# SOLVING WITHOUT STOPPING: ON-POLICY DISTILLATION AT SMALL SCALE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy distillation, where a student learns from a stronger teacher's feedback on its own outputs, is a common way to pass reasoning to smaller models. We analyze what it transfers at small scale, distilling Qwen3-8B into Qwen3 4B, 1.7B and 0.6B students, in thinking mode (reason at length, then end the reasoning and answer) and, for comparison, in non-thinking mode (no separate reasoning phase). Long reasoning needs two abilities, solving a problem and knowing when it is solved, and we find that distillation transfers the first, but in thinking mode not the second. Solving improves at every size, up to two ceilings, which we measure comprehensively across both modes and all student sizes: a student's single attempt never exceeds what it could already reach in many attempts before training, and the smaller the student, the further it stays below the teacher. Stopping is where the modes part. In non-thinking mode every student keeps stopping; in thinking mode students stop ending their reasoning early in training, and the smaller the student, the less of this ability survives: the teacher signals a stop almost only where a student already ends its reasoning, so distillation teaches no new stops; it only keeps the student's existing stops that land on a right answer, and a weak student has few such stops. The smallest students often reach the right value but do not commit to it: they either rarely mark it or mark it and write past it. Together, these results describe how small students behave under on-policy distillation, and a diagnostic that separates answer marking, correctness and stopping.

## 1 INTRODUCTION

On-policy distillation trains a student on its own outputs: the student writes a response, a stronger teacher scores every token of it, and the student moves toward the teacher's distribution (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). It has become a standard way to give small models the reasoning of larger ones, including from teachers that think at length before they answer. Small models are where this matters most: they are the ones deployed where compute is scarce, and their reasoning comes almost entirely from distillation. Yet what a small student actually takes from a thinking teacher is not well understood. Reasoning at length rests on two abilities, solving the problem and recognizing that it is solved, and prior work has studied the second as a problem of its own, from overthinking (Chen et al., 2025) to knowing when to stop (Zhang et al., 2025; Yang et al., 2026). We ask whether on-policy distillation passes both to a small student. End-point scores alone cannot answer this: a student that stops too rarely looks like one that cannot solve. We therefore follow solving and stopping together.

We distill Qwen3-8B (Yang et al., 2025) into Qwen3 4B, 1.7B and 0.6B students, so that teacher and students all sit at small scale. We compare two modes. In thinking mode the model reasons at length before it answers; non-thinking mode, which answers directly, serves as a reference. In thinking mode, before distillation, each student first learns the thinking format by supervised fine-tuning, and we use two sources for that data: short external solutions, or the teacher's own long solutions. Controls vary the rollout budget, start from the officially post-trained model and repeat runs. We follow every student both at evaluation and through training, and diagnose solving and stopping together, separating whether a student answers, whether its answer is right, and whether it stops.



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

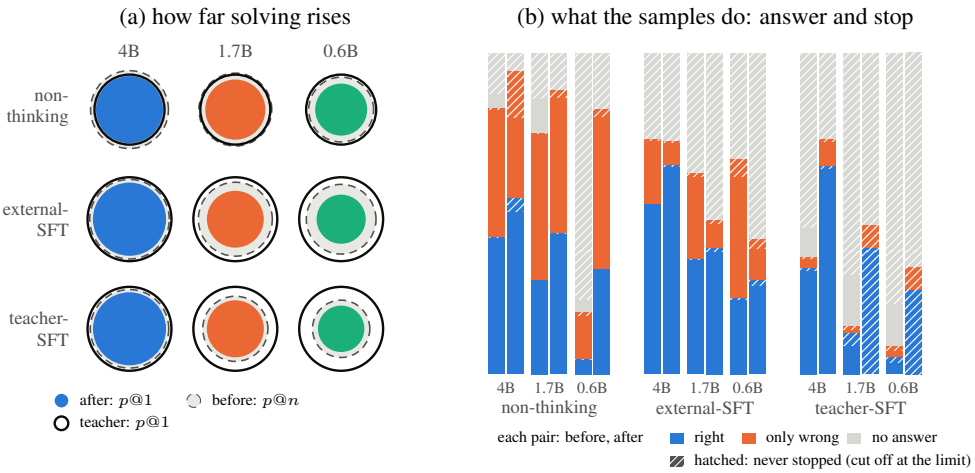



**Figure 1: Overview.** Qwen3-8B distilled into Qwen3 4B, 1.7B and 0.6B, on three routes; three-benchmark mean, lenient scoring. (a) Circle areas are proportional to accuracy. After distillation, a student's single-attempt accuracy (colored disk) stays inside what the start reached in many attempts (dashed), and the smaller the student, the further inside the teacher's (black ring). (b) Every sample, before and after distillation, by its marked answer and by whether it stopped. Without thinking, students keep stopping and wrong answers turn right; in thinking mode the small students trade finished wrong answers for samples that never stop and mark nothing.

The student acquires the first ability and, the smaller it is, less of the second (Figure 1). First, solving has a ceiling (Section 3). A distilled student's single-sample accuracy rises toward the accuracy it could already reach in many samples before training, but never past it (Figure 1a): one distilled attempt is worth a few attempts of the starting student. The gain comes from problems the student could already solve sometimes, it arrives early in training, and the set of problems a student can reach at all barely grows. The share of the teacher's accuracy that a student reaches falls with its size. When the student gives an answer, that answer is somewhat more often right after distillation at every size. Second, small students do not learn to stop (Section 4). Early in thinking-mode distillation from the external-SFT start, every student stops closing its reasoning, in the same step in which its responses fill the length budget. What follows works like a filter. Along a response that runs on, the teacher gives almost no signal to stop, and distillation keeps the stops that land on a correct answer and removes the others. Before training, the three students stop at similar rates that do not follow their size; the difference is where they stop. When the 4B stops, its answer is right about half the time; when the 0.6B stops, almost never. The filter therefore leaves the 4B with some of its stops and the 0.6B with almost none. The small student has not forgotten how to stop; it was never taught where to stop, and the stops it had were the wrong ones. Third, we look inside the samples that do not stop (Section 5). Before distillation, the 0.6B often gives up with a wrong answer; after it, it often keeps reasoning until it is cut off at the length limit and writes no answer at all. Samples with only wrong answers shrink sharply, and samples with no answer grow just as sharply. They are often not unsolved: the correct value appears in them early, several times more often than by chance, and the student keeps writing past it. From the teacher's own solutions, the small students do mark answers, but overwrite correct ones as they keep writing. Either way, they reach the right value but do not commit to it.

Our contributions are the following. **A comprehensive analysis** of on-policy distillation at small scale, across three student sizes, thinking and non-thinking mode, and two sources of fine-tuning data, followed both at evaluation and through training. **Two findings about what transfers**: solving rises only up to two ceilings, the student's own reach and the teacher's, systematically measured here for on-policy distillation in both thinking and non-thinking mode and across student sizes; and in thinking mode distillation filters stops rather than teaching them, which leaves the smallest students almost unable to stop. **A diagnostic** that separates whether a student answers, whether its answer is right, and whether it stops, so that a change in score can be traced to one of them; the main comparisons carry paired intervals.

## 2 SETUP

The teacher is Qwen3-8B (Yang et al., 2025). In thinking mode it reasons inside `<think>...</think>` and then answers; in non-thinking mode its chat template pre-fills an empty thinking block and it answers directly. The students are Qwen3 4B, 1.7B and 0.6B and always run in the teacher's mode. For thinking mode each student first learns the format by supervised fine-tuning, either on 5,878 short external solutions (the *external-SFT start*) or on 1,524 of the teacher's own correct solutions (the *teacher-SFT start*), and is then distilled for 142 steps. As a reference, the pretrained-only students are distilled from the non-thinking teacher for 212 steps (the *non-thinking route*); Appendix A lists every run. Distillation samples one response $y$ per prompt $x$, the teacher scores the same tokens, and each token receives a clipped one-sample estimate of the reverse KL divergence as its advantage in a PPO-style update, with no task reward, following Lu & Thinking Machines Lab (2025) (see also Agarwal et al., 2024; Gu et al., 2024):

$$A_t = -\operatorname{clip}\big(\log \pi_{\text{old}}(y_t \mid x, y_{<t}) - \log \pi_T(y_t \mid x, y_{<t}),\ -1,\ 1\big). \tag{1}$$

The token `</think>` is scored like any other, so the only pressure to stop comes from how likely the teacher is to stop at the point the student has reached.

Training uses 13,597 DAPO-Math-17k problems (Yu et al., 2025), 128 per step, and a training response may use at most 7,168 tokens (the *rollout budget*). We evaluate on AIME 2026, AMC23 and GSM8K-200 with 48, 48 and 32 samples per problem at temperature 0.6, and a limit of 30,720 tokens (8,192 for GSM8K-200); during training we validate every 20 steps on AIME 2025 and AMC23. Distillation converges early in both modes (Appendix Figure 9): after the first steps neither validation p@1 nor p@8 changes clearly on any run. A *marked answer* is a value the model itself presents as its answer, by writing it in `\boxed{}` or on a line that begins with `Answer:` or `The final answer is`; it may appear anywhere in a sample, inside the reasoning or after it, and a sample may mark several. A sample is correct under *lenient* scoring, our main score, if any of its marked answers is correct. A value that only appears in the working, for instance as an intermediate result such as "so $x = 12$", is not a marked answer and does not count, even if it happens to equal the correct answer. p@1 is the average over samples and p@n the fraction of problems solved in at least one of the $n$ samples. Within a benchmark, lenient p@1 is exactly the *marked-answer rate*, the fraction of samples with a marked answer, times the accuracy of the samples that have one. We report two behaviors beside the score, not inside it: the *closure rate*, the fraction of samples that close their reasoning with `</think>`, and *truncation*, the fraction that reach the token limit. Appendix B defines two stricter scorings; GPQA-Diamond results for the non-thinking route are in Appendix C. All intervals are paired 95% bootstrap intervals over problems (Appendix B). When we compare accuracy given a marked answer between two models, we use only the problems both answered.

## 3 THE CEILING OF ON-POLICY DISTILLATION

On-policy distillation raises every student's mean single-attempt accuracy, but only up to two ceilings, and no benchmark breaks the first (Figure 1a; per benchmark in Figures 2 and 3; values in Table 1). The first ceiling is the student itself. In all 27 combinations of route, size and benchmark, the distilled lenient p@1 stays below the p@n of the student it started from, the fraction of problems that student solved at least once in all the samples we drew. Over the three benchmarks the distilled p@1 reaches between 0.50 and 0.87 of the starting p@n, with every paired difference below zero (Appendix Table 8). The second ceiling is the teacher in the student's mode and the smaller the student, the farther below it the student stays. Over the three benchmarks the distilled p@1 reaches 0.922, 0.739 and 0.551 of the non-thinking teacher's from 4B to 0.6B, and 0.760, 0.458 and 0.343 of the thinking teacher's from the external-SFT start. The gap is widest on the hardest benchmark: on AIME 2026 the 1.7B and 0.6B distilled in thinking mode reach a tenth of the teacher's p@1 or less. In coverage the students come closer, and on GSM8K-200 the distilled p@n comes close to the teacher's at every size. The thinking teacher is also much stronger, 0.857 against 0.595 in p@1, so a smaller share of it is not a lower score.

The top of the range itself barely moves (Figure 4). The three-benchmark p@n changes by about 0.1 or less for every student distilled from the non-thinking teacher or from the external-SFT start, and grows a little more from the teacher-SFT start, by $+0.134$ at 4B and $+0.103$ at 1.7B, as those

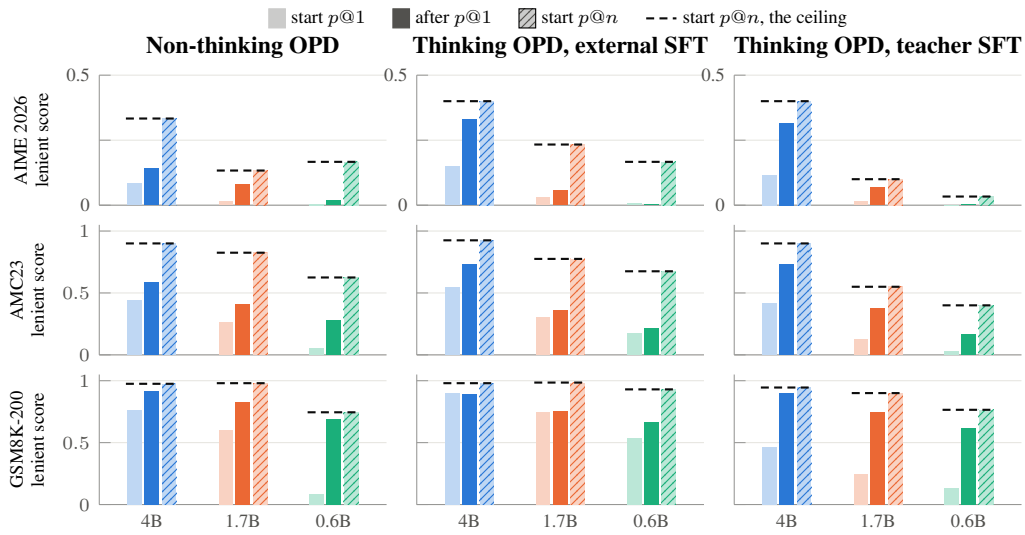



**Figure 2: After distillation, a single attempt never reaches what the student could reach in many attempts before.** Lenient scoring on each benchmark; $p@n$ over every sample we drew ($n = 48$ on AIME 2026 and AMC23, 32 on GSM8K-200). In every one of the 27 groups the distilled $p@1$ (dark solid bar) stays below the dashed line, the starting student's $p@n$ (hatched). How $p@n$ itself changes: Figure 4. Values: Table 5.

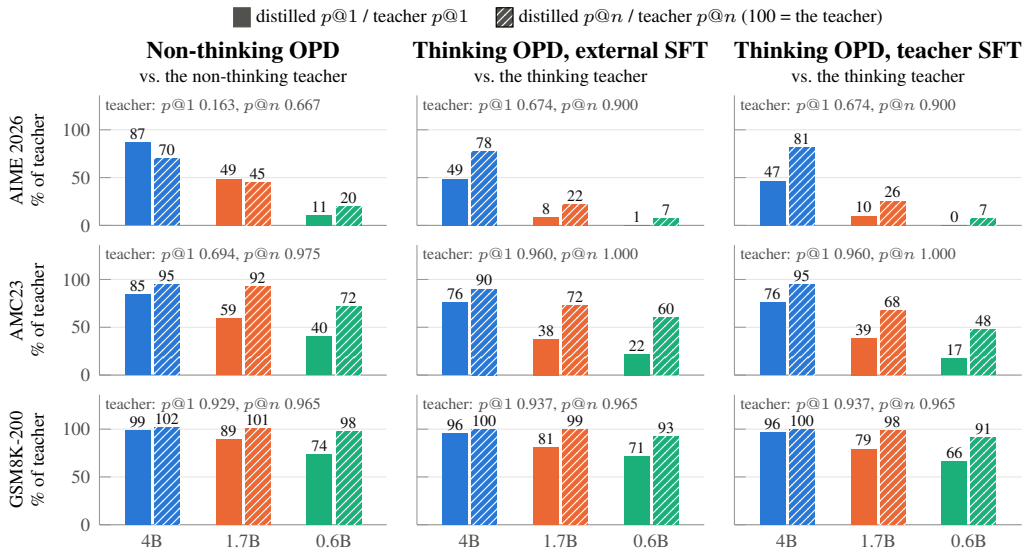



**Figure 3: The smaller the student, the farther it stays below the teacher.** The distilled student's lenient $p@1$ and $p@n$ on each benchmark as a percentage of the teacher's in the student's own mode (100: the teacher, whose own values are printed in each panel). On GSM8K-200 the distilled $p@n$ comes close to the teacher's at every size; on AIME 2026 the 1.7B and 0.6B in thinking mode reach a tenth of the teacher's $p@1$ or less.

students learn to write their answers down. The clear expansions are the 4B in thinking mode on AIME 2026, the hardest benchmark, where p@n rises from 0.400 to 0.700, and to 0.733 from the teacher-SFT start, and the pretrained-only 0.6B on GSM8K-200, from 0.745 to 0.945, which learns to write its answer down. At 1.7B and 0.6B from the external-SFT start, p@n ends slightly lower, by 0.036 and 0.070.

The gain comes from problems the student could already solve (Figure 5). Problems the start never solved in its $n$ samples account for 0.23 to 0.60 of the three-benchmark mean, yet they supply at most an eighth of the p@1 gain: after distillation they are solved in between 0.002 and 0.183 of attempts, and from the external-SFT start in at most 0.046. At 4B from the external-SFT start, the problems solved in fewer than half of the samples rise from 0.204 to 0.496; at 1.7B and 0.6B the

**Table 1: Main results, three-benchmark mean.** Lenient scoring; $p@n$ over every sample drawn ($n = 48$ on AIME 2026 and AMC23, 32 on GSM8K-200). The distilled $p@1$ stays below the start's $p@n$ in every row. From the teacher-SFT start, the 1.7B and 0.6B mark many answers per sample; Appendix B gives their scores without this. Per benchmark: Table 5.

| | Start | | After distillation | | After $p@1$ as a share of | |
|---|---|---|---|---|---|---|
| Student | $p@1$ | $p@n$ | $p@1$ | $p@n$ | start $p@n$ | teacher $p@1$ |
| *Non-thinking OPD* (teacher: $p@1$ 0.595, $p@n$ 0.869) | | | | | | |
| ● 4B | 0.427 | 0.736 | 0.549 | 0.791 | 0.75 | 0.922 |
| ● 1.7B | 0.294 | 0.646 | 0.440 | 0.725 | 0.68 | 0.739 |
| ● 0.6B | 0.048 | 0.512 | 0.328 | 0.593 | 0.64 | 0.551 |
| *Thinking OPD, external SFT* (teacher: $p@1$ 0.857, $p@n$ 0.955) | | | | | | |
| ● 4B | 0.532 | 0.768 | 0.651 | 0.855 | 0.85 | 0.760 |
| ● 1.7B | 0.358 | 0.664 | 0.392 | 0.628 | 0.59 | 0.458 |
| ● 0.6B | 0.238 | 0.591 | 0.294 | 0.521 | 0.50 | 0.343 |
| *Thinking OPD, teacher SFT* (teacher: $p@1$ 0.857, $p@n$ 0.955) | | | | | | |
| ● 4B | 0.330 | 0.748 | 0.649 | 0.883 | 0.87 | 0.757 |
| ● 1.7B | 0.128 | 0.517 | 0.394 | 0.619 | 0.76 | 0.460 |
| ● 0.6B | 0.055 | 0.399 | 0.262 | 0.474 | 0.66 | 0.306 |

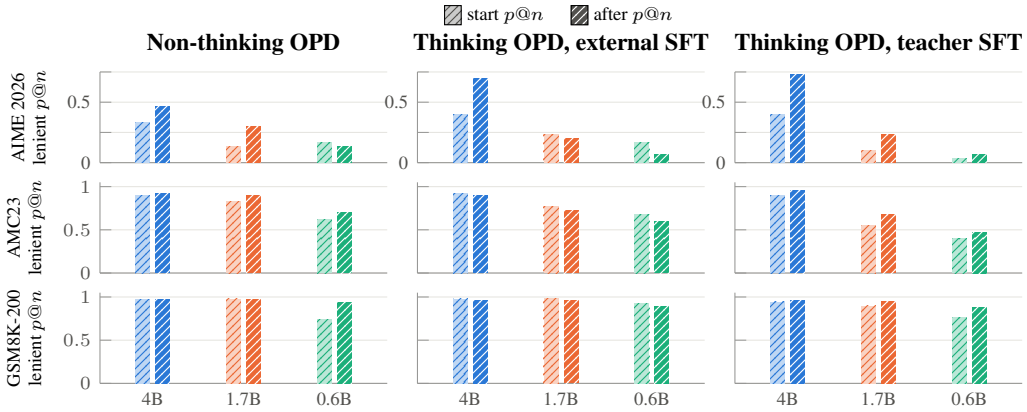



**Figure 4: What a student can reach at all barely changes.** Lenient $p@n$ on each benchmark before and after distillation, over every sample drawn. It grows clearly for the 4B in thinking mode on AIME 2026 and for the pretrained-only 0.6B on GSM8K-200.

same group rises much less, from 0.200 to 0.268 and from 0.192 to 0.301, and more problems leave the reachable set than enter it (11 against 3, and 22 against 9; Appendix Table 7).

What limits the small students in thinking mode is not the quality of their answers but their number. On the non-thinking route the students keep stopping (Figure 6), and lenient p@1 rises at every size, by $+0.121$, $+0.146$ and, clearly, $+0.280$. In thinking mode from the external-SFT start, answers are right more often at every size: on the problems both the starting and the distilled student answered, by $+0.190$, $+0.106$ and $+0.115$, all with intervals above zero (Appendix Table 9). The 4B marks answers as often as before ($-0.006$, within noise), so its lenient p@1 rises by $+0.119$. The 1.7B and 0.6B mark answers much less often ($-0.145$ and $-0.251$), and their lenient p@1 rises only by $+0.034$ and $+0.056$. Their missing answers belong to samples that never stopped, which the next two sections explain.

## 4 Why small students do not learn to stop

In thinking mode, distillation filters a student's stops rather than teaching them: it keeps the stops that land on a correct answer and removes the others, so a student whose stops were rarely right is left almost unable to stop. In every run from the external-SFT start at the 7,168-token budget, the fraction of training responses that close their reasoning drops sharply at step 6 (Figure 6), in the same step in which the median response reaches the rollout budget (Appendix Figure 11): between steps 1–5 and steps 6–10 it falls from 0.617 to 0.211 at 4B, from 0.445 to 0.128 at 1.7B and from 0.583 to 0.113 at 0.6B. Nothing we varied within thinking-mode training prevents this. Two runs of the same

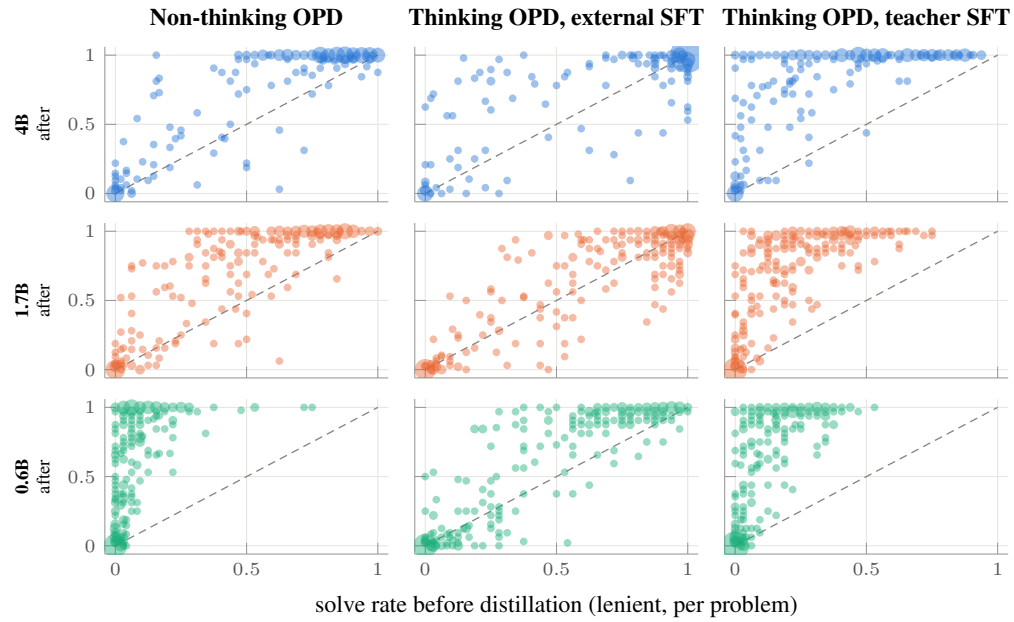



**Figure 5: Distillation works on problems the student could already solve.** Every problem of the three benchmarks is a point: how often the starting student solved it ($x$) against how often the distilled student does ($y$); larger marks stand for several problems at the same place. Above the dashed diagonal a problem got easier. Problems the start never solved ($x = 0$) mostly stay at the bottom; the gains come from problems it solved sometimes. Small students from the external-SFT start barely lift that middle band. Per-group numbers: Table 7.

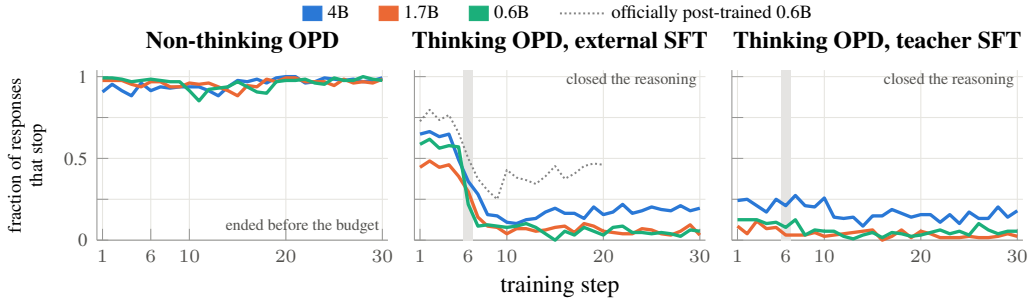



**Figure 6: Only thinking-mode students lose their stopping.** Training responses that stop, 7,168-token budget. Non-thinking OPD: almost every response ends before the budget throughout. Thinking OPD from the external-SFT start: closure collapses at step 6 (gray) at every size; afterwards the 4B keeps some, the 1.7B and 0.6B almost none (dotted: the officially post-trained 0.6B, 16,384 tokens). From the teacher-SFT start the students close little from the first step, because their responses already fill the budget. Response length, stop reliability and marked answers: Figure 11.

configuration at 4B and at 0.6B collapse at the same step; the 0.6B collapses under budgets of 7,168 and 16,384 tokens, filling the larger one a step later, and at 1,024 tokens fills the budget from the first step; the students from the teacher-SFT start fill the budget from the first step; and the officially post-trained Qwen3-0.6B, a much stronger starting point, also drops at step 6 (Appendix D).

Nothing brings the stops back because of where the teacher asks for one. We had the teacher score the students' first training responses (Figure 7; numbers in Appendix Table 13). Where a student closed its reasoning, the teacher agrees: its probability of `</think>` at that point is 0.976 at 4B, 0.764 at 1.7B and 0.673 at 0.6B. Along a typical response that ran to the budget, the teacher's largest probability of stopping at any point is at most about $10^{-7}$: along most such responses the teacher does not point to a better place to stop; it points nowhere. A stop is therefore reinforced where the student already stops, and nothing pulls a response that runs on toward a stop.

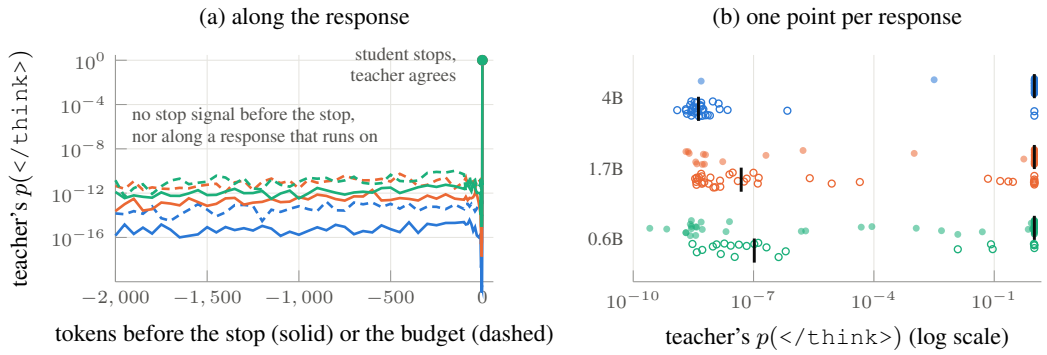



**Figure 7: The teacher agrees to stop where the student stopped, and almost nowhere along a response that never stops.** The students' first training responses (external-SFT start, 7,168-token budget), scored token by token by the thinking teacher; color: ●4B, ●1.7B, ●0.6B. (a) The teacher's probability of `</think>`, median over responses, at each position up to 2,000 tokens before the point where the student closed its reasoning (solid) or before the end of a response that ran to the budget (dashed). (b) Filled: the probability at the position where the student closed. Open: the largest probability anywhere along a response that ran to the budget. Black ticks: medians. Numbers: Appendix Table 13.

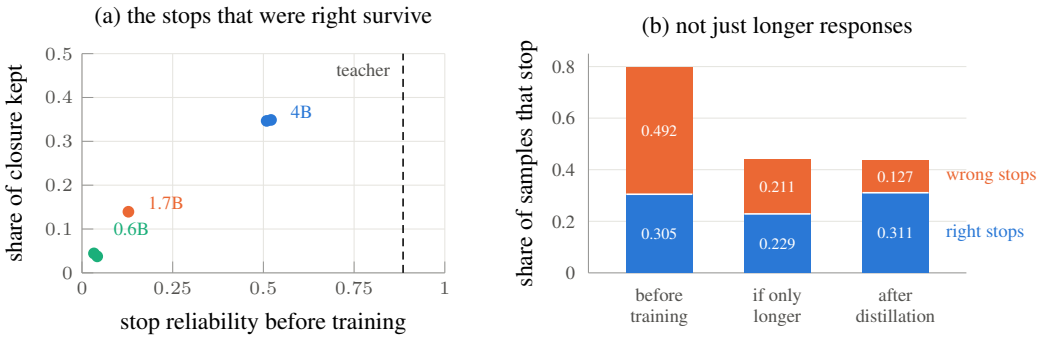



**Figure 8: Distillation filters stops: it keeps the ones that were right.** (a) Each run from the external-SFT start: the fraction of its stops that landed on a correct answer over the first four steps, against the share of its early closure it keeps late in training (color: size; the teacher's stop reliability dashed). Runs from the teacher-SFT start, which close little from the first step, and the 20-step runs of the officially post-trained 0.6B are in Appendix Table 12. (b) Could the lost stops simply be responses grown too long for the limit? Each bar splits the officially post-trained 0.6B's samples that stop into those whose answer is right (blue) and wrong (orange). Left: before distillation. Right: after 20 thinking-mode steps. Middle: what lengthening alone would give, if every starting response were stretched by one factor, chosen so that as many samples stop as after distillation. Lengthening alone would cut right and wrong stops alike; distillation instead keeps the right stops and removes mostly the wrong ones (Appendix Table 15).

What survives the filter is the correct stops (Figure 8), so the collapse is not the budget simply cutting responses short. Stretching every response of the officially post-trained 0.6B before distillation by one common factor, chosen to match the observed closure, predicts that correct stops fall from 0.305 to 0.229, but they stay at 0.311, while most of the wrong stops disappear. The stop token itself is unchanged: on identical prefixes, the officially post-trained 0.6B ranks `</think>` first wherever it had closed, both before and after 20 steps of distillation, and tens of thousands of ranks deep everywhere else, again in both (Appendix D.1). From the external-SFT start, what the filter leaves therefore depends on how many of a student's stops were right to begin with. Before training the students close their reasoning at rates that do not follow their size (0.648, 0.459 and 0.611 from 4B to 0.6B), but their *stop reliability*, the fraction of stops that land on a correct answer, is 0.521 at 4B, 0.128 at 1.7B and 0.042 at 0.6B, against 0.885 for the teacher (Figure 8a). Late in training they keep their closure in the same order, 0.226, 0.064 and 0.023, and the stops that remain are more often right where the student can solve: the accuracy of closed training responses rises at 4B from about one half to at most 0.698 within 30 steps, reaches 0.256 at 1.7B and stays near zero at 0.6B.

**Table 2: From the external-SFT start, the unfinished samples rarely mark an answer, but the correct value often appears in them, early; from the teacher-SFT start they mark answers but never finish.** Samples truncated at the cap (30,720 tokens; GSM8K-200 8,192), benchmarks pooled. Value present: the correct value occurs in the text (answers of $\geq$3 digits not in the statement); chance: another problem's answer; the dot plot shows the two (dot: present, tick: chance). First at, characters after it: medians. Loop: a line repeated $\geq$5 times at the end. Details: Table 17.

| | All samples | Truncated samples | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Truncated | With a marked answer | Correct value present | Chance | present vs. chance | First at | Chars. after it | Loop |
| Teacher (thinking) | 0.045 | 0.173 | 0.184 | 0.063 | | 0.48 | 23,492 | 0.077 |
| *External-SFT start, before and after thinking-mode distillation* | | | | | | | | |
| ● *4B start* | 0.274 | 0.035 | 0.067 | 0.028 | | 0.08 | 33,181 | 0.115 |
| after | 0.284 | 0.066 | 0.211 | 0.049 | | 0.21 | 29,796 | 0.084 |
| ● *1.7B start* | 0.387 | 0.047 | 0.126 | 0.024 | | 0.07 | 25,693 | 0.163 |
| after | 0.540 | 0.055 | 0.298 | 0.043 | | 0.14 | 26,362 | 0.197 |
| ● *0.6B start* | 0.392 | 0.178 | 0.169 | 0.036 | | 0.06 | 23,474 | 0.202 |
| after | 0.627 | 0.107 | 0.240 | 0.037 | | 0.21 | 20,757 | 0.276 |
| *Teacher-SFT start, after distillation* | | | | | | | | |
| ● 1.7B | 1.000 | 0.640 | 0.614 | 0.053 | | 0.07 | 23,486 | 0.650 |
| ● 0.6B | 1.000 | 0.506 | 0.506 | 0.045 | | 0.07 | 22,972 | 0.567 |

## 5 What the samples that do not stop contain

The small students often reach the right value but do not commit to it. From the external-SFT start they rarely mark it; from the teacher-SFT start they mark it and write past it. Below 4B, distillation trades wrong answers for no answer. We sort every evaluation sample by what it marks (Figure 1b; per size in Appendix Figure 12). From the external-SFT start, samples whose marked answers are all wrong shrink at every size, for instance from 0.433 to 0.127 of all samples at 0.6B. At 4B their place is taken by right answers, and samples without any marked answer stay where they were. At 1.7B and 0.6B their place is taken mostly by samples that mark no answer, which rise from 0.373 to 0.518 and from 0.328 to 0.579, while right answers rise only a little; every such sample ran to the evaluation limit.

From the external-SFT start, these unfinished samples, the ones cut off at the length limit, almost never mark an answer, yet many of them already contain the correct value, early. Among the samples that reach the limit, only 0.066, 0.055 and 0.107 mark any answer at 4B, 1.7B and 0.6B (Table 2), so lenient scoring cannot credit the rest, whatever their reasoning contains. We search the text of each unfinished sample for the correct value, on problems whose answer has at least three digits and does not occur in the statement, and compare with how often the answer of a different problem appears. The correct value appears in 0.211, 0.298 and 0.240 of these samples, four to seven times the chance rate, and more often than before distillation. It first appears around the first fifth of the text, and the student then writes tens of thousands of further characters. The unfinished training responses look the same: after the collapse, the correct value appears in 0.309 of the 4B's, against 0.019 by chance.

Forcing a stop points the same way. When we append `</think>` to the unfinished samples of the officially post-trained 0.6B and let it answer (Muennighoff et al., 2025), it answers correctly in 0.364 of them on problems its starting model had solved three or four times in four, and in 0.049 on problems the starting model never solved (Appendix E): where the student can solve, much of what these samples lack is the stop. The smallest students' unfinished samples also often end in repetition: 0.084 at 4B, about as many as for the teacher (0.077), but 0.197 and 0.276 at 1.7B and 0.6B. From the teacher-SFT start the failure takes another form. The 1.7B and 0.6B distilled from it never finish a sample at evaluation, yet they mark 185.5 and 150.3 answers per sample on average; they mark a first answer halfway through, at 0.44 and 0.50 of the output against 0.85 for the teacher, and then overwrite it: when a correct answer was marked, the last marked answer is wrong in 0.437 and 0.180 of those samples, against 0.004 for the teacher (Appendix E).

## 6 Related work

On-policy distillation trains the student on its own samples (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025) rather than on the teacher's outputs (Hinton et al., 2015; Kim & Rush, 2016). Recent variants change the divergence (Jin et al., 2026; Fu et al., 2026) or the credit assigned over time (Liu et al., 2026), cut rollouts where the teacher's signal weakens (Zhou et al., 2026; Xin et al., 2026), or mask the stop token's advantage to contain length (He et al., 2026). Analyses relate success to compatible thinking patterns (Li et al., 2026), report length exploitation (Wang et al., 2026), find that distillation improves sampling efficiency more than coverage (Ge et al., 2026), and describe outcome-confounded supervision (Ma, 2026) and privileged teachers that harm thinking students (Kaur et al., 2026). We hold one recipe fixed across student sizes and locate the loss of stopping in the states the student reaches, not in the token. For reinforcement learning with verifiable rewards, single-sample gains have been found to stay within what the base model reaches in many samples (Yue et al., 2025). For on-policy distillation, how far a student can rise, in thinking and in non-thinking mode and as a function of its size, has not been measured; Section 3 does so.

Prior work treats stopping as a problem of its own, separate from solving. Long reasoning overspends on easy problems (Chen et al., 2025) and abandons promising lines too early (Wang et al., 2025); models carry a signal that they have already answered correctly (Zhang et al., 2025), and inference-time methods impose a stop through budget forcing (Muennighoff et al., 2025), early exit (Yang et al., 2026) or abstention (Davidov et al., 2026). These methods decide when a trained model should stop; we ask what training teaches about stopping. We find that on-policy distillation filters a student's stops rather than teaching them.

Small dense models learn long reasoning from a strong model's traces (Guo et al., 2025), but the smallest learn better from short chains than from long ones (Li et al., 2025). In imitation learning, a learner that lacks information the expert acts on cannot recover the expert's decisions: an imitation gap remains (Weihs et al., 2021), and learners trained off-policy latch onto correlates (de Haan et al., 2019; Swamy et al., 2022). Stopping is such a decision: the teacher stops at the end of a finished derivation, which the small student rarely reaches.

## 7 Conclusion

We analyzed on-policy distillation at small scale across three student sizes, both modes and two sources of fine-tuning data. It transfers solving but, in thinking mode, not stopping. Solving improves at every size up to two ceilings, the student's own reach before training and a share of the teacher that shrinks with the student, which we measured across both modes and all sizes. In non-thinking mode every student keeps stopping; in thinking mode distillation filters stops rather than teaching them, so the smaller the student, the less stopping survives, and the smallest students reach the right value without committing to it. Our diagnostic, which separates answering, correctness and stopping, is what makes this visible: an end-point score alone would read the small students' lost stops as lost solving. These results locate where methods for small students should act: along the responses that run on, where the teacher gives no signal to stop, rather than at the stop token itself. We hope they offer guidance for distilling small reasoning models: how far solving can rise, why stopping fails in thinking mode, and what the unfinished outputs of the small students contain.

## AI use statement

We used generative AI tools mainly for two purposes: to improve the writing of this paper, and to help write analysis and plotting scripts. The authors designed the study, ran all training and evaluation, and checked every analysis, number and claim against the recorded logs and data. Generative AI was not used to generate data or to run the experiments; translation and dataset cleaning are not applicable to this work. We have reviewed all AI-assisted content and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This study distills mathematical reasoning between openly released language models and evaluates it on public benchmarks. It involves no human subjects, no personal data and no new data collection: the training prompts are a public competition-mathematics set, and the cold-start solutions come from a public corpus or from the teacher itself. We report how a widely used recipe can fail on small students so that the failure can be detected, and we see no direct route to misuse.

## REPRODUCIBILITY STATEMENT

Section 2 and Appendix A specify the teacher, the students, the objective, the hyperparameters, the data pipeline, every training run and the evaluation protocol. The anonymized supplementary archive, described in Appendix F, contains the code, including the scorer shared by training and evaluation and every analysis script; the derived data, from the per-problem counts and per-step aggregates to the probe summaries and a fact sheet that names the source file of every number in the text; the training, cold-start and evaluation datasets; the configuration of every run, the training and evaluation scripts and the patch to verl; and the paper source. From it, `make verify` rebuilds every interval, the fact sheet, every table and figure offline on a CPU, checks that each is byte-identical to the shipped copy, and builds the PDF. Model weights are not included; the teacher and the pretrained-only and officially post-trained students are public checkpoints, and the other students can be retrained from the included configurations and data.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499, 2025.

Hen Davidov, Nachshon Cohen, Oren Kalinsky, Yaron Fairstein, Guy Kushilevitz, Ram Yazdi, and Patrick Rebeschini. Knowing when to quit: A principled framework for dynamic abstention in LLM reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. In *Conference on Language Modeling (COLM)*, 2026.

Xinmu Ge, Zizhuo Zhang, Yu Huang, Jianing Zhu, Lin Yuan, Wanli Gu, Weichang Wu, Weiran Huang, Xiaolu Zhang, Bo Han, Jun Zhou, and Jiangchao Yao. Towards understanding on-policy distillation through the lens of test-time scaling. *arXiv preprint arXiv:2608.11829*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Haonan He, Haodi Lei, Yun Luo, Haoran Zhang, Shunkai Zhang, Yizhuo Li, Shengji Tang, Zhilin Wang, Runzhe Zhan, Lei Bai, Ganqu Cui, Fangchen Yu, Yafu Li, Peng Ye, Ning Ding, and Yu Cheng. SimpleOPD: Simple tokenizer-agnostic on-policy distillation for long-context reasoning. *arXiv preprint arXiv:2608.14277*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning (ICML)*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1317–1327, 2016.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

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

Shiqi Liu, Zeyu He, Letian Tao, Guojian Zhan, Jiaxin Gao, Feihong Zhang, Jingliang Duan, Wei Xiong, Kehua Sheng, Bo Zhang, Yang Guan, and Shengbo Eben Li. Beyond token-local imitation: Reward-compatible temporal credit assignment for on-policy distillation. *arXiv preprint arXiv:2609.16937*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. Thinking Machines Lab: Connectionism, `https://thinkingmachines.ai/blog/on-policy-distillation/`, October 2025. doi:10.64434/tml.20251026.

Guoqing Ma. Outcome-confounded local supervision in on-policy distillation. *arXiv preprint arXiv:2607.23731*, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 20275–20321, 2025.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Sequence model imitation learning with unobserved contexts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17665–17676, 2022.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of long reasoning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. Bridging the imitation gap by adaptive insubordination. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 19134–19146, 2021.

Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the KL agreement trap in on-policy distillation. *arXiv preprint arXiv:2606.09471*, 2026.

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

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 113222–113244, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Conference on Language Modeling (COLM)*, 2025.

Ziheng Zhou, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

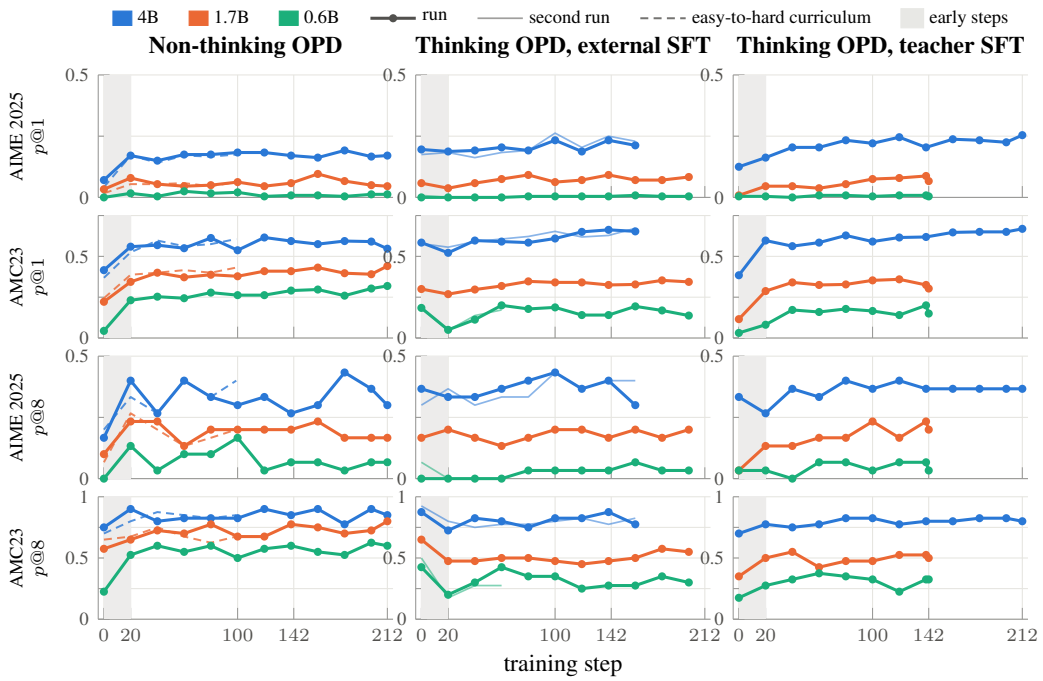



**Figure 9: Thinking or not, distillation converges early.** Lenient $p$@1 and $p$@8 on validation problems held out from training, AIME 2025 (30 problems) and AMC23 (40), 8 samples each, 16,384-token limit, every 20 steps; ticks at 142 and 212 mark the evaluated steps. After the shaded first steps no curve rises clearly. In thinking mode from the external-SFT start the curves dip early, when the students stop closing their reasoning (Section 4). Thin: a second run of the same setting; dashed: non-thinking runs with an easy-to-hard curriculum (99 steps).

## A Experimental details

**Training problems.** We started from the 17,398 problems of DAPO-Math-17k (Yu et al., 2025). Cleaning left 15,827; removing exact duplicates (924), near duplicates with a 13-gram Jaccard similarity of at least 0.7 (167), and problems that also occur in AIME 2024–2026, AMC23 or MATH (1,139) left 13,597. Every prompt ends with the instruction to reason step by step and to put the final answer in `\boxed{}`.

**The two cold starts.** The short external solutions are all 5,878 solutions of at most 2,048 tokens (mean 1,530) in a quality-filtered 45k subset of OpenR1-Math-220k (Apache-2.0), written as a thinking block followed by a boxed answer. The teacher's own solutions are, for each of 5,000 training problems, the shortest correct thinking-mode sample among the teacher's four (temperature 0.6), kept if it has at most 7,000 tokens and is not predominantly Chinese: 1,524 solutions with a mean of 4,809 tokens. The same fine-tuning recipe is used at every size (Table 4).

## B Scoring and statistics

**Marked answers.** A marked answer is the content of a `\boxed{}` or of an `Answer:` / `final answer is` line. It is compared with the reference after LaTeX normalization and, for numbers, by numeric value. The same matcher scores training rewards, evaluation and every analysis in this paper.

**Three scorings.** *Lenient*: at least one marked answer in the sample is correct, wherever it appears. *Strict*: the sample also finished, by closing its reasoning (thinking mode) or by stopping before the token limit (non-thinking mode), and the answer after the close is correct. *Committed*: the last marked `\boxed{}` is correct; majority votes count one committed answer per sample. A sample with no marked answer is wrong under all three.

**Table 3: Training runs used in this paper.** All runs distill Qwen3-8B on the same 13,597 problems; the student runs in the teacher's mode. *Steps*: the last training step whose responses were stored. *Evaluated*: the step whose checkpoint was evaluated on the benchmarks. Runs 1 and 2 of one configuration differ only by being run twice over the same data order.

| Student start | Teacher | Budget | Steps | Evaluated | Used in |
|---|---|---|---|---|---|
| *Non-thinking route* | | | | | |
| 4B pretrained-only | non-thinking | 7,168 | 212 | 212 | Section 3 |
| 1.7B pretrained-only | non-thinking | 7,168 | 212 | 212 | Section 3 |
| 0.6B pretrained-only | non-thinking | 7,168 | 212 | 212 | Section 3 |
| 4B pretrained-only, curriculum | non-thinking | 7,168 | 99 | – | Figure 1 |
| 1.7B pretrained-only, curriculum | non-thinking | 7,168 | 99 | – | Figure 1 |
| *Thinking mode* | | | | | |
| 4B external-SFT start, run 1 | thinking | 7,168 | 170 | – | Section 4 |
| 4B external-SFT start, run 2 | thinking | 7,168 | 167 | 142 | Sections 3–5 |
| 1.7B external-SFT start | thinking | 7,168 | 212 | 142 | Sections 3–5 |
| 0.6B external-SFT start, run 1 | thinking | 7,168 | 80 | – | Section 4 |
| 0.6B external-SFT start, run 2 | thinking | 7,168 | 212 | 142 | Sections 3–5 |
| 4B teacher-SFT start | thinking | 7,168 | 212 | 142 | Sections 3–5 |
| 1.7B teacher-SFT start | thinking | 7,168 | 142 | 142 | Sections 3–5 |
| 0.6B teacher-SFT start | thinking | 7,168 | 142 | 142 | Sections 3–5 |
| *Controls* | | | | | |
| 0.6B external-SFT start | thinking | 16,384 | 7 | – | budget, Section 4 |
| 0.6B external-SFT start | thinking | 1,024 | 212 | – | budget, Section 4 |
| officially post-trained 0.6B, run 1 | thinking | 16,384 | 20 | – | start, Section 4 |
| officially post-trained 0.6B, run 2 | thinking | 16,384 | 20 | 20 | Sections 4–5, probes |

**Table 4: Settings.** Distillation, the supervised cold starts, evaluation and in-training validation.

| | |
|---|---|
| *On-policy distillation (verl)* | |
| prompts per step; updates per step | 128; 4 mini-batches of 32 |
| optimizer | AdamW, learning rate $3 \times 10^{-6}$, 10 warm-up steps, then constant; weight decay 0.01; gradient clip 1.0 |
| advantage | clipped per-token reverse KL, Eq. (1), clip at $\pm 1$; no task reward |
| loss aggregation | sum over tokens, mean over sequences, divided by a constant of 7,168 |
| rollouts | 1 per prompt, temperature 1.0, budget 7,168 tokens (controls: 1,024 and 16,384) |
| length | 2 epochs $=$ 212 steps; thinking mode evaluated at step 142 |
| *Supervised cold start (same recipe at every size)* | |
| data | 5,878 short external solutions, or 1,524 of the teacher's own correct solutions |
| optimization | batch 16, learning rate $2 \times 10^{-5}$, cosine to $0.1\times$, warm-up ratio 0.01, 2 epochs |
| maximum length | 4,096 tokens (short solutions), 10,240 (teacher solutions) |
| *Evaluation* | |
| benchmarks (problems, samples, token limit) | AIME 2026 (30, 48, 30,720); AMC23 (40, 48, 30,720); GSM8K-200 (200, 32, 8,192) |
| non-thinking route only | GPQA-Diamond (198, 8, 16,384) |
| sampling | temperature 0.6, top-$p$ 0.95, top-$k$ 20 |
| in-training validation | every 20 steps; AIME 2025 and AMC23, 8 samples, 16,384 tokens, temperature 0.6 |
| *Compute* | |
| hardware | 3 nodes of 4 A100-40GB: the teacher on one node (vLLM), the student on two (FSDP) |

**Samples with several marked answers.** Lenient scoring would reward a sample that lists several different answers. This credit is at most 0.025 of the three-benchmark mean for every model except the 1.7B and 0.6B distilled from the teacher-SFT start, which mark over a hundred answers per sample (Section 5); removing it moves their lenient p@1 from 0.394 to 0.256 and from 0.262 to 0.231, close to their committed scores of 0.254 and 0.225.

**The correct-value diagnostic.** In Section 5 we ask whether the correct value appears anywhere in a sample that did not finish, marked or not. Small numbers occur in almost any derivation, so we use only problems whose reference answer is an integer of at least three digits that does not occur in the problem statement, and we report beside it how often the answer of a different such problem appears in the same samples. It is not part of any score.

**GPQA-Diamond.** GPQA-Diamond has four options, so a sample that names several options would collect lenient credit by listing them. We therefore evaluate it only for the non-thinking route, score it by the committed answer (strict beside it), and never average it with the other benchmarks (Table 10). We read the leading option letter of each marked answer, so that forms such as `\boxed{\text{A}}` are credited.

**Intervals.** Every interval is a paired bootstrap over problems with 10,000 resamples: within each benchmark, problems are drawn with replacement, the same draw is used for both models of a comparison, the metric is computed per benchmark and the benchmarks are averaged without weights, as in the text. We report 95% percentile intervals. We read a difference whose interval contains zero as no difference, and call a change of about 0.1 or less small even when its interval excludes zero.

## C Supplementary results: how far solving rises

**Scores per benchmark.** Table 5 gives lenient p@1 and p@n on every benchmark for the teacher and for every starting and distilled student; Table 1 in the main text gives their three-benchmark means.

**Table 5: Lenient $p@1$ and $p@n$ on every benchmark.** Distillation raises every student's mean lenient $p@1$, but on no benchmark past the $p@n$ of the student it started from. Lenient scoring; $p@n$ over $n = 48$ samples (AIME 2026, AMC23) or 32 (GSM8K-200), every sample we drew; mean: unweighted over the three benchmarks. Gray: the student before distillation. The two ceilings: Figures 2 and 3; paired intervals: Table 8.

| | | AIME 2026 | | AMC23 | | GSM8K-200 | | Mean | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | $p@1$ | $p@48$ | $p@1$ | $p@48$ | $p@1$ | $p@32$ | $p@1$ | $p@n$ |
| Teacher (8B) | thinking | 0.674 | 0.900 | 0.960 | 1.000 | 0.937 | 0.965 | 0.857 | 0.955 |
| | non-thinking | 0.163 | 0.667 | 0.694 | 0.975 | 0.929 | 0.965 | 0.595 | 0.869 |
| *Non-thinking route: from the pretrained-only model, non-thinking teacher* | | | | | | | | | |
| ● 4B | start | 0.083 | 0.333 | 0.438 | 0.900 | 0.761 | 0.975 | 0.427 | 0.736 |
| | after | 0.141 | 0.467 | 0.587 | 0.925 | 0.919 | 0.980 | 0.549 | 0.791 |
| ● 1.7B | start | 0.014 | 0.133 | 0.266 | 0.825 | 0.601 | 0.980 | 0.294 | 0.646 |
| | after | 0.079 | 0.300 | 0.410 | 0.900 | 0.830 | 0.975 | 0.440 | 0.725 |
| ● 0.6B | start | 0.004 | 0.167 | 0.053 | 0.625 | 0.087 | 0.745 | 0.048 | 0.512 |
| | after | 0.018 | 0.133 | 0.278 | 0.700 | 0.689 | 0.945 | 0.328 | 0.593 |
| *Thinking mode, from the external-SFT start, thinking teacher* | | | | | | | | | |
| ● 4B | start | 0.148 | 0.400 | 0.548 | 0.925 | 0.899 | 0.980 | 0.532 | 0.768 |
| | after | 0.329 | 0.700 | 0.728 | 0.900 | 0.896 | 0.965 | 0.651 | 0.855 |
| ● 1.7B | start | 0.029 | 0.233 | 0.299 | 0.775 | 0.747 | 0.985 | 0.358 | 0.664 |
| | after | 0.057 | 0.200 | 0.362 | 0.725 | 0.757 | 0.960 | 0.392 | 0.628 |
| ● 0.6B | start | 0.006 | 0.167 | 0.173 | 0.675 | 0.536 | 0.930 | 0.238 | 0.591 |
| | after | 0.004 | 0.067 | 0.212 | 0.600 | 0.667 | 0.895 | 0.294 | 0.521 |
| *Thinking mode, from the teacher-SFT start, thinking teacher* | | | | | | | | | |
| ● 4B | start | 0.116 | 0.400 | 0.412 | 0.900 | 0.461 | 0.945 | 0.330 | 0.748 |
| | after | 0.316 | 0.733 | 0.728 | 0.950 | 0.903 | 0.965 | 0.649 | 0.883 |
| ● 1.7B | start | 0.013 | 0.100 | 0.129 | 0.550 | 0.243 | 0.900 | 0.128 | 0.517 |
| | after | 0.067 | 0.233 | 0.372 | 0.675 | 0.743 | 0.950 | 0.394 | 0.619 |
| ● 0.6B | start | 0.001 | 0.033 | 0.029 | 0.400 | 0.135 | 0.765 | 0.055 | 0.399 |
| | after | 0.003 | 0.067 | 0.165 | 0.475 | 0.620 | 0.880 | 0.262 | 0.474 |

**Pass@$k$ and the problems behind the gain.** Figure 10 shows, for every route and size, how often the starting and the distilled student solve a problem at least once in $k$ samples; Table 6 gives the values and the $k^\star$ of every distilled student, and Table 7 the numbers behind Figure 5. On AMC23 and GSM8K-200 the two curves close in as $k$ grows for every student; on AIME 2026, the hardest benchmark, the distilled curve pulls clearly ahead for the 4B in thinking mode and falls behind at

large $k$ for the 0.6B from the external-SFT start. The curves also give an exchange rate: one distilled attempt is worth the start's best of 4 attempts at 4B and of only 2 at 1.7B and 0.6B from the external-SFT start, and of 8 to 15 only where the start rarely marked an answer, as the pretrained-only 0.6B (0.194 of its samples) and the teacher-SFT starts (0.366, 0.151 and 0.089) did.

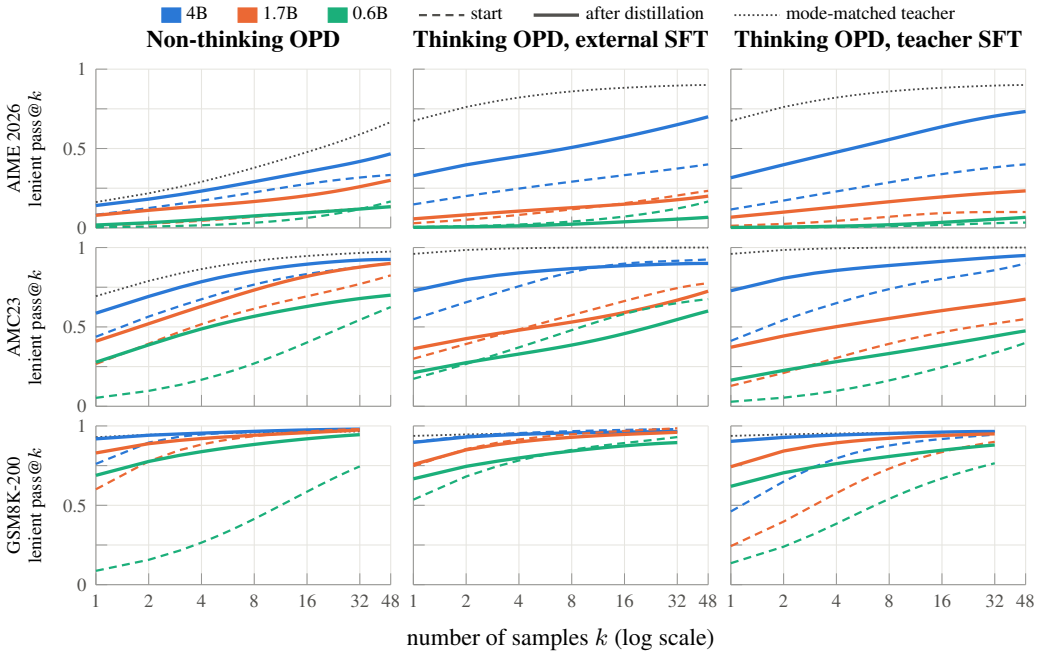



**Figure 10: Lenient pass@$k$ per benchmark, before (dashed) and after (solid) distillation.** Each curve runs to the number of samples drawn, 48 on AIME 2026 and AMC23 and 32 on GSM8K-200; its right end is the $p@n$ of Table 5. On AMC23 and GSM8K-200 the two curves close in as $k$ grows in every panel; on AIME 2026, the hardest benchmark, the distilled curve pulls clearly ahead for the 4B in thinking mode and falls behind at large $k$ for the 0.6B from the external-SFT start. Three-benchmark values and $k^\star$: Table 6.

**Table 6: Lenient pass@$k$ (Figure 10).** Three-benchmark mean of the unbiased per-problem estimate; beyond $k = 32$ GSM8K-200 (32 samples) enters with its $p@32$, so $k = 48$ is the mean $p@n$. $k^\star$: the smallest $k$ at which the starting model's pass@$k$ reaches the distilled model's $p@1$. Starting points in gray.

| Model | $k=1$ | $k=2$ | $k=4$ | $k=8$ | $k=16$ | $k=32$ | $k=48$ | $k^\star$ |
|---|---|---|---|---|---|---|---|---|
| Teacher (thinking) | 0.857 | 0.898 | 0.923 | 0.938 | 0.948 | 0.954 | 0.955 | |
| Teacher (non-thinking) | 0.595 | 0.651 | 0.703 | 0.751 | 0.796 | 0.840 | 0.869 | |
| 4B pretrained-only | 0.427 | 0.528 | 0.597 | 0.651 | 0.693 | 0.723 | 0.736 | |
| + non-thinking route | 0.549 | 0.605 | 0.657 | 0.703 | 0.741 | 0.773 | 0.791 | 3 |
| 4B external-SFT start | 0.532 | 0.597 | 0.653 | 0.701 | 0.735 | 0.757 | 0.768 | |
| + thinking-mode distillation | 0.651 | 0.709 | 0.746 | 0.777 | 0.807 | 0.838 | 0.855 | 4 |
| 4B teacher-SFT start | 0.330 | 0.455 | 0.558 | 0.634 | 0.686 | 0.728 | 0.748 | |
| + thinking-mode distillation | 0.649 | 0.711 | 0.758 | 0.799 | 0.837 | 0.869 | 0.883 | 10 |
| 1.7B pretrained-only | 0.294 | 0.399 | 0.482 | 0.540 | 0.585 | 0.624 | 0.646 | |
| + non-thinking route | 0.440 | 0.506 | 0.563 | 0.613 | 0.660 | 0.704 | 0.725 | 3 |
| 1.7B external-SFT start | 0.358 | 0.433 | 0.493 | 0.546 | 0.596 | 0.644 | 0.664 | |
| + thinking-mode distillation | 0.392 | 0.453 | 0.495 | 0.529 | 0.562 | 0.602 | 0.628 | 2 |
| 1.7B teacher-SFT start | 0.128 | 0.211 | 0.308 | 0.398 | 0.465 | 0.507 | 0.517 | |
| + thinking-mode distillation | 0.394 | 0.461 | 0.509 | 0.547 | 0.579 | 0.606 | 0.619 | 8 |
| 0.6B pretrained-only | 0.048 | 0.087 | 0.149 | 0.239 | 0.350 | 0.469 | 0.512 | |
| + non-thinking route | 0.328 | 0.398 | 0.459 | 0.509 | 0.548 | 0.581 | 0.593 | 15 |
| 0.6B external-SFT start | 0.238 | 0.320 | 0.392 | 0.456 | 0.515 | 0.567 | 0.591 | |
| + thinking-mode distillation | 0.294 | 0.342 | 0.380 | 0.417 | 0.457 | 0.499 | 0.521 | 2 |
| 0.6B teacher-SFT start | 0.055 | 0.099 | 0.162 | 0.238 | 0.311 | 0.377 | 0.399 | |
| + thinking-mode distillation | 0.262 | 0.312 | 0.351 | 0.387 | 0.423 | 0.459 | 0.474 | 11 |

**Table 7: Where the $p@1$ gain comes from (Figure 5).** Problems of the three benchmarks grouped by the starting model's lenient solve rate over its $n$ samples. Per group: number of problems, mean solve rate at the start → after distillation, and the group's contribution to the three-benchmark $p@1$ gain (the three contributions add up to the gain). Enter / leave: problems never solved at the start and solved at least once after, and the reverse.

| | | Never solved at the start | | | Solved in fewer than half | | | Solved in at least half | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Route | Size | problems | after | contrib. | problems | start → after | contrib. | problems | start → after | contrib. | enter / leave |
| non-thinking | 4B | 29 | 0.025 | +0.004 | 42 | 0.219 → 0.437 | +0.048 | 199 | 0.819 → 0.967 | +0.070 | 8 / 2 |
| | 1.7B | 37 | 0.017 | +0.005 | 86 | 0.250 → 0.523 | +0.080 | 147 | 0.747 → 0.946 | +0.061 | 11 / 4 |
| | 0.6B | 91 | 0.183 | +0.034 | 175 | 0.098 → 0.736 | +0.243 | 4 | 0.633 → 1.000 | +0.002 | 49 / 7 |
| external SFT | 4B | 25 | 0.046 | +0.013 | 36 | 0.204 → 0.496 | +0.086 | 209 | 0.952 → 0.953 | +0.021 | 9 / 4 |
| | 1.7B | 35 | 0.002 | +0.000 | 61 | 0.200 → 0.268 | +0.024 | 174 | 0.862 → 0.869 | +0.010 | 3 / 11 |
| | 0.6B | 52 | 0.019 | +0.003 | 94 | 0.192 → 0.301 | +0.021 | 124 | 0.776 → 0.909 | +0.032 | 9 / 22 |
| teacher SFT | 4B | 33 | 0.149 | +0.020 | 121 | 0.262 → 0.822 | +0.204 | 116 | 0.694 → 0.990 | +0.095 | 16 / 0 |
| | 1.7B | 65 | 0.069 | +0.016 | 184 | 0.225 → 0.764 | +0.232 | 21 | 0.602 → 0.970 | +0.018 | 22 / 3 |
| | 0.6B | 100 | 0.072 | +0.015 | 169 | 0.164 → 0.724 | +0.191 | 1 | 0.531 → 1.000 | +0.001 | 34 / 7 |

**The ceiling per benchmark.** No distilled student's p@1 exceeds its starting p@n on any benchmark (Table 8a). The four cases whose interval reaches zero are all on AIME 2026: the 4B from both thinking starts, and the 1.7B and 0.6B from the teacher-SFT start, where both scores are close to zero. The starting p@n itself moves by 0.2 or more only for the 4B in thinking mode on AIME 2026 and for the pretrained-only 0.6B on GSM8K-200, which before distillation often gave no answer at all (Table 8b).

**Table 8: The ceiling per benchmark.** (a) Distilled lenient p@1 minus the starting model's lenient p@n; no difference is above zero. (b) Change in lenient p@n from start to end. Paired 95% bootstrap intervals.

| Route | Size | AIME 2026 | AMC23 | GSM8K-200 | Mean |
|---|---|---|---|---|---|
| *(a) distilled p@1 minus starting p@n* | | | | | |
| non-thinking route | 4B | −0.192 [-0.328, -0.075] | −0.314 [-0.421, -0.210] | −0.056 [-0.084, -0.032] | −0.187 [-0.246, -0.133] |
| | 1.7B | −0.054 [-0.125, -0.002] | −0.415 [-0.538, -0.296] | −0.150 [-0.188, -0.114] | −0.206 [-0.255, -0.160] |
| | 0.6B | −0.149 [-0.283, -0.032] | −0.347 [-0.473, -0.222] | −0.056 [-0.106, -0.008] | −0.184 [-0.247, -0.126] |
| external-SFT start | 4B | −0.071 [-0.176, +0.023] | −0.197 [-0.294, -0.112] | −0.084 [-0.114, -0.056] | −0.117 [-0.165, -0.074] |
| | 1.7B | −0.176 [-0.311, -0.063] | −0.412 [-0.550, -0.282] | −0.228 [-0.270, -0.187] | −0.272 [-0.335, -0.212] |
| | 0.6B | −0.163 [-0.296, -0.033] | −0.463 [-0.602, -0.328] | −0.263 [-0.315, -0.214] | −0.296 [-0.365, -0.234] |
| teacher-SFT start | 4B | −0.084 [-0.185, +0.003] | −0.172 [-0.265, -0.090] | −0.042 [-0.074, -0.011] | −0.099 [-0.146, -0.058] |
| | 1.7B | −0.033 [-0.102, +0.019] | −0.178 [-0.289, -0.082] | −0.157 [-0.194, -0.121] | −0.123 [-0.166, -0.083] |
| | 0.6B | −0.031 [-0.094, +0.002] | −0.235 [-0.359, -0.124] | −0.145 [-0.191, -0.100] | −0.137 [-0.186, -0.092] |
| *(b) change in p@n, start to end* | | | | | |
| non-thinking route | 4B | +0.133 [+0.033, +0.267] | +0.025 [+0.000, +0.075] | +0.005 [-0.015, +0.030] | +0.054 [+0.014, +0.102] |
| | 1.7B | +0.167 [+0.033, +0.300] | +0.075 [-0.025, +0.175] | −0.005 [-0.025, +0.015] | +0.079 [+0.024, +0.139] |
| | 0.6B | −0.033 [-0.167, +0.100] | +0.075 [-0.075, +0.225] | +0.200 [+0.145, +0.260] | +0.081 [+0.008, +0.151] |
| external-SFT start | 4B | +0.300 [+0.133, +0.467] | −0.025 [-0.075, +0.000] | −0.015 [-0.035, +0.000] | +0.087 [+0.031, +0.147] |
| | 1.7B | −0.033 [-0.100, +0.000] | −0.050 [-0.175, +0.075] | −0.025 [-0.055, +0.000] | −0.036 [-0.084, +0.008] |
| | 0.6B | −0.100 [-0.200, +0.000] | −0.075 [-0.250, +0.075] | −0.035 [-0.075, +0.005] | −0.070 [-0.137, -0.007] |
| teacher-SFT start | 4B | +0.333 [+0.167, +0.500] | +0.050 [+0.000, +0.125] | +0.020 [+0.005, +0.040] | +0.134 [+0.075, +0.198] |
| | 1.7B | +0.133 [+0.033, +0.267] | +0.125 [+0.000, +0.250] | +0.050 [+0.015, +0.090] | +0.103 [+0.047, +0.165] |
| | 0.6B | +0.033 [+0.000, +0.100] | +0.075 [-0.050, +0.200] | +0.115 [+0.065, +0.170] | +0.074 [+0.025, +0.127] |

**What distillation changes.** Table 9 splits each route's change into the marked-answer rate and the accuracy of the marked answers. Accuracy rises on every route at every size. The marked-answer rate rises on the non-thinking route and from the teacher-SFT start, whose students rarely marked an answer before distillation, and falls from the external-SFT start at 1.7B and 0.6B.

**GPQA-Diamond.** On the non-thinking route every pretrained-only start is below chance, because it often runs to the limit without an answer. Distillation raises all three sizes, but only the distilled 4B ends above chance (Table 10).

**Majority vote.** A vote over the $n$ committed answers (Wang et al., 2023) tells whether a student's samples agree on the right answer. Distillation raises the 4B's vote from 0.656 to 0.792 from the external-SFT start. At 1.7B the thinking routes tie the non-thinking route under the vote (0.511 and 0.548 against 0.519), although they fall below it one sample at a time; at 0.6B the vote does not change with distillation (0.400 to 0.398).

**Table 9: What distillation changes, start to end, on the three-benchmark mean.** Differences with paired 95% bootstrap intervals. *Answer accuracy*: accuracy of the samples that mark an answer, on the problems both models answered.

| Route | Size | marked-answer rate | answer accuracy | lenient p@1 | lenient p@n |
|---|---|---|---|---|---|
| non-thinking route | 4B | +0.115 [+0.088, +0.143] | +0.063 [+0.041, +0.088] | +0.121 [+0.097, +0.149] | +0.054 [+0.014, +0.102] |
| | 1.7B | +0.134 [+0.106, +0.161] | +0.080 [+0.054, +0.111] | +0.146 [+0.118, +0.177] | +0.079 [+0.024, +0.139] |
| | 0.6B | +0.633 [+0.598, +0.667] | +0.092 [+0.065, +0.118] | +0.280 [+0.249, +0.310] | +0.081 [+0.008, +0.151] |
| external-SFT start | 4B | −0.006 [-0.043, +0.033] | +0.190 [+0.142, +0.243] | +0.119 [+0.083, +0.158] | +0.087 [+0.031, +0.147] |
| | 1.7B | −0.145 [-0.180, -0.110] | +0.106 [+0.072, +0.145] | +0.034 [+0.013, +0.057] | −0.036 [-0.084, +0.008] |
| | 0.6B | −0.251 [-0.290, -0.210] | +0.115 [+0.089, +0.142] | +0.056 [+0.035, +0.077] | −0.070 [-0.137, -0.007] |
| teacher-SFT start | 4B | +0.368 [+0.334, +0.403] | +0.104 [+0.063, +0.147] | +0.319 [+0.284, +0.357] | +0.134 [+0.075, +0.198] |
| | 1.7B | +0.314 [+0.281, +0.349] | +0.061 [+0.018, +0.110] | +0.266 [+0.232, +0.301] | +0.103 [+0.047, +0.165] |
| | 0.6B | +0.244 [+0.214, +0.276] | +0.087 [+0.051, +0.124] | +0.207 [+0.180, +0.237] | +0.074 [+0.025, +0.127] |

**Table 10: GPQA-Diamond, non-thinking route only, scored by the last marked answer.** With four options, chance is 0.250. Differences with paired 95% bootstrap intervals.

| Model | committed p@1 start → end | change | end minus chance | strict p@1 end | truncation start → end |
|---|---|---|---|---|---|
| Teacher (non-thinking) | 0.499 | – | +0.249 [+0.200, +0.300] | 0.497 | 0.026 |
| 4B | 0.157 → 0.397 | +0.239 [+0.198, +0.282] | +0.146 [+0.099, +0.195] | 0.361 | 0.462 → 0.125 |
| 1.7B | 0.191 → 0.275 | +0.084 [+0.050, +0.117] | +0.025 [-0.011, +0.062] | 0.270 | 0.181 → 0.111 |
| 0.6B | 0.049 → 0.234 | +0.185 [+0.155, +0.215] | −0.016 [-0.047, +0.017] | 0.233 | 0.581 → 0.154 |

# D SUPPLEMENTARY RESULTS: WHY SMALL STUDENTS DO NOT LEARN TO STOP

**Training dynamics in full.** Figure 11 adds to the closure rate of Figure 6 the median response length relative to the budget, the accuracy of the closed responses and the fraction of responses with a marked answer, over the same steps. The responses fill the budget in the step in which closure collapses, and the marked answers disappear with it.

**The collapse in every thinking-mode run.** Table 11 lists the closure of every thinking-mode run over the first 30 steps. Each run from the external-SFT start drops at step 6 at the 7,168-token budget and at step 7 at 16,384 tokens; the officially post-trained 0.6B drops at step 6 and partly recovers afterwards. The runs from the teacher-SFT start and the 1,024-token run close little from the first step on, because their responses already reach the budget.

**Table 11: Every thinking-mode run loses closure early.** Closure rate of the training responses, averaged over step windows, and the first step at which it falls below three quarters of its steps 1–4 level. Runs from the teacher-SFT start and the 1,024-token run begin with little closure to lose.

| Run | steps 1–5 | 6–10 | 11–20 | 21–30 | first step below 0.75× |
|---|---|---|---|---|---|
| *External-SFT start, 7,168 tokens* | | | | | |
| 4B, run 1 | 0.617 | 0.211 | 0.155 | 0.189 | 6 |
| 4B, run 2 | 0.633 | 0.208 | 0.170 | 0.212 | 6 |
| 1.7B | 0.445 | 0.128 | 0.071 | 0.051 | 6 |
| 0.6B, run 1 | 0.605 | 0.145 | 0.058 | 0.042 | 6 |
| 0.6B, run 2 | 0.583 | 0.113 | 0.055 | 0.052 | 6 |
| *Other rollout budgets* | | | | | |
| 0.6B, 16,384 tokens | 0.608 | 0.297 | – | – | 7 |
| 0.6B, 1,024 tokens | 0.125 | 0.105 | 0.080 | 0.064 | 5 |
| *Teacher-SFT start, 7,168 tokens* | | | | | |
| 4B | 0.225 | 0.231 | 0.145 | 0.148 | 11 |
| 1.7B | 0.078 | 0.033 | 0.039 | 0.023 | 6 |
| 0.6B | 0.117 | 0.070 | 0.032 | 0.055 | 6 |
| *Officially post-trained 0.6B, 16,384 tokens* | | | | | |
| run 1 | 0.736 | 0.372 | 0.410 | – | 6 |
| run 2 | 0.731 | 0.366 | 0.438 | – | 6 |

**Stop reliability and the closure each run keeps.** Table 12 gives, for every thinking-mode run, how often its closed responses were correct over the first four steps and how much of its early

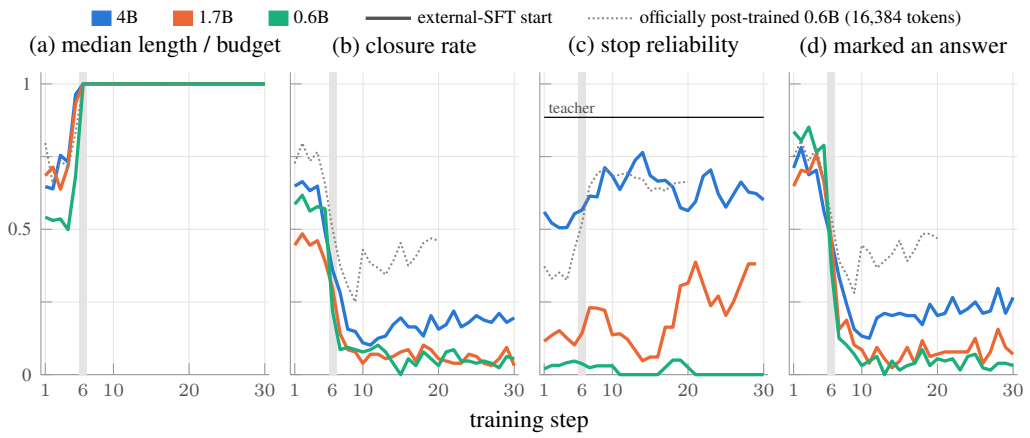



**Figure 11: In thinking-mode training, the responses fill the budget at step 6, closure collapses in the same step, and the students stop marking answers.** Training responses, 7,168-token budget unless noted. (c) Stop reliability, the accuracy of closed responses: 3-step centered moving average, omitted below five closed responses; black line: the thinking teacher, 0.885.

closure it keeps late in training. Among the runs from the external-SFT start, both follow size. The officially post-trained 0.6B, whose stops were right in about a third of cases, keeps more than half of its closure over its 20 steps.

**Table 12: Stop reliability before training, and the closure each run keeps.** *Stop reliability*: the fraction of closed training responses whose answer is correct over steps 1–4; the thinking teacher reaches 0.885 on its own samples. *Late closure*: the closure rate over the last stored steps of the run (steps 11–20 for the 20-step runs). *Kept*: late closure divided by the closure over steps 1–4.

| Run | stop reliability | closure, steps 1–4 | late closure | kept |
|---|---|---|---|---|
| 4B external-SFT start, run 1 | 0.521 | 0.648 | 0.226 | 0.349 |
| 4B external-SFT start, run 2 | 0.509 | 0.652 | 0.226 | 0.347 |
| 1.7B external-SFT start | 0.128 | 0.459 | 0.064 | 0.139 |
| 0.6B external-SFT start, run 1 | 0.042 | 0.611 | 0.023 | 0.038 |
| 0.6B external-SFT start, run 2 | 0.033 | 0.586 | 0.026 | 0.044 |
| 4B teacher-SFT start | 0.259 | 0.219 | 0.211 | 0.963 |
| 1.7B teacher-SFT start | 0.025 | 0.078 | 0.003 | 0.038 |
| 0.6B teacher-SFT start | 0.000 | 0.119 | 0.002 | 0.017 |
| officially post-trained 0.6B, run 1 | 0.349 | 0.756 | 0.410 | 0.542 |
| officially post-trained 0.6B, run 2 | 0.362 | 0.750 | 0.438 | 0.584 |

The 4B also stops finishing its training responses, but at evaluation, where the limit is 30,720 tokens, its longer reasoning still ends in time, which is why its marked-answer rate there does not fall.

### D.1 PROBES

**The teacher's stop signal (Table 13, Figure 7).** For each student from the external-SFT start we took its responses at the first training step, about a hundred per student at the 7,168-token budget, and had the thinking teacher score every token. For responses that closed, we read the teacher's probability of `</think>` at the position where the student closed. For responses that ran to the budget, we read the largest probability the teacher gave `</think>` at any position and report its median over responses.

**The stop token on identical prefixes (Table 14).** We cut prefixes from the officially post-trained 0.6B's own outputs at four kinds of position (where it closed, just after a correct or a wrong marked answer, and at token 500) and asked the model before and after 20 steps of thinking-mode distillation for its probability of `</think>` as the next token. Where the starting model had closed, both models rank `</think>` first. Everywhere else the token lies tens of thousands of ranks deep in both, so the changes there, although their intervals exclude zero, move a probability that is negligible in both models.

**Table 13: The teacher endorses the student's own stops and gives almost no stop signal along a truncated trajectory.** The thinking teacher scores each student's step-1 rollouts (7,168-token budget); $p$ is its probability of `</think>`. Advantage: what the objective gives the stop token. Stop reliability: accuracy of the rollouts that closed, steps 1–4. Details: Appendix D.1.

| | Samples the student stopped | | | Truncated samples | | | Before training |
|---|---|---|---|---|---|---|---|
| Student | $n$ | $p$ at the stop | Stop-token advantage | $n$ | Ever $p > 0.1$ | Median of max $p$ | Stop reliability |
| ● 4B external-SFT start | 83 | 0.976 | +0.027 | 36 | 0.000 | $4.4 \times 10^{-9}$ | 0.521 |
| ● 1.7B external-SFT start | 57 | 0.764 | +0.032 | 40 | 0.175 | $5.8 \times 10^{-8}$ | 0.128 |
| ● 0.6B external-SFT start | 75 | 0.673 | −0.062 | 23 | 0.043 | $1.0 \times 10^{-7}$ | 0.042 |
| *Teacher (thinking), own samples* | | | | | | | 0.885 |

**Table 14: On identical prefixes, distillation leaves the stop token where it was.** The officially post-trained 0.6B before and after 20 steps of thinking-mode distillation reads the same prefixes, cut from the starting model's own outputs, and we compare the log-probability each assigns to `</think>` as the next token. Two disjoint prefix sets. $\Delta$: after minus before, with a 95% bootstrap interval. Rank 0: `</think>` is the most likely next token.

| | | log $p$(`</think>`) | | | | median rank |
|---|---|---|---|---|---|---|
| Position in the prefix | $n$ | before | after | $\Delta$ | 95% interval | before → after |
| *Prefix set 1* | | | | | | |
| where the starting model closed | 205 | −0.000 | −0.001 | −0.001 | [−0.003, −0.000] | 0 → 0 |
| just after a correct marked answer | 71 | −25.865 | −27.235 | −1.370 | [−1.591, −1.157] | 64,401 → 67,323 |
| just after a wrong marked answer | 135 | −23.113 | −24.032 | −0.919 | [−1.089, −0.753] | 48,935 → 38,821 |
| at token 500 | 256 | −31.530 | −36.190 | −4.660 | [−4.953, −4.360] | 79,233 → 80,691 |
| *Prefix set 2* | | | | | | |
| where the starting model closed | 203 | −0.000 | −0.001 | −0.000 | [−0.000, −0.000] | 0 → 0 |
| just after a correct marked answer | 86 | −25.913 | −27.273 | −1.360 | [−1.572, −1.158] | 63,738 → 65,571 |
| just after a wrong marked answer | 118 | −24.355 | −25.397 | −1.042 | [−1.220, −0.869] | 53,955 → 43,624 |
| at token 500 | 256 | −31.874 | −36.346 | −4.472 | [−4.754, −4.197] | 77,146 → 77,038 |

**A lengthening-only null model (Table 15).** If distillation only made every response longer, closes would disappear wherever the longer response no longer fits the limit, whether the answer was right or wrong. Stretching the starting model's responses by one factor, fitted so that closure matches the observed value, removes most of the wrong closes that were lost, but also a quarter of the correct ones, which the distilled model in fact keeps.

**Table 15: Lengthening alone explains most lost wrong stops but not the kept correct ones.** 128 problems, 4 samples each, from the officially post-trained 0.6B before and after 20 thinking-mode steps. The null model stretches every starting sample's closing position by one factor $k$, chosen so that its closure matches the observed one, and keeps the close only if it still falls within the 16,384-token limit. Fractions of all samples.

| | closed | closed and correct | closed and wrong |
|---|---|---|---|
| before training | 0.797 | 0.305 | 0.492 |
| null model, $k = 1.71$ | 0.439 | 0.229 | 0.211 |
| after 20 thinking-mode steps | 0.438 | 0.311 | 0.127 |

# E SUPPLEMENTARY RESULTS: WHAT THE SAMPLES THAT DO NOT STOP CONTAIN

**The answer cases per benchmark.** Table 16 splits every model's samples, benchmark by benchmark, into those with no marked answer, those whose marked answers are all wrong, and those with a correct one. From the external-SFT start the share of only-wrong samples falls on every benchmark and at every size; at 1.7B and 0.6B the share with no marked answer rises on every benchmark. From the teacher-SFT start distillation mostly adds right answers, because those starts rarely marked an answer at all.

**Table 16: The answer cases per benchmark.** Fraction of samples with no marked answer, with only wrong marked answers, and with a correct marked answer (lenient p@1). Starting points in gray.

| | AIME 2026 | | | AMC23 | | | GSM8K-200 | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | none | wrong | right | none | wrong | right | none | wrong | right |
| Teacher (thinking) | 0.092 | 0.234 | 0.674 | 0.014 | 0.026 | 0.960 | 0.007 | 0.056 | 0.937 |
| 4B external-SFT start | 0.537 | 0.315 | 0.148 | 0.242 | 0.210 | 0.549 | 0.020 | 0.081 | 0.899 |
| + distillation | 0.535 | 0.135 | 0.329 | 0.230 | 0.042 | 0.728 | 0.050 | 0.054 | 0.896 |
| 4B teacher-SFT start | 0.819 | 0.065 | 0.116 | 0.565 | 0.023 | 0.412 | 0.516 | 0.023 | 0.461 |
| + distillation | 0.531 | 0.153 | 0.316 | 0.228 | 0.044 | 0.728 | 0.038 | 0.059 | 0.903 |
| 1.7B external-SFT start | 0.640 | 0.331 | 0.029 | 0.391 | 0.310 | 0.300 | 0.089 | 0.164 | 0.747 |
| + distillation | 0.842 | 0.101 | 0.057 | 0.540 | 0.097 | 0.363 | 0.172 | 0.071 | 0.757 |
| 1.7B teacher-SFT start | 0.961 | 0.026 | 0.013 | 0.852 | 0.019 | 0.129 | 0.733 | 0.025 | 0.243 |
| + distillation | 0.861 | 0.071 | 0.067 | 0.554 | 0.074 | 0.372 | 0.189 | 0.068 | 0.743 |
| 0.6B external-SFT start | 0.485 | 0.509 | 0.006 | 0.331 | 0.496 | 0.173 | 0.169 | 0.296 | 0.536 |
| + distillation | 0.851 | 0.145 | 0.003 | 0.644 | 0.144 | 0.212 | 0.242 | 0.092 | 0.667 |
| 0.6B teacher-SFT start | 0.968 | 0.031 | 0.001 | 0.930 | 0.041 | 0.029 | 0.835 | 0.030 | 0.135 |
| + distillation | 0.933 | 0.065 | 0.003 | 0.748 | 0.087 | 0.165 | 0.319 | 0.061 | 0.620 |

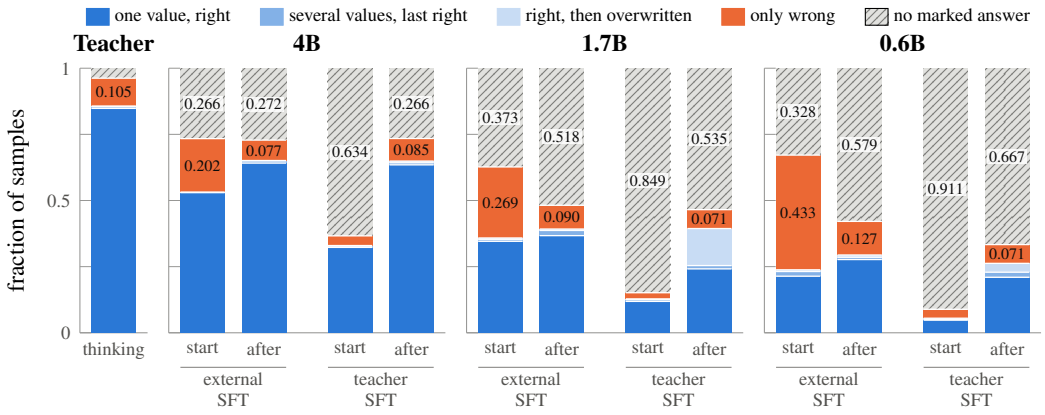



**Figure 12: Below 4B, distillation trades wrong answers for no answer at all.** From the external-SFT start, the share of samples with only wrong answers shrinks at every size; at 4B right answers take its place, at 1.7B and 0.6B mostly samples that mark no answer. From the teacher-SFT start, whose students rarely marked an answer, it mostly adds right answers. Three-benchmark mean; the blue cases sum to lenient $p@1$; after distillation every no-answer sample is truncated. Per benchmark: Table 16.

**The unfinished samples per benchmark.** Table 17 repeats Table 2 benchmark by benchmark for the distilled students from the external-SFT start and for the teacher. For the students the correct value is found above chance on every benchmark, most clearly on GSM8K-200.

**Table 17: The unfinished samples per benchmark.** Samples that reach the evaluation limit. *Correct value present*: the correct value occurs anywhere in the text, on problems whose integer answer has at least three digits and does not occur in the statement (*eligible*); *chance*: the answer of another such problem occurs. *Repeated ending*: some line occurs at least five times in the last 2,000 characters.

| Model | Benchmark | unfinished | marked an answer | eligible | correct value present | chance | repeated ending |
|---|---|---|---|---|---|---|---|
| Teacher (thinking) | AIME 2026 | 154 | 0.143 | 143 | 0.147 | 0.070 | 0.058 |
| | AMC23 | 36 | 0.250 | 1 | 0.000 | 0.000 | 0.111 |
| | GSM8K-200 | 58 | 0.207 | 30 | 0.367 | 0.033 | 0.103 |
| 4B, distilled | AIME 2026 | 798 | 0.034 | 666 | 0.141 | 0.051 | 0.060 |
| | AMC23 | 460 | 0.039 | 61 | 0.197 | 0.000 | 0.041 |
| | GSM8K-200 | 380 | 0.166 | 108 | 0.648 | 0.065 | 0.184 |
| 1.7B, distilled | AIME 2026 | 1,242 | 0.024 | 938 | 0.137 | 0.036 | 0.179 |
| | AMC23 | 1,090 | 0.049 | 127 | 0.386 | 0.039 | 0.149 |
| | GSM8K-200 | 1,215 | 0.093 | 437 | 0.618 | 0.059 | 0.256 |
| 0.6B, distilled | AIME 2026 | 1,273 | 0.037 | 931 | 0.044 | 0.016 | 0.306 |
| | AMC23 | 1,355 | 0.088 | 165 | 0.145 | 0.012 | 0.450 |
| | GSM8K-200 | 1,863 | 0.170 | 626 | 0.557 | 0.075 | 0.129 |

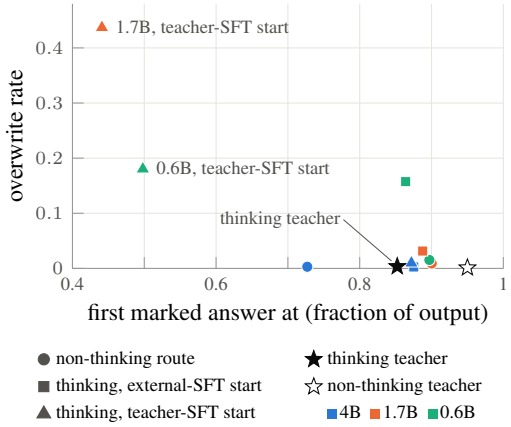



**Figure 13: The two small students distilled from the teacher-SFT start answer early, then overwrite the answer.** One point per evaluated model, three-benchmark mean. Overwrite rate: how often the last marked answer is wrong although a correct one was marked. Every other student answers late; from the external-SFT start, overwriting still rises as size falls.

**Forcing a stop.** We appended `</think>` to every unfinished sample of the officially post-trained 0.6B on 128 training problems and let the model continue (Table 18). Before and after distillation, a forced stop recovers a correct answer mostly on problems the starting model could solve, and rarely on problems it never solved. Many continuations run out of their 1,024 tokens without answering, so these rates are lower bounds.

**Table 18: Forcing a stop recovers answers mainly where the model could already solve.** Every unfinished sample of the officially post-trained 0.6B on 128 problems (4 samples each) is given back to the model that wrote it with `</think>` appended, and the model continues for up to 1,024 tokens. *Recovered*: the continuation's answer is correct. Problems are grouped by how many of the starting model's four samples solved them.

| | all | | solved 0/4 | | solved 1–2/4 | | solved 3–4/4 | |
|---|---|---|---|---|---|---|---|---|
| Model | $n$ | recovered | $n$ | recovered | $n$ | recovered | $n$ | recovered |
| before training | 104 | 0.067 | 84 | 0.036 | 19 | 0.158 | 1 | 1.000 |
| after 20 thinking-mode steps | 288 | 0.101 | 206 | 0.049 | 60 | 0.183 | 22 | 0.364 |

**Where the first marked answer appears.** Figure 13 places every evaluated model by where its first marked answer appears in the output and by how often it overwrites a correct marked answer. The teacher answers late and almost never overwrites (0.85 and 0.004). The 1.7B and 0.6B distilled from the teacher-SFT start answer at about half of the output and overwrite in 0.437 and 0.180 of the samples in which a correct answer appeared. From the external-SFT start the same sizes answer late (0.89 and 0.86), and they overwrite more as they get smaller.

**How repeated endings are counted.** A sample ends in repetition if some non-empty line occurs at least five times in its last 2,000 characters. Samples that mark many answers repeat their answer lines, which is why the students from the teacher-SFT start score high on this measure.

## F Data notes and reproducibility

**Data notes.** 856 of the 13,597 training problems (6.3%) have statements cut short at a blank line by an upstream extraction step; every run in this paper except the two easy-to-hard curriculum runs trains on the full set. GSM8K was not among the benchmarks we decontaminated against. AMC23 is used both for validation during training and for evaluation, but every evaluated checkpoint is taken at a fixed step, so no checkpoint was chosen on it. The two runs of one configuration differ only by being run twice over the same data order.

**Reproducibility package.** The anonymized supplementary archive contains the analysis code, every derived table and probe summary, the per-problem counts from which all evaluation numbers are recomputed, the training and evaluation scripts with the configuration of every run and our patch to verl, the datasets, and the LaTeX source of this paper. `make verify` recomputes every interval, table and figure from these files, checks that they are identical to the ones shipped, checks every number in the text against a fact sheet, and builds the PDF, on a CPU in a few minutes. Model

weights and the raw generations are not included; the students can be retrained and re-evaluated with the included scripts, configurations and data.