Revisiting the Capacity Gap in Chain-of-Thought Distillation from a Practical Perspective

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) distillation transfers reasoning behaviors from a strong teacher to a smaller student, but prior work reports a capacity gap: distillation may fail when the teacher–student capability mismatch is large. We revisit the capacity gap from a practical perspective by re-examining commonly used experimental settings. Notably, we find that CoT distillation often degrades performance compared to the student's pre-distillation baseline, and that some commonly used settings, while suitable for establishing the capacity gap as a phenomenon, do not reflect realistic deployment scenarios. Complementing prior work that establishes the capacity gap, we evaluate its practical impact under more realistic settings and find that it does not consistently dominate; stronger teachers tend to be preferable when candidate teachers differ substantially in performance. Our results offer practical guidance for selecting teacher–student pairs in CoT distillation.

1 Introduction

Large language models (LLMs) demonstrate strong reasoning when prompted to generate intermediate steps (**chain-of-thought (CoT)** prompting; Wei et al., 2022; Kojima et al., 2022), but the resulting long outputs raise deployment costs. **CoT distillation** mitigates this by training a smaller student to mimic the reasoning of a larger teacher (Ho et al., 2023; Li et al., 2023; Fu et al., 2023; Hsieh et al., 2023; Magister et al., 2023; Shridhar et al., 2023).[1]

A widely cited concern is the **capacity gap**: distillation may fail when the teacher–student capability mismatch is large (Li et al., 2025; Chen et al., 2025), implying that the strongest teacher is not always optimal and that practitioners face a costly hyperparameter search over candidate teachers.

[1]CoT distillation typically targets task-specific reasoning transfer, as opposed to large-scale pretraining-time distillation, which aims at broad capability acquisition.

In this paper, we revisit the capacity gap from a practical perspective. **Our aim is not to dispute the existence of the capacity gap, but to provide a complementary assessment of how much it matters in a realistic scenario**: given a task where the small, efficient student cannot adequately solve it without additional training, which teacher should be selected to distill into the student? We identify two protocol-level issues that obscure this practitioner's question in prior work: (1) configurations are compared only *after* distillation, without checking whether distillation improves over the pre-distillation baseline, and we find that it often *degrades* performance; and (2) practices such as cross-teacher data filtering and inclusion of larger-student settings strip stronger teachers of their natural advantages.

By adding pre-distillation baselines and removing these constraints, we find that the capacity gap does not consistently dominate across tasks and settings, and that when candidate teachers differ substantially in performance, the benefits of the stronger teacher outweigh capacity-gap effects.

Our contributions are as follows:

- We identify pitfalls in existing evaluation protocols from a practical perspective, most notably the missing comparison against the pre-distillation baseline, which obscures cases where distillation *degrades* performance.
- Through systematic experiments across model families under a more realistic protocol, we show that capacity-gap effects do not consistently dominate across tasks and settings.
- We derive conditional guidelines for practitioners: first verify that distillation improves over the pre-distillation baseline, and *when* this is confirmed *and* candidate teachers differ substantially in performance, prefer the higher-performing teacher.

2 Preliminaries

2.1 CoT Distillation

CoT distillation aims to transfer the reasoning capabilities of a large *teacher* model to a smaller, more efficient *student* model. The general procedure consists of two stages.

Rationale Generation. Given an original training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with ground-truth labels y_i, the teacher model $\mathcal{T}$ generates reasoning chains (rationales) $\{\hat{r}_i\}_{i=1}^{N}$ along with predicted answers $\{\hat{y}_i\}_{i=1}^{N}$. To ensure the quality of reasoning chains, only samples where the predicted answer matches the ground-truth label are retained:

$$\mathcal{D}_{\text{CoT}} = \{(x_i, \hat{r}_i, \hat{y}_i) \mid (x_i, y_i) \in \mathcal{D},\ \hat{y}_i = y_i\}. \tag{1}$$

Student Fine-tuning. The student model $\mathcal{S}$ is fine-tuned on the distillation dataset $\mathcal{D}_{\text{CoT}}$ to learn to produce similar reasoning chains. Formally, the student is trained to minimize the negative log-likelihood:

$$\mathcal{L} = -\sum_{(x,\hat{r},\hat{y}) \in \mathcal{D}_{\text{CoT}}} \log P_{\mathcal{S}}(\hat{r}, \hat{y} \mid x), \tag{2}$$

where $P_{\mathcal{S}}$ denotes the probability distribution of the student model.

This hard distillation can be viewed as a special case of sequence-level knowledge distillation (Kim and Rush, 2016), where the teacher distribution is approximated via sampling rather than used directly as a soft distribution (Hinton et al., 2015). Hard distillation is the dominant paradigm in CoT distillation because teacher logits are often unavailable from proprietary models, and tokenizer mismatches across model families preclude logit-level transfer (Ho et al., 2023; Li et al., 2023; Hsieh et al., 2023; Magister et al., 2023; Shridhar et al., 2023).

2.2 Capacity Gap

The *capacity gap* refers to the empirical observation that distillation may yield suboptimal results when there is a substantial capability mismatch between the teacher and student models. Such mismatch can arise from various sources, including differences in model size or in reasoning style, e.g., short vs. long chain-of-thought (Li et al., 2025).

This phenomenon has been reported in conventional soft knowledge distillation (Lopez-Paz et al., 2016; Cho and Hariharan, 2019; Mirzadeh et al., 2020; Jafari et al., 2021). CoT distillation lies on the same spectrum as soft knowledge distillation but at the hard-label end (Section 2.1). Capacity gap analyses developed for soft knowledge distillation therefore do not straightforwardly transfer, requiring independent empirical investigation. Prior work has conducted such investigation and reports the same phenomenon in CoT distillation (Li et al., 2025; Chen et al., 2025), suggesting that using CoT from the strongest teacher may not always be optimal.

However, the practical impact of the capacity gap under realistic deployment conditions remains unexamined. Our work is therefore complementary to these prior studies: rather than disputing the existence of the gap, we focus on the hard CoT distillation setting and examine whether commonly used evaluation protocols reflect realistic deployment scenarios, thereby assessing how much the capacity gap actually matters in practice.

3 Pitfalls in Existing Evaluation Protocols

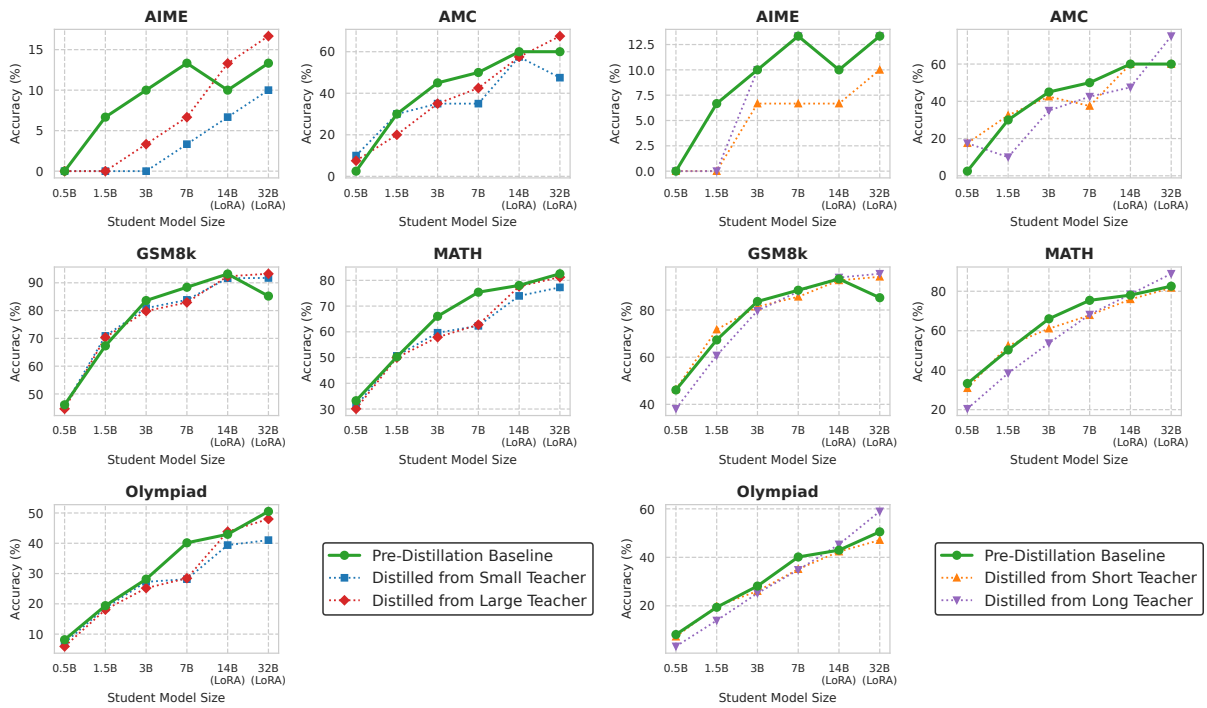
We critically examine experimental protocols commonly used in prior capacity gap studies (Li et al., 2025; Chen et al., 2025) and identify three pitfalls that may lead to conclusions that do not directly translate to practical deployment scenarios.

3.1 Pitfall 1: Lack of Pre-Distillation Baseline Comparison

Prior studies on the capacity gap (Li et al., 2025; Chen et al., 2025) compare different distillation configurations only *after* distillation, without examining whether distillation itself improves upon the student model's pre-distillation performance. This omission is problematic because recent LLMs already possess substantial reasoning capabilities and may already perform well on target tasks without fine-tuning. Therefore, verifying that distillation actually *improves* performance is essential for its practical adoption.

Experimental Setup. To examine whether prior experiments constitute a valid scenario where distillation is expected to be beneficial, we add comparison with pre-distillation baselines to the experimental settings of Li et al. (2025). Our experiments are based on their publicly available code and data to ensure consistent experimental settings.[2] Below, we briefly summarize the experimental setup; see Appendix A for more details.

[2] https://github.com/Small-Model-Gap/Small-Model-Learnability-Gap



(a) Results under the *small–large* setting.

(b) Results under the *short–long* setting.

Figure 1: Performance comparison on five math benchmarks (AIME, AMC, GSM8K, MATH, and OlympiadBench). We compare the pre-distillation baseline with distilled students. (a) *Small–large* setting: students distilled from Qwen2.5-3B-Instruct (Small Teacher) and Qwen2.5-72B-Instruct (Large Teacher). (b) *Short–long* setting: students distilled from Qwen2.5-32B-Instruct (Short Teacher) and QwQ-32B-Preview (Long Teacher).

Training Data and Teacher Models. We use the distillation datasets publicly released by Li et al. (2025). These datasets are constructed by generating CoT rationales for the MATH training set (Hendrycks et al., 2021) and filtering to retain only examples where both teachers in each comparison produce correct answers (see Section 3.2 for a discussion of the implications of this filtering). Two teacher comparisons are considered: the **small–large** setting, which compares teachers of different model sizes, specifically CoT from Qwen2.5-3B-Instruct with Qwen2.5-72B-Instruct (Qwen et al., 2025); and the **short–long** setting, which compares teachers that generate reasoning chains of different lengths, specifically short CoT from Qwen2.5-32B-Instruct with long CoT from QwQ-32B-Preview.

Evaluation. We evaluate student models on five benchmarks of varying difficulty: MATH, GSM8K (Cobbe et al., 2021), AMC 2023, AIME 2024, and OlympiadBench (He et al., 2024). To ensure a fair comparison, we apply the same prompt to all models, including the pre-distillation baseline.

Student Models and Training. We use Qwen2.5-{0.5B, 1.5B, 3B, 7B, 14B, 32B}-Instruct as student models and fine-tune them using `LLaMA-Factory` (Zheng et al., 2024). We use the same hyperparameters as Li et al. (2025).

Results. Figures 1a and 1b show the results in the small–large and short–long settings, respectively. We found a striking result: in most configurations, CoT distillation *degrades* performance compared to the student's pre-distillation baseline. A plausible interpretation is that recent instruction-tuned LLMs already possess strong reasoning capabilities acquired through extensive pretraining, and fine-tuning on teacher-generated rationales does not necessarily improve, and can degrade, these existing capabilities. We treat the specific mechanism as an open question; candidates include forgetting of pretrained knowledge, rationale-style mismatch, distribution shift, and teacher-CoT quality lower than the student's instruction-tuning data. This indicates that the experimental setting does not represent a practical scenario where distillation is beneficial; comparing distillation strategies then amounts to comparing degrees of degradation rather than identifying effective approaches.[3]

[3]We additionally found that such an impractical setting can lead to misleading artifacts: a strategy proposed in prior work

3.2 Pitfall 2: Data Filtering that Neutralizes Teacher Strength

When comparing different teacher models, Li et al. (2025) filter the training data to include only examples where *both* teachers produce correct answers. This is a reasonable methodological choice for isolating the capacity gap: by controlling the training data, one can attribute performance differences purely to the quality of reasoning chains. In this sense, prior work provides valid evidence that a capacity gap can exist under controlled conditions. However, our goal is to assess *how much the capacity gap matters in practice*. A practitioner would use all examples that the teacher solves correctly, and stronger teachers then provide two advantages: more training examples due to higher accuracy, and coverage of harder cases that the other teacher fails to solve. Cross-teacher filtering eliminates both.[4]

3.3 Pitfall 3: Inclusion of Larger-Student Settings

Some prior experiments include configurations where the student model is *larger* than the teacher model (Li et al., 2025). While such settings are valid for demonstrating the existence of the capacity gap, they do not align with the practical motivation for CoT distillation, whose primary goal is to reduce deployment costs by transferring knowledge to a smaller model. Since our focus is on assessing the practical impact of the capacity gap, we restrict experiments to settings where the student is strictly smaller than the teacher.

4 Re-Evaluation with Practical Protocols

We now propose a revised evaluation protocol that addresses the pitfalls in Section 3 and re-examine whether the capacity gap remains a concern under realistic conditions.

to mitigate the capacity gap appeared effective, but this was largely due to hyperparameter discrepancies that reduced the number of parameter updates rather than achieving genuine improvements (see Appendix B for details).

[4]Importantly, cross-teacher filtering should not be confused with curating high-quality training data, which has proven effective for improving LLM reasoning (Ye et al., 2025). Quality-based selection retains examples based on their instructional value, whereas cross-teacher filtering merely restricts data to the intersection of what both teachers solve correctly, discarding correct examples without any quality-based rationale. Indeed, our ablation study (Appendix C) shows that *applying cross-teacher filtering to our BBH experiments degrades overall distillation performance* and disproportionately penalizes stronger teachers, confirming that this filtering reduces useful supervision rather than improving data quality.

4.1 Proposed Evaluation Protocol

We propose a practical evaluation protocol consisting of the following modifications to existing experimental settings.

Modification 1: Task Selection Based on Expected Distillation Benefits. We first identify tasks where CoT distillation is likely to yield performance improvements, ensuring that we study scenarios where distillation is practically meaningful rather than comparing degrees of degradation.

We expect distillation to be most effective when the task requires knowledge beyond what pretraining provides. Although many-shot in-context learning (many-shot ICL; Agarwal et al., 2024) and distillation differ in mechanism, both aim to acquire task-specific knowledge from numerous examples; we therefore hypothesize that tasks where many-shot ICL is effective are also likely to benefit from distillation. Honda et al. (2025) report that many-shot ICL yields no gains on mathematical benchmarks with recent strong LLMs, but **BIG-Bench Hard** (**BBH**; Suzgun et al., 2023; Srivastava et al., 2023) remains an exception, likely because BBH tasks require task-specific patterns not inferable from prior knowledge alone. Based on this reasoning, we select BBH as our benchmark, expecting that distillation will similarly benefit from the availability of numerous task-specific examples.

BBH consists of 23 diverse tasks, and we further narrow down to tasks where distillation has substantial room to improve the student. Specifically, we measure the few-shot ICL performance of all teacher models and all pre-distillation student models on each task, and then compute the average performance within each group (i.e., the average across teacher models and the average across student models). We then retain tasks where the difference between these averages exceeds a threshold. This selection criterion ensures that there is sufficient room for the student to learn from the teacher, yielding tasks well-suited for examining CoT distillation in realistic scenarios.

Modification 2: Removal of Cross-Teacher Data Filtering. Instead of filtering training data to the intersection of examples solved correctly by both weak and strong teachers (an analytical convention in prior work), we use each teacher's full set of correct examples, reflecting the natural choice a practitioner would make and letting stronger teachers benefit from their higher accuracy. We acknowl-

edge that this protocol does not fully isolate the capacity gap, since reasoning quality and data quantity vary together across teachers. However, it suits our purpose of assessing the overall practical impact of teacher selection rather than its mechanistic breakdown. As a complementary analysis, Appendix C applies cross-teacher filtering and finds that the gains from stronger teachers are driven primarily by increased data quantity rather than by improved reasoning quality.

Modification 3: Restriction to Efficiency-Motivated Settings. We restrict our evaluation to configurations where the student model is smaller than the teacher model, aligning with the efficiency-driven motivation for CoT distillation. This allows us to derive actionable insights for practical deployment. The settings where knowledge is transferred from a smaller to a larger model, while theoretically interesting, fall outside the scope of this study.

4.2 Experimental Setup

Our experimental setup follows that of Section 3.1, with modifications according to the protocol described in Section 4.1. We select BBH tasks where the average few-shot accuracy gap between teacher and student model groups exceeds 30 percentage points, yielding 15 tasks (Appendix A.4). For each task, we split examples into training and test sets in a 3:2 ratio. To ensure sufficient student model sizes while keeping students smaller than teachers, we use Qwen2.5-14B-Instruct and Qwen2.5-72B-Instruct as small and large teachers in the *small–large* setting; for *short–long*, we use Qwen2.5-32B-Instruct and QwQ-32B-Preview as in Section 3.1. We do not apply cross-teacher filtering; instead, for each teacher, we retain all examples where that teacher produces a correct answer. For fair comparison, we apply the same few-shot CoT prompt to all models, including the pre-distillation baseline. We also conduct cross-family experiments using Gemma-2 teachers with Qwen2.5 students to validate that our findings generalize beyond a single model family (Appendix D).

4.3 Results

Figures 2 and 3 present the results under our practical evaluation protocol. We discuss the results further in Section 5. We additionally verify that these findings hold with Gemma-2 teachers in a cross-family setting (Appendix D).

Distillation Efficacy. In most configurations, distilled models outperform the pre-distillation baseline, confirming that many of the selected tasks represent valid distillation scenarios. However, some tasks show limited or negative distillation effects: in the small–large setting, four tasks (Formal Fallacies, Logical Deduction (5) and (7), and Penguins in a Table), and in the short–long setting, five tasks (with the addition of Logical Deduction (3)). We focus our subsequent analysis on the remaining tasks where distillation consistently improves performance (11 tasks in small–large, 10 tasks in short–long), and defer discussion of these problematic tasks to the end of this section.

Capacity Gap Effects. Among tasks where distillation is effective, capacity gap effects are observed in some cases but their impact varies across tasks and settings. We identify capacity gap effects using the following *diagnostic pattern*: if the capacity gap were a dominant factor, less capable teachers would yield better students at the smallest student sizes, with this advantage diminishing or reversing as student size increases.[5] Using this criterion, in the small–large setting, we observe such effects in eight of the eleven distillation-effective tasks, while Date Understanding, Dyck Languages, and Tracking Shuffled Objects (7) show no such effects. In the short–long setting, capacity gap effects appear only in Temporal Sequences and Tracking Shuffled Objects (7), while the remaining eight tasks show no such effects.

Effect of Teacher Performance Gap. Although capacity gap effects are present in some configurations, we identify a consistent practical pattern: when there is a substantial performance gap between teacher models, as observed in Dyck Languages and Word Sorting, higher-performing teachers produce better or comparable distilled students. This finding suggests that, even when capacity gap effects exist, the benefits of stronger teachers outweigh these effects when the teacher performance difference is large.

Analysis of Ineffective Distillation Tasks. For the tasks where distillation shows limited effective-

[5] Because we report single-run results on relatively small per-task test sets (Limitations section), per-task judgments should be interpreted with care. However, this diagnostic is asymmetric in a way that protects our main claim: detecting a capacity gap requires the weaker teacher to *actively* outperform the stronger one at small student sizes, so noise in marginal cases biases toward an inconclusive "no clear gap" reading rather than a false positive for capacity gap dominance.

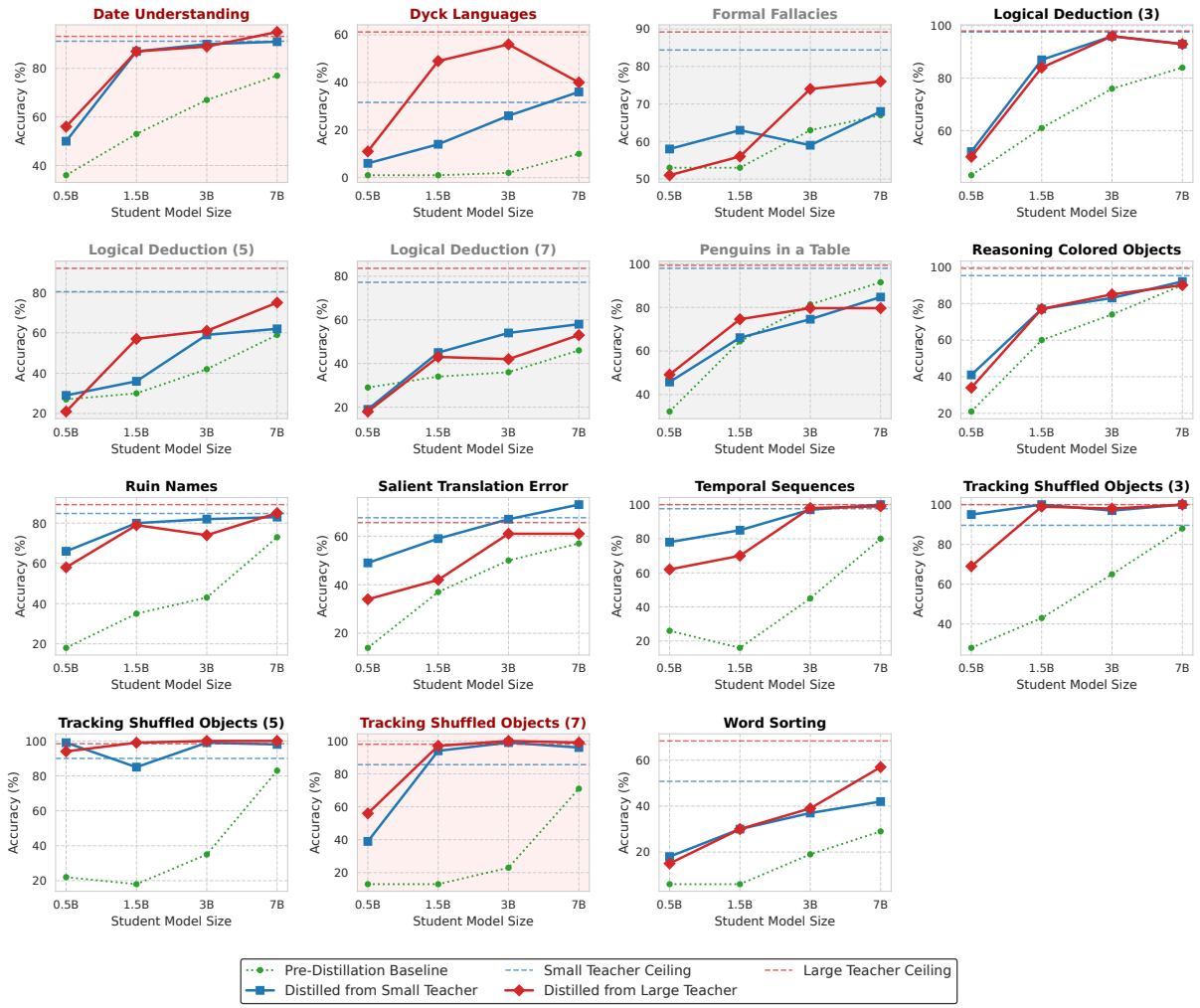


Figure 2: Results under the *small–large* setting on 15 selected BBH tasks with our practical evaluation protocol. We compare the pre-distillation baseline with students distilled from Qwen2.5-14B-Instruct (Small Teacher) and Qwen2.5-72B-Instruct (Large Teacher). Teacher Ceiling indicates the few-shot performance of each teacher model. Tasks with a gray background indicate cases where at least one distilled model underperformed the pre-distillation baseline. Tasks with a red background indicate cases where the results did not follow the capacity gap hypothesis.

ness (Formal Fallacies, Logical Deduction, Penguins in a Table), capacity gap patterns appear inconsistently across settings, and even for Formal Fallacies, where the teacher performance gap is relatively large, stronger teachers do not yield a consistent advantage. Thus, when distillation itself is ineffective for a task, neither capacity gap considerations nor teacher strength provide reliable guidance for teacher selection. However, since practitioners should not apply distillation to such tasks in the first place, this absence of guidance is not a practical concern.

5 Discussion

In this section, we synthesize our results into actionable recommendations for practitioners considering CoT distillation.

Guideline 1: Verify distillation efficacy against the pre-distillation baseline. Unlike earlier models that required task-specific training, recent LLMs possess substantial zero-shot and few-shot reasoning capabilities and may already perform well on target tasks without any fine-tuning. Consequently, distillation is not guaranteed to improve performance and may even cause degradation. Before deploying a distilled model or comparing distillation strategies, practitioners should confirm that distillation yields genuine improvements over the student's pre-distillation performance.

Since verifying this through actual fine-tuning is costly, a lightweight proxy is desirable. Our results suggest that few-shot ICL performance gaps between teacher and student models may serve this purpose: in our BBH experiments, tasks selected

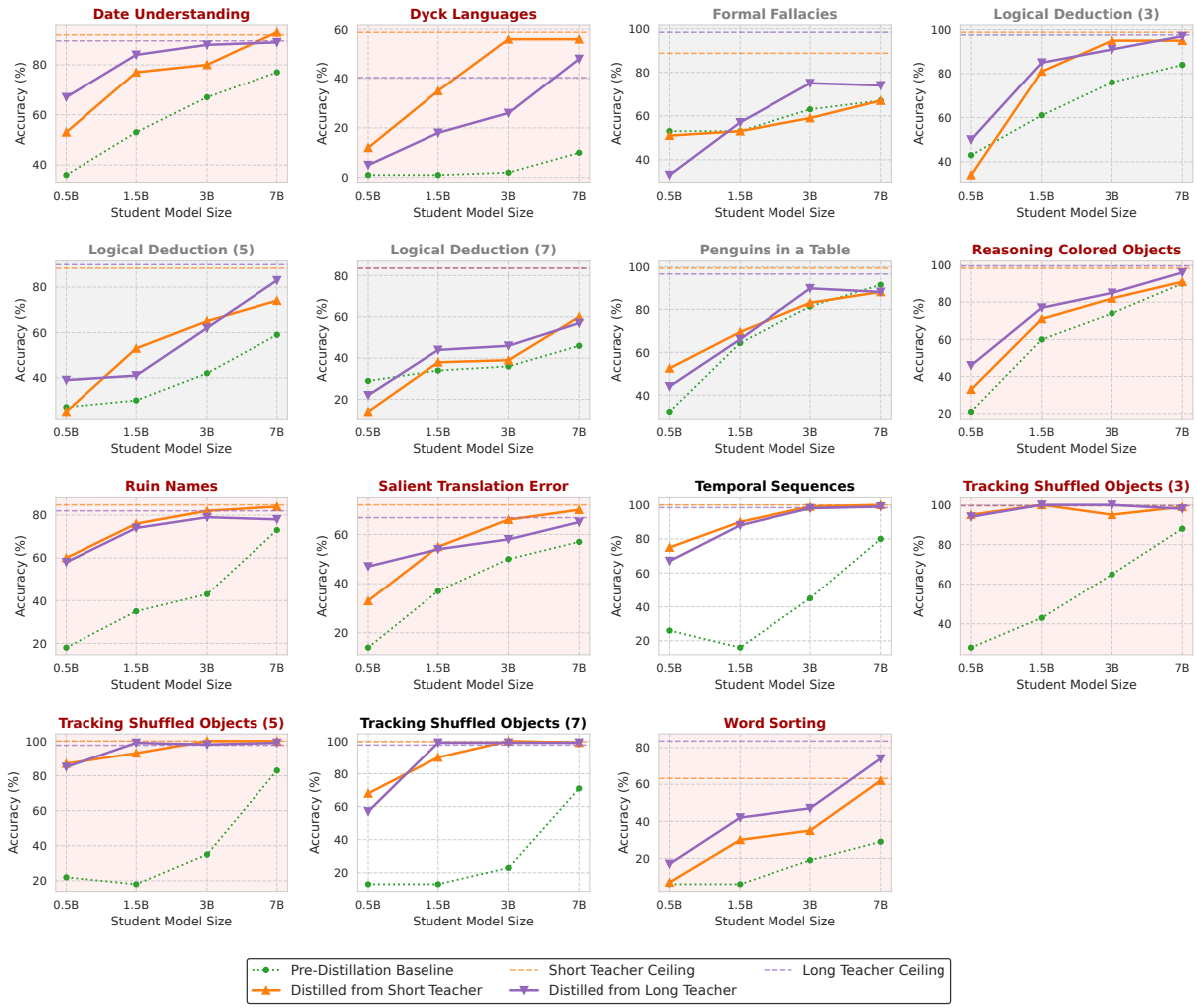


Figure 3: Results under the *short–long* setting on 15 selected BBH tasks with our practical evaluation protocol. We compare the pre-distillation baseline with students distilled from Qwen2.5-32B-Instruct (Short Teacher) and QwQ-32B-Preview (Long Teacher). Teacher Ceiling indicates the few-shot performance of each teacher model. Tasks with a gray background indicate cases where at least one distilled model underperformed the pre-distillation baseline. Tasks with a red background indicate cases where the results did not follow the capacity gap hypothesis.

based on large ICL gaps showed effective distillation in most configurations, indicating that this approach can potentially reduce the cost of identifying suitable tasks for CoT distillation.[6]

Guideline 2: If distillation is beneficial *and* teacher performance differs substantially, prefer the higher-performing teacher. Our results indicate that when there is a substantial performance gap between candidate teachers, higher-performing teachers produce better or comparable distilled students, regardless of student model size. We therefore state this as an explicitly conditional recommendation: *if* distillation has been verified to improve over the pre-distillation baseline (Guideline 1) *and* candidate teachers differ substantially in performance, *then* practitioners should select the higher-performing teacher. Where distillation itself is ineffective for a task, neither capacity gap considerations nor teacher strength provide reliable guidance for teacher selection, and the present guideline does not apply (Section 4.3).

This practical benefit of stronger teachers has multiple components, of which increased training data, a natural consequence of higher teacher accuracy, is a major one. Our ablation with cross-

[6]Additional evidence supports the generality of this proxy across teacher families and data domains: applying the same 30 percentage-point threshold to our Gemma-2 cross-family experiments (Appendix D) selects only tasks where distillation indeed improves over the baseline, and applying it to the math tasks in Section 3.1 yields no tasks above the threshold, consistent with our finding that distillation does not help there. Broader validation across more benchmarks and model families, as well as exploration of alternative selection criteria, remain important directions for future work.

teacher filtering (Appendix C) equalizes data sizes across teachers and shows that, in the two tasks where the stronger teacher provides substantially more data (Dyck Languages and Word Sorting, 1.3–1.9×), the stronger-teacher advantage largely disappears, while tasks with at most 1.1× difference are essentially unchanged. A data size ratio of around 1.3× may thus serve as a rough reference for when the stronger teacher's data-quantity advantage becomes practically decisive, though this is derived from a limited number of tasks and should be treated as an approximate heuristic rather than a rigid criterion. When the performance gap is small, practitioners may consider other factors such as computational cost or reasoning chain length, as shorter chains reduce inference costs.

6 Related Work

6.1 Knowledge Distillation

CoT distillation builds upon knowledge distillation (Hinton et al., 2015) and sequence-level knowledge distillation (Kim and Rush, 2016). As discussed in Section 2.1, hard distillation—supervised fine-tuning on teacher-generated sequences (Ho et al., 2023; Li et al., 2023; Hsieh et al., 2023; Magister et al., 2023; Shridhar et al., 2023)—is the dominant paradigm.[7] It has also been widely adopted for transferring general knowledge (West et al., 2022; Mukherjee et al., 2023) or reasoning traces for test-time scaling (Muennighoff et al., 2025).

6.2 Capacity Gap in Knowledge Distillation

The capacity gap is also observed in conventional soft knowledge distillation (Lopez-Paz et al., 2016; Cho and Hariharan, 2019; Mirzadeh et al., 2020; Jafari et al., 2021). These analyses, however, rely on access to the teacher's full output distribution (e.g., logit-based KL objectives or temperature annealing) and do not directly transfer to the hard CoT distillation studied here, where the student is trained only on sampled sequences (Section 2.1); the bridge is nonetheless tight, as Kim and Rush (2016) place sequence-level hard knowledge distillation on the same spectrum as soft knowledge distillation by approximating the teacher distribution via sampling, and the same phenomenon has been empirically reported in this regime (Li et al., 2025; Chen et al., 2025). Zhang et al. (2025) study the capacity gap in task-agnostic LM distillation, but their controlled setting, where the student is initialized by pruning the teacher, does not reflect practical scenarios involving independently pretrained models. Our contribution is an empirical re-evaluation of the capacity gap in hard CoT distillation under practical conditions, *complementary* to Li et al. (2025) and Chen et al. (2025): prior work establishes the gap under controlled protocols, whereas we assess how much it matters once realistic deployment factors are allowed in.

6.3 Advances beyond Teacher Selection

Beyond teacher selection, other work addresses the challenge that student models cannot perfectly replicate teacher reasoning. Approaches include improving faithfulness of distilled rationales (Wang et al., 2023), curriculum learning (Feng et al., 2024), augmenting students with external symbolic knowledge bases (Liao et al., 2025), multi-CoT consistent distillation (Chen et al., 2023), and self-guided rationale selection (Yan et al., 2025). These efforts are orthogonal to our focus on teacher selection: they address the separate problem of bridging the imitation gap given a fixed teacher–student pair.

7 Conclusion

We revisited the capacity gap in CoT distillation from a practical perspective, focusing on deployment scenarios that genuinely call for distillation: cases where a small, efficient student cannot adequately handle the target task from pretraining alone, such as tasks involving proprietary or domain-specific data with task-specific patterns unavailable from the student's pretraining. We identified pitfalls in existing evaluation protocols, most notably the lack of comparison against pre-distillation baselines, and proposed a more realistic protocol. Complementary to prior work that establishes the capacity gap under controlled conditions, our experiments across model families (Qwen2.5 in the main experiments and Gemma-2 in Appendix D) show that its practical impact is not universal and can be outweighed by the overall benefits of stronger teachers, of which increased training data availability is a major component. Based on these findings, we recommend two conditional guidelines: always verify that distillation improves upon the pre-distillation baseline, and if this is confirmed and candidate teachers differ substantially in performance, prefer the higher-performing teacher.

[7] While Fu et al. (2023) partially employs soft distillation via KL divergence minimization, it falls back to hard distillation depending on tokenizer alignment.

Limitations

Benchmark and Model Scope. Our practical evaluation focuses exclusively on BBH as the benchmark where CoT distillation demonstrates genuine improvements over pre-distillation baselines. While this limits the generalizability of our findings, we note that BBH comprises diverse tasks spanning various reasoning types, and our task selection procedure specifically targeted scenarios where the capacity gap is most likely to manifest. Regarding model scope, we provide initial cross-family validation using Gemma-2 teachers in Appendix D, but further validation across diverse model families remains important. We believe BBH represents an appropriate testbed for examining the practical impact of the capacity gap; nevertheless, future work should extend this evaluation to other domains where CoT distillation proves effective, such as code generation and instruction following.

Confounded Factors in Practical Evaluation. By removing cross-teacher data filtering (Section 4.1), our protocol does not allow us to disentangle the individual contributions of increased data quantity and improved reasoning quality that stronger teachers provide. However, this is an intentional design choice: our goal is to assess the overall practical impact of using stronger teachers under realistic deployment conditions, rather than to isolate the capacity gap as a phenomenon. While this approach limits our ability to provide mechanistic explanations, it directly addresses the practitioner's question of which teacher to select given real-world constraints. We conduct additional experiments with cross-teacher filtering in Appendix C, which confirm that filtering disproportionately penalizes stronger teachers, that capacity gap effects remain inconsistent even under equalized data sizes, and that the advantage of stronger teachers is largely driven by increased data *quantity*, more than by differences in CoT *quality*.

Scope of Claims Regarding the Capacity Gap. Our findings do not refute the existence of the capacity gap; indeed, both prior work (Li et al., 2025; Chen et al., 2025) and our own results confirm that it manifests in some tasks and settings. Rather, our contribution is to show that when candidate teachers differ substantially in performance, the benefits of stronger teachers outweigh capacity gap effects. When candidate teachers perform similarly, capacity gap effects may play a larger role, and other factors such as computational cost or reasoning chain length may guide teacher selection. Future work should develop more fine-grained selection criteria for such cases.

Statistical Reliability. Due to the high computational cost of distillation across configurations, we report results from a single run, following Li et al. (2025), and cannot report confidence intervals or variance estimates. The BBH test sets are also relatively small (approximately 100 examples per task after splitting), which may amplify noise in per-task results. As noted in Section 4.3, our capacity gap diagnostic is asymmetric, so noise in marginal cases biases toward inconclusive readings rather than false positives for capacity gap dominance; in addition, our conclusions are drawn from patterns observed across many tasks and settings rather than from individual task-level judgments, further mitigating this limitation.

Task Selection Threshold. Our task selection criterion (a 30 percentage-point gap between teacher and student ICL performance averages) is not derived from theoretical considerations and may require adjustment for other benchmarks or model families. This threshold does not guarantee that distillation will be effective; indeed, some of the selected tasks turned out to show limited or negative distillation effects (Section 4.3). Nevertheless, the criterion relies only on information available before distillation (few-shot ICL performance), making it a practical proxy for identifying tasks where a practitioner would realistically consider applying distillation. Moreover, the inclusion of both effective and ineffective distillation scenarios reflects the realistic uncertainty practitioners face, providing a practical testbed for examining the capacity gap.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 76930–76966. Curran Associates, Inc.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. MCC-KD: Multi-CoT consistent knowledge distillation. In *Findings of the Association for Computational Linguistics:*

EMNLP 2023, pages 6805–6820, Singapore. Association for Computational Linguistics.

Xinghao Chen, Zhijing Sun, Guo Wenjin, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. 2025. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15094–15119, Vienna, Austria. Association for Computational Linguistics.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802.

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:2110.14168v2.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. 2024. Keypoint-based progressive chain-of-thought distillation for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13241–13255. PMLR.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Ukyo Honda, Soichiro Murakami, and Peinan Zhang. 2025. Distilling many-shot in-context learning into a cheat sheet. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17158–17178, Suzhou, China. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models

struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.

Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. 2025. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24567–24575.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. *Preprint*, arXiv:1511.03643v3.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, Suzhou, China. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707v1.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115v2.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118v3.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

JianZhi Yan, Le Liu, Youcheng Pan, Shiwei Chen, Yang Xiang, and Buzhou Tang. 2025. Towards efficient CoT distillation: Self-guided rationale selector for better performance with fewer rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7818–7835, Suzhou, China. Association for Computational Linguistics.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: Less is more for reasoning. In *Second Conference on Language Modeling*.

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yao Hu. 2025. Towards the law of capacity gap in distilling language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22504–22528, Vienna, Austria. Association for Computational Linguistics.

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

Hyperparameter	Value
Method	
Stage	SFT
Finetuning type	Full
DeepSpeed config	ds_z3_config.json
Flash attention	FA2
Dataset	
Template	qwen
Max samples	10,000
Cutoff length	8,192
Preprocessing workers	32
Training	
Batch size (per device)	2
Gradient accumulation steps	1
Learning rate	1.0e-5
Number of epochs	2
LR scheduler	Cosine
Precision	BF16
Evaluation	
Validation size	0.01
Eval batch size (per device)	1
Eval strategy	Steps
Eval steps	100

Table 1: Hyperparameters for full fine-tuning of smaller student models.

Hyperparameter	Value
Method	
Stage	SFT
Finetuning type	LoRA
Lora Target	All
DeepSpeed config	ds_z3_config.json
Flash attention	FA2
Dataset	
Template	qwen
Max samples	10,000
Cutoff length	8,192
Preprocessing workers	32
Training	
Batch size (per device)	1
Gradient accumulation steps	1
Learning rate	1.0e-4
Number of epochs	2
LR scheduler	Cosine
Precision	BF16
Warmup ratio	0.03
Evaluation	
Validation size	0.01
Eval batch size (per device)	1
Eval strategy	Steps
Eval steps	100

Table 2: Hyperparameters for LoRA fine-tuning of larger student models (14B and 32B).

A Experimental Setup Details

This section provides a detailed description of the experimental setup. All experiments are conducted on four NVIDIA A100 GPUs, each with 80GB of memory. Due to the high computational cost of distillation across various configurations, we report results from a single run, following Li et al. (2025). The datasets used in our experiments are all in English and publicly available for research use.

A.1 Models

We use models from the Qwen2.5 family in our experiments. For student and teacher models, we use Qwen2.5-{0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B}-Instruct, available at `https://huggingface.co/collections/Qwen/qwen25`. For the long-CoT teacher in the short–long setting, we use QwQ-32B-Preview, available at `https://huggingface.co/Qwen/QwQ-32B-Preview`.

A.2 Hyperparameters

Following the codebase of Li et al. (2025), we use `LLaMA-Factory` (Zheng et al., 2024) for fine-tuning student models. Table 1 summarizes the hyperparameters used in our experiments. For larger student models (14B and 32B), we employ LoRA (Hu et al., 2022) following Li et al. (2025). The LoRA-specific hyperparameters are provided in Table 2.

A.3 MATH Data

Table 3 summarizes the distillation datasets used in Section 3.1. These datasets were publicly released by Li et al. (2025). They were constructed by generating CoT rationales for the MATH training set (Hendrycks et al., 2021) and retaining only examples where both teachers in each setting (small–large or short–long) produce correct answers. Datasets within the same setting have identical sizes due to this cross-teacher filtering (see Section 3.2).

Table 4 summarizes the evaluation benchmarks used in Section 3.1. MATH and GSM8K (Cobbe et al., 2021) are standard mathematical reasoning benchmarks; we use their test and validation splits, respectively. AMC 2023, AIME 2024, and

Distillation Data	Size	Source
MATH w/ Small Teacher	4,539	https://huggingface.co/datasets/UWNSL/MATH_training_split_distill_small_teacher
MATH w/ Large Teacher	4,539	https://huggingface.co/datasets/UWNSL/MATH_training_split_distill_large_teacher
MATH w/ Short Teacher	5,383	https://huggingface.co/datasets/UWNSL/MATH_training_split_short_cot
MATH w/ Long Teacher	5,383	https://huggingface.co/datasets/UWNSL/MATH_training_split_long_cot

Table 3: Distillation datasets for MATH experiments in Section 3.1.

Evaluation Data	Size	Source
MATH	5,000	https://huggingface.co/datasets/EleutherAI/hendrycks_math
GSM8K	1,319	https://huggingface.co/datasets/openai/gsm8k
AMC 2023	40	https://github.com/Small-Model-Gap/Small-Model-Learnability-Gap/tree/main/lm-evaluation-harness/math_eval_data/amc23
AIME 2024	30	https://github.com/Small-Model-Gap/Small-Model-Learnability-Gap/tree/main/lm-evaluation-harness/math_eval_data/aime24
OlympiadBench	675	https://github.com/Small-Model-Gap/Small-Model-Learnability-Gap/tree/main/lm-evaluation-harness/math_eval_data/olympiadbench

Table 4: Evaluation benchmarks for MATH experiments in Section 3.1.

OlympiadBench (He et al., 2024) are competition-level benchmarks designed to assess advanced mathematical problem-solving abilities. Following Li et al. (2025), we conduct evaluations using the evaluation script from their official implementation along with `lm-evaluation-harness` (Gao et al., 2024).

A.4 BBH Data

We use 15 tasks from BBH (Suzgun et al., 2023; Srivastava et al., 2023) based on the selection criterion described in Section 4.1. Table 5 provides brief descriptions of the selected tasks. Table 9 presents the complete few-shot ICL performance of the pre-distillation models.

Each task contains 250 examples, except for Penguins in a Table, which contains 146 examples. For each task, we split examples into training and test sets in a 3:2 ratio. For distillation, we prompt each teacher model to generate CoT rationales for the training examples and retain only those where the teacher produces a correct final answer. Unlike the MATH experiments, we do not apply cross-teacher filtering; each teacher's distillation dataset consists of all examples that the respective teacher solves correctly, as described in Section 4.1. All evaluations are conducted using `lm-evaluation-harness`, which loads the BBH data and uses few-shot CoT prompts.

B Artifacts from Comparing Degradation

As demonstrated in Section 3.1, existing experimental settings exhibit an undesirable situation where distillation *degrades* performance compared to the pre-distillation baseline. This is problematic not only because it undermines the practical relevance of the experiments, but also because it can lead to misleading evaluations of proposed strategies. When distillation consistently harms performance, any approach that appears to improve results may simply be mitigating the degree of degradation rather than achieving genuine improvements. We illustrate this concern with a concrete example from prior work.

Li et al. (2025) proposed *mix distillation*, which combines rationales from different teachers in a fixed ratio to mitigate the capacity gap. However, the original implementation uses different hyperparameters from the baseline: mix distillation uses a batch size of 5 with 4 gradient accumulation steps, while standard CoT distillation uses a batch size of 2 with 1 accumulation step. This discrepancy appears to stem from the mixing ratio: rationales are combined in a 4:1 ratio, requiring larger batches and more accumulation steps to ensure each parameter update reflects rationales from both teachers. Since the number of epochs is fixed at 2 and other hyperparameters such as learning rate remain the same, mix distillation performs 10 times fewer parameter updates than the baseline.[8]

To isolate the contribution of the data construction strategy, the core component of mix distillation, we aligned the effective batch size (batch size $\times$ gradient accumulation steps) across all methods. Specifically, we trained students (Qwen2.5-

[8] Mix distillation performs $\frac{N}{5\times4} \times 2 = \frac{N}{10}$ updates, while standard distillation performs $\frac{N}{2\times1} \times 2 = N$ updates, where N is the dataset size.

Task	Description
Date Understanding	Read dates from context and answer questions involving format conversion or duration calculation.
Dyck Languages	Complete a parenthesis sequence by supplying the missing closing brackets.
Formal Fallacies	Determine the logical validity of an informal argument.
Logical Deduction {3, 5, 7}	Given clues about the relative positions of 3, 5, or 7 objects, identify the correct position of a specified object.
Penguins in a Table	Answer questions about specific penguins or their counts based on a table of penguin attributes.
Reasoning Colored Objects	Given a description of object arrangements, identify the color of a specified object.
Ruin Names	Alter exactly one character of a given artist, band, or movie name to produce a humorous variant.
Salient Translation Error	Compare a German source sentence with an erroneous English translation and classify the type of translation error such as proper nouns, numbers, or negation.
Temporal Sequences	Determine the time period during which a specific event occurred, given a person's daily schedule.
Tracking Shuffled Objects {3, 5, 7}	Track a series of swap operations on 3, 5, or 7 objects and identify the final owner of each object.
Word Sorting	Sort a list of words into alphabetical order.

Table 5: Descriptions of the 15 BBH tasks used in our experiments.

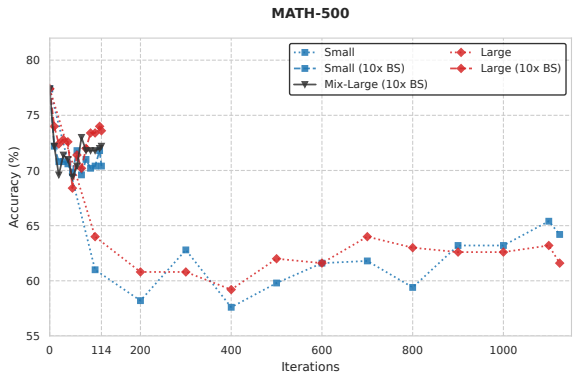


Figure 4: Effect of effective batch size alignment on distillation performance (MATH-500, small–large setting). (10 × BS) indicates configurations with 10 times the original effective batch size (batch size × gradient accumulation steps).

Teacher	No Filter	Filtered	Diff.
Small	67.49	65.70	−1.78
Large	67.95	64.34	−3.61
Short	67.75	64.92	−2.83
Long	69.09	66.11	−2.97

Table 6: Average accuracy across 15 BBH tasks (averaged over four student sizes) **with and without cross-teacher filtering**. Diff. = Filtered − No Filter.

7B-Instruct) using small-teacher and large-teacher distillation data with the same effective batch size as mix distillation (10 × the original). We used MATH-500 as the test set and evaluated at fine-grained iteration intervals to examine the effect of the number of update steps. Figure 4 presents the results. When batch size and gradient accumulation are aligned, standard distillation with large-teacher data achieves comparable or better performance than mix distillation (i.e., Large (10 × BS) matches or outperforms Mix (10 × BS)). This confirms that in this setting, fewer update steps, and thus smaller deviations from the original model, become the dominant factor affecting performance, making it difficult to properly compare distillation strategies.

C Effect of Cross-Teacher Filtering on BBH

To disentangle the effects of data quantity and reasoning quality, we repeated our BBH experiments (Section 4) with the cross-teacher filtering protocol used in prior work, where training data is restricted to examples that both teachers solve correctly.

Table 6 reports the average accuracy across all 15 BBH tasks (averaged over four student sizes: 0.5B, 1.5B, 3B, 7B). Cross-teacher filtering consistently degrades average accuracy compared to the unfiltered setting. Notably, the degradation is larger for the more capable teachers (Large: −3.61; Long: −2.97) than for the less capable ones (Small: −1.78; Short: −2.83), confirming that filtering disproportionately penalizes more capable teachers by discarding their additional correct examples.

The effect of filtering is most visible in Dyck Languages and Word Sorting, where the stronger teacher provided substantially more data in the unfiltered setting (approximately 1.3–1.9× more examples). Without filtering, the stronger teacher consistently produced better students regardless of student size in these tasks. Equalizing data sizes through filtering substantially weakened this pattern. Specifically, in Dyck Languages, the average absolute accuracy gap between students of the two teachers (across student sizes) shrank from 18.50 points (unfiltered) to 0.25 points (filtered) in the small–large setting, and from 15.50 to 0.75 points in the short–long setting. In Word Sorting, the gap

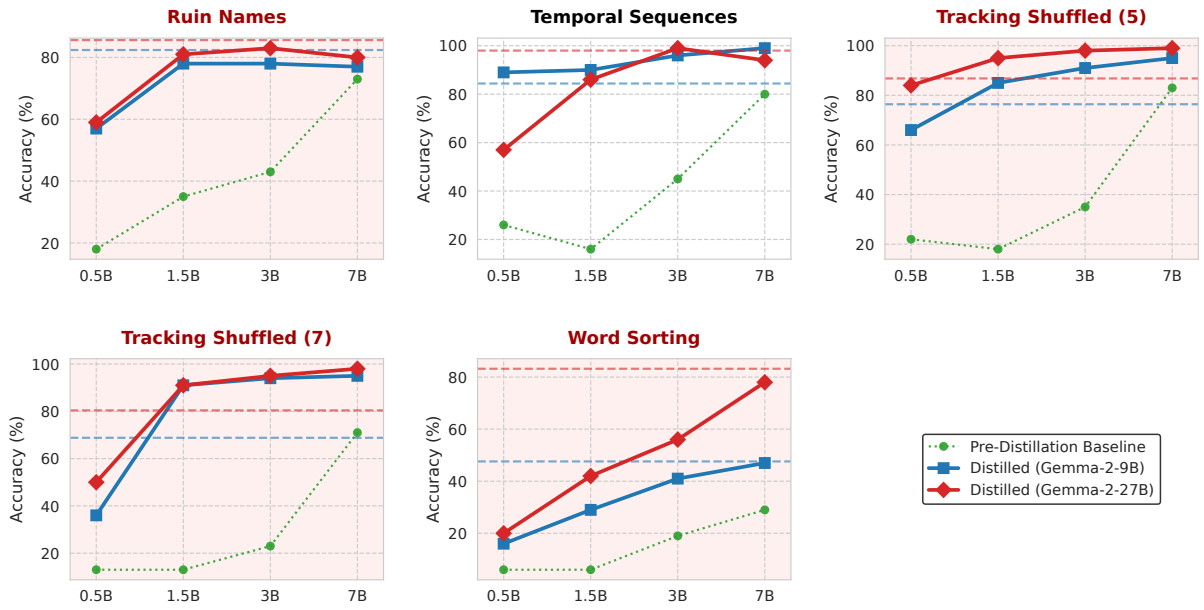


Figure 5: Results on 5 selected BBH tasks with **Gemma-2** teachers and Qwen2.5 students. We compare the pre-distillation baseline with students distilled from Gemma-2-9B-it (Small Teacher) and Gemma-2-27B-it (Large Teacher). Teacher Ceiling indicates the few-shot performance of each teacher model. Tasks with a red background indicate cases where the results did not follow the capacity gap hypothesis.

shrank from 3.50 to 1.25 points in the small–large setting, and from 14.00 to 4.75 points in the short–long setting. These results provide direct evidence that the advantage of stronger teachers is largely driven by increased data *quantity*, more than by differences in CoT *quality* between the two teachers.

Tables 7 and 8 present the complete per-task results under cross-teacher filtering for the small–large and short–long settings, respectively.

D Generalization to a Different Model Family

Our main experiments use the Qwen2.5 family for both teachers and students. To examine whether our findings generalize beyond a single model family, we conduct additional experiments using Gemma-2-{9B, 27B}-it (Team et al., 2024) as teachers while keeping the same Qwen2.5 students. This cross-family setting follows Li et al. (2025), who also employed Gemma-2-9B-it and Gemma-2-27B-it as small and large teachers in their small–large setting. As in Li et al. (2025), we evaluate only the small–large setting since the Gemma-2 family does not include a long-CoT teacher.

Applying the same 30 percentage-point gap threshold used in Section 4.1, we select 5 tasks for distillation: Ruin Names, Temporal Sequences, Tracking Shuffled Objects (5), Tracking Shuffled Objects (7), and Word Sorting. The pre-distillation performance of all models is reported in Table 10.

Figure 5 presents the results. Among the 5 selected tasks, distillation is effective in all of them, consistently improving over the pre-distillation baseline. Regarding capacity gap effects, only 1 out of 5 tasks (Temporal Sequences) exhibits such effects, where the smaller teacher (9B) yields better students at smaller student sizes. In the remaining 4 tasks, the larger teacher (27B) consistently produces better or comparable students across all student sizes. These results are consistent with our main findings from the Qwen-only experiments (Section 4), confirming that the impact of capacity gap effects does not consistently dominate across tasks, even when a different teacher model family is used.

	Small Teacher (14B)				Large Teacher (72B)			
Task	**0.5B**	**1.5B**	**3B**	**7B**	**0.5B**	**1.5B**	**3B**	**7B**
Date Understanding	51.0	81.0	91.0	92.0	55.0	85.0	80.0	91.0
Dyck Languages	3.0	12.0	28.0	37.0	1.0	12.0	23.0	43.0
Formal Fallacies	52.0	48.0	65.0	71.0	53.0	59.0	67.0	77.0
Logical Deduction (3)	45.0	76.0	92.0	95.0	39.0	83.0	93.0	96.0
Logical Deduction (5)	30.0	45.0	55.0	74.0	28.0	42.0	56.0	59.0
Logical Deduction (7)	17.0	37.0	49.0	60.0	13.0	40.0	41.0	58.0
Penguins in a Table	35.6	64.4	83.1	88.1	32.2	66.1	78.0	81.4
Reasoning about Colored Objects	35.0	74.0	85.0	90.0	25.0	76.0	84.0	87.0
Ruin Names	59.0	76.0	76.0	81.0	40.0	77.0	74.0	80.0
Salient Translation Error Detection	31.0	46.0	62.0	64.0	34.0	43.0	58.0	67.0
Temporal Sequences	82.0	96.0	97.0	98.0	65.0	96.0	99.0	100.0
Tracking Shuffled Objects (3)	85.0	99.0	98.0	99.0	83.0	99.0	99.0	100.0
Tracking Shuffled Objects (5)	85.0	94.0	97.0	100.0	77.0	94.0	99.0	99.0
Tracking Shuffled Objects (7)	41.0	94.0	99.0	98.0	43.0	93.0	100.0	99.0
Word Sorting	9.0	28.0	35.0	52.0	8.0	25.0	36.0	50.0

Table 7: Per-task accuracy on BBH under **cross-teacher filtering** in the *small–large* setting (see Figure 2 for unfiltered results).

	Short Teacher (32B)				Long Teacher (QwQ)			
Task	**0.5B**	**1.5B**	**3B**	**7B**	**0.5B**	**1.5B**	**3B**	**7B**
Date Understanding	52.0	80.0	83.0	92.0	55.0	75.0	85.0	84.0
Dyck Languages	1.0	7.0	24.0	38.0	1.0	10.0	31.0	25.0
Formal Fallacies	50.0	45.0	62.0	65.0	42.0	51.0	54.0	70.0
Logical Deduction (3)	48.0	73.0	88.0	96.0	47.0	82.0	93.0	96.0
Logical Deduction (5)	29.0	47.0	56.0	69.0	29.0	46.0	54.0	74.0
Logical Deduction (7)	10.0	41.0	38.0	62.0	18.0	36.0	47.0	67.0
Penguins in a Table	37.3	64.4	74.6	83.1	39.0	71.2	89.8	84.7
Reasoning about Colored Objects	31.0	72.0	85.0	90.0	41.0	77.0	86.0	94.0
Ruin Names	48.0	64.0	78.0	85.0	45.0	69.0	74.0	81.0
Salient Translation Error Detection	33.0	47.0	64.0	69.0	51.0	50.0	60.0	64.0
Temporal Sequences	81.0	81.0	95.0	97.0	71.0	72.0	100.0	100.0
Tracking Shuffled Objects (3)	95.0	100.0	100.0	100.0	94.0	99.0	99.0	98.0
Tracking Shuffled Objects (5)	82.0	98.0	99.0	99.0	89.0	99.0	99.0	99.0
Tracking Shuffled Objects (7)	63.0	99.0	99.0	100.0	64.0	94.0	97.0	99.0
Word Sorting	9.0	28.0	30.0	59.0	11.0	33.0	41.0	60.0

Table 8: Per-task accuracy on BBH under **cross-teacher filtering** in the *short–long* setting (see Figure 3 for unfiltered results).

	Student					Teacher					
Task	**0.5B**	**1.5B**	**3B**	**7B**	**Avg. (S)**	**14B**	**32B**	**72B**	**QwQ**	**Avg. (T)**	**Gap (T−S)**
Boolean Expressions	75.2	88.0	88.8	96.4	87.1	99.2	100.0	100.0	96.8	99.0	11.9
Causal Judgement	48.7	59.9	54.5	60.4	55.9	65.8	69.0	71.7	67.9	68.6	12.7
Date Understanding	30.4	53.6	66.4	82.4	58.2	91.2	92.0	93.2	89.6	91.5	33.3
Disambiguation QA	31.2	49.2	63.6	69.2	53.3	82.4	79.2	82.8	74.4	79.7	26.4
Dyck Languages	0.0	0.0	3.6	11.2	3.7	31.6	58.8	61.2	40.4	48.0	44.3
Formal Fallacies	53.2	52.0	62.4	70.4	59.5	84.4	88.8	89.2	98.4	90.2	30.7
Geometric Shapes	22.0	28.8	49.2	66.4	41.6	54.4	58.4	54.8	70.8	59.6	18.0
Hyperbaton	55.2	72.8	74.4	90.0	73.1	99.6	98.8	99.2	99.6	99.3	26.2
Logical Deduction (3)	37.6	59.2	70.0	86.8	63.4	97.6	98.8	98.0	97.6	98.0	34.6
Logical Deduction (5)	22.4	33.2	45.6	59.2	40.1	80.4	88.4	92.0	90.0	87.7	47.6
Logical Deduction (7)	22.0	34.4	32.8	45.6	33.7	77.2	83.6	83.6	83.6	82.0	48.3
Movie Recommendation	37.2	48.4	74.0	66.4	56.5	74.0	78.0	80.4	87.6	80.0	23.5
Multistep Arithmetic Two	30.8	60.8	91.2	94.8	69.4	97.6	96.8	100.0	98.4	98.2	28.8
Navigate	58.4	73.6	90.0	95.2	79.3	98.0	98.8	99.2	97.2	98.3	19.0
Object Counting	53.6	58.8	63.2	78.4	63.5	92.4	90.0	97.6	89.6	92.4	28.9
Penguins in a Table	32.9	61.6	79.5	94.5	67.1	97.9	99.3	99.3	96.6	98.3	31.2
Reasoning about Colored Objects	20.8	62.0	75.6	90.0	62.1	95.2	98.4	99.2	99.6	98.1	36.0
Ruin Names	24.0	36.0	42.8	66.0	42.2	84.8	84.8	89.2	82.0	85.2	43.0
Salient Translation Error Detection	12.0	33.2	46.0	54.8	36.5	67.6	72.0	65.6	66.8	68.0	31.5
Snarks	46.6	57.3	64.6	79.8	62.1	82.3	87.1	91.6	93.3	88.6	26.5
Sports Understanding	48.8	73.2	82.0	87.2	72.8	90.4	90.4	94.8	91.2	91.7	18.9
Temporal Sequences	24.8	15.2	45.6	81.6	41.8	97.6	100.0	100.0	98.4	99.0	57.2
Tracking Shuffled Objects (3)	32.4	41.2	67.6	88.8	57.5	89.6	99.6	100.0	99.6	97.2	39.7
Tracking Shuffled Objects (5)	15.6	18.8	39.6	77.2	37.8	90.0	100.0	98.4	97.6	96.5	58.7
Tracking Shuffled Objects (7)	17.2	16.4	26.4	65.6	31.4	85.6	99.6	98.0	97.6	95.2	63.8
Web of Lies	59.6	80.8	100.0	99.2	84.9	99.6	100.0	100.0	100.0	99.9	15.0
Word Sorting	4.4	6.8	16.0	29.2	14.1	50.8	63.2	68.4	83.6	66.5	52.4

Table 9: Results of few-shot ICL performance on BBH for *pre-distillation* models (Qwen2.5-{0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B}-Instruct and QwQ-32B-Preview). **Avg.** denotes the group average for student and teacher models, respectively. Tasks with a performance gap (**Gap =** Teacher Avg. − Student Avg.) exceeding 30 percentage points are highlighted in gray; these tasks are selected for our distillation experiments.

	Student					Teacher (Gemma-2)			
Task	**0.5B**	**1.5B**	**3B**	**7B**	**Avg. (S)**	**9B**	**27B**	**Avg. (T)**	**Gap (T−S)**
Boolean Expressions	75.2	88.0	88.8	96.4	87.1	91.6	91.2	91.4	4.3
Causal Judgement	48.7	59.9	54.5	60.4	55.9	63.1	66.8	65.0	9.1
Date Understanding	30.4	53.6	66.4	82.4	58.2	81.6	83.6	82.6	24.4
Disambiguation QA	31.2	49.2	63.6	69.2	53.3	75.6	82.4	79.0	25.7
Dyck Languages	0.0	0.0	3.6	11.2	3.7	9.2	9.2	9.2	5.5
Formal Fallacies	53.2	52.0	62.4	70.4	59.5	60.8	71.6	66.2	6.7
Geometric Shapes	22.0	28.8	49.2	66.4	41.6	52.0	54.8	53.4	11.8
Hyperbaton	55.2	72.8	74.4	90.0	73.1	77.2	98.4	87.8	14.7
Logical Deduction (3)	37.6	59.2	70.0	86.8	63.4	82.8	91.6	87.2	23.8
Logical Deduction (5)	22.4	33.2	45.6	59.2	40.1	57.2	61.2	59.2	19.1
Logical Deduction (7)	22.0	34.4	32.8	45.6	33.7	50.4	54.8	52.6	18.9
Movie Recommendation	37.2	48.4	74.0	66.4	56.5	59.6	73.2	66.4	9.9
Multistep Arithmetic Two	30.8	60.8	91.2	94.8	69.4	83.6	90.0	86.8	17.4
Navigate	58.4	73.6	90.0	95.2	79.3	58.8	98.0	78.4	0.9
Object Counting	53.6	58.8	63.2	78.4	63.5	88.0	95.6	91.8	28.3
Penguins in a Table	32.9	61.6	79.5	94.5	67.1	93.8	93.8	93.8	26.7
Reasoning about Colored Objects	20.8	62.0	75.6	90.0	62.1	79.6	88.4	84.0	21.9
Ruin Names	24.0	36.0	42.8	66.0	42.2	82.4	85.6	84.0	41.8
Salient Translation Error Detection	12.0	33.2	46.0	54.8	36.5	63.2	60.0	61.6	25.1
Snarks	46.6	57.3	64.6	79.8	62.1	80.9	87.1	84.0	21.9
Sports Understanding	48.8	73.2	82.0	87.2	72.8	92.4	94.4	93.4	20.6
Temporal Sequences	24.8	15.2	45.6	81.6	41.8	84.4	98.0	91.2	49.4
Tracking Shuffled Objects (3)	32.4	41.2	67.6	88.8	57.5	67.6	85.2	76.4	18.9
Tracking Shuffled Objects (5)	15.6	18.8	39.6	77.2	37.8	76.4	86.8	81.6	43.8
Tracking Shuffled Objects (7)	17.2	16.4	26.4	65.6	31.4	68.8	80.4	74.6	43.2
Web of Lies	59.6	80.8	100.0	99.2	84.9	100.0	100.0	100.0	15.1
Word Sorting	4.4	6.8	16.0	29.2	14.1	47.6	83.2	65.4	51.3

Table 10: Results of few-shot ICL performance on BBH for *pre-distillation* models (Qwen2.5-{0.5B, 1.5B, 3B, 7B}-Instruct as students and **Gemma-2-{9B, 27B}-it** as teachers). **Avg.** denotes the group average for student and teacher models, respectively. Tasks with a performance gap (**Gap =** Teacher Avg. − Student Avg.) exceeding 30 percentage points are highlighted in gray; these tasks are selected for our distillation experiments.