MOVE SELECTION UPSTREAM: CHOOSING WHAT THE TEACHER TRAINS ON IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own generated responses under teacher supervision, and existing data-selection methods choose the student's prompts for a given teacher. Yet even a well-trained teacher performs unevenly across prompts, and selecting student prompts leaves it unchanged. When the teacher can also be trained, can selecting its training data produce a better final student under the same training budget? We select teacher-training data that targets the initial student's weaknesses, prioritizing data that is challenging for the initial student but demonstrably solvable. Our theoretical analysis gives sufficient conditions for teacher-data selection to outperform any selection of the student's prompts at the same training budget: the teacher must improve sufficiently and the student must train long enough. Empirically, teacher-data selection improves final-student performance in instruction following and mathematics, with absolute gains of 6.9% and 4.1% over uniformly sampled teacher-training data. It also outperforms the evaluated student-prompt selection baselines under matched training budgets. In replicated multi-teacher experiments, selecting the instruction-following teacher's data yields absolute gains of 2.2% to 5.2% over uniform sampling and student-prompt selection baselines on two public instruction-following benchmarks. Improving the teacher through data selection can thus be more effective than selecting the student's prompts.

1 INTRODUCTION

The scaling of large language models (LLMs) increasingly relies on model-generated supervision. In on-policy distillation (OPD), the student policy learns from teacher supervision on its self-generated responses (Agarwal et al., 2024; Gu et al., 2024). However, generating and learning from these responses is computationally expensive, making data selection important for compute-bounded training. This selection manifests at two stages: (1) curating data to train the teacher and (2) selecting informative prompts to query the student.

Existing data selection methods largely focus on the student's prompts (Figure 1, bottom), seeking data that helps the student learn from a given teacher (Foster et al., 2025; Cheng et al., 2025). Yet even a well-trained teacher may perform unevenly across prompts (Zhang et al., 2026; Lu et al., 2026). In multi-teacher pipelines such as Open-MOPD (Gao et al., 2026), the teachers are themselves trained for the purpose, as domain specialists produced from a shared base and then distilled into one student, so what a teacher is trained on is a choice with a budget of its own. Selecting the teacher's training data (Figure 1, top) offers a different opportunity: directing teacher training toward tasks on which the student needs better supervision. We investigate whether teacher-data selection produces a better final student than student-prompt selection under matched training budgets.

To address this question, we select teacher-training data that is challenging for the initial student but demonstrably solvable. We identify such data using an existing model that already performs the task well, which we call the *reference model*. We rank prompts by the gap between the reference's and the initial student's success rates over sampled responses. The reference can be an already-released model: we need only its sampled responses and a way to judge them, without further training. Using an existing model only to score data is established practice in data selection for pretraining and

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.

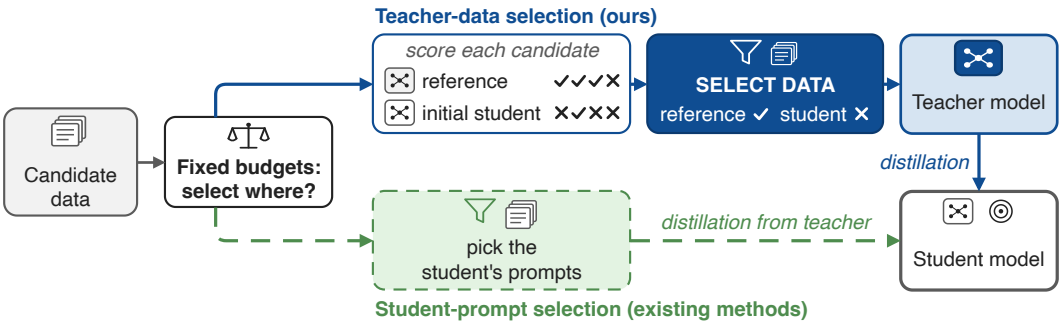


Figure 1: **Teacher-data selection and student-prompt selection.** The top pipeline selects data for training the teacher; the bottom pipeline selects prompts for training the student. Teacher and student training budgets are matched separately across pipelines. Our method prioritizes prompts on which the reference model produces correct responses more often than the initial student.

for reasoning fine-tuning (Lin et al., 2024; Muennighoff et al., 2025); those methods select data for the model being trained, whereas we use the score to choose what a separate teacher is trained on in OPD. In our experiments, we use the publicly released instruction-following specialist from Open-MOPD (Gao et al., 2026). Selection is stratified by response length.

Our theoretical analysis gives sufficient conditions for teacher-data selection to outperform any selection of the student's prompts at the same training budget, provided it sufficiently improves the teacher and the student trains long enough. The reason is that selecting student prompts changes which tasks the student learns from its teacher, whereas selecting teacher-training data can improve how well the teacher performs those tasks. A better teacher is not sufficient on its own: the benefit also depends on how much of it the student absorbs within its budget.

Empirically, we show that teacher-data selection improves student performance in instruction following and mathematics. With a single teacher, five independent data selections yield a 6.9% absolute improvement in instruction following over uniformly sampled teacher data. Separate comparisons show gains of 3.5% to 3.9% over student-prompt selection baselines at matched teacher and student training budgets. In mathematics, the gains are 4.1% over uniform teacher data and 3.8% over student-prompt selection, with improvements in all five training replications. We find no clear difference from direct reference distillation on the main score. Direct distillation scores higher when answers shorter than five words count as incorrect. With multiple teachers, selecting the instruction-following teacher's data improves the student's instruction following while leaving the mathematics and code teachers unchanged. On public benchmarks, the gains over uniform sampling and student-prompt selection baselines are 2.2% to 3.6% on IFEval and 3.5% to 5.2% on IFBench.

We make three contributions: (i) a teacher-data selection pipeline for OPD, which selects teacher-training data by the success-rate gap between a reference model and the initial student; (ii) a sufficient condition for teacher-data selection to outperform student-prompt selection at matched budgets, and a constructed example in which limited student training reverses the comparison; (iii) evidence that selecting the teacher's data improves the final student with its training prompts unchanged, reproduces across independent probe draws, and exceeds student-prompt selection at matched budgets in our comparisons.

2 Related Work

On-policy distillation. Building on knowledge distillation (Hinton et al., 2015), OPD supervises the student's own generations to reduce the mismatch between training data and student outputs (Gu et al., 2024; Agarwal et al., 2024). Multi-teacher methods combine domain specialists (Ma et al., 2026; Gao et al., 2026), adapt domain sampling (Sun et al., 2026), or balance specialization with general capabilities (Chen et al., 2026; Liu et al., 2026). Other work examines teacher–student mismatch and length exploitation (Wang et al., 2026), top-k gradient errors (Zhu et al., 2026), and prompt diversity versus repeated rollouts (Li et al., 2026b). These studies motivate how to use teacher supervision; we ask how selecting the teacher's training data can improve that supervision.

Data selection for distillation and reinforcement learning. Prompt-selection criteria include student learnability (Foster et al., 2025), teacher reliability (Zhang et al., 2026), coverage beyond earlier teachers (Lu et al., 2026), and intermediate difficulty (Cheng et al., 2025). Related approaches modify prefixes, tokens or rollouts (Zhao et al., 2026b; Ziheng et al., 2026; Shen et al., 2026; Li et al., 2026c; Lee et al., 2026; Zheng et al., 2026; Akhondzadeh et al., 2026; Yu et al., 2026). DAPO's dynamic sampling (Yu et al., 2025) instead replaces prompts with uniformly correct or incorrect rollouts, whose within-group reward advantages vanish in GRPO (Shao et al., 2024). Instead, we select the teacher's training data before optimization, targeting the student's subsequent learning.

Data selection beyond OPD. Data-mixture methods use pilot runs (Liu et al., 2025; Zhao et al., 2026a; Liang et al., 2025), teacher–student gaps (Liu et al., 2024), capability constraints (Corrado et al., 2025; Gualdoni et al., 2026), model scale (Li et al., 2026a), or distribution alignment (Xie et al., 2023). Our score relates to hard-example mining (Shrivastava et al., 2016) and reference-based loss differences (Mindermann et al., 2022; Lin et al., 2024), but uses differences in verified success. Pass-rate filters that use a stronger model only to judge the solvability of reasoning data (Muennighoff et al., 2025; Ye et al., 2025) share this access pattern but select data for the learner itself. Selecting training data to benefit another model has precedent (Fan et al., 2018; Ren et al., 2023); we study it in OPD, comparing teacher-data selection with student-prompt selection by their effects on the final student.

3 Framework

This section develops a framework for comparing teacher-data selection with student-prompt selection. Section 3.1 defines the training pipeline, budgets, and OPD objective. Section 3.2 presents our teacher-data selection method. Section 3.3 gives a sufficient condition under which teacher-data selection outperforms student-prompt selection.

3.1 Problem setup

Preliminaries. On-policy distillation (Agarwal et al., 2024; Gu et al., 2024; Yang et al., 2025) trains a student S, with policy π_S, toward a teacher T, with policy π_T, on the student's own samples. The teacher's weights stay frozen while the student trains. The student has a sequence of training prompts, written $\mathcal{D}_{\text{down}}$ because the student is the downstream stage of the pipeline. For a prompt x from that sequence the student draws a response $y \sim \pi_S(\cdot \mid x)$. Each token of that response is scored against the teacher's distribution over the next token given the same partial response. A common choice is to minimize the reverse KL divergence:

$$\min_{\pi_S} \mathbb{E}_{x\sim\mathcal{D}_{\text{down}}} \mathbb{E}_{y\sim\pi_S(\cdot|x)} \sum_t \Big[\log \pi_S(y_t \mid x, y_{<t}) - \log \pi_T(y_t \mid x, y_{<t})\Big].$$

Other distribution-matching objectives can also be used in OPD.

Problem formulation. We seek teacher-training data that maximizes final-student performance under separate teacher and student training budgets, B_T and B_S. The teacher T is trained first and then remains frozen while the student S learns from it through OPD. We call S the *initial student* before distillation and the *final student* afterward.

Both stages generate their own training responses, so selecting data here means selecting prompts. These prompts form the teacher's training sequence $\mathcal{D}_{\text{up}}$ and the student's training sequence $\mathcal{D}_{\text{down}}$; the subscripts indicate the upstream and downstream stages. A selected prompt may occur more than once. A *prompt presentation* is one use of a prompt in a training update. We fix the number of presentations separately at B_T for teacher training and B_S for student distillation, counting repeated uses each time. Matching these budgets does not equalize realized tokens or the cost of computing selection scores; Section 4 reports both.

Formally, let $\mathcal{A} \subseteq \mathcal{D}$ be the selected teacher-training data and $\pi_{\mathcal{A}}$ the final-student policy obtained after teacher training and distillation. With J measuring performance on the target tasks, we seek

$$\max_{\mathcal{A}\subseteq\mathcal{D}} \mathbb{E}[J(\pi_{\mathcal{A}})] \quad \text{subject to} \quad |\mathcal{D}_{\text{up}}(\mathcal{A})| = B_T, \quad |\mathcal{D}_{\text{down}}| = B_S. \tag{1}$$

Here $\mathcal{D}$ is the teacher's candidate pool, and $\mathcal{D}_{\text{up}}(\mathcal{A})$ is the training sequence constructed from $\mathcal{A}$. The training procedures, initial models, and student-prompt sampling rule are fixed; the expectation

is over training randomness. The constraints fix training use, not the number of distinct selected prompts. Section 3.2 gives a score-based approach to this objective.

We ask two questions: **First, does teacher-data selection improve final-student performance over uniform sampling at both stages? Second, at the same training budgets B_T and B_S, does it improve final-student performance more than student-prompt selection?**

3.2 TEACHER-DATA SELECTION

We prioritize prompts that are challenging for the initial student but demonstrably solvable. Directly evaluating Equation 1 for each candidate set would require training a teacher and distilling a student for each set. We instead score individual prompts and select among them while preserving the pool's coarse response-length composition. To assess solvability, we use a *reference model* R: an existing model that performs the task well. We do not train it further. Scoring requires only sampled responses and verifier judgments, not token probabilities or a tokenizer shared with the student.

Separate models also guide pretraining and reasoning-data selection (Mindermann et al., 2022; Lin et al., 2024; Muennighoff et al., 2025; Ye et al., 2025). Here, the reference–student gap selects data for the teacher rather than for the student itself. This separates the reference's ability to solve prompts from the teacher's ability to supervise the student: stronger teachers do not always produce better students (Busbridge et al., 2025), and preparing a teacher for the student can improve transfer over direct distillation from the available model (Mirzadeh et al., 2020; Li et al., 2026b). Section 4 names the reference used.

Data-selection score. Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ be the candidate pool and $r(x, y) \in \{0, 1\}$ indicate whether the task verifier judges response y to prompt x correct; it is the same reward the teacher is trained with. For each $x \in \mathcal{D}$ we draw k responses from the initial student and k from the reference, $y_i^S \sim \pi_0(\cdot \mid x)$ and $y_i^R \sim \pi_R(\cdot \mid x)$ for $i = 1, \ldots, k$, where π_0 is the initial student's policy, before any distillation, and π_R the reference's. We call these responses the *probes*. They give two empirical success rates and, as their gap, the *data-selection score* $u(x)$,

$$p^S(x) = \frac{1}{k}\sum_{i=1}^{k} r(x, y_i^S), \qquad p^R(x) = \frac{1}{k}\sum_{i=1}^{k} r(x, y_i^R),$$
$$u(x) = p^R(x) - p^S(x) \in \{-1, -1 + \tfrac{1}{k}, \ldots, 1\}. \tag{2}$$

A high u means the reference model produces correct responses more often than the initial student. The gap combines difficulty for the student with evidence of solvability: selecting only student failures could retain prompts that the reference also fails, while selecting only reference successes could retain prompts the student already solves. We use this gap as a selection heuristic. Proposition 1 gives a conditional justification; it does not establish end-to-end optimality for the implemented training pipeline.

Length-controlled selection. Ranking by $u(x)$ alone may shift the selected data toward unusually short or long responses. Prior work documents length bias in gradient-based selection (Xia et al., 2024) and uses explicit length controls in preference-data selection (Deshpande et al., 2025). Let $c(x)$ be the mean token length of the initial student's probe responses. Ranking by $u(x)/c(x)$ trades score against estimated cost, but does not preserve length composition; it is distinct from the gradient-norm normalization used by LESS. We instead partition the pool into length groups $\mathcal{Q}_1, \ldots, \mathcal{Q}_B$ by quantiles of c and select the highest-scoring prompts within each group in proportion to its size, up to integer rounding. This preserves coarse length composition, not exact training-token consumption, since response lengths can change during training. Appendix D.3.3 compares this length control with score normalization.

Selection rule. For the fixed-size variant, we use the summed score as a surrogate for the final-student objective in Equation 1. Let m_b be quotas proportional to group sizes up to integer rounding, with $\sum_b m_b = m$, the number of prompts to select. The selected set solves

$$\mathcal{A}^\star \in \arg\max_{\mathcal{A} \subseteq \mathcal{D}} \sum_{x \in \mathcal{A}} u(x) \quad \text{subject to} \quad |\mathcal{A} \cap \mathcal{Q}_b| = m_b,\ b = 1, \ldots, B,$$

which separates across groups: take the m_b highest-scoring prompts in each, breaking ties by a fixed rule. This exactly solves the surrogate problem. The selected set supplies the teacher's B_T

training presentations; its size and composition constraints are separate from the training budgets in Equation 1. Appendix C.2 gives the quotas, tie rule, quantile cut points, and a separate variant that allocates each group's share of probe tokens instead of prompts; Section 4 gives the experimental settings.

When does the score predict student benefit? A larger reference–student gap is useful when the teacher can learn the indicated improvement and the student can inherit it. We model teacher learning up to an attainable success rate for each prompt. If the reference identifies that rate, and initial success, learning dynamics, student exposure and evaluation weight are shared within each group, larger gaps give at least as much student improvement. Appendix B.1.1 allows these quantities to differ. The proposition below gives the general guarantee: selection remains near-optimal when the score closely tracks student gains, probe errors are small, and interactions between selected prompts are limited.

Proposition 1 (Conditional guarantee for fixed-size selection)**.** *Let $\bar{u}_i = \mathbb{E}[u(x_i)]$ and $|u(x_i) - \bar{u}_i| \le e_i$. Suppose that for every quota-feasible set $\mathcal{A}$, its expected final-student performance is $F(\mathcal{A}) = C + \sum_{i\in\mathcal{A}} g_i + \mathrm{rem}(\mathcal{A})$, where g_i is independent of the other selected prompts and $|\,\mathrm{rem}\,| \le \varepsilon$. In each group let a nondecreasing, L_b-Lipschitz function ψ_b satisfy $|g_i - \psi_b(\bar{u}_i)| \le d_b$. For the selected set $\mathcal{A}^\star$ and any quota-feasible $\mathcal{A}$, let $h_b = |(\mathcal{A} \setminus \mathcal{A}^\star) \cap \mathcal{Q}_b|$. Then*

$$F(\mathcal{A}) - F(\mathcal{A}^\star) \le \sum_b \left[2h_b d_b + L_b \sum_{i\in(\mathcal{A}\triangle\mathcal{A}^\star)\cap\mathcal{Q}_b} e_i \right] + 2\varepsilon. \tag{3}$$

In particular, exact probes, exact gain alignment and zero interactions imply optimality within this fixed-size model. (Proof in Appendix B.1.)

Appendix B.1.1 relates these errors to differences in learning dynamics, reference calibration and probe noise; grouping by length alone does not control them. Appendix B.3 treats the probe-token-budget variant separately.

3.3 Theoretical analysis

Improving the teacher can help more than changing the student's prompts once the student has enough training to absorb the improvement. We establish this result in a model with one success probability per prompt, comparing a baseline teacher T_0 with a selected-data teacher $T_\dagger$ trained under the same teacher budget. The argument first bounds what student-prompt selection can achieve with T_0, then asks when training with $T_\dagger$ exceeds that bound at the same student budget.

Definition 1 (Proximal distillation)**.** *For each prompt $j = 1, \dots, P$, we represent student and teacher success by Bernoulli distributions with probabilities p_j and t_j in $(0, 1)$. Performance is $J(\pi) = P^{-1}\sum_j p_j$ and $J(T) = P^{-1}\sum_j t_j$. A presentation updates only that prompt: $\pi^+ = \arg\min_\pi\{\eta\,\mathrm{KL}(\pi\|T) + (1-\eta)\,\mathrm{KL}(\pi\|\pi^-)\}$, where $\eta \in (0, 1)$ and all distributions in this update refer to prompt j.*

An allocation $n \in \mathbb{Z}^P_{\ge 0}$ assigns n_j presentations to prompt j, with $\sum_j n_j = B_S$. Write π_n^T for the resulting student policy, starting from a common π_0. A balanced allocation gives each prompt $\lfloor B_S/P \rfloor$ presentations and distributes the remainder one each; it is a deterministic analogue of uniform sampling. The update has the exact solution

$$\mathrm{logit}\, p_j(n_j) = (1-\eta)^{n_j}\,\mathrm{logit}\, p_j(0) + \left[1 - (1-\eta)^{n_j}\right]\mathrm{logit}\, t_j, \tag{4}$$

where $\mathrm{logit}\, p = \log[p/(1-p)]$. Each student probability stays between its initial value and the teacher's value. Prompt selection can therefore preserve strengths the student already has, but cannot raise success above both values. Averaging this observation gives the following upper bound.

Proposition 2 (Upper bound for student-prompt selection)**.** *Let $\Omega_+(T) = P^{-1}\sum_j (p_j(0) - t_j)_+$, where $(z)_+ = \max(z, 0)$. Every allocation satisfies $J(\pi_n^T) \le J(T) + \Omega_+(T)$. If $\Omega_+(T) = 0$, every allocation also satisfies $J(\pi_n^T) \ge J(\pi_0)$.* (Proof in Appendix A.2.)

The term $\Omega_+(T)$ accounts for the initial student's strengths, so the teacher's mean performance alone is not the student's upper bound. To guarantee an advantage over every allocation under T_0, it suffices for the student trained with $T_\dagger$ to exceed this bound. The next theorem accounts for incomplete transfer from $T_\dagger$ and gives a sufficient teacher improvement.

Theorem 1 (Matched-budget comparison)**.** *Write* $\Delta_T = J(T_\dagger) - J(T_0)$ *and* $\Lambda = \max_j |\operatorname{logit} p_j(0) - \operatorname{logit} t_j^\dagger|$. *For the balanced student* $\pi^{T_\dagger}$ *trained with* B_S *presentations,*

$$J(\pi^{T_\dagger}) - \max_{\substack{n \in \mathbb{Z}_{\geq 0}^P \\ \sum_j n_j = B_S}} J(\pi_n^{T_0}) \;\geq\; \Delta_T - \tfrac{1}{4}(1-\eta)^{\lfloor B_S/P \rfloor}\Lambda - \Omega_+(T_0). \tag{5}$$

(Proof in Appendix A.3.)

The theorem gives a sufficient condition: teacher improvement must exceed the *transfer cost* of incomplete student training and the *initial student advantage* over the baseline teacher. The transfer cost decreases geometrically with presentations per prompt. For fixed P, $\Delta_T > \Omega_+(T_0)$ therefore guarantees a better student with enough training, even against the best student-prompt allocation under T_0.

The remaining question is when the selection score produces the teacher improvement this condition requires. Under the learning assumptions in Section 3.2, Appendix B.1.2 derives that improvement from the score, allowing differences in learning dynamics and reference error. A sufficiently large resulting teacher advantage yields a better student than uniform teacher-data sampling on average, even when each baseline student receives its best prompt allocation.

When student training is limited, selecting useful prompts can help more than improving the teacher, because the student may not have enough training to benefit from the better supervision. Appendix B.6 illustrates this case. Appendices B.4 through B.7 examine finite training budgets, learning that transfers across prompts, and evaluation on a separate distribution. Appendix B.9 clarifies how the theoretical assumptions relate to the experimental measurements.

4 Experiments

We evaluate teacher-data selection in single-teacher and multi-teacher distillation, then examine the selection score and transfer to the student.

4.1 Experimental setup

Models and training. Our single-teacher experiments use SmolLM3-3B (Bakouch et al., 2025) teachers and students initialized from the same supervised checkpoint released by Open-MOPD (Gao et al., 2026). Its instruction-following and mathematics specialists provide the reference models. We train teachers with GRPO (Shao et al., 2024) and students with OPD. The main instruction-following comparisons fix teacher training at $5{,}120$ rollouts (640 prompt presentations with eight responses each) and student training at $2{,}560$ prompt presentations. These budgets match training use at each stage, not total compute including selection.

Baselines and controls. Uniform sampling at both training stages is the baseline. We compare selecting teacher data with selecting student prompts, keeping the other stage uniformly sampled. Student-prompt baselines include a static learnability filter (Foster et al., 2025; Yang et al., 2025) and our own gap score. We also test published selection criteria adapted to teacher data and additional student-prompt methods in multi-teacher distillation. Direct distillation from the reference model tests whether training a teacher is needed; it uses the same student-training budget without new teacher training.

Data and evaluation. Training prompts come from Open-MOPD. We evaluate instruction following on three separate 750-prompt sets: a development set, the original evaluation set, and a later holdout unused by this study before its first evaluation. We also evaluate IFEval (Zhou et al., 2023) and IFBench (Pyatkin et al., 2025). Mathematics uses $2{,}000$ reserved prompts; AIME tests mathematical performance of the instruction-following students. Multi-teacher experiments combine instruction-following, mathematics and code specialists. The in-domain instruction-following score, IF Primary, requires a response to satisfy all prompt constraints and pass the format checks. Raw-5 measures strict correctness without those format checks and assigns zero to responses shorter than five words. We identify separately the five-word checks that retain the format guard. An unresolved comparison does not meet its stated significance threshold; this does not imply equivalence. Scores are percentages; differences are absolute unless labeled relative. Bold identifies the highest score

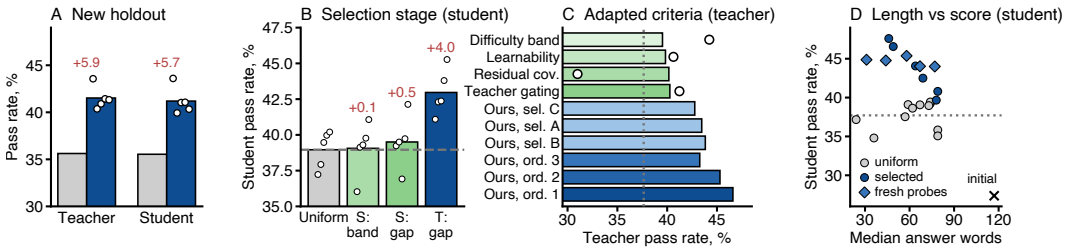


Figure 2: **Single-teacher comparisons.** A: five fresh-probe selections and eleven uniform draws on the new holdout; dots show selected-data runs and grey bars show uniform means. B: student-prompt (S) and teacher-data (T) selection on the original holdout; dots show teacher-level student scores. C: adapted teacher-data criteria; bars and open circles use different tie-breaking keys. D: student scores and response lengths; diamonds mark fresh-probe selections. Reference lines show uniform means.

within each comparison, not statistical significance. Appendix C documents study designs, baseline implementations, compute accounting and pre-registration history; Appendix D reports full results, uncertainty estimates and robustness checks.

4.2 Single-teacher results

Instruction-following gains reproduce across independent selections. Five independent rounds of probing, data selection and training produce students that score 44.61%, compared with 37.71% for students of uniformly trained teachers (Table 1). The gain is 6.90% (two-sided $p < .001$). It reproduces the original 5.82% gain even though the new selections retain only 26% to 28% of the original selected prompts. Thus the gain reproduces when candidate responses are regenerated and the selected corpus changes, not only when one corpus is retrained. The public-benchmark differences are $+1.31\%$ on IFEval, $+2.80\%$ on IFBench and -2.39% on AIME; none is statistically resolved.

Table 1: **Single-teacher instruction following.** Scores are final-student percentages; IF Primary is the in-domain holdout score. Student replicates are averaged within each trained teacher. The original selection has three training orders of one corpus; fresh probes produce five independent selections. Both comparisons use the same eleven uniform controls.

Teacher data	IF Primary	IFEval	IFBench	AIME
Uniform	37.71	72.68	22.78	**10.89**
Selected, original	43.54	73.96	23.24	9.72
Selected, fresh probes	**44.61**	**73.99**	**25.58**	8.50

The gain also holds on 750 previously unused prompts: the same five students score 41.2% against 35.5% for the eleven uniform controls, a 5.7% gain. The gain is 5.4% on Raw-5, but is unresolved when a correct answer must reach 25% or 50% of the initial student's response length on that prompt, as on the earlier holdouts. Median responses average 63 words versus 71 for uniform controls. These prompts were unused by this study, but occur in the reference's reported training corpus. This evaluates existing students on new prompts, not new training replications.

Teacher-data selection exceeds the student-prompt baselines. At matched teacher and student training budgets, teacher-data selection exceeds the static learnability filter by 3.9% and the same gap score applied to student prompts by 3.5% on the original holdout (Table 2). Applying the same score at either stage shows that the difference cannot be attributed to the scoring formula alone. These gains pass the registered Welch tests but not the predictive tests, and become inconclusive under the five-word floor.

On the new holdout, fixed-size selection improves over student-prompt gap selection by 5.7% and the static filter by 6.3% (both Holm-adjusted $p < .001$), following development gains of 6.7% to 6.8%. With a five-word minimum on the guarded score, the gains remain 5.5% and 6.1%, but are unresolved under relative-length thresholds. The follow-up was pre-registered before the student-prompt selectors were evaluated; the selected-teacher and uniform scores were already known.

The probe-token-budget variant also improves the primary score: it reaches 40.1% on the new holdout, versus 36.6% for student-prompt gap selection and 36.9% for the static filter. The gains of 3.5% and

Table 2: **Selection-stage comparisons.** Final-student IF Primary (%). Token-quota selection uses five teachers per condition. Fixed-size selection combines three orders of one corpus and five fresh selections, against five teachers per student-prompt baseline. The new-holdout uniform mean uses five controls, not the eleven in Table 1. Static filter: static learnability filter. $*$ denotes results not included in that column's comparison; token-quota follow-ups use separate controls.

Teacher data	Student prompts	Original holdout	Development	New holdout
Uniform	Uniform	39.0	37.7	35.2
Uniform	Static filter	39.1	37.3	34.8
Uniform	Gap score	39.5	37.4	35.4
Selected, token quota	Uniform	**43.0**	$*$	$*$
Selected, fixed size	Uniform	$*$	**44.1**	**41.1**

3.2% pass the paired tests across five blocks, but both become unresolved on Raw-5 (1.3% each). This length check uses raw correctness, whereas the fixed-size comparison above retains the format guard.

The gain extends to mathematics. Students of selected-data teachers improve accuracy by 4.10% over uniform teacher data and 3.83% over student-prompt selection (Table 3). Both differences are positive in all five paired training blocks and pass Holm correction ($p < .01$). These blocks repeat training for one selected corpus, rather than repeat data selection.

Table 3: **Single-teacher mathematics.** Final-student accuracy (%), averaged over five paired training blocks.

Teacher data	Student prompts	Math accuracy (%)
Uniform	Uniform	27.08
Uniform	Static filter	27.35
Selected (ours)	Uniform	**31.18**

4.3 Multi-teacher distillation

The instruction-following benefit extends to multi-teacher distillation. Teacher-data selection gains 2.2% to 3.6% on IFEval and 3.5% to 5.2% on IFBench over uniform sampling and student-prompt selection with DSIR (Xie et al., 2023), adapted IFD (Li et al., 2024) and adapted LESS (Xia et al., 2024) (Figure 3). Only the instruction-following teacher's data is selected; the mathematics and code teachers are unchanged. Student initialization, training budgets and domain loss weights are shared across configurations. Each has three training runs with its selected data held fixed.

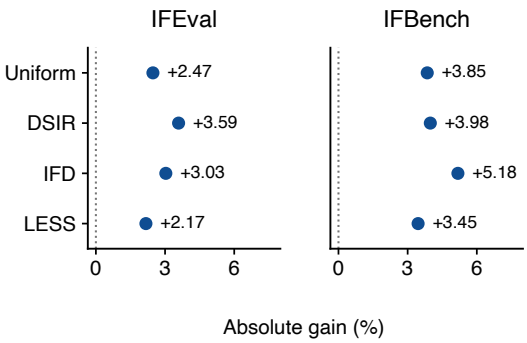


Figure 3: **Multi-teacher instruction following.** Each point shows the mean student-score difference between teacher-data selection and the named baseline, in absolute %, over three training runs per configuration. IFD and LESS are adapted implementations.

The selected-data configuration reaches 75.3% on IFEval and 23.2% on IFBench, compared with 72.9% and 19.4% under uniform sampling. Improving one specialist therefore benefits the student's corresponding capability even while it learns from several teachers.

All eight comparisons on these secondary public benchmarks have unadjusted two-sided $p < .05$. On the pre-declared panel, evaluated with different decoding settings, two of eight comparisons meet that threshold. Earlier replicated comparisons also show instruction-following gains, while their primary mathematics comparisons remain unresolved and AIME does not improve. Appendices D.5.2 and D.5.4 report the complete comparisons.

4.4 Analysis of data selection and transfer

Students retain the teacher's improvement. On the new holdout, the five fresh-probe teachers score 41.5% against 35.6% for uniform-data teachers. Their 5.9% gain becomes a 5.7% student gain, a descriptive transfer ratio of 0.96 (Figure 2A). Both gains persist with a five-word minimum, but are unresolved under relative-length thresholds.

The original comparison shows a similar pattern. Uniform training improves the teacher by 10.2% over its initialization; selecting its data adds another 5.7% at the same training budget. Selection therefore contributes more than half the improvement obtained from uniform teacher training. The final student retains this additional gain, rather than the benefit stopping at the teacher.

Teacher-data selection also helps when student prompts are selected. Selecting both stages gains 5.9% over student-prompt selection alone on development data and passes Holm correction. Adding student-prompt selection to an already selected teacher gives an unresolved 0.55% gain.

Selected prompts retain a large success gap on fresh responses. The reference outperforms the initial student by 47.6% on selected prompts, compared with 14.3% on comparison prompts. This supports using the score to identify data that the reference can solve more often than the student. Sampling above the score cutoff also improves final-student performance by 5.0% over uniform data, with no resolved difference from full ranking.

Subtracting student success avoids selecting prompts the student already solves. In a descriptive ablation with one teacher and student per rule, gap-based selection yields 45.5% student accuracy, versus 40.3% using reference success alone. The initial student succeeds on 65.9% of probes in the reference-only selection, compared with 8.8% under gap selection. The student term therefore directs selection toward tasks where the student has more to learn.

In the small development comparison of selection criteria, our gap score gives the highest observed student mean: 45.45%, compared with 42.10% for the strongest adapted criterion (Table 4).

Table 4: **Teacher-data selection criteria, earlier development study.** Mean IF Primary (%), using shared probes. Uniform: eleven teachers; adapted criteria: two tie-breaking orders each; ours: three training orders of one corpus. These differ from Table 2's later fixed-size runs. Appendix D.3 gives individual results.

Teacher-data rule	Teacher	Student
Uniform sampling	37.64	37.75
Learnability (Foster et al., 2025), adapted	40.23	40.52
Difficulty band (Cheng et al., 2025), adapted	41.88	41.82
Teacher gating (Zhang et al., 2026), adapted	40.75	42.10
Residual coverage (Lu et al., 2026), adapted	35.59	34.04
Reference-student gap (ours)	**45.06**	**45.45**

Teacher choice and training budget. Against direct reference distillation, the student differences are 0.50% across five paired blocks and 0.29% on the new holdout. Neither is statistically resolved, and both favor direct reference distillation under the five-word floor.

More training can also outweigh selection at the original budget. Uniform teacher data trained for 120 steps yields a student score of 51.1%, above 43.5% for selected data at 40 steps. At 120 steps, expanding the selected corpus gives 51.5%, whereas repeating the original corpus gives 48.1%; the expanded-corpus difference from uniform is unresolved. Probing costs about sixteen times a uniform teacher's training tokens, so matched-training gains do not establish a total-compute advantage. Appendices D.1.4 and D.1.13 detail these alternatives.

5 Conclusion

Teacher-training data is a useful selection target for OPD. Our analysis gives conditions under which improving the teacher outperforms reallocating student prompts. Experiments show final-student gains in instruction following and mathematics at matched training budgets, with independent-probe replication in instruction following. Multi-teacher experiments also improve the selected teacher's domain. These findings support considering the teacher's data alongside the student's prompts.

REPRODUCIBILITY STATEMENT

The appendix documents the models, data, training and evaluation settings, selection rules, and baseline implementations. The supplementary material provides training configurations, selection manifests, pre-registrations, and analysis scripts to support reproduction of the reported results.

STATEMENT ON THE USE OF AI

We used large language models to polish writing, identify potentially missing related work, and assist with drafting and reviewing manuscript passages. Under the authors' direction, AI agents also assisted with implementation, experiment execution and analysis, and developing and checking mathematical arguments. The authors are responsible for verifying the citations, proofs and reported results, and for the final manuscript.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation, 2026. URL https://arxiv.org/abs/2607.04037.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5977–6045. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/busbridge25a.html.

Tianlei Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. Counteraction-aware multi-teacher on-policy distillation for general capability recovery with domain preservation, 2026. URL https://arxiv.org/abs/2605.27115.

Jorge (Zhoujun) Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Taylor W. Killian, Haonan Li, Mikhail Yurochkin, Eric Xing, Zhengzhong Liu, and Zhiting Hu. Revisiting reinforcement learning for LLM reasoning from a cross-domain perspective. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4479. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c2f71567cd53464161cab3336e8fc865-Paper-Datasets_and_Benchmarks_Track.pdf.

Nicholas E. Corrado, Julian Katz-Samuels, Adithya M Devraj, Hyokun Yun, Chao Zhang, Yi Xu, Yi Pan, Bing Yin, and Trishul Chilimbi. AutoMixAlign: Adaptive data mixing for multi-task preference optimization in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 20234–20258, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.990. URL https://aclanthology.org/2025.acl-long.990/.

Vijeta Deshpande, Debasmita Ghose, John D Patterson, Roger E. Beaty, and Anna Rumshisky. Diverse, not short: A length-controlled data selection strategy for improving response diversity of language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33917–33938, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1721. URL https://aclanthology.org/2025.emnlp-main.1721/.

Yang Fan, Fei Tian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. Learning to teach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJewuJWCZ.

Thomas Foster, Anya Sims, Johannes Forkel, and Jakob Foerster. LILO: Learning to reason at the frontier of learnability. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48941–48974. Curran Associates, Inc., 2025. doi: 10.52202/085713-1634. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/45d5823950cab9a88e10c6e8dbff351c-Paper-Conference.pdf.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-MOPD: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation, 2026. URL https://arxiv.org/abs/2608.19098.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Eleonora Gualdoni, Sonia Laguna, Louis Bethune, Joao Monteiro, Pierre Ablin, and Marco Cuturi. DynaMiCS: Fine-tuning LLMs with performance constraints using dynamic mixtures, 2026. URL https://arxiv.org/abs/2605.10770.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-aware on-policy distillation via entropy-guided supervision, 2026. URL https://arxiv.org/abs/2606.28562.

Jingwei Li, Xinran Gu, and Jingzhao Zhang. Capacity-aware mixture law enables efficient LLM data optimization, 2026a. URL https://arxiv.org/abs/2603.08022.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.421. URL https://aclanthology.org/2024.naacl-long.421/.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026b. URL https://arxiv.org/abs/2604.13016.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation, 2026c. URL https://arxiv.org/abs/2606.02684.

Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. MoDoMoDo: Multi-domain data mixtures for multimodal LLM reinforcement learning, 2025. URL https://arxiv.org/abs/2505.24871.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 29029–29063. Curran Associates, Inc., 2024. doi: 10.52202/079017-0914. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3322a9a72a1707de14badd5e552ff466-Paper-Conference.pdf.

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. DDK: Distilling domain knowledge for efficient large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98297–98319. Curran Associates, Inc., 2024. doi: 10.52202/079017-3119. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b206d54ffbb803b5c51d85f405d422e4-Paper-Conference.pdf.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38305–38339, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5f67d864aae6115374fed7beddd119e0-Paper-Conference.pdf.

Ziyuan Liu, Jiao Ou, Jian Liang, Ruiming Tang, and Cheng Luo. Preserving general capabilities during domain specialization with uncertainty-calibrated MOPD, 2026. URL https://arxiv.org/abs/2608.26735.

Songshuo Lu, Zhi Chen, and Yaohua Tang. Beyond the best teacher: Expanding and compressing the reasoning solution manifold, 2026. URL https://arxiv.org/abs/2607.27770.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training, 2026. URL https://arxiv.org/abs/2606.30406.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15630–15649. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i04.5963. URL http://dx.doi.org/10.1609/aaai.v34i04.5963.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. URL https://arxiv.org/abs/2507.02833.

Yuxin Ren, Zihan Zhong, Xingjian Shi, Yi Zhu, Chun Yuan, and Mu Li. Tailoring instructions to student's learning levels boosts knowledge distillation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1990–2006, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.111. URL https://aclanthology.org/2023.acl-long.111/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zhanming Shen, Jiaqi Hu, Zeyu Qin, Hao Chen, Wentao Ye, Zenan Huang, yihong zhuang, Guoshan Lu, Junlin Zhou, and Junbo Zhao. Training-trajectory-aware token selection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=G6PSKuGEEe.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Zechen Sun, Zhiwei Zhang, Fei Zhao, Juntao Li, Mu Chuan, Huayu Deng, Guojian Zhan, Wenliang Chen, Yao Hu, and Min Zhang. D^3-MOPD: Dynamic domain ScheDuling for efficient multi-teacher distillation, 2026. URL https://arxiv.org/abs/2608.24987.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations, 2026. URL https://arxiv.org/abs/2607.13399.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34201–34227. Curran Associates, Inc., 2023. doi: 10.52202/075280-1482. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6b9aa8f418bde2840d5f4ab7a02f663b-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2505.09388.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=T2TZ0RY4Zk.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026. URL https://arxiv.org/abs/2605.12652.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation, 2026. URL https://arxiv.org/abs/2609.02998.

Kaiyan Zhao, Zhongtao Miao, Akiko Aizawa, and Yoshimasa Tsuruoka. RegMix-D: Dynamic data mixing via proxy training trajectories, 2026a. URL https://arxiv.org/abs/2606.18663.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-guided on-policy distillation: Mining golden trajectories from rollouts, 2026b. URL https://arxiv.org/abs/2606.21994.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting, 2026. URL https://arxiv.org/abs/2604.10688.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes, 2026. URL https://arxiv.org/abs/2605.11182.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation, 2026. URL https://arxiv.org/abs/2605.27028.

A PROOFS OF THE MAIN RESULTS

The proofs follow the argument in Section 3.3. The update first gives an exact student trajectory. That trajectory bounds what student-prompt selection can achieve; comparing this upper bound with a lower bound for the improved teacher's student gives the matched-budget result. Extensions, constructed examples and the scope of the results are in Appendix B.

A.1 THE LOG-ODDS IDENTITY

Proof of the log-odds identity, Equation 4. The minimiser of $\eta\,\mathrm{KL}(\pi\|T) + (1-\eta)\,\mathrm{KL}(\pi\|\pi^-)$ over distributions is $\pi^+ \propto (\pi^-)^{1-\eta}T^\eta$, by writing the objective as $\mathrm{KL}(\pi\|(\pi^-)^{1-\eta}T^\eta/Z) - \log Z$, where Z is the normalizing constant. For two outcomes the normaliser cancels in the ratio, so $\operatorname{logit}\pi^+ = (1-\eta)\operatorname{logit}\pi^- + \eta\operatorname{logit}T$. Iterating n_j times on prompt j gives Equation 4.

A.2 UPPER BOUND FOR STUDENT-PROMPT SELECTION

The identity keeps each student probability between its initial value and the teacher's value. Taking the larger endpoint on each prompt therefore bounds every allocation.

Proof of Proposition 2. The identity writes $\operatorname{logit}p_j(n_j)$ as a convex combination of $\operatorname{logit}p_j(0)$ and $\operatorname{logit}t_j$ with weights $(1-\eta)^{n_j}$ and $1-(1-\eta)^{n_j}$, both in $[0,1]$ for $\eta \in (0,1)$ and $n_j \geq 0$. A convex combination lies between its endpoints, so $p_j(n_j)$ lies between $p_j(0)$ and t_j; in particular $p_j(n_j) \leq \max(p_j(0), t_j) = t_j + (p_j(0) - t_j)_+$. Averaging over prompts gives $J(\pi_n) \leq J(T) + \Omega_+$ for every n, so the supremum obeys the same bound, and when $\Omega_+ = 0$ the same two-sided statement gives $p_j(n_j) \geq p_j(0)$. □

A.3 MATCHED-BUDGET COMPARISON

The preceding bound applies even to the best baseline allocation. It remains to bound how far the improved teacher's student falls short of its teacher, then subtract the two bounds.

Proof of Theorem 1. Proposition 2 applied to T_0 gives $\sup_n J(\pi_n^{T_0}) \leq J(T_0) + \Omega_+(T_0)$. For the balanced allocation each prompt receives at least $\lfloor B_S/P \rfloor$ presentations, so by Equation 4 $|\operatorname{logit}p_j - \operatorname{logit}t_j^\dagger| \leq (1-\eta)^{\lfloor B_S/P \rfloor}\Lambda$ for every j. The logistic map is $\frac{1}{4}$-Lipschitz, hence $|p_j - t_j^\dagger| \leq \frac{1}{4}(1-\eta)^{\lfloor B_S/P \rfloor}\Lambda$ and the same bound holds for the average, giving $J(\pi^{T_\dagger}) \geq J(T_\dagger) - \frac{1}{4}(1-\eta)^{\lfloor B_S/P \rfloor}\Lambda$. Subtracting the two displays and using $\Delta_T = J(T_\dagger) - J(T_0)$ gives Equation 5. □

Aggregate scores do not determine the two per-prompt terms in the bound. Convexity gives $\Omega_+(T_0) \geq \big(J(\pi_0) - J(T_0)\big)_+$, so an aggregate initial-student lead is a lower bound on the initial student advantage. It does not preclude certification: the selected teacher's improvement can still exceed this advantage and the transfer cost.

B ADDITIONAL THEORY

This appendix connects the data-selection score to student improvement under explicit learning assumptions, allows differences between prompts, and gives a sufficient selection margin for the matched-budget comparison. It also bounds the score-optimization error of the probe-token-budget variant. Further results cover exact finite-budget comparisons, student budgets smaller than the prompt pool, constructed examples, and models with shared student parameters or a separate evaluation distribution. The binary placement proofs are in Appendix A; the additional results are proved below.

B.1 FROM CORRECTNESS GAPS TO STUDENT GAINS

Ranking by the correctness gap helps when larger gaps lead to larger student gains. We first bound the loss from imperfect agreement between these rankings, then give learning dynamics under which their agreement follows rather than being assumed. The measured score $u_i = u(x_i)$ is the difference of the empirical probe success rates in Equation 2; $\bar{u}_i = \mathbb{E}[u_i]$ is the corresponding population gap. The following statements condition on the chosen groups and compare sets with the same feasible quotas, $0 \leq m_b \leq |\mathcal{Q}_b|$ and $\sum_b m_b = m > 0$. They do not assume that length groups are homogeneous in learnability or student exposure.

Proof of Proposition 1. In each group, pair an item $i \in \mathcal{A} \setminus \mathcal{A}^\star$ with an item $j \in \mathcal{A}^\star \setminus \mathcal{A}$. Equal quotas permit this pairing. Top-score selection gives $u_j \geq u_i$, including ties, hence $\bar{u}_i - \bar{u}_j \leq e_i + e_j$. By monotonicity and Lipschitz continuity,

$$g_i - g_j \leq \psi_b(\bar{u}_i) - \psi_b(\bar{u}_j) + 2d_b \leq L_b(\bar{u}_i - \bar{u}_j)_+ + 2d_b \leq L_b(e_i + e_j) + 2d_b.$$

Common coefficients cancel. Summing the exchanged pairs and adding $\mathrm{rem}(\mathcal{A}) - \mathrm{rem}(\mathcal{A}^\star) \le 2\varepsilon$ proves the result, including when $\mathcal{A}$ maximizes F. Exactness also holds for a selection-independent remainder, which cancels. A bounded but nonconstant remainder gives near-optimality, not exact optimality. □

For finite nonempty groups, the smallest monotone approximation error without a prescribed Lipschitz constant is

$$d_b^{\mathrm{iso}} = \tfrac{1}{2} \max_{i,j \in \mathcal{Q}_b:\, \bar{u}_i \le \bar{u}_j} (g_i - g_j)_+.$$

Every monotone fit requires $g_i - g_j \le 2d$ for an ordered pair. Conversely, at an observed population-score level v, the midpoint of $\max_{\bar{u}_i \le v} g_i$ and $\min_{\bar{u}_i \ge v} g_i$ is nondecreasing and within d_b^{iso} of each gain. Linear interpolation between distinct levels supplies a finite Lipschitz constant, not necessarily a small one. Population ties with different gains contribute to this error. Ties in the measured scores can instead reflect probe error; they do not imply equal population scores. The bound is deterministic on the event $|u_i - \bar{u}_i| \le e_i$ and supplies no confidence level or measured error constants by itself.

A sufficient learning-and-transfer model. The connection follows when a higher attainable teacher success rate leads to at least as much teacher improvement, which the student can then inherit. The model below isolates this mechanism by holding the other learning conditions fixed within each group. Within a group, let the teacher and initial student have common correctness $p_0 \in (0,1)$. Prompt i has an attainable teacher level $\tau_i \in [p_0, \tau_{\max}]$, with $\tau_{\max} < 1$. Consider the deterministic teacher recursion

$$T_0(\tau) = p_0, \qquad T_{n+1}(\tau) = \min\{\tau, f(T_n(\tau))\},$$

where the common map $f : [p_0, 1) \to [p_0, 1)$ is nondecreasing and $f(t) \ge t$. Induction gives $T_n(\tau) = \min\{\tau, f^{\circ n}(p_0)\}$: the teacher follows its uncapped trajectory until it reaches the cap, then remains there. Thus teacher correctness is nondecreasing and 1-Lipschitz in the attainable level. After a common n_S student presentations, Equation 4 gives

$$S(T) = \sigma(a \operatorname{logit} p_0 + (1-a) \operatorname{logit} T), \qquad a = (1-\eta)^{n_S}, \quad \sigma(z) = (1+e^{-z})^{-1}.$$

For a common evaluation weight $w \ge 0$, the gain $w[S(T_n(\tau_i)) - p_0]$ is nondecreasing in τ_i. Common mixtures of teacher and student exposure preserve this ordering. If the reference's population success equals τ_i, then $\bar{u}_i = \tau_i - p_0$, so the gain is a nondecreasing function of the population gap. This derives gap–gain alignment from the specified dynamics; it does not assume that a larger observed score directly causes a larger gain. It also need not give a strict advantage: before any cap binds, different attainable levels can produce identical gains.

At finite student exposure, $S'(T) = (1-a)S(T)(1-S(T))/[T(1-T)]$. Consequently a valid cap-to-gain Lipschitz bound is

$$L_{\mathrm{cap}} = \frac{w(1-a)}{4 \min\{p_0(1-p_0), \tau_{\max}(1-\tau_{\max})\}}.$$

It need not be at most w. If reference calibration satisfies $|p_R(i) - \tau_i| \le \beta$, then $\bar{u}_i \le \bar{u}_j$ implies $(\tau_i - \tau_j)_+ \le 2\beta$, giving $d_b^{\mathrm{iso}} \le L_{\mathrm{cap}}\beta$. This argument stays within the actual cap interval; it does not evaluate the learning map at an out-of-domain reference probability. Common initial correctness alone is insufficient: with $p_0 = 0.2$, trained teacher levels $(0.8, 0.9)$, student exposures $(4, 1)$ and $\eta = 1/2$, the per-prompt gains are approximately $(0.571, 0.400)$. The larger correctness gap then has the smaller student benefit because exposure differs.

B.1.1 Allowing differences in learning and reference accuracy

Exact equality of prompt difficulty and learning dynamics is not necessary for an approximate guarantee. The following extension bounds how much their differences can change the ranking. It also separates reference calibration from probe noise, rather than assuming the observed score orders student gains correctly.

Consider P prompts with the fixed group quotas of Section 3.2. Fix $I = [\ell, 1-\ell]$, $0 < \ell < 1/2$. On prompt i, the teacher and initial student share success probability $p_i \in I$. Teacher learning has an attainable level $\tau_i \in [p_i, 1-\ell]$ and a nondecreasing map $f_i : I \to I$ satisfying $f_i(t) \ge t$. Every selected prompt receives q_T teacher presentations, so $B_T = mq_T$ for every feasible selection.

Unselected prompts remain at p_i. The capped recursion above gives $t_i = \min\{\tau_i, v_i\}$, where $v_i = f_i^{\circ q_T}(p_i)$. For fixed student exposure n_i and coefficient $\eta_i \in (0,1)$, write $a_i = (1-\eta_i)^{n_i}$ and

$$s_i = \sigma(a_i \operatorname{logit} p_i + (1-a_i)\operatorname{logit} t_i)\,, \qquad g_i = w_i(s_i - p_i), \qquad w_i \geq 0.$$

Exposure and weights are independent of the selected set. Prompt-locality therefore gives $F(\mathcal{A}) = C + \sum_{i\in\mathcal{A}} g_i$, with $C = \sum_i w_i p_i$. We may also allow a remainder uniformly bounded by ε over feasible sets.

Let r_i be the reference's population success rate, $\bar{u}_i = r_i - p_i$, and $u_i = u(x_i)$. Assume $|r_i - \tau_i| \leq \beta_i$ and $|u_i - \bar{u}_i| \leq e_i$. The first condition relates reference performance to what the teacher can learn; the second controls the error from sampled responses. These are deterministic conditions, not an automatic confidence statement about the probes.

In each group choose comparison values $p_b \leq v_b$ in I, $a_b \in [0,1]$ and $w_b \geq 0$. Define $d_{p,i} = |p_i - p_b|$, $d_{v,i} = |v_i - v_b|$, $d_{a,i} = |a_i - a_b|$ and $d_{w,i} = |w_i - w_b|$. With clip_I denoting projection onto I, the comparison gain is

$$\phi_b(u) = w_b\left[\sigma(a_b \operatorname{logit} p_b + (1-a_b)\operatorname{logit}\min\{v_b, \operatorname{clip}_I(p_b+u)\}) - p_b\right].$$

It is nondecreasing in u because both teacher learning and student transfer are monotone. Set $L_\ell = [\ell(1-\ell)]^{-1}$, $R_\ell = 2\log[(1-\ell)/\ell]$, and

$$E_i = d_{w,i} + w_b\left\{d_{p,i} + \frac{L_\ell}{4}\left[a_b d_{p,i} + (1-a_b)\max\{d_{p,i} + \beta_i + e_i, d_{v,i}\}\right] + \frac{R_\ell}{4} d_{a,i}\right\}. \quad (6)$$

Then $|g_i - \phi_b(u_i)| \leq E_i$, and top-score selection $\mathcal{A}^\star$ satisfies

$$F(\mathcal{A}) - F(\mathcal{A}^\star) \leq \sum_{i\in\mathcal{A}\triangle\mathcal{A}^\star} E_i + 2\varepsilon \quad \text{for every quota-feasible } \mathcal{A}. \quad (7)$$

Thus the score is optimal in the exact, calibrated, common-dynamics case and approximately optimal when the stated deviations are small. The bound need not be informative when they are large.

Proof. Projection and the scalar minimum are nonexpansive, giving $|t_i - \widetilde{t}_i| \leq \max\{d_{p,i} + \beta_i + e_i, d_{v,i}\}$ for $\widetilde{t}_i = \min\{v_b, \operatorname{clip}_I(p_b + u_i)\}$. The difference of student logits equals

$$a_b(\operatorname{logit} p_i - \operatorname{logit} p_b) + (1-a_b)(\operatorname{logit} t_i - \operatorname{logit}\widetilde{t}_i) + (a_i - a_b)(\operatorname{logit} p_i - \operatorname{logit} t_i).$$

On I, logit is L_ℓ-Lipschitz, its range has width R_ℓ, and sigmoid is $1/4$-Lipschitz. Subtracting the initial probabilities and allowing the weight difference gives Equation 6. Pair omitted and admitted prompts within each group. The admitted score is no smaller, so monotonicity of ϕ_b bounds each gain difference by the two item errors. Common items cancel, and the remainder adds at most 2ε. □

The endpoint deviation $d_{v,i}$ can itself be bounded from the learning maps. If f_b is L_f-Lipschitz, $v_b = f_b^{\circ q_T}(p_b)$ and $\sup_{t\in I}|f_i(t) - f_b(t)| \leq \delta_{f,i}$, then

$$d_{v,i} \leq L_f^{q_T} d_{p,i} + \delta_{f,i}\sum_{r=0}^{q_T-1} L_f^r.$$

At each update the distance is at most L_f times the previous distance plus $\delta_{f,i}$; iteration proves the bound. These conditions concern learning dynamics, not an assumed ordering of final-student gains.

When the advantage is strict. Near-optimal selection need not improve on uniform sampling: all feasible sets could have the same value. The bound guarantees a strict advantage when the score-predicted difference exceeds the learning and estimation errors. Let $\mathcal{U}$ sample m_b prompts uniformly from each group and set $\rho_b = m_b/|\mathcal{Q}_b|$ and $z_i = \mathbf{1}[i \in \mathcal{A}^\star] - \rho_b$. The score-derived margin $M = \sum_i z_i \phi_b(u_i)$ is nonnegative. Canceling common contributions before applying the item bounds gives

$$F(\mathcal{A}^\star) - \mathbb{E}_{\mathcal{U}} F(\mathcal{U}) \geq M - \sum_i |z_i| E_i - 2\varepsilon.$$

A positive right-hand side guarantees improvement over uniform sampling with the same quotas. Without errors, M is the exact expected improvement. It can still be zero: if no attainable level limits teacher learning, different reference gaps can give identical gains.

B.1.2 From the selection score to the selection-stage comparison

A selection rule that improves the teacher can also beat student-prompt selection if the student absorbs enough of that improvement. We now combine the score-derived teacher gain with Theorem 1, giving each baseline student its best allocation of the same budget. Here performance is mean success over the P prompts and the teacher dynamics are exactly prompt-local. Define

$$\chi_b(u) = \min\{v_b, \mathrm{clip}_I(p_b + u)\} - p_b, \qquad E_i^T = d_{p,i} + \max\{d_{p,i} + \beta_i + e_i, d_{v,i}\},$$

and $M_T = P^{-1}\sum_i z_i \chi_b(u_i)$, $H_T = P^{-1}\sum_i |z_i| E_i^T$. The projection argument above gives $|(t_i - p_i) - \chi_b(u_i)| \le E_i^T$, hence $J(T_{\mathcal{A}^\star}) - \mathbb{E}_{\mathcal{U}} J(T_{\mathcal{U}}) \ge M_T - H_T$. Now train the selected teacher's student with a balanced allocation of B_S presentations and the common coefficient η of Definition 1. For each uniform teacher, allow its student the best integer allocation of the same budget. Then

$$J(\pi^{T_{\mathcal{A}^\star}}) - \mathbb{E}_{\mathcal{U}}\left[\max_{\substack{n \in \mathbb{Z}_{\ge 0}^P \\ \sum_i n_i = B_S}} J(\pi_n^{T_{\mathcal{U}}})\right] \ge M_T - H_T - \tfrac{1}{4}\Lambda_{\mathcal{A}^\star}(1-\eta)^{\lfloor B_S/P \rfloor}, \tag{8}$$

where $\Lambda_{\mathcal{A}^\star} = \max_i |\operatorname{logit} p_i - \operatorname{logit} T_{\mathcal{A}^\star}(i)|$.

Proof. Every teacher in this model is at least as accurate as the initial student on every prompt, since $\tau_i, v_i \ge p_i$. Thus $\Omega_+(T_{\mathcal{U}}) = 0$, and Proposition 2 bounds every baseline student by its teacher's mean success. The balanced student's transfer cost is bounded as in Theorem 1. Subtract, average over $\mathcal{U}$, and use the teacher-improvement bound above. □

If $M_T > H_T$, enough student training makes the lower bound positive. This conclusion holds in expectation over uniform teacher data, not against every uniform draw. Both procedures use mq_T teacher presentations and B_S student presentations. The learning assumptions are essential: high-reference-success prompts can be poor choices if they learn too slowly, and length grouping does not itself control these differences. A finite-budget example in Appendix B.6 shows that the sufficient condition can be positive.

Fixed budgets and repeated presentations. One analytical protocol allocates $B_T = qm + \rho$, $0 \le \rho < m$, as q presentations to each selected prompt plus one extra to a uniformly chosen ρ-subset. If the expected utility conditional on teacher multiplicities is $C + \sum_{i \in \mathcal{A}} G_i(n_i) + \mathrm{rem}(\mathcal{A}, n)$, with $G_i(0) = 0$ and a uniform bound $|\,\mathrm{rem}\,| \le \varepsilon$, then

$$g_i = (1 - \rho/m)G_i(q) + (\rho/m)G_i(q+1)$$

is independent of the other selected prompts, by linearity of expectation. Any student randomness included in G_i must have the same relevant conditional law across the compared sets and teacher-count conditioning. Equal unconditional exposure averages alone do not ensure this.

The experimental fill instead cycles the stored support, selects a content-seeded subset of positions and shuffles it. Order seeds do not resample the multiplicities. To show the distinction, consider only the special randomized law with $B_T = m$, cycle length $L = m + s > 1$, $0 \le s \le m$, and a uniformly random retained m-subset of positions. If $H_s(\mathcal{A})$ is the first s stored prompts, then

$$\mathbb{E}F(\mathcal{A}) = C + \sum_{i \in \mathcal{A}} k_1 G_i(1) + \sum_{i \in H_s(\mathcal{A})} \Delta_i + \overline{\mathrm{rem}}(\mathcal{A}),$$

where $k_1 = m/L$, $k_2 = m(m-1)/[L(L-1)]$, $p_1 = 2ms/[L(L-1)]$ and $\Delta_i = k_2 G_i(2) + (p_1 - k_1)G_i(1)$. These are the retention probabilities for one copy, both copies and exactly one of two copies. The identity follows by summing their expected gains. If $G_i(2) \ge G_i(1) \ge 0$ for every candidate, then $\Delta_i \ge 0$ and the added regret from this schedule term is at most $s \max_{i \in \mathcal{D}} \Delta_i$. The maximum must include omitted candidates, not just the selected set. For $m = 1, s = 0$ there is no schedule correction. This calculation does not bound a fixed content-hashed realization, nor cover arbitrary support sizes in the probe-token-budget variant.

The capped model uses deterministic prompt-local updates and a common training/evaluation population. It is not the expectation of a stochastic neural GRPO trajectory, and it does not identify held-out gains from reference correctness. Cross-prompt transfer, unequal exposure and the realized fill therefore remain substantive issues when applying the conditional result.

B.2 TEACHER IMPROVEMENT UNDER OTHER LEARNING DYNAMICS

The preceding result depends on how teacher learning responds to the reference gap. Under other learning dynamics, a larger gap need not imply a larger teacher gain. To make this dependence explicit, write $G_x^T(n)$ for the expected gain in the teacher's own success on prompt x from n presentations of x, and suppose the teacher's mean success over the pool decomposes prompt-locally, $J(T_{\mathcal{A}}) = J(T_{\text{init}}) + N^{-1}\sum_{x\in\mathcal{A}} G_x^T(n_x) + \text{rem}_T(\mathcal{A})$ with $|\,\text{rem}_T\,| \le \varepsilon_T$. Consider an idealized schedule with $B_T = m$ that presents each selected prompt exactly once. A uniform draw $\mathcal{U}$ of size m includes each prompt with probability m/N. The equality $B_T = m$ alone does not imply this schedule: the experimental fill can omit selected prompts and repeat others.

Proposition 3 (Teacher improvement at one presentation per prompt)**.** *Under this decomposition, for the selected set $\mathcal{A}^\star$ and a uniform draw $\mathcal{U}$,*

$$\mathbb{E}_{\mathcal{U}}\big[J(T_{\mathcal{A}^\star}) - J(T_{\mathcal{U}})\big] = \frac{1}{N}\Big[\sum_{x\in\mathcal{A}^\star} G_x^T(1) - \frac{m}{N}\sum_{x\in\mathcal{D}} G_x^T(1)\Big] + r, \qquad |r| \le 2\varepsilon_T.$$

If the scores are exact, $u = \bar{u}$, and within each group $G_x^T(1)$ is a nondecreasing function of $\bar{u}$, the selected set maximizes the bracket among fixed-quota sets. With noisy scores, the exchange bound of Proposition 1 instead applies when its alignment and error assumptions hold for these teacher gains.

Proof. Linearity of expectation and the inclusion probability m/N give the identity; the exchange argument of Proposition 1, applied to the teacher-endpoint gains, gives the maximization. □

Under this schedule, the expected improvement is m/N times the difference between the selected set's and the pool's mean one-presentation teacher gain, up to interactions. A larger correctness gap alone does not establish the required ordering of gains. For example, consider a Bernoulli policy-gradient surrogate with $K \ge 2$ responses, sample-standard-deviation-normalized advantages, and zero update for constant-reward groups. Its expected scalar log-odds increment is $\eta_T \kappa_K(p)$, where $\kappa_K(p) = \sqrt{(K-1)/K}\,\mathbb{E}_{M\sim\text{Bin}(K,p)}[\sqrt{M(K-M)}]/K$. For an interior initial success probability and a nonbinding attainable cap, the first-order success gain is $\eta_T p(1-p)\kappa_K(p) + O(\eta_T^2)$, which depends on initial success rather than on the cap revealed by the reference. This scalar calculation is not the update rule of a neural GRPO model.

In this surrogate the gain vanishes at a true success probability of zero, but zero successes in four or eight probes do not establish that condition. Substituting empirical probe rates into the surrogate gives a diagnostic, not a proof that the observed teacher gains require cross-prompt transfer. Gains on prompts omitted from training provide more direct evidence of generalization (Appendix B.9), without identifying the causal contribution of each selected prompt. Reference information can also reveal prompt-specific learnability in a local model with heterogeneous dynamics. Thus the matched-budget theorem alone takes teacher improvement as an input. Appendix B.1.2 supplies one set of learning assumptions under which the score yields such an improvement; other dynamics need not preserve that conclusion.

B.3 THE PROBE-TOKEN-BUDGET SELECTOR OPTIMIZES A DIFFERENT PROXY PROBLEM

A token constraint changes the selection problem: a high-scoring prompt may displace several lower-scoring prompts with a larger combined score. Unlike the fixed-size rule, the score-ordered scan need not maximize their sum. We bound its shortfall using the observed scores and costs. Condition on observed scores u_i, positive finite probe costs c_i, group membership and the realized tie order. The token variant scans each group in decreasing score order, accepts a whole prompt if it fits the remaining capacity C_b, and otherwise continues to the next prompt. It can admit negative scores. Let its selected set be $\mathcal{A}_G$ and $U(\mathcal{A}) = \sum_{i\in\mathcal{A}} u_i$. The feasible family is $\mathcal{K} = \{\mathcal{A} : \sum_{i\in\mathcal{A}\cap\mathcal{Q}_b} c_i \le C_b\ \forall b\}$, not the fixed-size family of Proposition 1.

For eligible items $E_b = \{i \in \mathcal{Q}_b : c_i \le C_b\}$ define

$$D_b = \min_{\lambda\ge 0}\left[\lambda C_b + \sum_{i\in E_b}(u_i - \lambda c_i)_+\right].$$

Let K_b be the largest affordable number of positive-score eligible items, found by sorting their costs increasingly, and let V_b sum their largest K_b scores ($V_b = 0$ if $K_b = 0$). For the score-optimal

token-feasible set $\mathcal{A}_U$, the unchanged scan satisfies

$$0 \le U(\mathcal{A}_U) - U(\mathcal{A}_G) \le \sum_b \min\{D_b, V_b\} - U(\mathcal{A}_G).$$

To prove this, any feasible binary selection z_i obeys $\sum_i u_i z_i \le \lambda C_b + \sum_i (u_i - \lambda c_i)_+$ for $\lambda \ge 0$. Minimizing gives the fractional-knapsack upper bound D_b. A feasible selection has at most K_b positive items, so V_b is another upper bound. The constraints separate across groups; summing the smaller bound and subtracting the scan's score proves the certificate. Both bounds are computable from the observed score/cost pool, without modifying the deployed selector.

The scan need not maximize even this proxy: costs $(6, 5, 5, 6, 6, 6, 6)$ with scores $(1, 3/4, 3/4, 0, 0, 0, 0)$ give the actual quarter-cost capacity 10. The scan earns 1, whereas the two cost-five prompts earn $3/2$; the upper certificate is $8/5$. The certificate concerns the summed observed scores, not repeated training exposure or final-student performance, and it does not compare with a different global-cap uniform rule. No production-pool certificate is claimed here.

B.4 Exact comparison at a finite budget

The upper bound in Proposition 2 can be loose at a finite budget. The strongest student-prompt allocation can instead be computed exactly by dynamic programming. Within Definition 1, let $\sigma(z) = (1 + e^{-z})^{-1}$ and write

$$f_j^0(n) = \sigma\big(\operatorname{logit} t_j^0 + (1-\eta)^n \big[\operatorname{logit} p_j(0) - \operatorname{logit} t_j^0\big]\big)\,.$$

Define $V_j(b)$ as the maximum contribution to mean success from the first j prompts using exactly b presentations. Set $V_0(0) = 0$ and $V_0(b) = -\infty$ for $b > 0$. For integer $b = 0, \dots, B_S$,

$$V_j(b) = \max_{n \in \{0,\dots,b\}} \left\{ V_{j-1}(b-n) + \frac{f_j^0(n)}{P} \right\}. \tag{9}$$

Then $V_P(B_S) = \max_{\sum_j n_j = B_S} J(\pi_n^{T_0})$, with nonnegative integer allocations. A student of $T_\dagger$ therefore beats every baseline allocation at this budget exactly when its mean success exceeds $V_P(B_S)$.

Proof. Any allocation of b presentations among j prompts assigns some n to the last prompt and $b - n$ to the preceding prompts. The best preceding allocation contributes $V_{j-1}(b-n)$; maximizing over n gives the recurrence. Induction from the initialized empty allocation proves the identity. □

The recurrence requires $O(PB_S^2)$ time and $O(B_S)$ memory for values. A cheaper upper bound, valid for every $\lambda \in \mathbb{R}$, is

$$V_P(B_S) \le \lambda B_S + \sum_{j=1}^{P} \max_{n \in \{0,\dots,B_S\}} \left\{ \frac{f_j^0(n)}{P} - \lambda n \right\}.$$

To obtain it, subtract λn_j from each prompt's contribution, maximize each term independently, and use $\sum_j n_j = B_S$. It costs $O(PB_S)$ per value of λ and need not be tight. Both calculations assume the model's per-prompt probabilities are supplied; neither is an algorithm for selecting teacher-training data from the reference probes.

B.5 Student training with fewer presentations than prompts

For $B_S < P$ the balanced allocation presents a set S of B_S distinct prompts once and the others never, so $\lfloor B_S/P \rfloor = 0$ and the decay factor in Theorem 1 equals one. Equation 5 then reads $\Delta_T - \Lambda/4 - \Omega_+(T_0)$, which is nonpositive whenever $\Lambda \ge 4\Delta_T$; in the first constructed example of Appendix B.6, $\Lambda/4 \approx 1.13$. For the single-teacher comparisons using $B_S = 2{,}560$ and candidate pools of 8,000–10,000 prompts, this ratio is 0.26–0.32 if the entire candidate pool is identified with the model's prompt set. This identification is illustrative: the model's balanced allocation differs from a repeated or selected experimental sequence. The original theorem remains valid below one pass but

may be loose; the following bound offers an alternative. Write $\gamma_j(T) = p_j(1;T) - p_j(0)$ for the gain of one presentation toward teacher T, $\bar{\gamma}(T) = P^{-1}\sum_j \gamma_j(T)$ for its pool mean, and, for $k \le P$,

$$G_k(T_0) = \frac{1}{P}\max_{|K|=k}\sum_{j\in K}\big(p_j(B_S;T_0) - p_j(0)\big)_+,$$

an upper bound on the total gain the baseline teacher can deliver on k prompts within the budget; it allows each of those prompts the full budget separately and is at most the same sum of asymptotic headrooms $(t_j^0 - p_j(0))_+$.

Proposition 4 (Sub-one-pass comparison)**.** *Let $B_S \le P$ and let S be a uniformly random B_S-subset of the prompts, each presented once. Then*

$$\mathbb{E}_S\, J(\pi^{T_\dagger}) - \max_{\substack{n\in\mathbb{Z}^P_{\ge 0}\\ \sum_j n_j = B_S}} J(\pi_n^{T_0}) \;\ge\; \frac{B_S}{P}\,\bar{\gamma}(T_\dagger) - G_{B_S}(T_0), \tag{10}$$

and for any fixed S the same holds with $\frac{B_S}{P}\bar{\gamma}(T_\dagger)$ replaced by $P^{-1}\sum_{j\in S}\gamma_j(T_\dagger)$.

Proof. Each prompt lies in S with probability B_S/P, so $\mathbb{E}_S J(\pi^{T_\dagger}) = J(\pi_0) + (B_S/P)\bar{\gamma}(T_\dagger)$; for a fixed S the sum over S replaces the expectation. For any allocation with $\sum_j n_j = B_S$, at most B_S prompts have $n_j \ge 1$, and each has $n_j \le B_S$; by Equation 4 $p_j(n)$ moves monotonically from $p_j(0)$ toward t_j^0, so $p_j(n_j) - p_j(0) \le (p_j(B_S;T_0) - p_j(0))_+$, and unpresented prompts contribute nothing. Hence $J(\pi_n^{T_0}) \le J(\pi_0) + G_{B_S}(T_0)$ for every allocation. Subtracting the two displays proves Equation 10. □

Below one pass, the improved teacher's advantage reaches the student only on the presented fraction B_S/P of the pool in this prompt-local model, while student-prompt selection's advantage is concentration: it can improve at most B_S prompts, by no more than the baseline teacher's gain on each within the budget. A positive right-hand side in Equation 10 certifies an advantage; a nonpositive value does not predict its absence. Both terms are per-prompt quantities on the training prompts, which aggregate holdout scores do not supply; this is why Appendix B.9 gives no numerical certificate for the runs. The comparator term is attained only when the baseline student realizes each of those B_S gains, so the bound can be negative when the true difference is positive; the exact maximum is the dynamic programme of Appendix B.4. In 600 random instances with $P \le 7$ and $B_S < P$, solved exactly, Equation 10 held for the expectation and for every fixed subset, Equation 5 was negative in all 600 while remaining valid, and Equation 10 certified 24 of the 50 instances whose true difference was positive (`verify_subonepass.py`).

The quarter-budget comparison in Appendix D.2.7 gives student-score differences of -0.001, -0.005, and $+0.006$ between 640 and 2,560 presentations for the selected teachers (Table 33); uniform-teacher students gain 0.014 from the extra budget. These descriptive differences do not establish equivalence or identify a transfer mechanism. The measured scores use separate evaluation prompts, whereas the presented-fraction identity concerns expected performance on the model's own prompt set under uniform sampling without replacement. They are not a direct test of that identity.

B.6 Constructed examples and numerical verification

Gap-based selection can give a strict finite-budget advantage. For Equation 8, take one group of four prompts with initial success $p_i = 0.2$, attainable levels $(0.3, 0.5, 0.7, 0.9)$, and a quota of two prompts. Let $I = [0.1, 0.9]$, $f_i(t) = 0.9$, and give each selected prompt one teacher presentation. With a calibrated reference and exact scores, selection keeps the last two prompts. Its mean teacher improvement is 0.30, versus 0.20 in expectation for uniform selection, so $M_T = 0.10$ and $H_T = 0$. With $\eta = 1/2$ and $B_S = 16$, the transfer cost is $\log(36)/64$ and Equation 8 guarantees a student advantage of at least 0.0440. Exact allocation search gives an expected advantage of 0.0875. These are constructed model values, not empirical scores.

Prompt allocation can win at a small student budget. Take two prompts with initial success probabilities $p(0) = (0.10, 0.99)$, baseline teacher $t^0 = (0.90, 0.99)$, and improved teacher $t^\dagger = (0.91, 0.99)$. Set $\eta = 1/2$ and $B_S = 2$. Balanced training on the baseline teacher gives mean success 0.745. Using the improved teacher with the same allocation gives 0.752274, whereas assigning both

presentations to the first prompt under the baseline teacher gives 0.870. These values follow directly from Equation 4; the second prompt does not change because its initial and teacher probabilities coincide. Here $\Delta_T = 0.005$, $\Omega_+(T_0) = 0$, and $\Lambda = \mathrm{logit}(0.91) - \mathrm{logit}(0.10) \approx 4.511$. The lower bound in Equation 5 is approximately -0.559, so it does not certify a teacher-data-selection advantage.

The exact comparison can succeed when the lower bound is negative. With $p(0) = (0.2, 0.2)$, $t^0 = (0.8, 0.2)$, $t^\dagger = (0.9, 0.7)$, $\eta = 1/2$, and $B_S = 2$, balanced training under the improved teacher reaches 0.516515. The optimal baseline allocation assigns both presentations to the first prompt and reaches 0.433333. The exact margin is positive even though Equation 5 gives $0.3 - \frac{1}{8}\log(36) \approx -0.147940$. Thus the sufficient condition is conservative, and a small student budget does not by itself determine which selection stage wins.

Numerical verification. The accompanying script `verify_general.py` samples synthetic problem instances and checks the displayed inequalities, using exhaustive allocation search where needed. In its $20{,}000$ ceiling-check instances, the *optimal* allocation exceeds teacher mean success in $14{,}574$ cases; this is not a result about uniformly sampled allocations. No violation of the stated ceiling or matched-budget bound was found in their respective checks. These calculations check finite instances of the mathematics, not performance on the experimental benchmarks.

B.7 Shared-parameter placement on a separate evaluation distribution

The comparison can also allow updates to affect several prompts and evaluate the student on a different distribution. The sufficient condition has the same form: an advantage in the teacher-dependent target must exceed the errors from finite training, model mismatch and noise. Let $V_Q(\theta) = \mathbb{E}_{z\sim Q} v(\theta, z)$ be an L_V-Lipschitz student utility on a fixed evaluation distribution Q, which may differ from the training distribution. For teacher $k \in \{0, 1\}$, choose a fixed reference state β_k in the student's parameter space; it need not represent the teacher exactly. Suppose the student dynamics satisfy

$$\theta_{t+1} - \beta_k = M_t(\theta_t - \beta_k) + \alpha h_t + \alpha\zeta_t,$$

where $\alpha > 0$, M_t, h_t are measurable given the history $\mathcal{F}_t$ before the update, $\|M_t\|_2 \le \rho_k$, $\|h_t\| \le b_k$, $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[\|\zeta_t\|^2 \mid \mathcal{F}_t] \le \sigma_k^2$. The baseline constants must hold uniformly over the entire allowed student-prompt policy class $\mathfrak{P}$; the selected-teacher constants need only hold for its fixed student procedure. Set

$$R_k(H) = \sqrt{\rho_k^{2H}\|\theta_0 - \beta_k\|^2 + \alpha^2\sigma_k^2 \sum_{r=0}^{H-1} \rho_k^{2r}} + \alpha b_k \sum_{r=0}^{H-1} \rho_k^r.$$

At the same H student updates, these assumptions imply

$$\mathbb{E}V_Q(\theta_H^1) - \sup_{\mathcal{P}\in\mathfrak{P}} \mathbb{E}V_Q(\theta_H^{0,\mathcal{P}}) \ge V_Q(\beta_1) - V_Q(\beta_0) - L_V\big[R_1(H) + R_0(H)\big]. \tag{11}$$

Proof. For either teacher write $e_t = \theta_t - \beta_k$ and construct $y_0 = e_0$, $y_{t+1} = M_t y_t + \alpha\zeta_t$ on the same history. Since M_t, y_t are $\mathcal{F}_t$-measurable, the conditional cross term vanishes, giving $\mathbb{E}\|y_{t+1}\|^2 \le \rho_k^2\mathbb{E}\|y_t\|^2 + \alpha^2\sigma_k^2$. Iteration gives the squared term under the root in R_k. The difference satisfies $e_{t+1} - y_{t+1} = M_t(e_t - y_t) + \alpha h_t$, hence its norm is at most $\alpha b_k \sum_{r=0}^{H-1} \rho_k^r$ at time H. The triangle inequality and Cauchy–Schwarz yield $\mathbb{E}\|\theta_H - \beta_k\| \le R_k(H)$. Lipschitz continuity bounds the selected student's utility below and every allowed baseline student's utility above by the respective anchor utility plus or minus $L_V R_k$. Uniform baseline constants permit taking the supremum, proving Equation 11. Future adaptive matrices were not assumed independent of earlier noise. □

For example, shared-feature squared-loss distillation with teacher output $f_k(x)$ and student output $\psi(x)^\top\theta$ has $M_t = I - \alpha C_{\nu_t}$, where $C_\nu = \mathbb{E}_\nu[\psi\psi^\top]$, and approximation term $h_{k,\nu} = \mathbb{E}_\nu[\psi(f_k - \psi^\top\beta_k)]$. If every allowed prompt distribution obeys $\mu I \preceq C_\nu \preceq LI$, with $\mu > 0$ and $0 < \alpha \le 1/L$, then $\rho_k = 1 - \alpha\mu < 1$. A uniform residual bound $\mathbb{E}_\nu(f_k - \psi^\top\beta_k)^2 \le \epsilon_k^2$ gives $b_k \le \sqrt{L}\epsilon_k$ by Cauchy–Schwarz in each unit direction. Shared features permit effects across prompts, while approximation error allows a student that cannot exactly represent the teacher. Lipschitz utility is a separate assumption; it does not automatically hold for greedy-decoding accuracy.

This extension separates finite training, model mismatch and gradient noise. With $\rho_k < 1$, the initialization term decays, but at constant step size the latter two terms can persist. Requiring coverage $\mu > 0$ restricts the baseline selector class: the result is not a comparison with every unrestricted allocation. Taking $\rho_k = 1$ permits nonexpansive updates without that coverage requirement, but the error bound no longer decays. The result still assumes an improved anchor; it neither derives that improvement from u nor identifies useful bounds for the neural runs.

B.8 A LINEAR TRANSFER MODEL

A representable linear model separates the effects of changing the teacher and the prompt distribution. The student minimises $\frac{1}{2}(\theta-\phi)^\top C_v(\theta-\phi)$ for a teacher parameter ϕ and a prompt second moment C_v, and runs G gradient steps of size η. With $M = I-\eta C_v$ the trajectory is exactly $\theta_G = \phi+M^G(\theta_0-\phi)$. A change to the teacher transmits as $(I - M^G)\,\delta\phi$, which approaches $\delta\phi$ when $\|M\|_2 < 1$. For a perturbation $C_v + \epsilon D$, the derivative at $\epsilon = 0$ is $-\eta \sum_{r<G} M^{G-1-r} D M^r(\theta_0 - \phi)$, of norm at most $\eta G \rho^{G-1}\|D\|\,\|\theta_0 - \phi\|$ with $\rho = \|M\|_2$. Sensitivity to student-prompt selection carries the factor $\|\theta_0 - \phi\|$, the distance from the student's start to its teacher, whereas sensitivity to the teacher does not. Under contraction, the prompt-mixture sensitivity bound vanishes with training, while the teacher perturbation is eventually transmitted in full. This describes parameter sensitivity, not an exact crossing budget or an ordering of task-performance gains; those also depend on the perturbations' directions and magnitudes.

B.9 SCOPE AND RELATION TO THE EXPERIMENTS

Why the bounds do not certify the neural runs. The tabular model defines B_S/P, Δ_T, Λ and Ω_+ on a common set of prompts. Training-prompt counts and success rates from a separate holdout do not supply those quantities. In addition, the proximal coefficient η is not identified with the neural optimizer's learning rate. Identifying P with the full candidate pool in the 2,560-presentation comparisons gives $B_S < P$, so the geometric decay factor equals one. Theorem 1 remains valid, but it need not give a positive lower bound; Appendix B.5 supplies an alternative for uniformly chosen, unrepeated subsets. Neither allocation is automatically the experimental sequence. Finally, the binary model evaluates on its training prompts, whereas the student comparisons use separate evaluation sets. Appendix B.7 allows this separation under additional stability and noise assumptions that have not been verified for the neural runs.

What the bounds need and what the probes supply. Table 5 lists every term of the three bounds against the available measurements. Budgets and pool sizes are recorded, but the realized probe errors e_i are unknown. With independent Bernoulli probes, each model's empirical rate has standard deviation at most 0.25 for $k = 4$. Hoeffding's inequality and a union bound give a simultaneous gap-error radius $\sqrt{\log(2N/\delta)/k} \approx 1.80$ for $N = 10{,}000$ and $\delta = 0.05$. This conservative bound is too loose to establish useful selection regret here; it does not prove that every finite-probe guarantee must be uninformative. Teacher probes now supply descriptive estimates for some other terms, as reported below. They do not identify counterfactual final-student gains or uniform interaction bounds.

Table 5: **Bound terms and measurement requirements.** • Recorded budgets, estimated probabilities, and structural assumptions.

Term	Definition	Source
B_S/P	presentations divided by the model's prompt-set size	0.26–0.32 if P is the full pool in the 2,560-presentation comparisons
e_i	probe error $\lvert u_i - \bar{u}_i \rvert$	unknown; simultaneous 95% radius 1.80 under independent probes
d_b, L_b	alignment and smoothness of final-student gains versus $\bar{u}$	require counterfactual student outcomes and structural assumptions
ε, ε_T	uniform interaction remainders across feasible selections	not identified by probes of a few trained models
Δ_T on the pool	selected minus baseline teacher success on a common prompt set	estimated on an excluded-prompt panel, not the entire pool
Λ	largest initial-student–teacher log-odds gap	clipped plug-in diagnostic only; population maximum unknown
$\Omega_+(T_0)$, $G_{B_S}(T_0)$	initial student advantage and a budgeted-gain upper bound	nonlinear plug-in estimates sensitive to probe noise and panel choice
$\bar{\gamma}(T_\dagger)$	mean one-presentation gain in the proximal model	calculated from estimated rates at chosen, uncalibrated η values

Diagnostics from measured pass rates. We evaluated the initialisation at $k = 8$ and the three selected and three uniform teachers of the uniform-draw wave at $k = 4$, under the trainer reward at temperature 1, on their own 640 selected prompts and on a common panel of 640 pool prompts outside all four selected corpora. Table 6 reports descriptive rates and plug-in diagnostics: three training orders of one selected corpus against three uniform draws. The panel samples the remaining 7,667 candidates, not the entire pool or the student's actual training sequence. Positive parts and maxima are sensitive to probe noise; resampling under equal per-prompt success illustrates that sensitivity, but does not estimate the actual bias or supply a hypothesis test.

Table 6: **Teacher probes and bound diagnostics.** • Selected: three corpus orders; uniform: three draws. • Gains: teacher minus fresh initialization on the same prompts. • Noise: resampling around a common per-prompt rate; $\rho = B_S/P$. • Excluded-prompt panel; descriptive plug-in diagnostics, not certified bounds.

Term	Selected	Uniform
fresh $p_j(0)$ on the selected set, $k = 8$ (what the selector saw, $k = 4$)	0.272 (0.099)	panel 0.285 (0.288)
fresh reference success on the selected set, $k = 8$ (what the selector saw)	0.748 (0.924)	panel 0.428 (0.416)
measured score u, fresh (what the selector saw)	0.476 (0.825)	panel 0.143 (0.127)
own-prompt teacher gain, presented prompts	+0.49 / +0.53 / +0.39	+0.13 / +0.15 / +0.14
own-prompt gain, the ≈ 30 unpresented prompts	+0.43 / +0.53 / +0.27	+0.14 / −0.02 / +0.08
own-prompt gain where the fresh initial probe has zero successes	+0.68 / +0.76 / +0.52 (n=140)	+0.15 / +0.18 / +0.14 (n≈350)
panel success (initialisation 0.285)	0.456 / 0.477 / 0.388	0.390 / 0.394 / 0.403
Δ_T on the panel, mean (pairs)	+0.045 (−0.016 to +0.087)	
$\Omega_+(T_0)$ on the panel (noise at equal rates)	—	0.027 / 0.047 / 0.031 (0.056–0.060)
Λ, clipped plug-in value; value divided by 4	4.65; 1.16	—
$\rho\,\bar{\gamma}(T_\dagger)$, $\eta \in [0.1, 0.8]$, $\rho = 0.256$ (0.32)	0.003–0.030 (0.003–0.038)	—
$G_{B_S}(T_0)$ plug-in value (equal-rate simulation)	—	0.119 (0.058); 0.134 (0.058) at $\rho = 0.32$
plug-in right-hand side of Equation 10	negative for every pair on the tested η grid	

The plug-in calculations do not certify an advantage. The clipped log-odds statistic is 4.65, but clipping noisy rates does not give a lower bound on the unknown population Λ. The sub-one-pass expression is negative for all nine teacher pairs at $\eta \in \{0.1, 0.3, 0.5, 0.8\}$ and both displayed ratios. These calculations neither establish the sign of the population bound nor contradict a positive measured student difference. The comparator term G_{B_S} is itself an upper bound on the best allocation, not its exact value. Equal-rate resampling is a sensitivity analysis, not a valid bias subtraction or a test that Ω_+ equals zero.

The probes provide descriptive evidence of generalization to unpresented prompts. The selected teachers gain 0.27–0.53 on selected prompts omitted by the training fill and 0.10–0.19 on the common excluded-prompt panel. Their mean panel score exceeds the uniform teachers' mean by 0.045; this is not the full-pool Δ_T of the theorem. Zero successes in the initial eight-response probe do not imply zero population success, so gains on that subgroup alone cannot identify cross-prompt transfer. Nor do these measurements identify the causal contribution of individual training prompts or the final-student interaction remainder.

On the selected set, fresh $k = 8$ probes give an initial-student success rate of 0.272 and a reference success rate of 0.748, against the 0.099 and 0.924 the selector saw. Replacing both estimates reduces the mean measured gap from 0.825 to 0.476, by 0.35, almost equally through the two terms; on the panel the same replacement changes the mean gap by +0.02, so the change of decoding does not explain the shift. This is consistent with selection-induced optimism in both terms, and the selected prompts retain a larger mean gap than the panel, 0.476 against 0.143. The fresh rates are themselves noisy, and the reference was trained on the whole pool, so they are not held-out estimates. They do not measure the realized per-prompt errors e_i required by Proposition 1.

The models isolate how teacher targets and student allocation affect learning under specified dynamics. In the binary model, convergence to the teacher on every prompt requires every prompt to receive increasingly many presentations. In the linear model, convergence requires contraction on the relevant parameter space. Unvisited prompts and unobserved feature directions need not converge. If the student cannot represent its teacher, changing prompt weights can also change the limiting approximation, so the vanishing prompt-mixture effect is not universal.

The proximal model updates Bernoulli correctness distributions, not distributions over complete responses; its ceiling is specific to that abstraction. The binary and representable linear analyses use deterministic updates. The shared-parameter extension allows stochastic updates under the stated conditional noise bounds; none of these models derives the neural implementation's top-k truncation, clipping, or stop-gradient dynamics. The matched-budget theorem conditions on the trained teachers; Appendix B.1.2 derives a teacher improvement from the data-selection score under additional learning and calibration assumptions. The empirical comparisons in Section 4.2 are evaluated independently of these theoretical bounds; the fixed-size placement contrast of Appendix D.2.4 is the empirical counterpart of Theorem 1's comparison.

C Detailed experimental setup

This appendix documents the training configurations, selector and baseline implementations, resource use, and registered analyses. Table 7 summarizes the comparisons; Appendix D reports their results.

Table 7: **Study designs.** • Pool size: training candidates. • Replication units: training orders, independent selections, or paired blocks. • Within-study comparisons; no pooling across rows.

Study	Selection rule and candidate pool	Replication	Evaluation
IF, original uniform comparison	Fixed-size quotas; 10,000 teacher candidates	3 orders of one corpus; 11 uniform corpora	Original IF
IF, fresh probes	Same rule and pool; new selection per run	5 independent pipelines; same 11 controls	Original and new IF
IF, original selection-stage comparison	Probe-token budget over 2,000 teacher candidates; disjoint 8,000 student candidates	5 teachers per condition	Original IF
IF, fixed-size selection-stage follow-up	Fixed-size selection; pools in Appendix D.2.4	3 orders plus 5 fresh selections; 5 teachers per student baseline	Development and new IF
IF, token-budget follow-up	Original token-budget recipe; Appendix D.2.6	5 paired blocks; existing students	New IF
IF, direct reference	Selected teacher versus released reference; paired student prompts	5 paired blocks; one selected corpus	Original and new IF
Mathematics	One 640-prompt corpus from 9,996 teacher candidates; 8,997 student candidates	5 paired blocks; selection fixed	Reserved 2,000 prompts
Adapted teacher-data criteria	Four scores on the same 10,000 candidates	2 tie keys per criterion; 3 orders for ours	Development IF
Multi-teacher, replicated baselines	Shared initialization, domain weights and budget; Appendix D.5.4	3 students per configuration; selection fixed	Declared and public panels
Multi-teacher, other comparisons	Study-specific recipes; Appendix D.5.2	1 to 4 students per condition, stated per study	Study-specific panels

C.1 Shared configuration

The placement decision as a constrained objective. Let $G(\mathcal{D}_{\text{up}}, \mathcal{D}_{\text{down}}) \in \mathbb{R}^K$ be the capability vector of the final student, one coordinate per evaluation set, and let F be a coordinatewise nondecreasing scalar utility declared in advance. With B_T teacher and B_S student presentations, $\mathcal{P}$ a set of protected coordinates and τ their floors, the placement decision of Section 3.1 reads

$$\max_{\mathcal{D}_{\text{up}},\, \mathcal{D}_{\text{down}}} F\big(G(\mathcal{D}_{\text{up}}, \mathcal{D}_{\text{down}})\big) \quad \text{s.t.} \quad |\mathcal{D}_{\text{up}}| = B_T, \quad |\mathcal{D}_{\text{down}}| = B_S, \qquad G_k \geq \tau_k \;\; \forall k \in \mathcal{P}, \quad (12)$$

Here cardinalities count repeated prompt presentations. We report each capability in G separately, without choosing an aggregate utility F or protected-performance floors.

Table 8 gives the original single-teacher instruction-following configuration. Study-specific changes appear with the corresponding results.

Table 8: **Original single-teacher instruction-following configuration.** • Other configurations: specified in their respective studies.

Models	Teacher and student are both SmolLM3-3B from the same released mixed-SFT checkpoint (Gao et al., 2026); we do not train the initialisation. They share a base and a tokenizer. The reference R (Section 3.2) is a third model, probed and never trained further.
Teacher training	GRPO on an instruction-following pool drawn from the `nemotron_if_rl` source of the Open-MOPD RL prompt mix (Gao et al., 2026): 40 steps $\times$ 16 prompts $\times$ 8 rollouts $=$ 5,120 rollout sequences per condition; group-normalised advantages at the framework default; no KL penalty; no entropy bonus; learning rate 1.5×10^{-6}; 2,048-token response cap; temperature 1.
Support and repetition	Uniform-draw wave and development matrix: a fixed 640-prompt support from a 10,000-prompt pool; the 640 draws consume 610–613 distinct rows, a repetition factor of about 1.05, the same for selected and uniform conditions. Placement waves: a 2,000-prompt pool, where the quota of Section 3.2 was a probe-token budget rather than a fixed size: quartile b takes prompts in decreasing u, ties left to the builder's deterministic sort rather than to σ, while their total probe cost stays within a quarter of the quartile's own, $\sum_{x \in \mathcal{A} \cap \mathcal{Q}_b} c(x) \leq \frac{1}{4} \sum_{x \in \mathcal{Q}_b} c(x)$, skipping a prompt that would exceed it. This took roughly a quarter of the root, and each uniform draw was cut to the same probe-token budget. Repetition factor 1.30 for the selected condition against 1.23–1.32 for uniform draws.
Student training	On-policy distillation from exactly one teacher: 40 steps at batch $64 = 2{,}560$ prompt presentations per condition. A teacher's two fill shuffles are within-cluster replicates, never independent samples.
Selection probes	k=4 rollouts per candidate from the initial student and k=4 from the reference, giving p_i^S and p_i^R. The reference is the instruction-following RL specialist released with the initialisation (Gao et al., 2026), trained by its authors on the full instruction-following RL corpus, so p^R carries competence information from a run outside our matched budget. All 10,000 candidate prompts are rows of that corpus under an exact normalised-text match, so every selected and every uniform teacher corpus is drawn from it, as are both evaluation holdouts; whether p^R is higher on prompts the reference fits more strongly was not measured. The sealed static-band student-prompt comparator scores on p^S alone; the same-score comparator also uses p^R. Two of the four adapted teacher-data baselines use the same p^R.
Selector	Score $u = p^R - p^S$ from the probes; length groups are the four quartiles of the initial student's mean probe response length, with quotas proportional to group size. In the uniform-draw wave and development matrix the selected set holds 640 prompts and supplies the training sequence under the repetition rule above. The placement waves use the probe-token-budget variant above. Tie handling: Appendix C.2.
Evaluation	Confirmatory holdout and development holdout: disjoint instruction-following sets of 750 prompts each. Primary: the reward's strict criterion with its format guard, which revokes a pass whose response is empty, single-token or constraint-only. Raw base: the unguarded strict score. Length-filtered analyses specify whether they use raw strict correctness or retain the format guard. External panel: IFEval (Zhou et al., 2023) and IFBench for instruction following, AIME as a mathematics control channel.
Hardware	1–2 L40S GPUs per job on a shared university cluster. Run logs retain scheduling, preemption and completed-step records.

Pool statistics. Table 9 distinguishes the original selection probes from fresh responses on the selected data and an unselected comparison panel. Degenerate-group shares use the initial k=4 probes, not the k=8 teacher-training rollouts. Reference scores come from the released specialist; a reference trained on uniform data within the teacher budget was not tested. Probe matrices and response-token costs differ across conditions, as reported in Table 10.

Table 9: **Candidate pools and selected data.** • Original probes: $k = 4$; fresh selected-set and unselected-panel probes: $k = 8$. • Teacher/student candidate pools: 2,000/8,000 prompts from a 10,000-prompt pool.

Set	$\bar{p}^S$	$\bar{p}^R$	mean u	mean probe cost (tok.)
10,000-prompt pool (uniform-draw wave, development matrix)	—	0.429	+0.154	449
share of pool with $u > 0$ / $u = 0$	39.8% / 52%			
selected set (ours, 640 rows)	0.099	0.924	+0.825	444
selected set, fresh k=8 re-probe	0.272	0.748	+0.476	—
640-prompt panel outside every training set, fresh k=8	0.285	0.428	+0.143	—
2,000-prompt root (placement waves, teacher data)	0.281	0.435	+0.154	454
8,000-prompt root (placement waves, student prompts)	0.273	0.427	+0.154	447

C.2 Selector implementation

For the fixed-size variant, let $m \leq N$ be the number of distinct prompts to select, fixed in advance. The probe-token-budget variant instead fixes a total probe cost, so its selected-set size is determined by the admitted prompts (Appendix C.1). The score u takes only $2k + 1$ values, so we fix a tie-breaking order before selection. Let $\sigma : \mathcal{D} \to \{1, \ldots, N\}$ number the prompts in one random order, drawn from a hash of the probe scores, so that every selector scored on the same probes breaks ties identically and no prompt is favoured by its position in the pool. Write $x \succ x'$ when $u(x) > u(x')$, or when $u(x) = u(x')$ and $\sigma(x) < \sigma(x')$, and for $\mathcal{Q} \subseteq \mathcal{D}$ let $\mathrm{top}_\ell(\mathcal{Q})$ be the ℓ largest elements of $\mathcal{Q}$ under $\succ$. Let $c(x)$ be the cost proxy of Section 3.2 and $q_1 \leq q_2 \leq q_3$ the quartile cut points of c over $\mathcal{D}$, with $q_0 = -\infty$ and $q_4 = +\infty$; the four groups are $\mathcal{Q}_b = \{\, x \in \mathcal{D} : q_{b-1} \leq c(x) < q_b \,\}$ for $b = 1, \ldots, 4$, each holding about a quarter of the pool's prompts. The selector takes

$$\mathcal{A} = \bigcup_{b=1}^{4} \mathrm{top}_{m_b}(\mathcal{Q}_b), \qquad m_b = \big\lfloor m\,|\mathcal{Q}_b|/N \big\rfloor + \mathbf{1}[\,b \in \mathcal{L}\,], \tag{13}$$

where $\mathcal{L}$ contains the largest quartiles receiving the remaining slots after rounding down, so that $\sum_b m_b = m$. These quotas preserve the pool's quartile proportions up to rounding, rather than its exact mean length or training-token use. Table 9 reports mean probe costs; response lengths may change during training. The teacher's training data $\mathcal{D}_{\text{up}}$ is a list of B_T presentations constructed from $\mathcal{A}$ using the study's fixed repetition and training-order rules (Appendix C.1). The selection order $\succ$ determines membership, not the order of training presentations. In the selected-set analyses of Appendix D.3.3, every admission boundary fell inside a set of prompts of equal u, so about half of $\mathcal{A}$ was settled by σ rather than by the score. Appendix C.1 gives the variant used in the original selection-stage comparison, which fixes each quartile's share of probe tokens rather than its share of prompts and leaves ties to a deterministic sort rather than to σ.

C.3 Baseline implementations and cost accounting

We compare four published selection criteria under a common teacher-training configuration. These are adaptations of the criteria, not reproductions of the complete algorithms. Residual coverage follows Lu et al. (2026) with $\tau_{\text{drop}} = 0.5$ and no solvability floor. We use four probe responses instead of sixteen, uniformly subsample to the training budget, and run one selection round instead of four. Teacher gating follows Zhang et al. (2026) and retains prompts with a reference-model pass rate of at least two thirds. Learnability follows Foster et al. (2025) and difficulty banding follows Cheng et al. (2025). The sealed student-prompt comparator applies the learnable band as a fixed pre-filter in the manner of the Qwen3 report (Yang et al., 2025). The static teacher-data comparisons share our teacher-training configuration. Each criterion has two tie-breaking replications sharing one probe matrix; the resulting method comparisons are descriptive. The separately evaluated dynamic-sampling component uses a different generation schedule, as described below.

Dynamic sampling. The dynamic-sampling condition enables group filtering in the teacher's GRPO loop (Yu et al., 2025) with a generation batch of 32 prompts and up to four regenerations per step to fill a 16-prompt training batch, on the same uniform training rows as the uniform condition, for five seeds. The training-prompt budget and optimizer steps are matched, but generation costs differ

(Table 10). Computing our selection scores costs more than this oversampling for one teacher run. Reusing those scores across runs reduces the cost per run; the performance comparisons are not normalized by this amortized cost.

Table 10: **Cost accounting.** • Teacher-training rollouts and student-training prompts matched separately. • Reported costs: disclosed, not matched or used to normalize outcomes.

Quantity	Value
Teacher-presented prompt tokens, selected condition vs. uniform mean (placement wave)	$-149{,}402$ ($-4.03\,\mathrm{SD}_U$; below every uniform draw)
Teacher prompt tokens, $\mathrm{U}_1\mathrm{D}_0$ vs. $\mathrm{U}_0\mathrm{D}_1^z$	$1{,}113{,}896$ vs. $1{,}263{,}298$ (-11.8%)
Probe response tokens to compute selection scores, teacher data vs. student prompts	$6{,}064{,}503$ vs. $14{,}311{,}640$ minimal ($2.36\times$); $24{,}061{,}509$ stored ($3.97\times$)
Teacher-scored student tokens, $\mathrm{U}_1\mathrm{D}_0 - \mathrm{U}_0\mathrm{D}_1^z$ (the pre-registered supervision currency)	$-293{,}489$, SE $271{,}744$, $t = -1.08$: not distinguishable at 5 vs 5
Dynamic sampling, realised rollouts relative to a uniform run	$2.68\times$

C.4 Pre-registration and inference

Table 11 records the original instruction-following studies and their first follow-ups. The later holdout evaluations are summarized below; study-specific designs and analyses accompany the results in Appendix D. Registration versions and amendments are retained with their hashes, including superseded versions.

Table 11: **Original instruction-following studies and first follow-ups.** • Holdouts: initially sealed, subsequently reused, or development. • Read index: seven endpoint-level reads; stored-output rescoring excluded.

Wave	Holdout	Replication	Estimator, α	Frozen before	Seal read
Uniform-draw wave, teacher and student endpoints	sealed	teachers, 3 vs 11	prediction-interval t, one-sided	training	1, 2
Placement contrast, published student-prompt score	sealed	teachers, 5 vs 5	Welch, one-sided 0.05, own family	generation	3
Placement contrast, same score	sealed	teachers, 5 vs 5	Welch, one-sided 0.05, own family	generation	4
Same-score placement, development	development	teachers, 5 vs 5	Welch, two-look 0.005 / 0.0482	scoring	—
Era-matched development matrix	development	teachers, 1 each vs 11	placement in the uniform prediction interval	scoring	—
Selection-level replication	development	selections, 3 vs 11	prediction-interval t, one-sided	training (analysis script after)	—
Direct-reference comparison	exposed IF	paired blocks, 5	paired t, two-sided 0.05	new training	5
Fresh-probe replication, teacher and student	exposed IF	pipelines, 5 vs 11	Welch; student: two-sided 0.05, teacher: secondary	probing	6, 7

Original evaluation history. The original instruction-following holdout was first used for the uniform-data comparison at the teacher and student endpoints. The selected and uniform models were trained for that evaluation. Two subsequent comparisons tested teacher-data selection against student-prompt selection using models developed beforehand. Development scores were known when those protocols were frozen. The same-score comparison also reused the selected teachers' known holdout scores; only its student-prompt comparator was newly evaluated. The two comparisons share selected teachers and are not independent replications.

These original studies account for four endpoint-level reads. Direct reference distillation adds one, and fresh-probe replication adds two, at the teacher and student endpoints: seven reads in total, or six if the two fresh-probe endpoints are counted together. Rescoring stored responses does not add a

read. Later studies are prospectively registered follow-ups on a reused holdout, with study-specific rather than study-wide error control.

New-holdout evaluations. A third 750-prompt set excluded the candidate pool, earlier holdouts, study training sequences and probe panel. Its first registered evaluation compared the fresh-probe students with uniform controls (Appendix D.1.3). The fixed-size student-prompt comparison was registered next, after the selected and uniform scores were known but before generating the additional comparator responses (Appendix D.2.5). Three further questions evaluated token-budget selection, direct reference distillation and the cutoff control (Appendices D.2.6, D.1.4 and D.1.12). The cutoff comparison is descriptive. Four later questions evaluated the fresh-probe teachers, the token-budget static-filter comparator, longer teacher training and the reference model itself. The last two are descriptive. These are nine questions on the new set, not additional reads of the original set. Later comparisons were registered before their additional responses were generated, not before all earlier results were known. The three in-domain evaluation sets are excluded from study training, but occur in the reference's reported training corpus.

Cross-question sensitivity check. Six questions contain nine primary tests: the fresh-probe teacher and student comparisons, four fixed-size placement contrasts, two token-budget placement contrasts, and direct-reference distillation. In a retrospective Bonferroni check over all nine tests, both fresh-probe gains, both fixed-size placement gains and both token-budget placement gains remain below .05 (Table 12). This supplements rather than replaces each study's registered analysis. It covers the listed primary tests, not secondary endpoints or future reads, and cannot remove the effect of prior knowledge on the later study designs.

Table 12: **New-holdout multiplicity check.** Two-sided nominal p-values and retrospective Bonferroni correction, $\min(1, 9p)$, across the nine listed primary tests. The fixed-size study also has a registered four-test Holm analysis (Appendix D.2.5).

Comparison	Nominal p	Bonferroni p
Fresh-probe students vs uniform-data students	.0000544	.000489
Fresh-probe teachers vs uniform-data teachers	.00000909	.0000819
Fixed size: teacher data vs student gap score	.000329	.00296
Fixed size: teacher data vs student static filter	.0000294	.000265
Fixed size: student gap score vs uniform	.495	1
Fixed size: student static filter vs uniform	.476	1
Token quota: teacher data vs student gap score	.00190	.0171
Token quota: teacher data vs student static filter	.00339	.0306
Selected teacher vs direct reference	.579	1

Estimator sensitivity. The original studies registered different tests. Single-run prediction thresholds also follow their registered conventions: the original and development comparisons use one-sided 95% upper bounds, whereas the new-holdout teacher and student comparisons use upper endpoints of two-sided 95% intervals. These thresholds are not interchangeable; the result sections identify the convention used. Welch compares condition means while accounting for variation in both conditions. The predictive test compares a selected result with the distribution of a further uniform draw. Table 13 reports both tests and identifies the registered analysis for each study.

In the original uniform-data comparison, three training orders share one selected corpus. Including uncertainty in their mean in the predictive comparison gives the statistic $\Delta/\sqrt{\mathrm{SE}_{\mathrm{sel}}^2 + \mathrm{SD}_U^2(1 + 1/11)}$. With Welch–Satterthwaite degrees of freedom, the one-sided p-values are .037 for the student and .060 for the teacher; using the original ten degrees of freedom gives .028 and .050. By contrast, including this uncertainty in the development selection-level replication changes p only from .011 to .012. These checks concern the original studies. The independent-probe replication in Section 4.2 has its own registered analysis.

Table 13: **Estimator sensitivity.** • One-sided p in the registered direction; registered estimator identified per study. • Welch: difference of means; predictive: comparison with a further uniform draw.

Contrast	Effect	Welch, condition means	Prediction-interval t
Placement contrast, published student-prompt score, 5 vs 5, student endpoint			
primary	3.9%	**0.004** (reported)	0.064
raw base, no floor	4.7%	0.008	0.100
raw base, $\geq$ 5-word floor	2.6%	0.073	0.194
Placement contrast, same score, 5 vs 5, student endpoint			
primary	3.5%	**0.007** (reported)	0.082
raw base, no floor	4.5%	0.007	—
raw base, $\geq$ 5-word floor	1.9%	0.133	—
Uniform-draw wave, 3 vs 11, student endpoint			
primary	5.8%	0.045	**0.005** (reported)
same wave, teacher endpoint	5.7%	0.061	**0.012** (reported)
raw base, no floor	5.6%	0.054	**0.006** (reported)
raw base, $\geq$ 5-word floor	6.8%	0.021	**0.003** (reported)

D ADDITIONAL EXPERIMENTAL RESULTS

Tables distinguish uniform sampling at both stages, $\mathrm{U}_0\mathrm{D}_0$; teacher-data selection, $\mathrm{U}_1\mathrm{D}_0$; student-prompt selection, $\mathrm{U}_0\mathrm{D}_1$; and selection at both stages, $\mathrm{U}_1\mathrm{D}_1$. Superscripts z and u identify the static learnability filter and gap score, respectively.

We report comparisons with uniform sampling and student-prompt selection, followed by adapted selection criteria, design details, and multi-teacher evaluations. Unless a table explicitly reports percentages, effects are differences in pass rate on the $[0, 1]$ scale; p is one-sided in the pre-registered direction unless stated.

D.1 TEACHER-DATA SELECTION AGAINST UNIFORM SAMPLING

D.1.1 ORIGINAL COMPARISON WITH UNIFORM TEACHER DATA

Table 14: **Original sealed-holdout comparison.** • Teacher and student endpoints; three orders of one selected corpus versus eleven uniform draws. • Prediction-interval t, 10 df. • Primary: guarded; length-floor comparisons: raw strict base only.

Endpoint	Metric	Selected mean	Members	Uniform mean (SD)	Δ	Δ/SD_U	p
teacher	primary	**0.4332**	0.4463, 0.4650, 0.3883	0.3758 (0.0208)	+0.0574	2.77	0.0122
teacher	raw, no floor	—	—	—	+0.0570	—	0.0120
teacher	raw, $\geq$ 5 words	**0.4011**	0.4067, 0.4280, 0.3687	0.3334 (0.0209)	+0.0677	3.25	0.0055
student	primary	**0.4354**	0.4329, 0.4709, 0.4024	0.3771 (0.0175)	+0.0582	3.32	0.0049
student	raw, no floor	—	—	—	+0.0556	—	0.0060
student	raw, $\geq$ 5 words	**0.4015**	0.3978, 0.4315, 0.3752	0.3336 (0.0182)	+0.0679	3.74	0.0025

The student retains the teacher's improvement: the ratio of student to teacher gains is $0.0582/0.0574 = 1.014$. From an initial score of 0.2737, uniform teacher training adds 0.1021, and selection adds a further 0.0574, or 56% of the uniform-training gain. This is a comparison of means: four of the eleven uniform draws (0.3920, 0.3920, 0.3923, 0.4040) exceed the weakest selected teacher.

Response lengths are similar across conditions: selected teachers write a median of 63.0 words against 60.36 for uniform draws ($p = 0.440$), their students 64.17 against 62.18 ($p = 0.459$), and student response tokens are 252.0 against 293.09 ($p = 0.735$). External student-score differences are unresolved on the 0 to 100 scale: composite $+0.87$ ($p = 0.310$), IFEval $+1.29$ ($p = 0.305$), IFBench $+0.46$ ($p = 0.420$), AIME -1.17 ($p = 0.612$).

D.1.2 INDEPENDENT PROBE-TO-STUDENT REPLICATIONS

We repeat the complete selection pipeline five times: fresh four-response probes of both the initial student and reference, the frozen 640-prompt fixed-size selector, teacher training, and student distillation. The 10,000-prompt pool, training budgets and student prompt sequence are unchanged. Each selection has one teacher and one student. Configuration checks confirm the original training recipe, including two runs restarted after infrastructure failures. Rescoring reproduces all seventeen original student records. The comparison reuses the eleven existing uniform-data teachers' students.

All five replications specified before probing are completed. The original selected students are shown separately and are not pooled into the test. This registered follow-up uses the previously evaluated holdout, with study-specific error control.

Table 15: **Independent probe replications.** • Student scores (%); five selections; eleven uniform controls. • Overlap: prompts shared with the original 640-prompt corpus. • Raw-5: unguarded strict correctness, minimum five words; individual rows: descriptive.

Replication	Overlap	IF Primary	Raw-5	IFEval	IFBench	AIME
1	167	44.87	40.00	73.65	28.29	12.50
2	178	44.77	40.67	74.21	27.89	13.13
3	164	44.03	40.10	73.65	23.90	12.29
4	167	45.37	41.83	74.21	21.91	1.04
5	168	44.00	40.37	74.21	25.90	3.54
Uniform mean	—	37.71	33.36	72.68	22.78	**10.89**
Fresh-probe mean	—	**44.61**	**40.59**	**73.99**	**25.58**	8.50

The registered Welch comparison gives a 6.90% primary gain (two-sided $p = 1.97 \times 10^{-8}$). All five students exceed the uniform mean and its predeclared single-run prediction threshold of 41.03%. The Raw-5 difference is 7.24%. At the teacher endpoint, the five replications average 44.29% versus 37.58% for the eleven uniform-data teachers, a 6.71% gain. All five exceed the 41.51% single-run prediction threshold; this teacher-endpoint analysis is secondary, without a separately allocated significance level. Secondary external differences are +1.31 on IFEval, +2.80 on IFBench, and −2.39 on pooled AIME; none is distinguishable from zero, and no separate significance level is allocated to these outcomes. The five selections share only 164–178 prompts with the original corpus, but retain the reference-success/student-difficulty pattern. Their student SD is 0.59%, combining selection, teacher-training and student-training variation rather than isolating any one source. It is smaller than the 3.43% spread of the original corpus's three training orders, whose third order fell near the uniform bound. The follow-up holds the pool, initialization, budgets, teacher fill order and student prompt sequence fixed while varying probe seeds and retraining the teacher and student for each new selection.

D.1.3 EVALUATION ON A PREVIOUSLY UNUSED HOLDOUT

Teacher-data selection also improves performance on prompts unused by the study before this evaluation. After seven endpoint-level reads of the original holdout, we drew a third 750-prompt set on 25 September. Row and content checks exclude the 10,000-prompt candidate pool, both earlier holdouts, all study training sequences and the probe panel. The prompts occur in the reference's reported RL training corpus. Prefix caching is disabled, unlike the earlier evaluations, so comparisons use scores within this holdout.

The registered primary compares five fresh-probe students with eleven uniform-teacher students using a two-sided Welch test at $\alpha = .05$. Original-corpus students, length thresholds and response lengths are descriptive analyses. Before the first read, checks examined output counts and indices, but not responses or scores.

Table 16: **Teacher and student evaluation on the new holdout.** Scores (%); SDs across models in parentheses. Students and teachers were evaluated in separate registered reads. Raw-5 requires at least five words; Rel-25/50 require raw passes to reach 25%/50% of the initial checkpoint's per-prompt median length. Words are means of model-level medians.

Teacher data	Models	IF Primary (SD)	Raw-5	Rel-25	Rel-50	Words
Student endpoint						
Uniform teacher data	11	35.5 (1.8)	33.0	29.2	21.0	71
Selected, original corpus, three orders	3	40.4 (3.3)	37.9	**32.0**	**21.5**	66
Selected, five fresh probes	5	**41.2** (1.4)	**38.5**	30.8	20.1	63
Initial checkpoint, untrained	—	25.7	23.9	—	—	134
Teacher endpoint						
Uniform teacher data	11	35.6 (1.6)	33.0	29.2	**21.2**	70
Selected, five fresh probes	5	**41.5** (1.2)	**38.6**	**32.0**	20.7	64

The registered criterion is met: the fresh-probe students score 41.2% against 35.5% (+5.7%, Welch $t = 6.70$ on 10.0 degrees of freedom, two-sided $p < .0001$), and every one of the five lies above the upper endpoint of the two-sided 95% single-run prediction interval, 39.8%; the prediction-interval test agrees ($t = 5.72$ on 10 degrees of freedom, $p < .001$). The gain holds under the five-word floor (+5.4%) but is unresolved under the relative floors (+1.6% and −0.9%, both $p > .48$), as on the earlier holdouts; the fresh-probe students answer in 63 words against 71, and the tersest of them has the highest guarded score and the lowest score at the half-length floor. The original corpus's three orders score 40.4%, 4.8% above uniform data, with the same third-order dip as before. The read cost 4.8 GPU-hours.

The teacher improvement transfers to the student. A separately registered evaluation of the same five fresh-probe teachers and eleven uniform-data teachers gives 41.52% versus 35.62% (SDs 1.22% and 1.61%; Welch two-sided $p = .00000909$). Every selected teacher exceeds the upper endpoint of the corresponding two-sided 95% single-run prediction interval, 39.4%. The student gain divided by the teacher gain is 0.96, a descriptive ratio rather than a causal estimate of transfer. Under Raw-5 the teacher scores are 38.65% and 33.04%, a 5.61% gain. The teacher differences at the relative-length thresholds are +2.83% and −0.43%, both unresolved. This evaluation was registered after the student result was known and adds no trained models.

Reference-model context. A descriptive evaluation of the reference itself gives 39.23% on IF Primary, 37.53% on Raw-5, 37.20% on Rel-25 and 30.67% on Rel-50. Its median response has 102 words. Its primary score lies between the uniform and selected students, but this comparison does not test which model provides better distillation supervision; that question is evaluated directly below.

D.1.4 Direct distillation from the reference

We compare selected-teacher students with students distilled directly from the available instruction-following reference. Five blocks fix the initial student, student prompt sequence, training recipe, 2,560 student presentations and evaluation contract, changing the supervising teacher. The selected teachers are five training orders of one corpus, not five independent selections. One selected student is reused under the preregistered rule; the other nine students are newly trained from scratch. Analysis pairs the two students within each block.

Table 17: **Selected teacher versus direct reference.** • Final-student scores and absolute differences (%); five paired blocks. • Primary outcome: IF Primary. • Raw-1/5: common unguarded strict base, with one-/five-word minima.

Task / metric	Selected teacher	Direct reference	Difference
IF Primary	**43.91**	43.41	+0.50
IF Raw-1	**46.02**	43.45	+2.57
IF Raw-5	37.08	**40.77**	−3.69
IFEval	73.17	**74.43**	−1.26
IFBench	**27.81**	22.79	+5.02
AIME, pooled	4.53	**13.33**	−8.80

The paired primary differences are -2.03%, -0.30%, -0.50%, $+2.67\%$, and $+2.67\%$ (two-sided $p = .620$). The primary difference is unresolved. Selected-teacher students produce shorter responses, with the average of their per-model median lengths 64 words lower. Their Raw-1 advantage reverses under Raw-5. The external IFBench gain is descriptive, alongside the IFEval and AIME decreases. Direct-reference students also have higher AIME accuracy and longer, frequently capped responses.

The holdout is previously exposed, and all 750 evaluation prompts, as well as all 2,560 consumed prompts of every student fill, occur in the reference's reported training corpus. Corpus membership was verified; actual reference-training exposure relies on the publisher's description. Selected-teacher and student training sequences exclude the evaluation prompts. In the same scoring run the reference itself scores 74.21 on IFEval and 23.90 on IFBench, between the two student conditions on IFEval and below the selected-teacher students on IFBench. Direct use of the released reference requires no new teacher training; the five selected teachers cost 18.7 L40S-hours in the recorded accounting. Thus matched student training does not imply equal total compute.

Follow-up on the new holdout. The same five paired blocks were evaluated on the third 750-prompt holdout, with the analysis registered before their new responses were generated. The selected-teacher minus direct-reference differences are $+0.60\%$, -0.16%, -1.30%, $+1.46\%$ and $+0.83\%$, averaging $+0.29\%$ (paired t, two-sided $p = .579$). The primary comparison remains unresolved. Raw strict correctness favors selected-teacher students by 2.49%, but the five-word score favors direct-reference students by 3.65%. Selected-teacher students have per-model median lengths of 24 to 55 words, compared with 97 to 108 for direct-reference students. At the quarter- and half-length thresholds, the differences are -13.71% and -15.90%. This fourth registered question uses one reference teacher across all blocks, with the same reference-corpus exposure described above.

D.1.5 A CROSS-FAMILY REFERENCE

Replacing the reference with Qwen2.5-7B-Instruct (Team, 2024) gives an unresolved student improvement over uniform teacher data (Table 18). The registered comparison keeps the probing procedure, score, fixed-size selector, pool and training budgets unchanged. One selected corpus is trained in three orders, with one student each, and compared with eleven uniform-teacher students on the development holdout. Under the training verifier this reference is on average weaker than the student: it passes 26.0% of the pool against the student's 27.5%, exceeds the student on 19.4% of prompts, and scores 23.0% on the holdout against 26.4% for the untrained student. The selector still finds 640 prompts on which it is strong (85.6% against 17.6% for the student), but judged by the same-family reference that corpus is lower-scoring than the original ($u = 0.44$ against 0.83; 144 of 640 prompts shared). This comparison changes reference competence as well as model family.

Table 18: **Reference-model comparison.** • Development student scores (%); teacher SDs in parentheses. • Same-family teacher values: means of two students. • Raw: unguarded strict; Raw-5: minimum five words. • Words: mean of student medians; revoked: format-guard revocation rate.

Teacher data	IF Primary (SD)	Per teacher	Raw	Raw-5	Words	Revoked
Cross-family reference, three orders	39.0 (3.4)	37.2 / 36.9 / 43.0	43.7	30.9	18	10.8%
Same-family reference, three orders	43.5 (2.6)	43.8 / 46.0 / 40.8	44.2	40.7	61	1.6%
Same-family, five fresh probes	**44.4** (0.8)	43.7–45.5	**45.2**	**41.2**	53	1.8%
Uniform, eleven teachers	37.7 (1.7)	—	38.6	34.3	61	2.4%
Cross-family reference itself	23.0	—	23.7	20.1	60	3.1%
Initial checkpoint	26.4	—	26.6	24.3	121	0.8%

The registered contrast is not distinguishable from zero: cross-family minus uniform is $+1.3\%$ with a Welch test on 2.3 degrees of freedom (two-sided $p = .58$). On unguarded strict correctness the cross-family students lead by 5.1%, but two of the three answer in 12 and 11 median words and the guard revokes 14.7% and 14.6% of their passes against 0.9 to 8.0% for uniform students, and under the five-word floor the difference is -3.4%, the same pattern as direct distillation from the reference (Appendix D.1.4). One of the three students clears the uniform one-draw bar of 40.9; the between-order spread (3.4%) is twice the between-draw spread, as in the original comparison. The

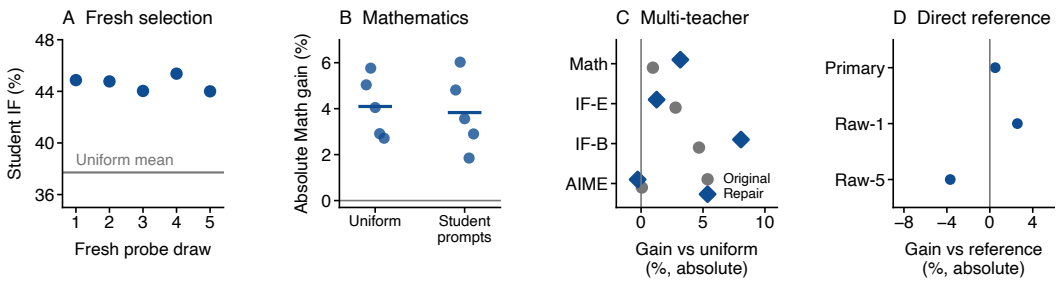


Figure 4: **Follow-up comparisons.** • A: five fresh-probe students; line: eleven reused uniform controls. • B: five paired mathematics differences against uniform and student-prompt selection. • C: four multi-teacher differences from uniform, original and rendering-matched runs. • D: paired direct-reference differences; primary test on Primary only. • Differences: absolute %.

secondary contrasts against the same-family selections are -4.5% and -5.4%, neither distinguishable from zero. The experiment uses 23.6 GPU-hours and does not evaluate the original holdout.

A pre-registered repeat with a stronger reference stopped at its strength gate: on a 1,000-prompt slice of the pool the 14B model of the same family passes 28.2% against 25.4% for the student on the same prompts and 42.0% for the same-family reference, a margin below the required 5% absolute, so no teacher was trained and a cross-family reference stronger than the student under this verifier remains untested.

D.1.6 Length sensitivity on the original comparisons

A CPU-only analysis reuses the original generated responses and fixed-denominator scores. Absolute floors require at least one or five words. Relative floors require at least 10%, 25% or 50% of the initial student's median response length for that prompt; they can reject valid concise answers and are sensitivity measures, not corrected truth labels. Exploratory standardization reweights length-bin pass rates to a common length distribution. Because length is affected by training, this is descriptive rather than a causal adjustment.

Table 19: **Length sensitivity of the original comparisons.** • Selected minus comparator (%). • Raw-1/5: common unguarded base. • Standardized: exploratory equal-group reference weights.

Comparator	Primary	Raw-1	Raw-5	Rel-10	Rel-25	Rel-50	Standardized
Uniform, original study	5.82	5.56	6.79	6.46	5.55	2.71	5.72
Static band on student prompts	3.91	4.74	2.63	0.85	−1.97	−4.02	3.50

The original selected-versus-uniform gain remains positive under every reported floor. The placement gain shrinks under absolute floors and reverses at the two larger relative floors, while exploratory standardization retains most of its primary point estimate. Figure 4 summarizes the independent replication and comparison results.

D.1.7 Length-controlled re-scoring of the development results

We re-score 26 registered development contrasts from stored responses, covering both-stage, cutoff-only, fixed-size, quarter-budget, cross-family and training-budget comparisons. The original primary scores reproduce within 5×10^{-4}. Five-word thresholds are applied to the guarded score, and relative-length thresholds to raw strict correctness. Both test directions are recorded; these sensitivity analyses receive no additional confirmatory significance level. Table 20 summarizes the main contrasts. Under the five-word floor the sign survives in 25 of 26 and the unadjusted 5% test decision is unchanged in 22. Under thresholds of 25% and 50% of the initial student's per-prompt median response length, the sign survives in 16 and 12 of 26, respectively.

Table 20: **Development results under length thresholds.** • Absolute differences (%); primary: guarded strict; w5: guarded strict, minimum five words. • Rel-25/50: raw strict, minimum 25%/50% of initial-student per-prompt median length; sensitivity analyses. • Δwords: difference in median length; *: unadjusted two-sided $p < .05$.

Contrast	Primary	w5	Rel-25	Rel-50	Δwords
Fresh probes minus uniform (five draws)	$+6.7^*$	$+7.0^*$	$+2.7$	-0.1	-8
Fixed-size placement, teacher data minus gap score on student prompts	$+6.7^*$	$+7.1^*$	$+3.2$	-0.1	-11
Fixed-size placement, teacher data minus static band	$+6.8^*$	$+7.0^*$	$+3.3$	$+0.0$	-13
Tie-band control minus uniform bar	$+5.0^*$	$+5.5^*$	$+4.2$	$+2.2$	$+2$
Quarter budget, teacher data minus uniform	$+6.8$	$+7.2^*$	$+3.9$	$+0.6$	-8
Both stages minus student prompts only	$+5.9^*$	$+3.5$	-1.9	-3.5	-23
Uniform at 120 steps minus uniform at 40 (students)	$+13.4^*$	$+9.2$	-6.4	-8.5^*	-40
Selected at 120 steps minus selected at 40	$+4.6$	$+0.4$	-7.6	-7.5	-32
Selected at 120 steps minus uniform at 120	-3.0	-2.3	$+3.9$	$+3.6$	$+8$
Cross-family reference minus uniform	$+1.3$	-3.8	-11.7	-11.0^*	-43

Fresh-probe and fixed-size selection gains retain their size and test decision with the five-word threshold, but shrink under relative-length thresholds. The 13.4% gain from training uniform-data teachers for 120 rather than 40 steps becomes unresolved with the five-word threshold and reverses under both relative thresholds. The selected corpus's gain from longer training also falls near zero with the five-word threshold. Both longer-training conditions produce responses about 30 to 40 words shorter. Cross-family-reference students answer 43 words more briefly and lose their unresolved lead under every threshold. The analysis uses no new GPU computation; its implementation was checked on the cross-family and quarter-budget comparisons before the full re-scoring.

D.1.8 Single-teacher mathematics: registered five-block comparison

Teacher-data selection improves mathematics accuracy over both comparators in all five training blocks (Table 21). The evaluation uses a reserved 2,000-prompt split, identified as `mathAntE` in the experiment records, rather than a public benchmark under that name. The registered construction takes 2,000 rows starting at offset 4,000 of the reserve partition of the mathematics candidate pool. Content-identity checks exclude teacher-training prompts, student-training prompts, the development split used to choose generation settings, and the separate confirmation split. Each prompt receives four sampled responses with a 4,096-token cap and thinking disabled. The frozen grader extracts the boxed answer and checks mathematical correctness, using the same rule as the selection probes. All fifteen student evaluations contain 8,000 responses and zero scorer faults. The mixed-SFT initialization scores 17.74%. The mathematics reference is the released RL-Math specialist. Selection uses a 9,996-prompt teacher candidate pool and one fixed 640-prompt selected corpus. Both student-prompt conditions draw from the same disjoint 8,997-prompt pool. The student-prompt filter retains prompts with $0 < p_S < 1$ over four probes; it is a static learnability-band baseline, not online DAPO.

Within each block, selected- and uniform-teacher students share their uniform student prompt sequence. The two conditions using a uniform teacher share that teacher and differ in student prompt selection. Teacher and student training budgets are matched separately, at 640 and 2,560 prompt presentations. Uniform teacher corpora vary across blocks; the selected corpus and probe remain fixed. Thus the five blocks measure training-outcome variation conditional on this selection, not variation over independently selected corpora.

Table 21: **Single-teacher mathematics.** • Student accuracy and absolute differences (%) on the reserved mathematics split. • Student-prompt baseline: static learnability filter with a uniform-data teacher. • Means and differences calculated before rounding.

Block	Selected teacher	Uniform teacher	Student prompts	Selected − uniform	Selected − student prompts
1	**30.90**	25.86	26.09	+5.04	+4.81
2	**31.47**	25.71	25.45	+5.76	+6.02
3	**31.21**	27.16	27.65	+4.05	+3.56
4	**30.98**	28.06	29.12	+2.91	+1.85
5	**31.31**	28.60	28.41	+2.71	+2.90
Mean	**31.175**	27.080	27.345	+4.095	+3.830

Both planned contrasts are positive in all five blocks. The registered two-sided paired t tests give $t(4) = 6.93$, $p = 0.0023$ against uniform teachers and $t(4) = 5.25$, $p = 0.0063$ against student-prompt selection. Both reject under Holm correction over the two comparisons at $\alpha = 0.05$ (adjusted $p = 0.0045$ and 0.0063). The selected students' response-cap rates range from 8.2% to 32.4%, compared with 2.3%–48.8% for uniform-teacher students and 2.1%–43.6% for student-prompt selection. These response-cap differences accompany the accuracy results. Infrastructure failures were recovered under the recorded amendments before evaluation; all five blocks are included.

The earlier teacher-level gate used three training orders of the selected corpus and four uniform teachers. Their means are 30.57% and 27.38%: a 3.19% difference, or 1.73 uniform-teacher SDs, exceeding the prespecified 1.0-SD threshold. Its one-sided prediction-interval test gives $p = 0.110$. Passing this design gate is not a significant teacher-endpoint result and is not the test used for the student comparison above.

D.1.9 Replication over tie-breaking selections

Table 22: **Tie-breaking replications.** • Three corpora; shared rule, pool, probes, support and quota; one teacher each. • Top: overlap with original corpus. • Middle: comparison with eleven uniform draws, prediction-interval t, 10 df. • Bottom: descriptive teacher-score dispersion, with corpus and training effects unseparated.

Selection	Overlap	Mean u	Cost	Primary	Raw, ≥ 5 words	Median words
A (k_{101})	478/640 (74.7%)	+0.8254	442	0.4347	0.4117	66
B (k_{102})	457/640 (71.4%)	+0.8254	439	0.4383	0.4077	62
C (k_{103})	468/640 (73.1%)	+0.8254	439	0.4277	0.3807	34
frozen corpus, 3 orders	—	+0.8254	444	0.4327–0.4660	0.3897–0.4080	24–46

Metric	Set mean	Uniform mean	Δ	Δ/SD_U	t	p
primary	**0.4336**	0.3764	+0.05716	2.81	2.6944	0.01126
raw base, no floor	0.4431	—	+0.05296	2.04	1.9542	0.03960
raw base, ≥ 5 words	0.4000	—	+0.06258	3.25	3.1098	0.00553

Source of variation in the teacher primary	SD	Comparison
between re-selected corpora (n=3)	0.00539	vs. orders: $F = 9.69$ on $(2, 2)$, $p = 0.094$
between training orders of one corpus (n=3)	0.01678	vs. draws: $F = 14.19$ on $(10, 2)$, $p = 0.068$
between uniform draws (n=11)	0.02030	

All three tie-reselected teachers improve on the uniform mean (Table 22). The seeds, estimator, primary outcome, test direction and length-check sign condition were fixed before training. The analysis script was written after scores existed, using the previously validated selection-stage analyzer. Rescoring reproduced all eleven uniform controls and three original training orders. This comparison uses the development holdout only.

Final-student transfer from the reselected corpora. We subsequently distill one student from each of these three teachers, using 40 updates and $2{,}560$ student prompt presentations. Evaluation uses the 750-prompt development holdout with four responses per prompt. The selected-student mean is 43.53%, exceeding the eleven historical uniform controls by 5.78% and a fresh student trained from one existing uniform teacher by 5.17% (Table 23). The guarded five-word floor retains positive differences of 5.73% and 4.67%, respectively.

Table 23: **Students of tie-reselected teachers.** • Development scores (%); one teacher and student per corpus. • Historical uniform: eleven teachers; fresh uniform student: one reused teacher. • Raw-5/Guarded-5: five-word minima on raw strict/Primary. • Words: individual student medians.

Student	Primary	Raw strict	Raw-5	Guarded-5	Loose	Median words
Reselected A	**44.83**	**45.83**	**42.37**	**41.90**	**52.30**	57
Reselected B	43.63	44.10	40.43	40.37	50.90	71
Reselected C	42.13	44.83	36.33	35.63	49.93	30
Reselected mean	**43.53**	**44.92**	**39.71**	**39.30**	**51.04**	—
Historical uniform mean	37.75	38.93	33.89	33.57	44.07	—
Fresh uniform-teacher student	38.37	38.93	34.83	34.63	44.57	61

The candidate pool, reference and probe matrix are unchanged. With one teacher and student per corpus, the observed variation combines selection and training effects. This development comparison is descriptive, with no new confirmatory test. Its mean Guarded-5 score equals the historical teacher-gating student's 39.30%.

D.1.10 Sensitivity to probe sampling

Repeating the four-response probe draw at the same reference checkpoint, tie key, support budget, and per-datum cost produces a selected corpus sharing 291 of 640 prompts with the original (45.5% overlap). Mean score is 0.756 versus 0.759, and mean probe cost is 439.5 versus 438.8. Training three orders per selected corpus yields a final-student difference of -2.62% absolute on the development instrument, unresolved conditional on these two draws. No external endpoint resolves a difference. There is only one probe draw per condition, so the training-order replicates do not provide independent probe-level replication. The comparison demonstrates sensitivity of selected-set membership to probe sampling, but leaves its effect on final-student performance unresolved.

Fresh responses retain a large reference-student gap. With eight new responses per prompt, the gap is 0.476 on the 640 selected prompts, compared with 0.143 on a 640-prompt panel outside all training sets. This is smaller than the original selection estimate of 0.825, but remains 0.33 above the panel. Both models contribute to the reduction: reference success falls from 0.924 to 0.748, while initial-student success rises from 0.099 to 0.272. These changes are 0.176 and 0.174, respectively.

The fresh measurements use the training verifier with prefix caching disabled; re-scoring the original responses reproduces the selection matrix exactly. The initial student's fresh mean is 0.272 on selected prompts and 0.285 on the panel. Selected prompts include fewer all-failure cases (22% versus 43%) and fewer all-success cases (0.6% versus 6.9%). In the original four-response selection probes, 62% of selected prompts had no student successes. Thus the selected set concentrates on prompts the student sometimes solves, rather than being uniformly harder than the pool. On three uniform draws and the panel, original and fresh student rates agree within 0.02.

For the reference, 27% of selected prompts pass all eight fresh responses, compared with 71% passing all four original responses. Its panel mean changes from 0.416 to 0.428, with per-prompt correlation 0.90. The original selected-set advantage over the panel was 0.70; the smaller fresh estimate accounts for selection on noisy probes.

The same evaluation measures three selected teachers and three uniform teachers. Their success rates are 0.65 to 0.80 versus 0.39 to 0.42 on their own training prompts, and 0.39 to 0.48 versus 0.39 to 0.40 on the comparison panel. These are descriptive measurements of three orders of one selected corpus and three uniform draws. Appendix B.9 relates them to the theoretical quantities (Table 6).

D.1.11 REFERENCE SUCCESS WITHOUT THE STUDENT GAP

Ranking by reference success alone gives a lower student score than ranking by the success gap between reference and student in one development comparison. Both selections use the same probe draw and length groups. The reference-only student scores 40.33%, versus 45.53% for the gap-score student and 37.71% for the eleven uniform controls. Raw-5 scores for the paired selections are 37.03% and 41.70%. Reference-only selection admits prompts the initial student already solves more often: its mean probed success is 65.9%, versus 8.8% under gap selection. The gap score therefore selects data with more measured room for student improvement. This development comparison has one teacher and student per selection, no confirmatory test, and different response lengths (65 versus 32 median words).

D.1.12 SAMPLING ABOVE THE SCORE CUTOFF WITHOUT RANKING

Sampling above the score cutoff improves student performance over uniform data, while its difference from full ranking is unresolved (Table 24). At $k = 4$, ties determine 310 of the 640 selected prompts. The registered control keeps the four length groups and their quotas but fills each group by a uniform draw from every prompt at or above that group's score boundary ($u \geq 0.50$ in the first group, ≥ 0.75 in the others), under three seeds, so the ranking and tie-break inside each group are removed. Teacher and student training match the original uniform-data comparison, with the trained teacher as the replication unit and development student performance as the outcome.

Table 24: **Cutoff-only selection control.** • Development student scores (%); teacher-level means and SDs. • Ranked selection: two students per teacher; other conditions: one. • Raw-5: raw strict, minimum five words; words: mean of student medians.

Teacher data	IF Primary (SD)	Per teacher	Raw-5	Words
Ranked selection, three orders	**43.5** (2.6)	43.8 / 46.0 / 40.8	**40.7**	61
Tie-band control, three seeds	42.7 (2.1)	42.7 / 40.6 / 44.8	39.8	63
Uniform, eleven teachers	37.7 (1.7)	—	34.3	61
Initial checkpoint	26.4	—	24.3	121

The two registered comparisons are Holm-corrected. The control clears the uniform prediction bar: its mean of three draws against the eleven uniform teachers gives $t = 4.62$ on 10 degrees of freedom, one-sided $p = .0005$, and two of its three teachers clear the one-draw bar of 40.9 on their own. Against the ranked selection the control is 0.8% lower (Welch $t = -0.43$, 3.8 degrees of freedom, two-sided $p = .69$). The detectable difference at 80% power is 7.2%, larger than the 5.8% ranked-selection gain over uniform data. The comparison therefore leaves the additional value of ranking unresolved. The five-word floor preserves the ordering.

Unlike ranking, the control may omit prompts strictly above the cutoff. Its mean score is 0.76 rather than 0.83, and its corpora share 357 to 365 of 640 prompts with the ranked corpus. Evaluation uses the development set on which the ranked rule was chosen. The models are those evaluated on the original holdout in Table 1, but these are separate evaluations on different prompts. Supplementary records give per-student scores and evaluation identities. The experiment uses 22.4 GPU-hours with the same hardware and training configuration as the original comparison.

Follow-up on the new holdout. A pre-declared descriptive evaluation of the same three cutoff-control students gives 40.58%, against 35.54% for the eleven uniform students and 40.37% for three ranked-selection students. The cutoff-control scores are 41.07%, 38.97% and 41.70%; two exceed the uniform single-run prediction threshold. Under the five-word floor, the means are 38.33%, 33.05% and 37.88%, respectively. Median response lengths average 73, 71 and 66 words. This fifth registered question on the new holdout assigned no significance test or alpha. The uniform and ranked-selection scores were already known before registration. These results reproduce the cutoff control's positive difference from uniform data without a confirmatory comparison against ranked selection.

D.1.13 UNIFORM TEACHER DATA WITH THREE TIMES THE TRAINING

Longer teacher training improves the final student with either uniform or selected data (Table 25). A registered control trains three uniform-data teachers for 120 steps on independent 1,920-prompt draws, keeping the remaining teacher and student settings unchanged. One student is trained per teacher and compared with the eleven uniform 40-step students on the development holdout. Selected-data students are included for context. The selection probes cost about 16 times a 40-step teacher's rollout tokens, so this longer-training control still uses fewer rollout tokens than probing for one selection.

Table 25: **Teacher-training budget and corpus-size comparisons.** • Development student scores (%); teacher-level means and SDs. • Selected-teacher values: means of two students. • Raw-5: raw strict, minimum five words; words: mean of student medians. • IFEval/IFBench: public strict prompt accuracy; AIME: accuracy over eight samples.

Teacher data	Student (SD)	Per teacher	Teacher	Raw-5	Words	IFEval	IFBench	AIME
Uniform, 120 steps (three teachers)	51.1 (1.2)	52.0 / 51.7 / 49.7	**50.6**	44.3	21	**77.2**	**34.1**	3.6
Selected, 1,920 prompts, 120 steps (three orders)	**51.5** (1.1)	51.8 / 50.3 / 52.5	—	**46.3**	28	—	—	—
Selected, 120 steps (three orders)	48.1 (4.5)	51.8 / 49.3 / 43.1	49.1	41.6	29	76.7	32.1	10.6
Selected, 40 steps (three orders)	43.5 (2.6)	43.8 / 46.0 / 40.8	43.3	40.7	61	74.0	23.2	9.7
Fresh-probe selections, 40 steps (five teachers)	44.4 (0.8)	43.7–45.5	—	41.2	53	74.0	25.6	8.5
Uniform, 40 steps (eleven teachers)	37.7 (1.7)	—	37.2	34.3	61	72.7	22.8	**10.9**
Initial checkpoint	26.4	—	—	24.3	121	70.9	15.5	12.9

Uniform-data students gain 13.4% from the longer teacher training (Welch $t = 15.3$, 4.2 degrees of freedom, two-sided $p < .001$), exceeding the 5.8% selection gain at 40 steps. Responses are shorter, with a median of 21 words rather than about 61. With the five-word threshold, scores are 44.3% for longer-trained uniform teachers and 40.7% for selected teachers at 40 steps; the comparison with uniform data at 40 steps is unresolved ($p = .064$, 2.2 degrees of freedom). Relative-length thresholds reverse that latter difference to -6.4% and -8.5% at 25% and 50%.

On IFEval and IFBench, the longer-training condition scores higher, but none of the four registered contrasts passes Holm correction (smallest two-sided $p = .013$, threshold .0125). Its AIME accuracy is 3.6%, compared with 9.7% for selected-data students and 10.9% for uniform-data students at 40 steps. AIME responses also shorten to a median of 49 words from about 102.

Repeating the original 640 selected prompts for 120 steps gives a student score of 48.1%, versus 51.1% for uniform data at the same budget. The -3.0% difference is unresolved (Welch $t = -1.13$, 2.3 degrees of freedom, two-sided $p = .36$; detectable effect at 80% power, 13%). IFEval and IFBench scores are 76.7% and 32.1%, versus 77.2% and 34.1% for uniform data, with neither contrast passing Holm correction. The gain over selected data at 40 steps is 4.6% ($p = .22$). Median responses shorten to 29 words, including one run at 8 words, while AIME accuracy remains 10.6%. This comparison changes corpus repetition as well as selection; the expanded-corpus comparison below matches the number of distinct prompts.

The uniform 120-step condition uses 3.0 times the GPU-hours and 1.7 times the response tokens of uniform 40-step training. Three rather than four planned teachers were completed within the budget, costing 39.2 GPU-hours; the repeated selected-data condition costs 33.0 GPU-hours. Hardware and resolved training configurations are otherwise matched. Selection used this development holdout during method development and incurs a separate probing cost, so the matched-training gains are not an equal-total-compute comparison.

With diversity matched as well, three orders of the 1,920 highest-scoring prompts trained for 120 steps score 51.5% against 51.1% for uniform data at 120 steps (+0.4%, Welch two-sided $p = .69$; five-word floor +1.9%, same sign; detectable effect at 80% power 3.6%), and +3.5% over the repeated 640-prompt corpus at 120 steps ($p = .31$). The earlier deficit is not reproduced, although both differences are unresolved and do not isolate the effect of repetition. The selected configuration also incurs the probing cost. The 1,920-prompt selection contains all 640 original prompts (mean score 0.83) and 1,280 further prompts averaging 0.54; its teachers are three orders of one selection against three independent uniform draws, and both 120-step conditions answer tersely (median 28 and 21 words). The condition cost 32.0 GPU-hours.

On the new holdout, the same three-run conditions score 49.69% for uniform data, 49.01% for the $1{,}920$-prompt selection and 47.30% for the repeated 640-prompt selection. Raw-5 scores are 42.91%, 43.84% and 41.73%, respectively. This later evaluation is descriptive; it does not establish an advantage for selection at the longer budget.

D.2 Placement contrasts: teacher-data against student-prompt selection

D.2.1 Sealed holdout

Table 26: **Sealed selection-stage contrasts.** • Student endpoint; five teacher means per condition; shared selected-teacher comparator. • Welch t; MDE_{80}: one-sided minimum detectable effect at 80% power under observed variance.

	vs. published score ($\mathrm{U_1D_0} - \mathrm{U_0D_1^z}$)	**vs. same score** ($\mathrm{U_1D_0} - \mathrm{U_0D_1^u}$)
Δ primary	$+0.039050$	$+0.034600$
SE	0.010984	0.010970
t (df)	3.5553 (7.840)	3.1540 (7.844)
p, one-sided	0.00385	0.00693
MDE_{80}	0.0300	0.0302
Δ raw base, no floor	$+0.0474$	$+0.0446$ ($p = 0.0066$)
Δ raw base, ≥ 5 words	$+0.0263$	$+0.0193$ ($p = 0.133$)

Table 27: **Sealed selection-stage scores.** • Student Primary, averaged within teacher. • Probe-token budget; $2{,}000$ teacher candidates, versus $10{,}000$ for fixed-size selection. • Uniform-teacher conditions share five teachers; comparisons with $\mathrm{U_0D_0}$ paired within teacher. • Words: mean student median; revoked: teacher-mean format-guard revocation rate.

Condition	**Mean (SD)**	**Members**	**vs.** $\mathrm{U_0D_0}$	**Med. words**	**Revoked**
$\mathrm{U_0D_0}$, select nowhere	0.3896 (0.0131)	0.3723, 0.3793, 0.3947, 0.3997, 0.4020	—	57.1	2.64%
$\mathrm{U_0D_1^z}$, static band on student prompts	0.3906 (0.0186)	0.3603, 0.3913, 0.3930, 0.3977, 0.4107	$+0.0010$	52.0	3.26%
$\mathrm{U_0D_1^u}$, gap score on student prompts	0.3950 (0.0185)	0.3923, 0.3950, 0.3692, 0.3973, 0.4213	$+0.0054$	52.5	—
$\mathrm{U_1D_0}$, gap score on teacher data	$\mathbf{0.4297}$ $\mathbf{(0.0161)}$	0.4109, 0.4230, 0.4237, 0.4380, 0.4527	$+0.0401$	30.1	4.96%

Student-prompt selection versus uniform prompts. The three uniform-teacher conditions share teachers d_{21}–d_{25}, each student-prompt-selection gain over $\mathrm{U_0D_0}$ is a within-teacher difference. Averaging each teacher's students first and pairing by teacher identity: published score $\mathrm{U_0D_1^z} - \mathrm{U_0D_0} = +0.0010$, SD of the five differences 0.0102, two-sided $p = 0.84$; same score $\mathrm{U_0D_1^u} - \mathrm{U_0D_0} = +0.0054$, $p = 0.31$. These were computed after the sealed reads from the frozen result files and are not pre-registered tests. Both student-prompt-selection differences are unresolved.

The published-score comparator is the learnable band $0 < p^S < 1$ applied as a fixed pre-filter on the $8{,}000$-prompt root: $4{,}235$ prompts have $p^S = 0$, 815 have $p^S = 1$, $2{,}950$ lie inside the band, and the condition draws $2{,}688$ of them in one fixed order, 91.1% of the band, of which the $2{,}560$ presentations consume all but a reserve block.

D.2.2 THE SAME SELECTION SCORE AT EITHER STAGE

Table 28: **Same-score selection-stage contrast.** • Development $\mathrm{U}_1\mathrm{D}_0 - \mathrm{U}_0\mathrm{D}_1^u$; five teacher means per condition; Welch test. • Pre-registered two-look design, one read. • External-panel scores: percentages.

Metric	Δ	t	df	p
primary	$+0.04774$	6.069	5.51	0.00062
raw base, no floor	$+0.06523$	4.344		0.00518
raw base, ≥ 5 words	$+0.02860$	2.489		0.01879
loose criterion	$+0.07431$	5.213		0.00164
external composite	$+3.152$	3.129		0.00864
IFBench	$+5.339$			
IFEval	$+0.965$			
AIME (control)	-3.750			

Applying the gap score to teacher data improves the student score by 4.77% compared with applying it to student prompts (Table 28). Selected-teacher students write a median of 27.5 words against 55.0 for the no-selection condition and 53.5 for the same-score student-prompt-selection condition, and the guard revokes 7.37% of their responses against 3.81% and 4.12%. These are condition means of per-student holdout medians, the basis used for length throughout.

D.2.3 BOTH STAGES SELECTED, DEVELOPMENT HOLDOUT

Teacher-data selection also improves performance when student prompts are already selected. A fourth condition, $\mathrm{U}_1\mathrm{D}_1$, uses the five selected teachers of the development comparison and the student prompts selected for $\mathrm{U}_0\mathrm{D}_1^u$. The registered analysis pairs two contrasts across five teacher blocks on the development holdout (750 prompts, four responses per student, guarded primary score). Rescoring both comparator conditions reproduces their recorded values.

Table 29: **Selection at one or both stages.** • Development student scores (%); SDs across five teacher blocks. • Raw-5: raw strict, minimum five words. • Words: mean student median; revoked: format-guard revocation rate.

Condition	**IF Primary (SD)**	**Raw-5**	**Words**	**Revoked**
$\mathrm{U}_1\mathrm{D}_1$: both stages selected	$\mathbf{44.3}$ (1.1)	$\mathbf{37.9}$	26.6	7.4%
$\mathrm{U}_1\mathrm{D}_0$: teacher data only	43.8 (0.6)	37.3	28.2	7.2%
$\mathrm{U}_0\mathrm{D}_1^u$: student prompts only	38.4 (1.7)	34.0	49.4	4.4%
Initial checkpoint	26.4	—	121	0.8%

Selecting both stages improves on student-prompt selection alone by 5.90% (paired t, 4 degrees of freedom, two-sided $p = .00013$, Holm .00026). With a five-word minimum on raw correctness, the difference is 3.89% ($p = .052$). Adding student-prompt selection to teacher-data selection gives 0.55% ($p = .22$, Holm .22). This smaller difference is unresolved: gains up to 1.6% remain compatible with the data, and the declared detectable effect at 80% power was 2.7%.

All five blocks share one student-prompt sequence selected by the probe-token-budget variant's global top-u rule. Replications therefore vary teachers, not student-prompt selections. The both-stage students were trained twelve days after their comparators with the same code and resolved settings. The experiment costs 13.9 GPU-hours, including generation retries. Appendix D.2.4 reports the fixed-size variant.

D.2.4 PLACEMENT UNDER THE FIXED-SIZE SELECTOR, DEVELOPMENT HOLDOUT

The fixed-size selector also favors teacher-data over student-prompt selection. This registered comparison uses the $10{,}000$-prompt candidate pool. Each of five uniform teachers d_{42} to d_{46} supervises two students: one uses prompts chosen by the fixed-size gap-score rule and one uses the static learnability filter. Both exclude that teacher's own 640 training prompts. The teacher-data

comparators are three training orders of the original selected corpus and five fresh-probe selections, with uniformly sampled student prompts. Checks reproduce both the original selected corpus and the earlier static-filter selection before training. The four registered contrasts use Holm correction and teacher-level student means (Table 30).

Table 30: **Fixed-size selection-stage comparison.** • Development student scores (%); SDs across teachers. • Three orders of one corpus, five fresh-probe corpora, or five uniform teachers. • Raw-5: raw strict, minimum five words; words: mean student median.

Condition	IF Primary (SD)	Raw-5	Words
Selected teacher data, one corpus in three orders	43.5 (2.6)	40.7	61
Selected teacher data, five fresh-probe corpora	**44.4** (0.8)	**41.2**	53
Uniform teacher data, gap score on student prompts	37.4 (2.0)	33.9	67
Uniform teacher data, static band on student prompts	37.3 (1.5)	34.0	69
Uniform at both stages	37.7 (1.4)	34.3	64
Initial checkpoint	26.4	24.3	121

Teacher-data selection exceeds the gap score on student prompts by 6.7% and the static band by 6.8% (Welch, eight selected teachers against five, two-sided $p = .0004$ and .00002, Holm .001 and .0001), and the gains do not shrink under the five-word floor (7.1 and 7.0); median response lengths differ by 11 to 13 words without a resolved difference. Taken separately, the three orders of the original corpus give 6.1% and 6.2% ($p = .031$ and .036) and the five fresh-probe corpora 7.0 and 7.1 ($p = .0007$ and .00007); the pooled eight mix order replicates with independent draws; the separate estimates identify these replication types. Relative to uniform student prompts, paired differences over five uniform teachers are -0.3% for the gap score and -0.4% for the static filter, both unresolved.

The student-prompt selections share 594 to 611 of the original selected corpus's 640 prompts, placing similar high-scoring data at different training stages. The five student sequences share about 2,500 of 2,688 prompts, so the replications mainly vary teachers. Uniform-at-both-stages controls use student prompts outside this pool; those paired contrasts change the source pool as well as the selection rule. The experiment costs 27.9 GPU-hours, including a student restarted after a node failure before its first update. Hardware and resolved training settings match the comparator runs.

D.2.5 Fixed-size selection on the new holdout

Teacher-data selection also exceeds both student-prompt selectors on the third, 750-prompt holdout (Table 31). This follow-up evaluates the existing models from Appendix D.2.4, without further training. The registered analysis pools three teacher-training orders of the original selected corpus and five fresh-probe selections. For each of the three original teachers, both students are averaged before testing, whereas Table 16 reports one student per original teacher. Including the five fresh selections gives eight teacher-level observations against five for each student-prompt selector. The four-test Holm family includes both teacher-data versus student-prompt contrasts and both student-prompt versus uniform contrasts.

Table 31: **Fixed-size selection on the new holdout.** • Student IF Primary (%). • Same models and pooling as the development comparison; no new training replications.

Teacher data	Student prompts	Teachers	IF Primary
Uniform	Uniform	5	35.16
Uniform	Gap score	5	35.41
Uniform	Static filter	5	34.83
Selected, fixed size	Uniform	8	**41.12**

The gains over gap-score and static-filter student selection are 5.71% and 6.29% (two-sided Welch $p = .00033$ and .000029; Holm-adjusted $p = .00099$ and .00012). With a five-word minimum on the guarded score, the gains remain 5.54% and 6.06%. Neither student-prompt selector differs detectably from uniform prompts: their paired differences are $+0.26\%$ and -0.32%, both Holm-adjusted $p = .95$.

This was the second registered question on the new holdout. Selected-teacher and uniform scores were already known from its first read; the protocol was frozen before generating the additional student-prompt-selector responses. All training prompts were checked for content overlap with the holdout. Prefix caching was disabled, and scores from reused outputs reproduced their first-read values. The original study's pool and student-fill differences still apply. The gain is not resolved under relative-length thresholds: retaining responses at least one-quarter of the initial student's median response length for each prompt gives differences of 1.95% and 2.11%; a one-half threshold gives -1.26% and -1.08% (all $p > .35$). Descriptively, the three original training orders alone give $p = .077$ and $.068$, whereas the five fresh selections give $p = .00049$ and $.000082$. The pooled eight-teacher comparison, not either subset, is the registered analysis.

D.2.6 Token-quota selection on the new holdout

The token-quota placement comparison also favors teacher-data selection on the third holdout. Five blocks compare selection at the teacher only, at the student only, and at both stages (Table 32). These are the existing models from Appendix D.2.3, evaluated without further training. The sole registered primary contrast is teacher-only minus gap-score student selection, paired by block; both-stage contrasts are descriptive. All five blocks share one selected student-prompt sequence. This third registered question was frozen before generating any of these responses on the new holdout.

Table 32: **Probe-token-budget selection on the new holdout.** • Mean student scores (%); five paired blocks. • Raw-5: raw strict, minimum five words; words: mean student median.

Teacher data	Student prompts	IF Primary	Raw-5	Words
Uniform	Gap score	36.63	32.99	59.6
Uniform	Static filter	36.95	33.03	58.8
Selected	Uniform	40.11	34.31	35.0
Selected	Gap score	**40.55**	**34.77**	33.2

Teacher-only selection improves over student-only selection by 3.48% (paired $t = 7.27$, 4 degrees of freedom, two-sided $p = .00190$). All five differences are positive. The five-word difference falls to 1.31% ($p = .50$), and is negative in three blocks. At the quarter- and half-length thresholds the differences reverse to -3.61% and -4.49%, both unresolved. Teacher-only students answer about 25 words more briefly. The primary result is positive, but the gain is not robust to response-length filtering.

A later registered comparison adds the static student-prompt filter, reusing the teacher-only students above. Teacher-data selection gains 3.16% across five paired blocks (two-sided $p = .00339$), with all five differences positive. The Raw-5 difference is 1.27% ($p = .50$); the relative-length differences are -3.65% and -4.84%, both unresolved. The teacher-only scores were known when this additional comparison was registered.

Selecting both stages adds 0.44% to teacher-only selection and 3.92% to student-only selection. Their nominal paired p-values are $.0082$ and $.0017$; neither was allocated confirmatory alpha in this follow-up. These contrasts share the primary comparison's students and are not independent replications. The fixed-size study in Appendix D.2.5 uses different teachers, corpora and student prompts; its values should not be pooled with these.

D.2.7 Placement at a quarter of the student budget

A pre-registered comparison reduces student training to 640 presentations, or ten steps, with other settings unchanged: the three selected teachers with uniform student prompts, the five uniform teachers d_{42}–d_{46} with uniform prompts, and the same five with the gap score's fixed-size selection on student prompts (a random 640 of the top-scoring $2{,}688$ prompts). Each quarter-budget student is the first ten steps of the corresponding full-budget student of the same teacher, so the full-budget values are shown beside them (Table 33).

Table 33: **Quarter-budget student training.** • Development student scores (%); teacher SDs in parentheses. • Raw-5: raw strict, minimum five words; words: mean student median. • Last column: same teachers' full-budget students, descriptive.

Condition	IF Primary (SD)	Per teacher	Raw-5	Words	Full budget
Selected teacher data, three orders	43.1 (3.7)	43.8 / 46.4 / 39.1	40.5	56	43.1
Uniform teacher data, gap score on student prompts	36.7 (1.9)	—	33.3	67	—
Uniform at both stages	36.3 (1.1)	—	33.0	63	37.7
Initial checkpoint	26.4	—	24.3	121	—

Teacher-data selection retains a 6.8% difference from uniform data at the smaller student budget, although the registered two-sided test is unresolved (Welch $p = .078$, 2.2 degrees of freedom). The three selected-corpus training orders have SD 3.7%, versus 1.1% for uniform teachers. Under the descriptive five-word check, the difference is 7.4%; median lengths are 56 versus 63 words, with revocation rates of 2.3% and 2.2%. Student-prompt selection adds an unresolved 0.4% over uniform prompts in the five paired teacher comparisons.

Selected-teacher students change little between 640 and 2,560 presentations (-0.1%, -0.5% and $+0.6\%$), while uniform-teacher students gain 1.4%. This budget comparison is descriptive, using separate scoring runs on the same holdout. The experiment uses 16.8 GPU-hours against a planned cap of 15, including generation retries.

D.3 Published selection criteria on the teacher's data

D.3.1 Development comparison of selection criteria

Relation to the later comparisons. The 45.45% student mean in Table 4 comes from teacher orders o31 to o33 and their students (46.03%, 46.00%, 44.33%). The later 43.5% development mean uses the same selected corpus but different teacher orders, o34 to o36, with two students averaged per teacher (43.75%, 46.00%, 40.83%; Table 24). The 44.1% development mean in Table 2 additionally includes five fresh-probe selections. These are different groups of trained models, not alternative scores for one group. The earlier criterion study likewise uses different uniform controls: its student mean is 37.75%, versus 37.71% in the later fixed-size study. Its teacher means and response lengths belong to these earlier development runs, not to the later teachers or the original holdout.

Table 34: **Development teacher scores by condition.** • 750 prompts; shared pool, budget rule and tie convention. • Primary: guarded; raw strict/loose: unguarded. • Uniform: eleven draws; single-draw prediction thresholds: 0.4148 (Primary), 0.3714 (five-word floor).

Family	Condition	Primary	Raw strict	Loose	Revoked	Median words	Tokens
initialisation	mixed-SFT	0.2643	0.2663	0.3193	0.8%	121	446
ours, training orders	order 1	**0.4660**	**0.4877**	**0.5507**	4.4%	24	153
	order 2	0.4530	0.4653	0.5253	2.7%	38	192
	order 3	0.4327	0.4437	0.4890	2.5%	46	195
ours, re-selected	selection A	0.4347	0.4417	—	—	66	—
	selection B	0.4383	0.4457	—	—	62	—
	selection C	0.4277	0.4420	—	—	34	—
published, transplanted	learnability	0.3983	0.4083	0.4757	2.4%	55	256
	residual coverage	0.4017	0.4293	0.4720	6.4%	27	179
	teacher gating	0.4027	0.4077	0.4667	1.2%	83	294
	difficulty band	0.3953	0.4050	0.4707	2.4%	68	276
uniform draws	d_{31}–d_{41}	0.3510–0.4130					

Across the seven selector conditions first scored (three orders and four published), Spearman's rank correlation between median words and the primary is -0.607, and with the raw strict score -0.857; including the initialisation, -0.738 and -0.905. Under the five-word floor the teacher-gating condition scores 0.3733 against the floored bar of 0.3714.

Table 35: **Tie-key replications of adapted criteria.** • Teacher and student endpoints; two selections per criterion; shared probes, pool and budget rule. • Uniform reference: eleven rescored conditions; single-draw threshold from $t_{0.95,10}$; †: above threshold. • w5: five-word floor; words: condition median; revoked: format-guard revocation rate.

		Teacher endpoint				Student endpoint			
Criterion	**Key**	**Primary**	**w5**	**Words**	**Revoked**	**Primary**	**w5**	**Words**	**Revoked**
learnability	first	0.3983	0.3543	55	2.4%	0.3920	0.3453	52	3.0%
	second	0.4063	0.3640	55	2.3%	0.4183†	0.3843†	58	2.6%
residual coverage	first	0.4017	0.3313	27	6.4%	0.3740	0.2943	26	10.1%
	second	0.3100	0.2247	13	15.1%	0.3067	0.2223	13	15.4%
teacher gating	first	0.4027	0.3733†	83	1.2%	0.4207†	0.3930†	80	0.9%
	second	0.4123	0.3783†	48	1.7%	0.4213†	0.3813†	42	2.5%
difficulty band	first	0.3953	0.3513	68	2.4%	0.4047	0.3587	69	3.3%
	second	0.4423†	0.3947†	44	1.8%	0.4317†	0.3813†	41	2.0%
ours, three orders	one	0.4327–0.4660†	0.3897–0.4080†			0.4433–0.4603†	0.3920–0.4097†		

All three training orders of our gap-score selection exceed the uniform prediction threshold at both endpoints. Independent probe replications are reported in Appendix D.1.2. The adapted criteria vary across tie keys. Residual coverage falls by 0.09 at the teacher endpoint and 0.07 at the student endpoint; its second selection produces short responses (median 13 words, 15% revoked). The difficulty band rises by 0.047 and 0.027 and exceeds both thresholds under the second key. Learnability exceeds the student threshold under the second key only, and teacher gating does so under both keys.

Tie keys change membership only among prompts with equal scores and share the same probe error. These two-key comparisons therefore remain descriptive rather than establishing an ordering of the criteria. The tables use the corrected threshold multiplier 1.8125 (10 degrees of freedom), replacing 1.796 (11 degrees of freedom); the correction changes thresholds by at most 0.0004 and no threshold-crossing decision. The experiment costs 24.5 GPU-hours, including one student restart.

D.3.2 Dynamic sampling during teacher training

We evaluate the dynamic-sampling component of GRPO described in Appendix C.3, rather than reproduce the complete DAPO algorithm. Table 36 reports its teacher and student results, generation cost and failures under the prespecified short-response criterion.

Table 36: **Dynamic sampling comparison.** • Teacher endpoint; earlier $2{,}000$-prompt development study with its own uniform controls. • Collapse: short-response share above the uniform maximum or median length below the uniform minimum.

Quantity	**Value**
Primary, ours (3 orders) − dynamic sampling (5 seeds); two-sided p	$+0.026$; $p = 0.35$
Collapse rule, both legs: dynamic sampling / ours, earlier trio / ours, replicated trio / uniform runs	$3/5$ / $1/3$ / $0/3$ / $0/11$
Raw base, $\geq$ 5 words, surviving runs only: ours vs. dynamic sampling (n=2 each)	0.3850 vs. 0.3828 ($+0.18$ SD_U)
Single-seed floored range of dynamic sampling	0.2740–0.3903
Realised rollouts relative to a uniform run	$2.68\times$
Student endpoint, dynamic sampling (5) vs. uniform, development holdout	$+0.0264$ ($p = 0.103$); floored $+0.0068$ ($p = 0.353$)

D.3.3 Selected-set statistics, tie resolution and selector variants

Length grouping preserves a coarse response-length distribution during selection. Both it and score normalization yield positive teacher-score differences from uniform data (Table 38). Their ordering depends on the comparison: grouping gives fewer revoked passes in the initial single-seed study, whereas u/c gives fewer revocations and a higher five-word score in the replicated study. The variants change both normalization and stratification, so this comparison does not isolate either component.

Table 37: **Selected-data statistics and teacher scores.** • Probe rates: selected-set means, $k = 4$, 10,000-prompt pool. • Outcome: teacher Primary from Table 34. • Degenerate groups: identical rewards across all rollouts.

Rule	Mean p^S	Mean p^R	Degenerate groups	Teacher primary
ours: p^R high and p^S low	0.099	0.924	62.2%, all failures	0.4327–0.4660, clears
residual coverage: $p^R \leq 0.5$ first	0.087	≈ 0.125	—	0.4017, inside
teacher gating: p^R high only	0.607	0.924	—	0.4027, inside
learnability band	—	—	0%, 2.6× our effective support	0.3983, inside
uniform draw	—	0.42–0.44	64.5%, mixed	0.3764 mean

The criteria select different types of prompts. The gap score selects prompts with high reference success and low initial-student success (Table 37). Teacher gating retains equally high reference success ($p^R = 0.924$), but also high student success ($p^S = 0.607$). Residual coverage prioritizes reference success at most 0.5; the selected set's low student success ($p^S = 0.087$) is an outcome, not its selection criterion.

The learnability filter has no identical-reward groups in the initial probes and 2.6× as many prompts with mixed rewards as our selection. Nevertheless, its teacher score remains within the uniform prediction interval in this comparison. Our selection and uniform draws have similar proportions of identical-reward probe groups, 62.2% and 64.5%, respectively. These are initial $k = 4$ probe statistics, not training gradients. The criteria also differ in thresholds, quotas and tie-breaking, so their outcomes do not isolate the contribution of either term in the gap score.

Tie-breaking controls. At $k = 4$, the score has nine possible values, and ties determine 48.4% of the selected set. Boundaries fall at $u = +0.75$ in three length quartiles and $+0.50$ in the fourth. A fixed, pre-registered order resolves ties without using source-corpus identity.

The learnability selection is entirely tie-decided. Across 400 alternative tie orders, its mean selected-data score has SD 0.0079 and its identical-reward-group fraction is unchanged. For teacher gating, the implemented selection is close to the center of 400 alternative orders: $z = +0.26$ for mean student success, $+0.52$ for mean reference success and $+0.30$ for identical-reward-group share. Its first-key five-word teacher score exceeds the uniform prediction threshold by 0.0019 (Table 35). A separate tie seed shares none of the original selection's positions, confirming that the randomization changes the order. These controls measure selected-data statistics; the alternative selections were not retrained.

Table 38: **Score variants and replication.** • Top: five forms, one seed each, development holdout. • Bottom: shuffled training orders against eleven uniform draws; scores in uniform SD units. • No equivalent baseline search; u/c changes both normalization and stratification.

Functional forms, one seed each Variant	Rule	Primary	Revoked	Median words
u over probe cost	greedy, exponent 1	0.4087	10.6%	16
u alone	greedy, exponent 0	0.4240	7.4%	20
u over $\sqrt{\text{cost}}$	greedy, exponent 0.5	0.4453	4.0%	40
cost-normalised u, cheapest quartile demoted	greedy	**0.4593**	3.7%	36
length-quartile quota (retained)	quartile quota	0.4527	2.7%	36
Replication with shuffled fills, teacher primary in uniform SD above eleven draws Variant	Runs	Primary	Raw, ≥ 5 words	Revoked per run
length-quartile quota	3	$+3.50$ SD$_U$	$+1.11$ SD$_U$	7.2, 7.9, 7.5%
u over probe cost	4	**$+3.57$ SD$_U$**	**$+2.30$ SD$_U$**	5.5, 3.1, 7.3, 1.8%

D.4 Interpreting the original comparisons

Selection stage and comparison design. The two original selection-stage comparisons share the same selected-data teachers. The static student-prompt filter uses only initial-student success, whereas teacher-data selection also uses reference responses. The same-score comparison holds the scoring formula fixed, but not the candidate pool or probe volume: teacher and student selectors use

disjoint pools of 2,000 and 8,000 prompts. Thus neither comparison varies selection stage alone. Appendix C.4 records which scores were known before registration and the subsequent evaluation history.

Evaluation and resource accounting. Length-filtered scores and their tests accompany each comparison; a positive difference alone need not be statistically resolved. Matched training budgets refer to prompt presentations and rollouts, with token use and probing costs reported separately. Adapted-criterion and dynamic-sampling comparisons retain their own replication units and generation budgets.

D.5 External panel and multi-teacher distillation

D.5.1 External panel, development students

The earlier development comparison gives mixed external results. One of three selected-data students exceeds the uniform instruction-following prediction threshold. Their mean AIME scores are close to uniform-data students, and trained students generally score below the initial model on AIME.

The panel uses 539 IFEval and 251 IFBench prompts under strict correctness, plus two AIME sets. Eighteen trained students and the initial model each have 2,761 evaluation rows. The registered primary pools the 790 instruction-following prompts. Uniform draws average 0.5631 (SD 0.0171), with a 95% single-run prediction threshold of 0.5954. Selected-data students score 0.6063, 0.5835 and 0.5835; one exceeds the threshold by 0.0109. The four adapted criteria score 0.5532 to 0.5886. IFBench alone has a threshold 0.0639 above the uniform mean, compared with a uniform SD of 0.0337.

On the two AIME sets, the initial model scores 0.1104 and 0.1542, versus trained-condition means of 0.0583 and 0.0718. Uniform draws average 0.0621 and 0.0769, and selected-data means are within 0.001 of these. Three uniform runs exceed the initial model on the first set (0.1146, 0.1208, 0.1156), and one on the second (0.1573); the matched selected-data run has the lowest second-set score. Trained students' response medians range from 24 to 98 words, compared with 121 for the initial model. The detectable effect exceeds the trained conditions' mean AIME scores, and these descriptive associations do not identify the cause of the decline. This evaluation uses development students, not the original holdout.

D.5.2 Original multi-teacher comparison and development screen

The original multi-teacher comparison improves instruction-following scores with teacher-data selection (Table 39). Mathematics and code teachers are fixed; only the instruction-following teacher's data changes. Each condition has one student, so the results are descriptive. The LiveCodeBench (Jain et al., 2024) v5 scores retain the designated evaluation despite timing sensitivity on one student-prompt control problem; diagnostic reruns do not replace it.

Table 39: **Original multi-teacher evaluation.** • Scores (%). • Uniform: no selection; other labels: selected training stage. • Separate evaluation from the 244-question screen in Table 40.

Endpoint	Uniform	Teacher data	Student prompts
IFEval, aligned strict	73.28	**75.14**	71.80
IFBench, aligned strict	20.72	**25.90**	20.32
AIME24	13.13	11.56	**14.17**
AIME25	17.81	17.08	**19.17**
LiveCodeBench v5	17.07	17.07	**17.31**
LiveCodeBench v6	19.60	**20.51**	19.43

The later development screen uses the same decoding recipe in both conditions: temperature 1, top-p 1, one response per question, and a 16,384-token response cap. Its six populations contain 60, 60, 30, 30, 32, and 32 questions for IFEval, IFBench, AIME24, AIME25, LiveCodeBench v5, and LiveCodeBench v6, respectively. Training uses twelve updates on 768 prompt presentations; the evaluation cap is not the training response cap. Both conditions use teacher-panel-based model merging, a 0.02 KL retention coefficient, and the same route-calibration procedure, but the realized

initializations and route weights differ. The fixed Math/Code/IF scale factors are $(1.477, 1.040, 0.484)$ for teacher-data selection and $(1.312, 0.924, 0.764)$ for student-prompt selection, applied to the route-loss weights with mean normalization. The student-prompt baseline trains the instruction-following teacher on uniform data and retains prompts with initial-student success strictly between zero and one over four probes, using a fixed hash order. Mathematics and code teachers, prompts and prompt order remain unchanged. This is a static learnability filter. Because student initializations and domain weights also differ, the study compares complete training pipelines rather than selection stage alone. Supplementary records identify the models, generation settings and scorers.

The teacher-data-selection recipe scores higher on the screened IFEval and IFBench sets by 5.00% and 6.67% absolute, respectively. It also scores higher on AIME25 and both LiveCodeBench subsets. Table 40 reports the complete development screen. The small endpoint samples and single trained model per condition do not support a method-level significance claim.

Table 40: **Multi-teacher selection-stage screen.** • Scores and absolute differences (%); one student per condition, one response per question. • IFEval/IFBench: screened strict; AIME: separate years; LiveCodeBench: subsets.

Endpoint	Questions	Teacher data	Student prompts	Difference
IFEval	60	**65.00**	60.00	+5.00
IFBench	60	**21.67**	15.00	+6.67
AIME24	30	20.00	**30.00**	−10.00
AIME25	30	**33.33**	20.00	+13.33
LiveCodeBench v5	32	**25.00**	15.63	+9.38
LiveCodeBench v6	32	**34.38**	31.25	+3.13

D.5.3 Replicated multi-teacher comparison and the rendering repair

Selecting instruction-following and mathematics teacher data raises mean student scores on those domains in both multi-teacher comparisons, but neither primary mathematics test is resolved (Table 41). The original study includes four uniform-teacher students, two selected-teacher students and four students using uniform teachers with student-prompt filtering. The code teacher is fixed. Each selected domain uses one corpus in two training orders. Students train for twelve updates of 64 prompts, with an $8{,}192$-token response cap. The primary mathematics evaluation uses the reserved split, a $4{,}096$-token cap and thinking disabled.

Table 41: **Multi-teacher results before and after the rendering repair.** • Scores (%); student runs count separately trained students. • Rerun: one student-prompt configuration excluded for failed export; AIME thinking-mode mismatch retained. • Between-study comparison: descriptive.

Study	Method	Student runs	IFEval	IFBench	Math	AIME
Original	Uniform	4	70.97	24.40	19.74	14.52
	Student prompts	4	72.40	24.20	18.98	**15.05**
	Selected	2	**73.75**	**29.08**	**20.69**	14.59
Rendering-matched	Uniform	4	71.94	23.80	27.50	**16.08**
	Student prompts	3	71.49	21.78	24.35	15.87
	Selected	2	**73.20**	**31.87**	**30.67**	15.81

A setup check found that mathematics teachers were trained with thinking disabled, while mathematics prompts enabled thinking during multi-teacher student training. Amendment 5 aligns these settings across conditions and recomputes the mathematics student-prompt filter. The rerun was specified after the original result but before any rerun scores, so the two studies are reported separately. Nine of ten configurations completed: four uniform, two selected and three student-prompt-selection students; one student-prompt model failed to export. Each admitted mathematics evaluation has $2{,}000$ prompts and $8{,}000$ responses, with no scoring faults; the instruction-following scorers report no exceptions.

The registered selected-versus-uniform mathematics differences are 0.95% in the original study ($p = .321$) and 3.17% after the repair ($p = .150$), using one-sided predictive tests. The rerun's IFBench difference is 8.06% (nominal one-sided $p = .026$, two-sided $.051$), a secondary outcome

without an allocated significance level. Against student-prompt selection, rerun differences are $+6.31$ on mathematics, $+1.70$ on IFEval, $+10.09$ on IFBench and -0.06 on AIME. Neither study meets the registered criterion of improving every task. Uniform IFBench scores span 9.16% in the original study. AIME reaches its $8{,}192$-token cap in 72.7% to 75.2% of responses, and its thinking-enabled evaluation remains different from mathematics training. The rerun changes the setup but does not identify the cause of the original null result.

Selecting both specialist teachers in a separate three-pair comparison. With IF and Math teacher-data selection and a $2{,}048$-token student response cap, paired Math differences against uniform teacher data are 0.79%, 0.91% and 2.38%, averaging 1.36 (two-sided paired $p = .117$). All three mathematics differences are positive. Mean differences are $+0.37$ on IFEval, $+2.12$ on IFBench and $+0.43$ on AIME, with two-sided $p = .815$, $.601$ and $.268$, respectively. IFBench differences vary across pairs ($+7.97$, $+2.39$, -3.99). All four task comparisons are unresolved.

D.5.4 Additional baselines and transfer-recipe checks

Additional student-prompt selectors give mixed results on the 244-question panel. Table 42 separates two independently trained sets of configurations, labeled A and B; their checkpoints and scores must not be pooled. Each row represents one trained student. The teacher-data-selection and student-prompt-selection configurations use teacher-panel-dependent initializations and route weights, so these are comparisons of complete recipes, not isolated selection-stage effects. Within configuration set A, DSIR and IFD share the same training configuration apart from the selected prompt sequence. IFD and LESS use the uniform teacher's generated answers because the candidate pool has no gold responses; constructing this bank costs 32.49 million response tokens. This adaptation and its cost are distinct from DSIR's inputs. Configuration set B uses a common $1{,}664$-token training-input allowance and the original 244-question evaluation protocol. Model, generation and scorer identities have been verified. Candidate-eligibility and near-duplicate checks remain incomplete, so these single-student comparisons are retained as pilot results.

Table 42: **Additional multi-teacher baselines.** • Scores (%); 244-question panel; one student per row. • IF-E/B: IFEval/IFBench; LCB5/6: LiveCodeBench v5/v6. • Student initialization and domain weights differ between selection stages.

Configuration	IF-E	IF-B	AIME24	AIME25	LCB5	LCB6
Configuration set A						
Teacher-data selection	61.67	**18.33**	16.67	16.67	9.38	**31.25**
Student-prompt DSIR	**70.00**	15.00	16.67	16.67	18.75	25.00
Student-prompt IFD	**70.00**	**18.33**	**20.00**	**23.33**	**21.88**	28.13
Configuration set B						
Teacher-data selection	**71.67**	**18.33**	**16.67**	**26.67**	**25.00**	**28.13**
Student-prompt static band	63.33	13.33	13.33	23.33	21.88	**28.13**
Student-prompt DSIR	63.33	13.33	**16.67**	20.00	21.88	**28.13**

Replicated comparisons favor teacher-data selection on public instruction-following benchmarks. With three independently trained students per configuration, teacher-data selection exceeds uniform sampling, DSIR, adapted IFD and adapted LESS by 2.2% to 3.6% on IFEval and 3.5% to 5.2% on IFBench (Table 43). All eight public-panel comparisons have unadjusted two-sided Welch $p < .05$.

These experiments share student initialization, domain loss weights and training budgets. Each selected set is held fixed, so replications vary training and generation rather than data selection. The declared primary panel has 244 questions, including 60 IFEval and 60 IFBench items, with one sampled response at temperature one. The secondary public panel has 539 IFEval and 251 IFBench prompts with greedy decoding. On the declared panel, two of eight instruction-following contrasts meet the unadjusted $.05$ threshold: IFBench against uniform ($+5.55\%$) and IFD ($+6.11\%$). DSIR, IFD and LESS do not differ detectably from uniform on either panel. The two panels are analyzed separately.

Table 43: **Replicated multi-teacher baseline comparison.** • Mean (SD), %; three students per configuration. • Left: declared 244-question panel; right: public panel. • DSIR and adapted IFD/LESS: 256 instruction-following student prompts; ours: instruction-following teacher data. • Shared student initialization, training budgets and domain weights.

	244-question panel		Public panel	
Configuration	IFEval	IFBench	IFEval	IFBench
Uniform at both stages	60.56 (3.47)	17.22 (0.96)	72.85 (0.57)	19.39 (1.00)
Student-prompt DSIR	65.00 (6.01)	16.67 (2.89)	71.74 (1.18)	19.26 (1.00)
Student-prompt IFD, adapted	66.11 (3.47)	16.67 (1.67)	72.29 (0.60)	18.06 (1.66)
Student-prompt LESS, adapted	61.11 (4.19)	18.33 (0.00)	73.16 (0.57)	19.79 (0.61)
Teacher-data selection (ours)	**67.22** (3.47)	**22.78** (2.55)	**75.33** (0.81)	**23.24** (1.40)

The larger evaluation has 1,192 distinct problems and 2,504 sampled continuations per checkpoint. It averages responses within each problem to estimate mean pass@1. Its primary response cap is 16,384 tokens; shorter prefixes are diagnostics of the same generated trajectories. Neither more problems nor more responses supplies additional training replication. The complete differences in Table 44 show that instruction-following gains coexist with losses on AIME24 and LiveCodeBench v5 under both recipes.

Table 44: **Expanded multi-teacher evaluation.** • Teacher-data minus student-prompt selection (%); one student per condition. • Separate AIME years and code-evaluation populations.

Endpoint	Problems	Hybrid	Scale-controlled
IFEval	539	+3.15	+1.95
IFBench	251	+3.98	+3.78
AIME24	30	−4.17	−5.00
AIME25	30	−0.83	+3.33
LiveCodeBench v5	167	−2.99	−1.20
LiveCodeBench v6	175	+1.43	+0.57

The transfer-recipe screen varies training response length and which tokens receive supervision (Table 45). Late-512 improves all six tasks in its initial run, but the pattern does not reproduce with a new student seed. The table retains both the development results and this follow-up.

Table 45: **Transfer-recipe screen and follow-up.** • Teacher-data minus student-prompt selection (%); 244 questions; one student per condition. • Last row: new student seed after development-based choice of late-512 configuration.

Variant	IF-E	IF-B	AIME24	AIME25	LCB5	LCB6
Longer 4K responses	+8.33	+6.67	−6.67	−13.33	−3.13	−3.13
Terminal-neutral	−5.00	+5.00	+10.00	−6.67	+12.50	+6.25
Early-512	+5.00	+5.00	−10.00	−3.33	0.00	0.00
Late-512	+1.67	+1.67	+10.00	+3.33	+6.25	+12.50
Support-mass	0.00	+6.67	+3.33	0.00	0.00	−6.25
Late-512, new seed	0.00	0.00	−3.33	−3.33	0.00	0.00