	-0.227	-1.83	-1.42	-5.00	-0.296	3.33	0.152	-0.646
Llama3.2-3B	2.08	-7.50	-0.888	0.00	1.67	-0.928	-0.14	10.0	-0.593	3.33	1.06	2.73
Llama3.1-8B	0.56	0.00	0.078	0.00	-0.516	0.0243	-2.18	7.50	2.67	0.00	-1.29	1.34
Llama3.1-70B	0.02	7.50	-0.741	10.0	0.152	3.39	2.72	17.5	5.48	6.67	0.986	6.67
Qwen2.5-0.5B	-4.56	0.00	0.741	0.00	0.592	-0.645	-1.88	0.00	0.185	0.00	-1.74	-0.688
Qwen2.5-1.5B	-1.20	2.50	-1.19	0.00	-0.986	-0.174	-1.48	5.00	-0.148	3.33	-1.14	1.11
Qwen2.5-3B	0.44	5.00	1.78	0.00	-0.758	1.29	-1.26	5.00	-0.741	-3.33	-1.29	-0.325
Qwen2.5-7B	0.22	5.00	1.04	-3.33	3.94	1.37	3.68	20.0	4.15	3.33	2.81	6.79
Qwen2.5-14B	1.32	2.50	-0.148	0.00	-0.986	0.537	2.18	0.00	0.445	3.33	-0.303	1.13
Qwen2.5-32B	0.10	2.50	1.48	3.44	1.36	1.78	2.72	-2.50	5.63	3.33	0.834	2.00

表 9: This table presents the performance of student models distilled from different teacher models, including Gemma-27B-it vs Gemma-9B-it and Llama3.1-72B-Instruct vs Llama3.1-8B-Instruct. We observe that small student models may experience degraded performance when distilled from a large teacher compared to a small teacher, whereas larger student models benefit more from the distilling a large teacher.

exhibit the largest rank shifts as the most shifted tokens. We choose Qwen2.5-3B-Instruct as the student model. We put the results of most shifted tokens after fine-tuning on long CoT data in Figure 9 and 10. The results of most shifted tokens after fine-tuning on large teacher CoT data are shown in Figure 11. Our analysis reveals that these tokens are predominantly associated with expressive and stylistic elements, such as “wait” , “But” , and “Let” .

D Examples of Various CoT Data

This section demonstrates examples of different CoT data including long CoT, short CoT, large teacher CoT and small teacher CoT. See the following examples.

Example of most shifted tokens after fine-tuning on long CoT data. (1/2)

So I’ve got this math problem here. It says: “Let (x, y) be an ordered pair of real numbers that satisfies the equation $x^2 + y^2 = 14x + 48y$. What is the maximum value of y ? Okay, let’s see how to approach this.

First, I need to understand what the equation is telling me. It’s an equation involving x and y , and it’s set equal to zero:

$$x^2 + y^2 - 14x - 48y = 0.$$

I need to find the maximum value of y for real numbers x and y that satisfy this equation.

Hmm, maybe I can rearrange this equation to make it easier to work with. Let’s see:

$$x^2 - 14x + y^2 - 48y = 0.$$

I can complete the square for both x and y terms.

For x : $x^2 - 14x$, to complete the square, I take half of -14 , which is -7 , and square it to get 49. So, $x^2 - 14x$ becomes $(x - 7)^2 - 49$.

For y : $y^2 - 48y$, half of -48 is -24 , squared is 576. So, $y^2 - 48y$ becomes $(y - 24)^2 - 576$.

Now, plugging these back into the equation:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0.$$

Simplify:

$$(x - 7)^2 + (y - 24)^2 - 625 = 0,$$

$$\text{So, } (x - 7)^2 + (y - 24)^2 = 625.$$

This looks like the equation of a circle with center at $(7, 24)$ and radius 25.

Wait, but the problem is asking for the maximum value of y , and I have a circle equation. So, the maximum y would be the y -coordinate of the highest point on the circle, which is the center plus the radius.

Wait, but the center is at $(7, 24)$, and the radius is 25, so the highest point would be at $(7, 24 + 25) = (7, 49)$.

But, let me double-check if this makes sense. The equation is

$$x^2 + y^2 = 14x + 48y,$$

and I’ve transformed it into

$$(x - 7)^2 + (y - 24)^2 = 625,$$

which is indeed a circle with center $(7, 24)$ and radius 25.

So, the maximum y would be $24 + 25 = 49$.

But, wait, the problem says “the maximum value of y ,” and I’ve found that to be 49. But let me consider if there are any other constraints or if I’ve missed something.

Let me think about the original equation:

$$x^2 + y^2 = 14x + 48y.$$

17

I can rearrange it to:

$$x^2 - 14x + y^2 - 48y = 0.$$

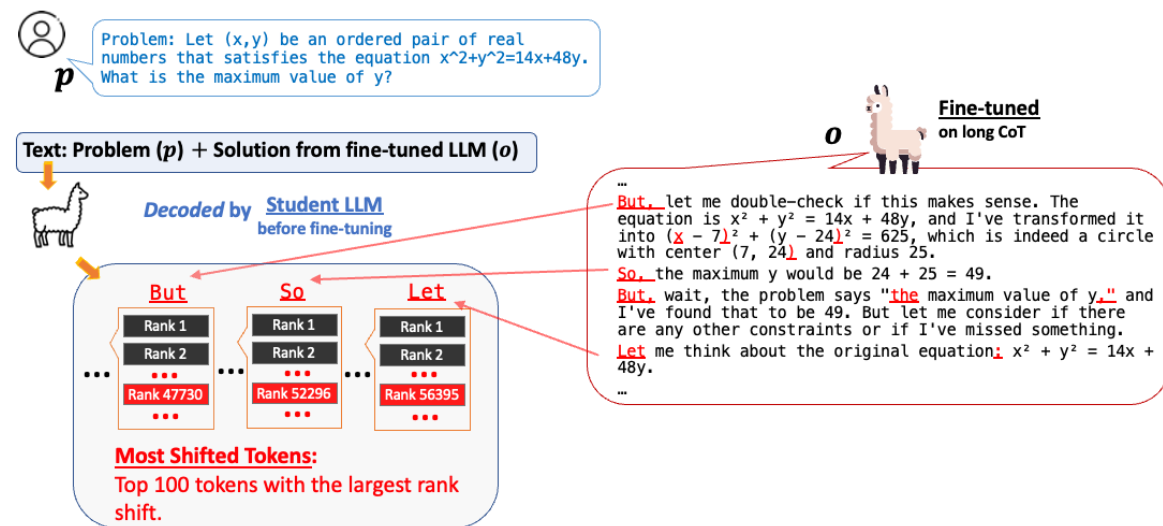


图 8: The process of calculating most shifted tokens. We decode each token generated by the fine-tuned LLM in the student model before fine-tuning. Then we calculate the rank shift in the student model for each token generated by the fine-tuned model. We annotate the tokens that exhibit the largest rank shifts as the most shifted tokens. We found that these tokens are predominantly associated with expressive and stylistic elements, such as “But” and “Let”.

Problem

Let (x, y) be an ordered pair of real numbers that satisfies the equation $x^2 + y^2 = 14x + 48y$. What is the maximum value of y ?

Example of most shifted tokens after fine-tuning on long CoT data (2/2)

Then, completing the square:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0,$$

$$(x - 7)^2 + (y - 24)^2 = 625.$$

Yes, that's a circle with center $(7, 24)$ and radius 25.

So, the maximum y is indeed $24 + 25 = 49$.

But, just to be thorough, let's consider if there are any points on the circle where y is greater than 49. The center is at $(7, 24)$, and the radius is 25, so the highest point is at $(7, 49)$, and the lowest point would be at $(7, -1)$.

But, the problem is asking for the maximum value of y , and I've found that to be 49.

Therefore, the final answer is 49.

Final Answer:

49

图 10: 大多数偏移的标记在长CoT数据上微调后 (2/2)。所有最偏移的标记都标为红色。

Example of **most shifted tokens** after fine-tuning on long CoT data. (1/2)

So I've got this math problem here. It says: "Let (x, y) be an ordered pair of real numbers that satisfies the equation $x^2 + y^2 = 14x + 48y$. What is the maximum value of y ? Okay, let's see how to approach this.

First, I need to understand what the equation is telling me. It's an equation involving x and y , and it's set equal to zero:

$$x^2 + y^2 - 14x - 48y = 0.$$

I need to find the maximum value of y for real numbers x and y that satisfy this equation.

Hmm, maybe I can rearrange this equation to make it easier to work with. Let's see:

$$x^2 - 14x + y^2 - 48y = 0.$$

I can complete the square for both x and y terms.

For x : $x^2 - 14x$, to complete the square, I take half of -14 , which is -7 , and square it to get 49 .

So, $x^2 - 14x$ becomes $(x - 7)^2 - 49$.

For y : $y^2 - 48y$, half of -48 is -24 , squared is 576 . So, $y^2 - 48y$ becomes $(y - 24)^2 - 576$.

Now, plugging these back into the equation:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0.$$

Simplify:

$$(x - 7)^2 + (y - 24)^2 - 625 = 0,$$

$$\text{So, } (x - 7)^2 + (y - 24)^2 = 625.$$

This looks like the equation of a circle with center at $(7, 24)$ and radius 25 .

Wait, but the problem is asking for the maximum value of y , and I have a circle equation. So, the maximum y would be the y -coordinate of the highest point on the circle, which is the center plus the radius.

Wait, but the center is at $(7, 24)$, and the radius is 25 , so the highest point would be at $(7, 24 + 25) = (7, 49)$.

But, let me double-check if this makes sense. The equation is

$$x^2 + y^2 = 14x + 48y,$$

and I've transformed it into

$$(x - 7)^2 + (y - 24)^2 = 625,$$

which is indeed a circle with center $(7, 24)$ and radius 25 .

So, the maximum y would be $24 + 25 = 49$.

But, wait, the problem says "the maximum value of y ," and I've found that to be 49 . But let me consider if there are any other constraints or if I've missed something.

Let me think about the original equation:

$$x^2 + y^2 = 14x + 48y.$$

19

I can rearrange it to:

$$x^2 - 14x + y^2 - 48y = 0.$$

Example of **Most Shifted Tokens** after fine-tuning on large model CoT

To solve the problem, we start with the given equation:

$$x^2 + y^2 = 14x + 48y$$

First, we rearrange the equation to bring all terms to one side:

$$x^2 - 14x + y^2 - 48y = 0$$

Next, we complete the square for both x and y .

For x :

$$x^2 - 14x$$

To complete the square, we add and subtract $\left(\frac{14}{2}\right)^2 = 49$:

$$x^2 - 14x + 49 - 49$$

$$(x - 7)^2 - 49$$

For y :

$$y^2 - 48y$$

To complete the square, we add and subtract $\left(\frac{48}{2}\right)^2 = 576$:

$$y^2 - 48y + 576 - 576$$

$$(y - 24)^2 - 576$$

Now, substitute these completed squares back into the equation:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0$$

$$(x - 7)^2 + (y - 24)^2 - 625 = 0$$

$$(x - 7)^2 + (y - 24)^2 = 625$$

This is the equation of a circle with center $(7, 24)$ and radius 25 .

To find the maximum value of y , we need to find the highest point on the circle. The highest point on the circle is directly above the center, which is at $y = 24 + 25 = 49$.

Thus, the maximum value of y is:

$$y = 49$$

Final Answer: 49

图 11: 在大型教师CoT上微调后最偏移的标记的示例。所有最偏移的标记均标为红色。

Example of **most shifted tokens** after fine-tuning on long CoT data (2/2)

Then, completing the square:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0,$$
$$(x - 7)^2 + (y - 24)^2 = 625.$$

Yes, that's a circle with center $(7, 24)$ and radius 25.

So, the maximum y is indeed $24 + 25 = 49$.

But, just to be thorough, let's consider if there are any points on the circle where y is greater than 49. The center is at $(7, 24)$, and the radius is 25, so the highest point is at $(7, 49)$, and the lowest point would be at $(7, -1)$.

But, the problem is asking for the maximum value of y , and I've found that to be 49. Therefore, the final answer is 49.

Final Answer:

49

Problem

What is the sum of the positive odd divisors of 60?

图 10: **Most shifted tokens** after fine-tuning on long CoT data (2/2). All the most shifted tokens are labeled as red.

Example of **Most Shifted Tokens** after fine-tuning on large model CoT

To solve the problem, we start with the given equation:

$$x^2 + y^2 = 14x + 48y$$

First, we rearrange the equation to bring all terms to one side:

$$x^2 - 14x + y^2 - 48y = 0$$

Next, we complete the square for both x and y .

For x :

$$x^2 - 14x$$

To complete the square, we add and subtract $\left(\frac{14}{2}\right)^2 = 49$:

$$x^2 - 14x + 49 - 49$$

$$(x - 7)^2 - 49$$

For y :

$$y^2 - 48y$$

To complete the square, we add and subtract $\left(\frac{48}{2}\right)^2 = 576$:

$$y^2 - 48y + 576 - 576$$

$$(y - 24)^2 - 576$$

Now, substitute these completed squares back into the equation:

$$(x - 7)^2 - 49 + (y - 24)^2 - 576 = 0$$

$$(x - 7)^2 + (y - 24)^2 - 625 = 0$$

$$(x - 7)^2 + (y - 24)^2 = 625$$

This is the equation of a circle with center $(7, 24)$ and radius 25.

To find the maximum value of y , we need to find the highest point on the circle. The highest point on the circle is directly above the center, which is at $y = 24 + 25 = 49$.

Thus, the maximum value of y is:

$$y = 49$$

Final Answer: 49

Long CoT Example

To find the sum of the positive odd divisors of 60, I first need to understand what the divisors of 60 are and then identify which of those are odd.

First, let's find all the positive divisors of 60. To do this, I can start by factoring 60 into its prime factors.

$$60 = 2^2 \times 3^1 \times 5^1$$

Now, to find all the positive divisors, I can take all possible products of these prime factors raised to their possible exponents. For 2, the exponents can be 0, 1, or 2; for 3, 0 or 1; and for 5, 0 or 1.

Listing them out:

$$1 = 2^0 \times 3^0 \times 5^0,$$

$$2 = 2^1 \times 3^0 \times 5^0,$$

$$4 = 2^2 \times 3^0 \times 5^0,$$

$$3 = 2^0 \times 3^1 \times 5^0,$$

$$6 = 2^1 \times 3^1 \times 5^0,$$

$$12 = 2^2 \times 3^1 \times 5^0,$$

$$5 = 2^0 \times 3^0 \times 5^1,$$

$$10 = 2^1 \times 3^0 \times 5^1,$$

$$20 = 2^2 \times 3^0 \times 5^1,$$

$$15 = 2^0 \times 3^1 \times 5^1,$$

$$30 = 2^1 \times 3^1 \times 5^1,$$

$$60 = 2^2 \times 3^1 \times 5^1.$$

So, the positive divisors of 60 are:

$$1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60.$$

Now, I need to identify which of these are odd. Odd numbers are not divisible by 2, so I can eliminate any divisors that are even.

The odd divisors are:

$$1, 3, 5, 15.$$

Now, I need to find the sum of these odd divisors: $1 + 3 + 5 + 15 = 24$. Alternatively, using the prime factorization approach: Since the odd divisors are those that do not include the factor 2, I can ignore the 2's in the prime factorization. The remaining odd part is:

$$3^1 \times 5^1.$$

The sum of the divisors of this odd part is:

$$(1 + 3)(1 + 5) = 4 \times 6 = 24.$$

This matches the previous result.

Therefore, for $60 = 2^2 \times 3^1 \times 5^1$, the sum of the odd divisors is:

图 11: Example of **most shifted tokens** after fine-tuning on large teacher CoT. All the most shifted tokens are labeled as red.

Problem

What is the sum of the positive odd divisors of 60?

Short CoT Example

To find the sum of the positive odd divisors of 60, we first need to factorize 60 into its prime factors. The prime factorization of 60 is:

$$60 = 2^2 \times 3 \times 5$$

Since we are only interested in the odd divisors, we can ignore the factor of 2. Therefore, we only consider the factors 3 and 5. The odd divisors of 60 are the divisors of $3 \times 5 = 15$. The divisors of 15 are 1, 3, 5, and 15.

Next, we sum these divisors:

$$1 + 3 + 5 + 15 = 24$$

Therefore, the sum of the positive odd divisors of 60 is:

$$\boxed{24}$$

图 13: Short CoT example

Problem

Find the greatest a such that

$$\frac{7\sqrt{(2a)^2 + (1)^2} - 4a^2 - 1}{\sqrt{1 + 4a^2} + 3} = 2.$$

Long CoT Example

To find the sum of the positive odd divisors of 60, I first need to understand what the divisors of 60 are and then identify which of those are odd.

First, let's find all the positive divisors of 60. To do this, I can start by factoring 60 into its prime factors.

$$60 = 2^2 \times 3^1 \times 5^1$$

Now, to find all the positive divisors, I can take all possible products of these prime factors raised to their possible exponents. For 2, the exponents can be 0, 1, or 2; for 3, 0 or 1; and for 5, 0 or 1. Listing them out:

$$1 = 2^0 \times 3^0 \times 5^0,$$

$$2 = 2^1 \times 3^0 \times 5^0,$$

$$4 = 2^2 \times 3^0 \times 5^0,$$

$$3 = 2^0 \times 3^1 \times 5^0,$$

$$6 = 2^1 \times 3^1 \times 5^0,$$

$$12 = 2^2 \times 3^1 \times 5^0,$$

$$5 = 2^0 \times 3^0 \times 5^1,$$

$$10 = 2^1 \times 3^0 \times 5^1,$$

$$20 = 2^2 \times 3^0 \times 5^1,$$

$$15 = 2^0 \times 3^1 \times 5^1,$$

$$30 = 2^1 \times 3^1 \times 5^1,$$

$$60 = 2^2 \times 3^1 \times 5^1.$$

So, the positive divisors of 60 are:

$$1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60.$$

Now, I need to identify which of these are odd. Odd numbers are not divisible by 2, so I can eliminate any divisors that are even.

The odd divisors are:

$$1, 3, 5, 15.$$

Now, I need to find the sum of these odd divisors: $1 + 3 + 5 + 15 = 24$. Alternatively, using the prime factorization approach: Since the odd divisors are those that do not include the factor 2, I can ignore the 2's in the prime factorization. The remaining odd part is:

$$3^1 \times 5^1.$$

The sum of the divisors of this odd part is:

$$(1 + 3)(1 + 5) = 4 \times 6 = 24.$$

This matches the previous result.

Therefore, for $60 = 2^2 \times 3^1 \times 5^1$, the sum of the odd divisors is:

Large Teacher CoT Example (1/2)

To solve the equation $\frac{7\sqrt{(2a)^2+1^2}-4a^2-1}{\sqrt{1+4a^2}+3} = 2$, we start by simplifying the expression inside the square root in the numerator. The equation becomes:

$$\frac{7\sqrt{4a^2+1}-4a^2-1}{\sqrt{1+4a^2}+3} = 2$$

Let's denote $x = \sqrt{1+4a^2}$. Then the equation transforms into:

$$\frac{7x-4a^2-1}{x+3} = 2$$

Next, we clear the fraction by multiplying both sides by $x+3$:

$$7x-4a^2-1 = 2(x+3)$$

Expanding and simplifying the right side, we get:

$$7x-4a^2-1 = 2x+6$$

Rearranging terms to isolate x and a^2 , we have:

$$7x-2x-4a^2-1-6=0 \implies 5x-4a^2-7=0 \implies 5x=4a^2+7 \implies x=\frac{4a^2+7}{5}$$

Recall that $x = \sqrt{1+4a^2}$. Therefore, we have:

$$\sqrt{1+4a^2} = \frac{4a^2+7}{5}$$

To eliminate the square root, we square both sides:

$$1+4a^2 = \left(\frac{4a^2+7}{5}\right)^2$$

Expanding the right side, we get:

$$1+4a^2 = \frac{(4a^2+7)^2}{25} = \frac{16a^4+56a^2+49}{25}$$

Multiplying both sides by 25 to clear the denominator, we obtain:

$$25+100a^2 = 16a^4+56a^2+49$$

Short CoT Example

To find the sum of the positive odd divisors of 60, we first need to factorize 60 into its prime factors. The prime factorization of 60 is:

$$60 = 2^2 \times 3 \times 5$$

Since we are only interested in the odd divisors, we can ignore the factor of 2. Therefore, we only consider the factors 3 and 5. The odd divisors of 60 are the divisors of $3 \times 5 = 15$. The divisors of 15 are 1, 3, 5, and 15.

Next, we sum these divisors:

$$1 + 3 + 5 + 15 = 24$$

Therefore, the sum of the positive odd divisors of 60 is:

$$\boxed{24}$$

图 13: Short CoT example

Problem

Find the greatest a such that

$$\frac{7\sqrt{(2a)^2 + (1)^2} - 4a^2 - 1}{\sqrt{1 + 4a^2} + 3} = 2.$$

Large Teacher CoT Example (2/2)

Rearranging all terms to one side, we get a standard polynomial equation:

$$16a^4 + 56a^2 + 49 - 100a^2 - 25 = 0 \implies 16a^4 - 44a^2 + 24 = 0$$

Let $y = a^2$. Then the equation becomes a quadratic in y :

$$16y^2 - 44y + 24 = 0$$

We solve this quadratic equation using the quadratic formula $y = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $a = 16$, $b = -44$, and $c = 24$:

$$y = \frac{44 \pm \sqrt{(-44)^2 - 4 \cdot 16 \cdot 24}}{2 \cdot 16} = \frac{44 \pm \sqrt{1936 - 1536}}{32} = \frac{44 \pm \sqrt{400}}{32} = \frac{44 \pm 20}{32}$$

This gives us two solutions for y :

$$y = \frac{64}{32} = 2 \quad \text{and} \quad y = \frac{24}{32} = \frac{3}{4}$$

Since $y = a^2$, we have $a^2 = 2$ or $a^2 = \frac{3}{4}$. The corresponding values for a are $a = \pm\sqrt{2}$ and $a = \pm\frac{\sqrt{3}}{2}$. The greatest value of a is $\sqrt{2}$.

Thus, the greatest a that satisfies the equation is:

$$\boxed{\sqrt{2}}$$

Large Teacher CoT Example (1/2)

To solve the equation $\frac{7\sqrt{(2a)^2+1^2-4a^2-1}}{\sqrt{1+4a^2}+3} = 2$, we start by simplifying the expression inside the square root in the numerator. The equation becomes:

$$\frac{7\sqrt{4a^2+1}-4a^2-1}{\sqrt{1+4a^2}+3} = 2$$

Let's denote $x = \sqrt{1+4a^2}$. Then the equation transforms into:

$$\frac{7x-4a^2-1}{x+3} = 2$$

Next, we clear the fraction by multiplying both sides by $x+3$:

$$7x-4a^2-1 = 2(x+3)$$

Expanding and simplifying the right side, we get:

$$7x-4a^2-1 = 2x+6$$

Rearranging terms to isolate x and a^2 , we have:

$$7x-2x-4a^2-1-6=0 \implies 5x-4a^2-7=0 \implies 5x=4a^2+7 \implies x=\frac{4a^2+7}{5}$$

Recall that $x = \sqrt{1+4a^2}$. Therefore, we have:

$$\sqrt{1+4a^2} = \frac{4a^2+7}{5}$$

To eliminate the square root, we square both sides:

$$1+4a^2 = \left(\frac{4a^2+7}{5}\right)^2$$

Expanding the right side, we get:

$$1+4a^2 = \frac{(4a^2+7)^2}{25} = \frac{16a^4+56a^2+49}{25}$$

Multiplying both sides by 25 to clear the denominator, we obtain:

$$25+100a^2 = 16a^4+56a^2+49$$

Small Teacher CoT Example (1/2)

To solve the equation $\frac{7\sqrt{(2a)^2+(1)^2-4a^2-1}}{\sqrt{1+4a^2}+3} = 2$, we start by simplifying the expression inside the square root and the equation itself.

First, simplify the expression under the square root in the numerator:

$$(2a)^2+1^2=4a^2+1$$

So the equation becomes:

$$\frac{7\sqrt{4a^2+1}-4a^2-1}{\sqrt{4a^2+1}+3} = 2$$

Next, let $x = \sqrt{4a^2+1}$. Then the equation transforms into:

$$\frac{7x-4a^2-1}{x+3} = 2$$

Since $x = \sqrt{4a^2+1}$, we can express $4a^2$ in terms of x :

$$4a^2 = x^2 - 1$$

Substitute $4a^2 = x^2 - 1$ into the equation:

$$\frac{7x-(x^2-1)-1}{x+3} = 2$$

Simplify the numerator:

$$7x-x^2+1-1=7x-x^2$$

So the equation becomes:

$$\frac{7x-x^2}{x+3} = 2$$

Multiply both sides by $x+3$ to clear the denominator:

$$7x-x^2=2(x+3)$$

Expand and simplify:

$$7x-x^2=2x+6$$

Rearrange all terms to one side:

$$-x^2+7x-2x-6=0$$

$$-x^2+5x-6=0$$

Multiply through by -1 to make the leading coefficient positive:

$$x^2-5x+6=0$$

Large Teacher CoT Example (2/2)

Rearranging all terms to one side, we get a standard polynomial equation:

$$16a^4 + 56a^2 + 49 - 100a^2 - 25 = 0 \implies 16a^4 - 44a^2 + 24 = 0$$

Let $y = a^2$. Then the equation becomes a quadratic in y :

$$16y^2 - 44y + 24 = 0$$

We solve this quadratic equation using the quadratic formula $y = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $a = 16$, $b = -44$, and $c = 24$:

$$y = \frac{44 \pm \sqrt{(-44)^2 - 4 \cdot 16 \cdot 24}}{2 \cdot 16} = \frac{44 \pm \sqrt{1936 - 1536}}{32} = \frac{44 \pm \sqrt{400}}{32} = \frac{44 \pm 20}{32}$$

This gives us two solutions for y :

$$y = \frac{64}{32} = 2 \quad \text{and} \quad y = \frac{24}{32} = \frac{3}{4}$$

Since $y = a^2$, we have $a^2 = 2$ or $a^2 = \frac{3}{4}$. The corresponding values for a are $a = \pm\sqrt{2}$ and $a = \pm\frac{\sqrt{3}}{2}$. The greatest value of a is $\sqrt{2}$.

Thus, the greatest a that satisfies the equation is:

$$\boxed{\sqrt{2}}$$

Small Teacher CoT Example (2/2)

Factor the quadratic equation:

$$(x - 2)(x - 3) = 0$$

So, the solutions are:

$$x = 2 \quad \text{or} \quad x = 3$$

Recall that $x = \sqrt{4a^2 + 1}$. We need to check both solutions:

1. For $x = 2$:

$$\sqrt{4a^2 + 1} = 2$$

Square both sides:

$$4a^2 + 1 = 4$$

$$4a^2 = 3$$

$$a^2 = \frac{3}{4}$$

$$a = \pm \frac{\sqrt{3}}{2}$$

2. For $x = 3$:

$$\sqrt{4a^2 + 1} = 3$$

Square both sides:

$$4a^2 + 1 = 9$$

$$4a^2 = 8$$

$$a^2 = 2$$

$$a = \pm\sqrt{2}$$

The greatest value of a among the solutions is $\sqrt{2}$.

Thus, the final answer is:

$$\boxed{\sqrt{2}}$$

Problem

What is the greatest common divisor of 128, 144 and 480?

Small Teacher CoT Example (1/2)

To solve the equation $\frac{7\sqrt{(2a)^2+(1)^2-4a^2-1}}{\sqrt{1+4a^2+3}} = 2$, we start by simplifying the expression inside the square root and the equation itself.

First, simplify the expression under the square root in the numerator:

$$(2a)^2 + 1^2 = 4a^2 + 1$$

So the equation becomes:

$$\frac{7\sqrt{4a^2+1}-4a^2-1}{\sqrt{4a^2+1}+3} = 2$$

Next, let $x = \sqrt{4a^2+1}$. Then the equation transforms into:

$$\frac{7x-4a^2-1}{x+3} = 2$$

Since $x = \sqrt{4a^2+1}$, we can express $4a^2$ in terms of x :

$$4a^2 = x^2 - 1$$

Substitute $4a^2 = x^2 - 1$ into the equation:

$$\frac{7x-(x^2-1)-1}{x+3} = 2$$

Simplify the numerator:

$$7x-x^2+1-1=7x-x^2$$

So the equation becomes:

$$\frac{7x-x^2}{x+3} = 2$$

Multiply both sides by $x+3$ to clear the denominator:

$$7x-x^2=2(x+3)$$

Expand and simplify:

$$7x-x^2=2x+6$$

Rearrange all terms to one side:

$$-x^2+7x-2x-6=0$$

$$-x^2+5x-6=0$$

Multiply through by -1 to make the leading coefficient positive:

$$x^2-5x+6=0$$

Large Teacher CoT Example

To find the greatest common divisor (GCD) of 128, 144, and 480, we can use the prime factorization method.

1. Prime Factorization:

$$128 = 2^7,$$

$$144 = 2^4 \times 3^2,$$

$$480 = 2^5 \times 3 \times 5.$$

2. Identify the common prime factors:

The common prime factor among 128, 144, and 480 is 2.

3. Determine the lowest power of the common prime factor:

The lowest power of 2 in the factorizations is 2^4 .

4. Calculate the GCD:

The GCD is:

$$2^4 = 16.$$

Therefore, the greatest common divisor of 128, 144, and 480 is:

$$\boxed{16}$$

Small Teacher CoT Example (2/2)

Factor the quadratic equation:

$$(x - 2)(x - 3) = 0$$

So, the solutions are:

$$x = 2 \quad \text{or} \quad x = 3$$

Recall that $x = \sqrt{4a^2 + 1}$. We need to check both solutions:

1. For $x = 2$:

$$\sqrt{4a^2 + 1} = 2$$

Square both sides:

$$4a^2 + 1 = 4$$

$$4a^2 = 3$$

$$a^2 = \frac{3}{4}$$

$$a = \pm \frac{\sqrt{3}}{2}$$

2. For $x = 3$:

$$\sqrt{4a^2 + 1} = 3$$

Square both sides:

$$4a^2 + 1 = 9$$

$$4a^2 = 8$$

$$a^2 = 2$$

$$a = \pm\sqrt{2}$$

The greatest value of a among the solutions is $\sqrt{2}$.

Thus, the final answer is:

$$\boxed{\sqrt{2}}$$

Small Teacher CoT Example

To find the greatest common divisor (GCD) of 128, 144, and 480, we start by finding the prime factorization of each number.

The prime factorization of 128 is:

$$128 \div 2 = 64$$

$$64 \div 2 = 32$$

$$32 \div 2 = 16$$

$$16 \div 2 = 8$$

$$8 \div 2 = 4$$

$$4 \div 2 = 2$$

$$2 \div 2 = 1$$

Thus,

$$128 = 2^7.$$

The prime factorization of 144 is:

$$144 \div 2 = 72$$

$$72 \div 2 = 36$$

$$36 \div 2 = 18$$

$$18 \div 2 = 9$$

$$9 \div 3 = 3$$

$$3 \div 3 = 1$$

Thus,

$$144 = 2^4 \times 3^2.$$

The prime factorization of 480 is:

$$480 \div 2 = 240$$

$$240 \div 2 = 120$$

$$120 \div 2 = 60$$

$$60 \div 2 = 30$$

$$30 \div 2 = 15$$

$$15 \div 3 = 5$$

$$5 \div 5 = 1$$

Thus,

$$480 = 2^5 \times 3 \times 5.$$

Now, we identify the common prime factors. The factorizations are:

$$128 = 2^7, \quad 144 = 2^4 \times 3^2, \quad 480 = 2^5 \times 3 \times 5.$$

The only common prime factor among all three numbers is 2.

Next, we determine the smallest power of the common prime factor. The lowest power of 2 among these factorizations is 2^4 .

Therefore, the GCD of 128, 144, and 480 is:

Problem

What is the greatest common divisor of 128, 144 and 480?

Large Teacher CoT Example

To find the greatest common divisor (GCD) of 128, 144, and 480, we can use the prime factorization method.

1. Prime Factorization:

$$\begin{aligned}128 &= 2^7, \\144 &= 2^4 \times 3^2, \\480 &= 2^5 \times 3 \times 5.\end{aligned}$$

2. Identify the common prime factors:

The common prime factor among 128, 144, and 480 is 2.

3. Determine the lowest power of the common prime factor:

The lowest power of 2 in the factorizations is 2^4 .

4. Calculate the GCD:

The GCD is:

$$2^4 = 16.$$

Therefore, the greatest common divisor of 128, 144, and 480 is:

16

Problem

Rounded to 2 decimal places, what is $\frac{7}{9}$?

Large Teacher CoT Example

To solve the problem, we need to divide 7 by 9 and then round the result to two decimal places.

$$\frac{7}{9} = 0.777\ldots$$

Rounded to two decimal places, this becomes 0.78.

Final Answer: 0.78

Small Teacher CoT Example

To solve the problem of finding the value of $\frac{7}{9}$ rounded to 2 decimal places, we can follow these steps:

1. **Perform the division:** Calculate $\frac{7}{9}$.
2. **Convert to decimal:** Perform the division to get the decimal equivalent.
3. **Round to 2 decimal places:** Round the decimal to 2 decimal places.

Let's start with the division:

$$\frac{7}{9} \approx 0.777\ldots$$

Next, we need to round this decimal to 2 decimal places. The digit in the third decimal place is 7, which is greater than or equal to 5. Therefore, we round the second decimal place up from 7 to 8. So, $0.777\ldots$ rounded to 2 decimal places is 0.78.

The final answer is:

0.78

Small Teacher CoT Example

To find the greatest common divisor (GCD) of 128, 144, and 480, we start by finding the prime factorization of each number.

The prime factorization of 128 is:

$$\begin{aligned} 128 \div 2 &= 64 \\ 64 \div 2 &= 32 \\ 32 \div 2 &= 16 \\ 16 \div 2 &= 8 \\ 8 \div 2 &= 4 \\ 4 \div 2 &= 2 \\ 2 \div 2 &= 1 \end{aligned}$$

Thus,

$$128 = 2^7.$$

The prime factorization of 144 is:

$$\begin{aligned} 144 \div 2 &= 72 \\ 72 \div 2 &= 36 \\ 36 \div 2 &= 18 \\ 18 \div 2 &= 9 \\ 9 \div 3 &= 3 \\ 3 \div 3 &= 1 \end{aligned}$$

Thus,

$$144 = 2^4 \times 3^2.$$

The prime factorization of 480 is:

$$\begin{aligned} 480 \div 2 &= 240 \\ 240 \div 2 &= 120 \\ 120 \div 2 &= 60 \\ 60 \div 2 &= 30 \\ 30 \div 2 &= 15 \\ 15 \div 3 &= 5 \\ 5 \div 5 &= 1 \end{aligned}$$

Thus,

$$480 = 2^5 \times 3 \times 5.$$

Now, we identify the common prime factors. The factorizations are:

$$128 = 2^7, \quad 144 = 2^4 \times 3^2, \quad 480 = 2^5 \times 3 \times 5.$$

The only common prime factor among all three numbers is 2.

Next, we determine the smallest power of the common prime factor. The lowest power of 2 among these factorizations is 2^4 .

Therefore, the GCD of 128, 144, and 480 is:

Problem

Rounded to 2 decimal places, what is $\frac{7}{9}$?

Large Teacher CoT Example

To solve the problem, we need to divide 7 by 9 and then round the result to two decimal places.

$$\frac{7}{9} = 0.777 \dots$$

Rounded to two decimal places, this becomes 0.78.

Final Answer: 0.78

Small Teacher CoT Example

To solve the problem of finding the value of $\frac{7}{9}$ rounded to 2 decimal places, we can follow these steps:

1. **Perform the division:** Calculate $\frac{7}{9}$.
2. **Convert to decimal:** Perform the division to get the decimal equivalent.
3. **Round to 2 decimal places:** Round the decimal to 2 decimal places.

Let's start with the division:

$$\frac{7}{9} \approx 0.7777 \dots$$

Next, we need to round this decimal to 2 decimal places. The digit in the third decimal place is 7, which is greater than or equal to 5. Therefore, we round the second decimal place up from 7 to 8.

So, $0.7777 \dots$ rounded to 2 decimal places is 0.78.

The final answer is:

0.78