FROM COMPLEMENTARITY TO INTEGRATION VALUE: A SUPPLY–CONVERSION FRAMEWORK FOR MULTI-MODEL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models have made multi-model integration an important approach to enhancing reasoning. Additional models can contribute through two broad modes: online integration, exemplified by inference-time collaboration, and offline transfer, exemplified by teacher-guided supervised fine-tuning (SFT) followed by standalone deployment. Existing methods improve partner complementarity or teacher–student compatibility within each mode. Beyond these within-mode advances, there remains limited understanding of how the same candidate models create different integration value across roles and what governs these differences. To this end, we introduce a *supply–conversion* framework that evaluates shared candidate pools across online and offline modes against their corresponding single-model alternatives. For online collaboration, we quantify complementary coverage and selection loss to derive an exact condition for outperforming the stronger candidate. For offline SFT, we measure realized student improvement and the additional value of combining teachers relative to the better single-teacher alternative. Across diverse reasoning tasks, we find that collaboration gains depend jointly on complementarity and selection quality, teacher value follows a different structure from inference-partner value, and the better single teacher provides a strong reference for uniform teacher mixtures. Building on this analysis, we combine task-local probes with cross-pair judge calibration to select between single-candidate inference and collaboration, achieving lower selection regret than the evaluated score-, size-, and diversity-based baselines.

1 INTRODUCTION

Recent advances in large language models (LLMs) have made leveraging existing model capabilities an increasingly important approach to improving reasoning. Model-generated answers and reasoning traces can support both the integration of alternative solutions and the construction of supervision for a target model (Jiang et al., 2023; Wang et al., 2025; Hsieh et al., 2023). As a result, models increasingly serve not only as standalone generators, but also as sources of candidate solutions, evaluation signals, and training supervision, allowing reasoning systems to exploit capabilities distributed across models.

These uses can be broadly organized into two modes. *Online integration*, exemplified by inference-time collaboration, retains access to candidate models when answering queries. *Offline transfer*, exemplified by teacher-guided supervised fine-tuning (SFT), incorporates supervision from candidate models into a student that is subsequently deployed on its own. The former exploits additional models through output integration at inference time, whereas the latter transfers their capabilities through training. Together, they represent two major routes for improving reasoning with external model capabilities. As the pool of available models continues to expand, effectively exploiting candidate models has therefore become an increasingly important problem in multi-model system design.

Existing work has substantially advanced both modes through improved selection and integration mechanisms. Online approaches rank and fuse candidate responses, coordinate multi-model discussions, and route queries according to quality–cost objectives (Jiang et al., 2023; Chen et al., 2024a; Ong et al., 2025). Offline approaches improve distillation objectives, model teacher–student compatibility, and allocate supervision across teachers (Gu et al., 2024; Xu et al., 2025; Panigrahi

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.

et al., 2026; Zhang et al., 2026). Structured distillation further connects multi-agent interactions to student learning (Chen et al., 2024b), while model-similarity studies investigate how relationships between models affect evaluation and supervision (Goel et al., 2025). Overall, these lines of work have developed increasingly sophisticated mechanisms for exploiting models within a particular integration role. However, there remains limited systematic understanding of how the same candidate models acquire different integration value across roles, and what determines these differences.

This motivates the central question of our study: *when the same candidate models are used as inference partners or training teachers, how are their available capabilities converted into realized gains through different integration mechanisms?* Because inference partners and training teachers contribute through fundamentally different processes, the same relationships among candidate models may translate into different forms of system value. We characterize this through two components: the useful capability made available by additional models, which we call *supply*, and the effectiveness with which a particular mechanism converts that capability into performance improvement, which we call *conversion*. Under this view, integration value emerges jointly from what candidate models can provide and how effectively a given role exploits it.

Building on this perspective, we introduce a **supply–conversion framework** for analyzing shared candidate pools across online collaboration and offline transfer. For collaboration, we relate realized gains to complementary capability and the effectiveness of the selection mechanism. For SFT, we examine the improvement transferred to the student under single- and multi-teacher supervision. This provides a common analytical view of the two modes while preserving their distinct mechanisms: online collaboration converts complementary candidate behavior at inference time, whereas offline transfer converts teacher-generated supervision into learned student capability.

We evaluate this framework across mathematics, science, coding, and reasoning tasks, using shared candidate pools as both inference partners and SFT teachers. Our results reveal substantial role dependence in integration value: collaboration depends strongly on both available complementarity and selection quality, while teacher value follows a different structure from inference-partner value. Under a fixed training budget, we also find that the better single teacher provides a strong reference for assessing uniform multi-teacher mixtures. Finally, we show that task-local probes, together with lightweight calibration of collaboration quality, can support deployment-conditioned decisions between single-model inference and collaboration, reducing selection regret relative to the evaluated capability-, size-, and diversity-based baselines.

Our contributions are threefold:

- **Cross-role analysis of shared candidate pools.** We evaluate the same candidate models as inference partners and SFT teachers against their corresponding single-model alternatives, revealing substantial role dependence in integration value.
- **A supply–conversion framework for model integration.** We separate available capability supply from mechanism-specific conversion, providing a unified view of online collaboration and offline transfer while characterizing their distinct realization mechanisms.
- **Probe-based collaboration selection.** We use task-local complementarity estimates and collaboration calibration to choose between single-candidate inference and collaboration, reducing selection regret relative to the evaluated score-, size-, and diversity-based baselines.

2 Problem Formulation

We consider how a target model should use additional models under a fixed task and deployment setting. Let S denote the target model, $\mathcal{C}$ a pool of candidate models, $\mathcal{D}$ the task domains, and Π the deployment constraint specifying whether external candidates can be accessed at inference time. We organize candidate use into two modes: *online integration*, which retains candidate models during inference, and *offline transfer*, which incorporates their supervision into the target before deployment.

Accordingly, we study four concrete actions. *Single* directly uses one candidate model. *Team* obtains independently generated answers from two candidates and lets a frozen judge select one without rewriting. *Distill* trains S on supervision from a single teacher, whereas *Mix* uses supervision from two teachers. The same checkpoint may appear in multiple roles, but these correspond to distinct operations. We also retain the unchanged target as a feasible alternative. When Π prohibits external

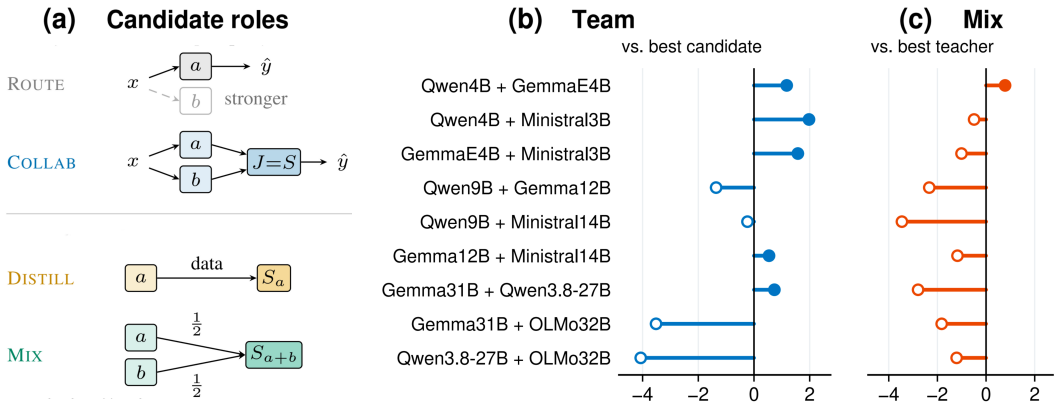


Figure 1: **Candidate roles and their integration value.** (a) Inference retains candidates online; training deploys the resulting student alone. (b,c) The same nine pairs are compared with their better fixed candidate and better single-teacher student, respectively. Filled/open markers denote gains/losses, not uncertainty. Pair scores and full checkpoint names are in Tables 2 and 4.

model calls, only S and its trained variants can be deployed; when online access is permitted, single-candidate inference and collaboration additionally become feasible.

Our objective is not simply to identify the strongest candidate, but to measure the *additional value created by integration*. Collaboration is therefore compared with using the better single candidate, while multi-teacher SFT is compared with the better single-teacher student. These role-specific references separate improvements obtained from accessing a stronger model from improvements that are actually created by the integration mechanism. Section 3 develops the supply–conversion framework used to analyze these gains.

3 Supply–Conversion Framework

Our supply–conversion framework distinguishes the opportunities supplied by candidate models from the gains realized by different integration mechanisms. Online collaboration must recover complementary successes without sacrificing the stronger candidate's correct answers, whereas offline transfer must convert supervision into behavior retained by the student. We derive an exact characterization for collaboration, measure realized transfer under a specified training recipe, and identify the signals available before role selection.

3.1 Complementary Coverage and Collaboration Gain

Fix a domain, input distribution, and candidate-generation protocol. Let $c_a, c_b \in \{0, 1\}$ indicate correctness, with $p_a = \mathbb{P}(c_a = 1) \geq p_b = \mathbb{P}(c_b = 1)$. Define

$$s = \mathbb{P}(c_a = c_b = 1), \qquad D = \mathbb{P}(c_a \neq c_b), \qquad \Delta = p_a - p_b. \tag{1}$$

Here, D is *correctness disagreement*: it measures the probability that exactly one candidate is correct, corresponding to cases where selection can recover a correct answer from the candidate pair. Oracle selection attains $u = s + D = p_a + p_b - s$. Its headroom beyond the stronger candidate is

$$H = u - p_a = p_b - s = \frac{D - \Delta}{2}. \tag{2}$$

The identity $D = 2H + \Delta$ separates complementary coverage from an advantage already supplied by the stronger model. Marginal accuracies cannot distinguish these contributions: equally accurate candidates can make identical errors or solve different examples. Measuring H requires paired correctness on the task. Self-pairs provide a reference for coverage available through resampling without introducing a different model.

For $D > 0$, let $\eta_{J,a,b,d}$ be the judge's probability of selecting the correct answer conditional on exactly one candidate being correct. The subscripts permit dependence on judge, pair, and domain; we abbreviate it as η. Both-correct and both-wrong events cannot distinguish valid selection rules.

Proposition 1 (Collaboration margin)**.** *If the judge returns one of the two fixed answers on every example, then* $A_{\text{col}} = s + \eta D$ *and* $G_{\text{col}} = A_{\text{col}} - p_a = H - (1-\eta)D = \frac{(2\eta-1)D-\Delta}{2}$*. For* $D > 0$*, positive gain occurs exactly when*

$$\eta > \rho \equiv \frac{1}{2} + \frac{\Delta}{2D}. \tag{3}$$

The proof counts successes on both-correct and exactly-one-correct examples, then subtracts the stronger candidate's score (Appendix B). In Equation 1, H is complementary supply and $(1-\eta)D$ is selection loss. Their difference is the *collaboration margin*. The boundary ρ is the stronger candidate's conditional accuracy on disagreements: the judge must beat this default, not merely beat chance.

Equal candidate accuracy reduces the requirement to beating chance when $D > 0$. Without weaker-only successes, even perfect selection can only tie the stronger candidate. When $D = 0$, selection cannot change accuracy and η is undefined. No independence of candidate errors is required. The identity applies to fixed-answer selection, not rewriting or additional sampling; invalid selections and domain aggregation are treated in Appendix B. A larger gap makes selection harder at fixed D, η, but neither gap alone nor the identity establishes a transferable decision threshold.

The joint events also fix the measurement unit. Candidate correctness, judge choices, and the resulting accuracy must refer to the same generated answers. Separately measured standalone scores remain useful references, but cannot be substituted into a matched-run identity. This distinction separates exact accounting from an empirical comparison across independently recorded runs.

3.2 Distillation and Transfer Metrics

For teacher t and recipe r, let $S_{t,r}$ be the trained student. Its primary outcome is $G_{\text{SFT}} = \text{Acc}(S_{t,r}) - \text{Acc}(S)$. When the teacher's macro accuracy exceeds the student's by at least one point, we additionally report

$$\tau_{t \to S;r} = \frac{\text{Acc}(S_{t,r}) - \text{Acc}(S)}{\text{Acc}(t) - \text{Acc}(S)}. \tag{4}$$

We therefore use τ as a descriptive measure of realized transfer efficiency rather than as a predictive score. Importantly, H and τ characterize different stages of integration: H measures recoverable online opportunity, whereas τ summarizes transfer realized after training. Pre-training teacher selection consequently requires separate observable signals.

For two teachers, the integration contrast is

$$G_{\text{mix}} = \overline{A}(S_{a+b,r}) - \max\{\overline{A}(S_{a,r}), \overline{A}(S_{b,r})\}, \tag{5}$$

where bars denote four-domain means. Beating the constituent average is weaker than preserving the better teacher's value. With a fixed prompt budget, uniform mixing replaces supervision rather than adding a second full dataset. The contrast measures allocation value, not its cause.

3.3 Probe-Based Selection

A task probe provides candidate accuracies and joint correctness, while calibration estimates conditional judge accuracy from other pairs in the same domain. The former measures the current opportunity, and the latter estimates its converter without using the predicted judge outcomes. Calibration reuse is an empirical hypothesis, not implied by the margin identity.

Algorithm 1 uses these inputs within two deployment branches. Online selection compares the predicted margin with zero and otherwise retains the probe-preferred candidate. Teacher selection uses pre-SFT task accuracy, never the candidate's post-training τ. Mixtures remain evaluation alternatives even though this rule does not choose them by default.

Conditional forecasting withholds a pair's judge outcomes while supplying its correctness statistics. Finite-probe evaluation additionally tests errors from estimating those inputs. These distinguish explanation of an observed score from prediction and decision evidence. Section 7 evaluates the two branches; global train–team selection also needs a common utility.

Algorithm 1 Role-aware selection within the evaluated deployment branches

Require: Candidate pair a, b, labeled task probe, separately calibrated $\widehat{\eta}_d$.
1: **if** external candidates may run online (T1) **then**
2: Estimate $\widehat{p}_a, \widehat{p}_b, \widehat{D}$ and $\widehat{\Delta} = |\widehat{p}_a - \widehat{p}_b|$ per domain.
3: Select TEAM if $[(2\widehat{\eta}_d - 1)\widehat{D} - \widehat{\Delta}]/2 > 0$; otherwise use the probe-preferred candidate.
4: **else if** a teacher must be chosen for target-only training (T2) **then**
5: Select the candidate with higher pre-SFT task accuracy; do not mix by default.
6: **end if**

Table 1: **Four-domain study design.** Each domain supplies 1,000 SFT prompts and 250 evaluation items. Sources and scoring details are in Appendix A.

Domain	SFT source	Evaluation	Metric
Math	GSM8K train	GSM8K (250)	Numeric EM
Science	ARC-C/E (500 each)	ARC-Challenge (250)	Option acc.
Code	Code Alpaca	HumanEval + MBPP (164 + 86)	pass@1
Reasoning	RuleTaker/ProofWriter (500 each)	RuleTaker (250)	Label acc.

4 EXPERIMENTAL SETUP

This section summarizes the data, study coverage, and evaluation protocol used to test the supply–conversion framework across collaboration and distillation.

Data and evaluation. We evaluate mathematics, science, coding, and reasoning using the balanced suite in Table 1. The SFT dataset contains 4,000 prompts, evenly split across the four domains, and all main configurations share this prompt pool. Evaluation uses 250 items per domain with domain-specific metrics. For uniform two-teacher SFT, each teacher receives half of the prompts in every domain, keeping the total example budget fixed.

Models and study coverage. The main SFT study contains 18 teachers and three students, yielding 54 teacher–student configurations. Nine candidate pairs are evaluated under both collaboration and uniform two-teacher SFT, allowing direct cross-role comparison. An expanded collaboration set of 23 retained pairings supports coverage, calibration, and finite-probe analyses. A separate fixed-student study examines sensitivity to supervision and optimization choices. Full model identities and training details are provided in Appendix A.

Evaluation metrics and references. Scores are percentages and gains are percentage points (pp), with the unweighted four-domain mean as the headline metric. For collaboration, the *better fixed candidate* has the higher four-domain mean, while the *per-domain reference* selects the stronger member within each domain. The same distinction applies to single-teacher students. These are retrospective references rather than deployed routers. Figure 1 and Table 2 use fixed references, whereas the collaboration identity is domain-specific. Section 7 evaluates decisions made from task-local probes.

5 COLLABORATION RESULTS AND ANALYSIS

5.1 CANDIDATE COMPLEMENTARITY AND COLLABORATION GAINS

Positive collaboration gains concentrate in balanced pairs, whereas larger-gap combinations include the largest losses (Figure 1b). The stronger per-domain reference reduces positive margins without changing their signs. The margin therefore depends jointly on candidate imbalance, disagreement, and judge efficiency, rather than on a universal accuracy-gap cutoff.

The expanded study separates these ingredients (Figure 2a). Pairs with comparable disagreement can leave different amounts of unique-correct coverage. Balanced pairs retain more headroom, while cross-scale pairs disagree frequently but add few successes beyond the stronger source. High disagreement is therefore not evidence that the selector has more useful information to recover. Much of it can simply reflect one model correcting the other's failures.

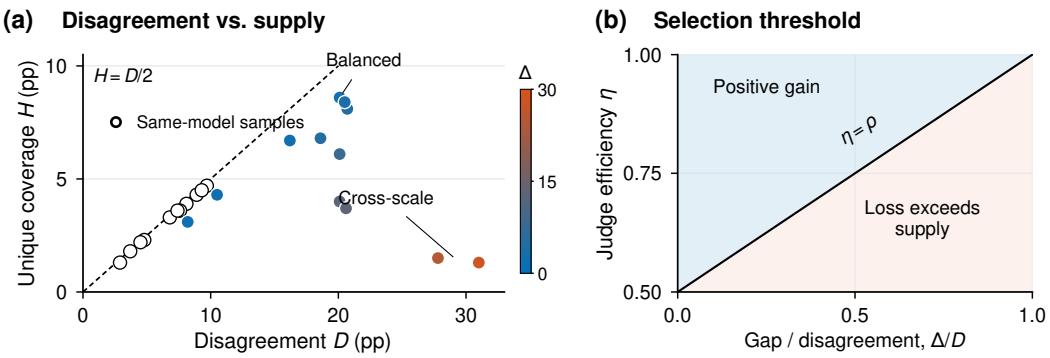


Figure 2: **Disagreement does not identify useful coverage.** (a) Paired-run headroom H versus disagreement D, colored by the accuracy gap Δ (pp); hollow markers are self-pairs. (b) Analytical selection boundary: the judge must exceed the stronger-candidate default. The aggregates in (a) are reported in Appendix D.

Self-pairs provide genuine sampling complementarity and an important reference for this interpretation. A second sample may recover errors without introducing a different model; heterogeneous coverage should not automatically be attributed to distinct expertise. At the opposite extreme, an unbalanced team can exceed the original target yet fall below its strongest member. These are not contradictory outcomes: candidate access improves the system's available answers, while imperfect selection gives back part of that advantage.

The margin distinguishes two bottlenecks. Small H leaves little recoverable benefit even for a reliable judge; larger H remains insufficient when selection errors consume it. An unsuccessful team therefore does not, by itself, demonstrate absent complementarity.

5.2 Judge Calibration and Cross-Pair Prediction

To test prediction rather than reconstruct an observed outcome, we estimate conditional judge accuracy from other pairs in the same domain, weighted by disagreement mass:

$$\widehat{\eta}_{-k,d} = \frac{\sum_{j \neq k} n_{j,d} D_{j,d} \eta_{j,d}}{\sum_{j \neq k} n_{j,d} D_{j,d}}, \qquad \widehat{A}_{k,d} = s_{k,d} + \widehat{\eta}_{-k,d} D_{k,d}. \tag{6}$$

For equal evaluation counts, $n_{j,d}$ cancels. Held-out pairs contribute correctness statistics but not judge outcomes, separating calibration transfer from finite-probe estimation of supply.

Judge efficiency differs across domains but varies comparatively little across the studied candidate groups (Figure 3a). Calibration from other pairs consequently tracks the collaboration margin and usually recovers the direction of improvement (Figure 3b). The break-even condition recovers 81 of 84 calibration-cell signs. Pair-specific opportunities therefore need task-local measurement, while selection behavior can be estimated from other combinations within this setting. These results support cross-pair reuse of judge calibration within the evaluated task and judge setting.

The gain coordinates make the decision boundary visible. Successful teams offer modest improvements, whereas harmful combinations can lose more accuracy. Avoiding unnecessary integration can therefore matter as much as identifying promising teams. The remaining question is whether this benefit survives uncertainty in the pair-specific inputs, as tested with finite probes in Section 7.

6 Distillation Results and Analysis

6.1 Single-Teacher Transfer

The complete matrix shows that single-teacher supervision usually helps, but also exposes weak gains and negative transfer (Figure 4a). Self-supervision provides little improvement, and another small candidate can harm the student. The strongest teacher is consistently effective across students and domains. A strong fixed teacher is therefore an essential baseline, rather than an outcome to remove before motivating a more elaborate selector.

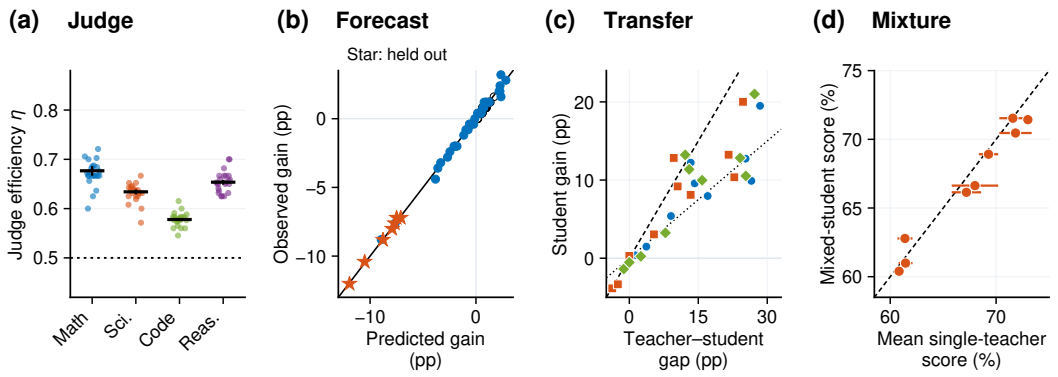


Figure 3: **Two roles impose different conversion requirements.** (a) Pair-level judge efficiency and pooled domain estimates. (b) Predicted versus measured team gain; stars mark held-out pairs. Forecast MAE is 0.14 pp in calibration and 0.10 pp on held-out pairs. (c) Student gain versus teacher–student gap; blue circles, orange squares, and green diamonds denote Qwen4B, GemmaE4B, and Ministral3B. Lines show $\tau = 1$ and $1/2$. (d) Mixture scores; segments span single-teacher scores, not uncertainty, and the diagonal marks their mean.

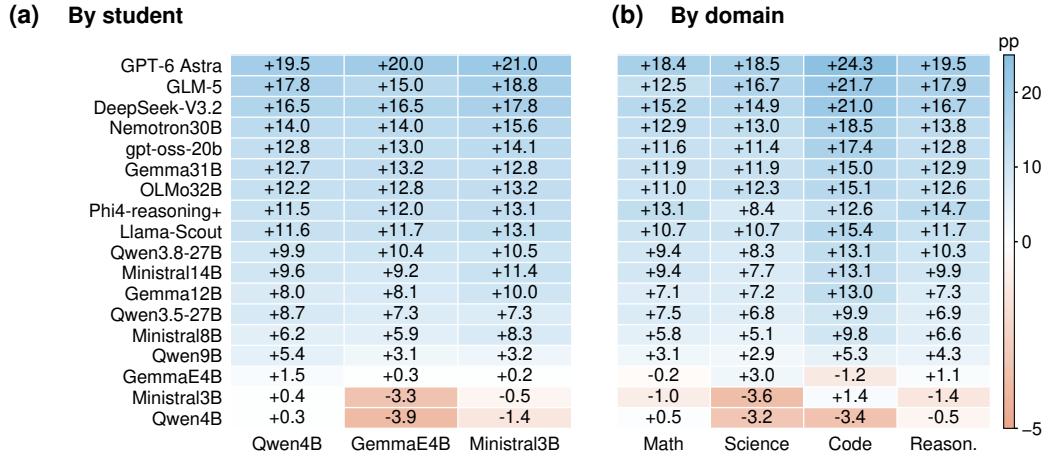


Figure 4: **Stable teacher rankings coexist with uneven transfer.** (a) All 54 teacher–student macro gains. (b) The same teachers' domain gains, averaged over the three students. Both panels share the row order and color scale; every value is a change from the corresponding untrained student in pp. Panel (b) aggregates the original runs, not additional training. Full domain scores are in Appendix C.

Teacher rankings are closely aligned across students under the shared recipe, but consistency across students does not imply agreement with standalone ability. OLMo-32B is weaker on its own than Qwen3.8-27B, yet produces the stronger Qwen3.5-4B student in every domain. The gap–gain comparison (Figure 3c) shows why a single transfer fraction is inadequate: a smaller teacher advantage can be converted more effectively than a larger one.

The domain view qualifies the stable macro rankings (Figure 4b). Strong supervision improves performance across domains, especially Code, whereas near-peer supervision can help one domain and harm another. These cells describe the same shared training run, not independently trained specialists: macro transfer is not uniform task-level improvement.

The cross-role comparison further distinguishes useful supervision from useful additional answers. OLMo supervision helps students, yet its tested collaborations reduce stronger candidates' accuracy; balanced small-model pairs provide useful inference coverage while offering little training improvement. Partner value and teacher value concern different outcomes and references.

Table 2: **Same candidate identities, different roles.** Four-domain macro scores (%) for Qwen4B as judge/student. Single and Distill denote the better *fixed* member and constituent single-teacher student, respectively. Δ is the improvement over the corresponding reference in pp. The base student scores 60.52. Qwen4B/Qwen9B denote Qwen3.5; Qwen3.8 retains its version.

		TEAM			TRAIN		
ID	**Candidate pair**	**Single**	**Team**	Δ_{team}	**Distill**	**Mix**	Δ_{mix}
1	Qwen4B+GemmaE4B	64.23	65.40	+1.18	62.00	62.78	+0.78
2	Qwen4B+Ministral3B	61.73	63.70	+1.98	60.89	60.40	-0.49
3	GemmaE4B+Ministral3B	64.23	65.80	+1.58	62.00	60.99	-1.01
4	Qwen9B+Gemma12B	77.56	76.20	-1.36	68.48	66.14	-2.34
5	Qwen9B+Ministral14B	74.74	74.50	-0.24	70.09	66.63	-3.46
6	Gemma12B+Ministral14B	77.56	78.10	+0.54	70.09	68.92	-1.18
7	Gemma31B+Qwen3.8-27B	87.07	87.80	+0.73	73.26	70.46	-2.80
8	Gemma31B+OLMo32B	85.82	82.30	-3.52	73.26	71.43	-1.83
9	Qwen3.8-27B+OLMo32B	87.07	83.00	-4.07	72.75	71.54	-1.21

6.2 Uniform Teacher Mixing andTraining Sensitivity

Uniform mixing rarely improves on the better single teacher (Table 2). Figure 3d makes this comparison explicit: being near the mean of two single-teacher students does not preserve their strongest behavior. Most mixtures fall below that reference, though individual domains can benefit. The result concerns random prompt allocation at a fixed example budget (Zhang et al., 2026; Chen et al., 2024b).

Each teacher supplies fewer examples than in its single-teacher run, so these contrasts do not isolate knowledge conflict from replacement of useful supervision. They establish an operational requirement: a mixture must earn its inclusion relative to the better constituent, not merely relative to the untrained student. Oracle coverage cannot be assumed to survive distillation: data allocation and learning intervene before deployment.

The sensitivity study further reveals that teaching value is jointly determined by teacher identity and the transfer process. Output construction, training budget, and optimization alter the magnitude—and occasionally the ordering—of transfer gains across domains. Appendix F reports all eight groups. This complements the main matrix: rankings are highly structured under a fixed recipe, while realized teacher value remains conditional on how supervision is constructed and optimized.

7 Model Selection Evaluation

Algorithm 1 is evaluated in two settings. T1 chooses single-candidate use or collaboration when external calls are allowed, using a task probe and other pairs' judge calibration. T2 chooses a teacher for target-only deployment using pre-SFT task accuracy.

For domain d, $\mathcal{F}_d$ denotes the feasible alternatives and $\widehat{a}_d$ the chosen action. We measure regret as

$$R = \frac{1}{|\mathcal{D}|} \sum_{d} \left[\max_{a \in \mathcal{F}_d} A_d(a) - A_d(\widehat{a}_d) \right], \tag{7}$$

with averaging over candidate pools as well. T1 is scored on items not used as the current-task probe; T2 is compared with the two available single-teacher students and their mixture. Appendix G distinguishes training, calibration, probe, and evaluation data.

Finite probes test whether the collaboration margin remains useful when its pair-specific inputs must be estimated. More observations reduce collaboration regret (Figure 5). At the larger budget, the rule improves on generic-score and capability-profile rules despite their optimistic thresholds. Disagreement-only, size-based, and fixed policies incur greater regret (Table 3).

Low regret is not perfect action prediction. Mistakes near a tie can cost little, whereas avoiding a harmful team preserves substantial accuracy. Agreement and lost performance should be read together, retaining single-candidate use as an alternative to integration. Teacher selection exhibits a different pattern: task-local accuracy improves over generic capability proxies and unconditional mixing, while parameter count remains competitive in the restricted nine-pair evaluation. This contrast reinforces

Table 3: **Selection regret (pp; lower is better).** At 125 probe items per domain. T1: single candidate or team; T2: teacher choice. Agreement and protocols: Appendix G.

Selection rule	T1	T2
Task-local probe	**0.44**	1.11
Generic-score proxy	0.86	2.16
Capability profile	0.86	2.16
Disagreement rate	1.62	–
Parameter count	1.66	**0.84**
Always collaborate	1.57	–
Always one candidate	1.47	–
Always mix	–	2.01
Per-domain oracle	0.00	0.00

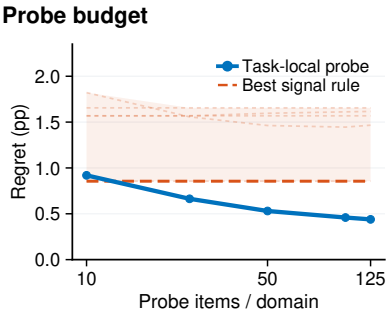


Figure 5: **Finite probes improve selection.** Shading spans signal baselines, not uncertainty. All five measured budgets are retained.

the central distinction of our framework: unlike paired correctness in collaboration, pre-training teacher attributes are only indirect proxies for realized transfer.

The two decisions use task evidence differently. Paired correctness enters the collaboration identity, whereas teacher accuracy only proxies the behavior a student may learn. Their different selection outcomes are thus consistent with the framework: conversion is role-specific, even when the candidate pool and task are shared. When deployment cost is part of the objective, the same comparison extends naturally to a utility criterion; Appendix H.2 derives the corresponding break-even condition.

8 Related Work

Inference-time integration. Inference-time integration spans sampling, ranking, fusion, multi-model discussion, and routing (Wang et al., 2023; Jiang et al., 2023; Chen et al., 2024a; Wang et al., 2025; Ong et al., 2025). Our exact margin concerns selecting between two fixed answers; rewriting, debate, and additional sampling can change the available answer set. The generation–verification gap also makes the selector's capability a separate object of study (Song et al., 2025). We measure it on correctness disagreements, where selection changes accuracy.

Distillation and teacher selection. Distillation transfers teacher behavior through objectives and supervision construction (Hinton et al., 2015; Kim & Rush, 2016; Gu et al., 2024; Hsieh et al., 2023). CAR and GRACE show why teacher suitability deserves measurements beyond standalone strength (Xu et al., 2025; Panigrahi et al., 2026). PerSyn allocates prompts across teachers, while MAGDi distills structured multi-agent interactions (Zhang et al., 2026; Chen et al., 2024b). We compare roles and a fixed-budget uniform mixture, rather than benchmark those selectors or objectives.

Model similarity and efficient evaluation. Model similarity and efficient evaluation provide another perspective. Goel et al. (2025) study similarity in model judging and weak-to-strong learning; Kim (2026) distinguishes capability-related diversity from ensemble gains. We add a selection-specific margin, shared-pool training comparisons, and decision evaluation. tinyBenchmarks estimates performance from fewer examples (Maia Polo et al., 2024); here, probes estimate joint correctness for a task. Estimating marginal performance and complementary coverage are distinct objectives.

9 Discussion and Conclusion

Candidate integration value is role-dependent: the same model can provide different value as an inference partner and a training teacher. Our supply–conversion framework explains this difference by separating capability made available by candidates from the mechanism that converts it into realized gain. For collaboration, the margin exactly characterizes how complementary coverage and selection quality determine improvement over the stronger candidate. For SFT, shared-pool experiments show that teacher value follows a different structure and that the better single teacher remains a strong reference for uniform mixtures. Cross-pair calibration and finite-probe evaluation further translate this analysis into collaboration decisions. Together, these results provide a common basis for analyzing candidate value across online integration and offline transfer.

Reproducibility Statement

The problem formulation, training configuration, and evaluation sources are described in Section 4 and Appendix A. Appendix B provides proofs and aggregation rules. Appendices C–G retain the complete supplied SFT, mixing, sensitivity, and decision summaries.

Ethics Statement

This study evaluates language models on standard academic benchmarks and uses model-generated supervision. It does not introduce human-subject experiments or private user data. Candidate models and datasets retain their original usage restrictions.

AI Use Statement

We used LLMs solely as writing aids to polish language (e.g., grammar, wording, and clarity) for text written by the authors, and to assist in generating synthetic training data. All suggestions and generated data were reviewed by the authors and accepted, modified, or discarded at our discretion. No passages were inserted without author verification, and all references and factual claims were curated and checked by the authors.

The authors take full responsibility for the contents of this paper, including any text refined with LLM assistance. LLMs are not eligible for authorship and are not listed as authors. No non-public or sensitive data were provided to the LLM service beyond manuscript text, and our usage adhered to relevant institutional and venue policies on LLM use.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Justin Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. MAGDi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7220–7235. PMLR, 2024b. URL https://proceedings.mlr.press/v235/chen24ah.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3882–3890, 2020. doi: 10.24963/ijcai.2020/537. URL https://www.ijcai.org/Proceedings/2020/537.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19621–19678. PMLR, 2025. URL https://proceedings.mlr.press/v267/goel25b.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Donghwan Kim. Are diversity metrics measuring diversity? a capability-controlled audit of majority-vote gain in LLM ensembles. *arXiv preprint arXiv:2607.20768*, 2026. URL https://arxiv.org/abs/2607.20768.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good GRACES: Principled teacher selection for knowledge distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59fa5d049b777d880fdcddba2b24738f-Abstract-Conference.html.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *The Thirteenth International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63943ee9fe347f3d95892cf87d9a42e6-Abstract-Conference.html.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y. Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcaf7fca52b6-Abstract-Conference.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Stronger models are not always stronger teachers for instruction tuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4392–4405, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.224. URL https://aclanthology.org/2025.naacl-long.224/.

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X. Chang, and Ngai Wong. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14619–14637, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.666. URL https://aclanthology.org/2026.acl-long.666/.

A A EXPERIMENTAL DETAILS

A.1 MODEL IDENTITIES AND EXPERIMENTAL UNITS

Table 4: Model labels in the main study. T: teacher; S: student; C: inference candidate.

Alias	Model checkpoint	Roles
Qwen4B	Qwen3.5-4B	T, S, C
Qwen9B	Qwen3.5-9B	T, C
Qwen3.5-27B	Qwen3.5-27B	T
GemmaE4B	gemma-4-E4B-it	T, S, C
Gemma12B	gemma-4-12B-it	T, C
Gemma31B	gemma-4-31B-it	T, C
Ministral3B	Ministral-3-3B-Instruct-2512	T, S, C
Ministral8B	Ministral-3-8B-Instruct-2512	T
Ministral14B	Ministral-3-14B-Instruct-2512	T, C
Qwen3.8-27B	Qwen3.8-27B	T, C
Llama	Llama-4-Scout-17B-16E-Instruct	T
gpt-oss-20b	gpt-oss-20b	T
OLMo32B	Olmo-3.1-32B-Instruct	T, C
Nemotron	NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16	T
Phi	Phi-4-reasoning-plus	T
DeepSeek-V3.2	DeepSeek-V3.2	T, C
GLM-5	GLM-5	T, C
GPT-6 Astra	GPT-6 Astra	T, C

The three students are Qwen4B, GemmaE4B, and Ministral3B. The nine-pair role comparison fixes Qwen4B as both the frozen judge and the trainable student, in separate experiments. Figure 3 labels the student columns Qwen4B, GemmaE4B, and Ministral3B and abbreviates the strongest teacher as GPT-6 Astra. Qwen3.5-27B (Qwen3.5-27B) is distinct from Qwen3.8-27B (Qwen3.8-27B). The 18-by-three matrix contains 54 training configurations evaluated on four domains, not 216 independent training runs.

The main nine pairs have both realized collaboration and uniform-mixture SFT scores. An expanded collaboration study starts from 26 pair configurations and retains 23 after a consistency screen, including self-pairs and cross-scale pairs, giving 84 calibration and eight held-out pair–domain cells.

The separate sensitivity study uses DeepSeek-R1-Distill-Qwen-1.5B. Qwen2.5-14B/32B denote Instruct models; R1-Distill14B denotes DeepSeek-R1-Distill-Qwen-14B. QwQ-32B, Qwen3-14B, and GPT-5.5-2026-04-24 retain their source labels.

A.2 DATA CONSTRUCTION AND SCORING

Math training draws 1,000 prompts from GSM8K train (Cobbe et al., 2021). Science draws 500 each from ARC-Challenge and ARC-Easy train (Clark et al., 2018). Code uses 1,000 prompts from `flwrlabs/code-alpaca-20k`. Reasoning uses 500 RuleTaker (Clark et al., 2020) and 500 ProofWriter OWA (Tafjord et al., 2021) prompts. The latter has True/False/Unknown labels.

Evaluation uses 250 GSM8K test items, 250 ARC-Challenge test items, 164 HumanEval (Chen et al., 2021) plus 86 MBPP-sanitized (Austin et al., 2021) items, and 250 RuleTaker items (125 entailment and 125 not-entailment). Metrics are numeric exact match, option accuracy, execution pass@1, and label accuracy. The scoring description treats failed generations, unparseable answers, and execution errors as incorrect.

Table 5: Training configuration shared by the main SFT matrix and nine uniform mixtures.

Setting	Value
Update	Full-parameter SFT
Training prompts	4,000 total; 1,000 per domain
Epochs; learning rate	3; 10^{-5}
Sequence cutoff	32,768 tokens
Per-device train/eval batch	1 / 1
Gradient accumulation	32 steps
Warmup; precision	0.05; bfloat16
Two-teacher allocation	Random 50/50 split within each domain
Main judge	Frozen pre-SFT Qwen3.5-4B; answer selection only

B Derivations and Aggregation

B.1 Joint correctness and tight bounds

Let $q_{ij} = \mathbb{P}(c_a = i, c_b = j)$. The four cells are nonnegative and sum to one. For $p_a \geq p_b$,

$$p_a = q_{11} + q_{10}, \qquad p_b = q_{11} + q_{01}, \tag{8}$$

$$s = q_{11}, \qquad u = q_{11} + q_{10} + q_{01}, \tag{9}$$

$$D = q_{10} + q_{01}, \qquad \Delta = q_{10} - q_{01}. \tag{10}$$

It follows that $q_{10} = (D + \Delta)/2$ and $q_{01} = (D - \Delta)/2 = H$. Nonnegativity gives $\max(0, p_a + p_b - 1) \leq s \leq p_b$. Substituting into $H = p_b - s$ yields

$$0 \leq H \leq \min(p_b, 1 - p_a), \qquad \Delta \leq D \leq \min(p_a + p_b, 2 - p_a - p_b). \tag{11}$$

Both endpoints are attainable by a valid joint distribution, so these bounds are tight.

For the illustrative case $p_a = p_b = 0.8$, identical errors give

$$(q_{11}, q_{10}, q_{01}, q_{00}) = (0.8, 0, 0, 0.2).$$

Disjoint errors instead give

$$(q_{11}, q_{10}, q_{01}, q_{00}) = (0.6, 0.2, 0.2, 0).$$

These two hypothetical distributions share the same marginal accuracies but have different headroom.

B.2 Gain and the default selection rule

Conditional on exactly one correct answer, always selecting a succeeds on the q_{10} events. Its conditional accuracy is $\rho = q_{10}/D = 1/2 + \Delta/(2D)$. A judge succeeds on ηD of all inputs in this category. Since both-correct and both-wrong categories do not distinguish valid selection rules,

$$A_{\text{col}} - p_a = \eta D - q_{10} = (\eta - \rho)D.$$

Equivalently, perfect selection attains u and the judge loses $(1 - \eta)D$, yielding Proposition 1. When $D = 0$, conditioning is undefined; accuracy remains $s = p_a = p_b$. When $H = 0$, even perfect selection can only tie the stronger candidate.

Corollary 1 (Feasibility from marginal accuracy)**.** *For a fixed hypothetical conditional accuracy $\eta > 1/2$, positive gain is possible for some joint distribution with given $p_a \geq p_b$ only if*

$$(2\eta - 1)\min(p_a + p_b, 2 - p_a - p_b) > \Delta.$$

This condition is also sufficient within the abstract binary-correctness model. If $\eta \leq 1/2$, positive gain over the stronger candidate is impossible.

Proof. For $\eta > 1/2$, Equation (1) increases with D. Substitute the maximum feasible D from Equation (11). The maximum is attainable by the minimum feasible s. If $\eta \leq 1/2$, both terms in $(2\eta - 1)D - \Delta$ are nonpositive. □

The corollary holds η fixed as a hypothetical quantity. It does not imply that a real judge maintains the same η when the candidate error distribution changes.

B.3 INVALID SELECTIONS

The exact identity assumes that the judge returns one candidate answer on every example. If an invalid output is scored wrong, let r_{11} be the probability of a valid selection conditional on both candidates being correct. Define η to count invalid outputs as failures on the exactly-one-correct category. Then

$$A_{\text{col}} = r_{11}s + \eta D, \qquad A_{\text{col}} - p_a = H - (1-\eta)D - (1-r_{11})s.$$

The additional term captures failures even when both answers are correct. A fallback that always selects one candidate restores the original accounting, but changes η and must be specified. The original prompt and fallback implementation must be retained when reproducing an empirical estimate.

B.4 DOMAIN AGGREGATION AND RETROSPECTIVE BASELINES

For weights $w_d \geq 0$, $\sum_d w_d = 1$, let $B_{\text{domain}} = \sum_d w_d \max(p_{a,d}, p_{b,d})$ and $B_{\text{fixed}} = \max(\sum_d w_d p_{a,d}, \sum_d w_d p_{b,d})$. Then $B_{\text{domain}} \geq B_{\text{fixed}}$ and

$$\bar{A}_{\text{col}} - B_{\text{fixed}} = (\bar{A}_{\text{col}} - B_{\text{domain}}) + (B_{\text{domain}} - B_{\text{fixed}}).$$

The second term is the advantage of retrospective domain choice, not a gain supplied by the judge. We use $w_d = 1/4$ throughout. The gain against the per-domain reference is

$$\bar{G}_{\text{col,domain}} = \sum_d w_d[H_d - (1-\eta_d)D_d].$$

If a pooled η is needed for this expression, its weights must be proportional to $w_d D_d$: $\eta_{\text{pool}} = \sum_d w_d D_d \eta_d / \sum_d w_d D_d$ when the denominator is positive. An unweighted average of conditional accuracies generally gives a different result.

The same maximum-before-or-after-averaging issue applies to constituent single-teacher scores. Table 6 reports both references.

Table 6: Alternative aggregation references. Domain oracle chooses the better constituent separately in each domain. Mix–mean compares with the arithmetic mean of constituent SFT macro scores. Gains are in pp. These are retrospective references.

ID	Pair	Cand. oracle	Team Δ	SFT oracle	Mix Δ	Mix–mean
1	Qwen4B+GemmaE4B	64.52	+0.89	62.25	+0.54	+1.39
2	Qwen4B+Ministral3B	62.57	+1.13	61.77	-1.37	-0.44
3	GemmaE4B+Ministral3B	64.88	+0.92	62.44	-1.45	-0.46
4	Qwen9B+Gemma12B	77.56	-1.36	68.48	-2.34	-1.05
5	Qwen9B+Ministral14B	74.74	-0.24	70.09	-3.46	-1.37
6	Gemma12B+Ministral14B	77.56	+0.54	70.32	-1.41	-0.37
7	Gemma31B+Qwen3.8-27B	87.31	+0.49	73.26	-2.80	-1.38
8	Gemma31B+OLMo32B	85.82	-3.52	73.34	-1.91	-1.57
9	Qwen3.8-27B+OLMo32B	87.07	-4.07	72.75	-1.21	-0.04

C COMPLETE STANDALONE AND SINGLE-TEACHER RESULTS

The following tables contain all main-study standalone scores and every cell of the 18-teacher by three-student SFT matrix. Macro scores are calculated from the four displayed domains, before rounding to two decimals.

Figure 4 in the main text summarizes these same runs by student and by domain.

Table 7: Standalone evaluation scores of the candidate models used in the main study.

Model	Math	Science	Code	Reasoning	Macro
Qwen3.5-4B	71.48	62.54	42.15	65.90	60.52
Gemma-E4B	70.32	69.57	47.76	69.25	64.23
Ministral-3B	69.78	61.93	50.37	64.82	61.73
Qwen3.5-9B	78.04	69.09	56.08	75.10	69.58
Gemma-12B	82.07	80.05	67.08	81.04	77.56
Ministral-14B	82.04	74.04	65.10	77.76	74.74
Gemma-31B	89.07	87.05	79.08	88.07	85.82
Qwen3.8-27B	91.07	86.07	82.07	89.06	87.07
OLMo-32B	82.06	74.04	59.37	80.04	73.88
GPT-6 Astra	94.24	92.43	76.81	92.35	88.96

Table 8: Complete single-teacher results for student Qwen3.5-4B. Percentages; Macro is the equal-domain average. Bold marks the best score in each column.

Teacher	Math	Science	Code	Reasoning	Macro
Base	71.48	62.54	42.15	65.90	60.52
Qwen3.5-4B	71.33	62.85	41.82	67.13	60.78
Qwen3.5-9B	75.35	66.17	50.47	71.67	65.92
Qwen3.5-27B	79.78	70.29	54.63	72.27	69.24
Gemma-E4B	70.34	65.79	44.26	67.61	62.00
Gemma-12B	77.52	70.16	53.84	72.39	68.48
Gemma-31B	81.20	74.65	59.31	77.86	73.26
Ministral-3B	70.58	62.00	45.76	65.23	60.89
Ministral-8B	76.39	66.82	53.17	70.64	66.76
Ministral-14B	79.71	69.24	56.58	74.83	70.09
Qwen3.8-27B	79.18	70.76	56.43	75.29	70.42
Llama-Scout	80.47	72.91	58.24	76.68	72.08
gpt-oss-20b	81.64	73.28	60.79	77.43	73.29
OLMo-32B	81.36	74.83	57.21	77.59	72.75
Nemotron-30B	82.73	75.19	61.86	78.42	74.55
Phi-4-reasoning+	83.20	69.74	55.43	79.81	72.05
DeepSeek-V3.2	84.68	77.25	64.91	81.37	77.05
GLM-5	85.42	79.87	65.23	82.64	78.29
GPT-6 Astra	**87.60**	**80.34**	**67.85**	**84.27**	**80.02**

Table 9: Complete single-teacher results for student Gemma-E4B. Percentages; Macro is the equal-domain average. Bold marks the best score in each column.

Teacher	Math	Science	Code	Reasoning	Macro
Base	70.32	69.57	47.76	69.25	64.23
Qwen3.5-4B	71.08	60.46	43.71	66.24	60.37
Qwen3.5-9B	73.41	71.86	51.29	72.54	67.28
Qwen3.5-27B	77.84	75.36	56.19	76.71	71.53
Gemma-E4B	70.83	70.05	46.80	70.42	64.53
Gemma-12B	78.67	76.25	56.93	77.48	72.33
Gemma-31B	84.42	80.73	62.18	82.54	77.47
Ministral-3B	69.81	60.47	47.35	65.92	60.89
Ministral-8B	75.31	73.89	55.46	75.72	70.10
Ministral-14B	79.36	76.67	59.12	78.49	73.41
Qwen3.8-27B	80.61	77.84	60.35	79.52	74.58
Llama-Scout	81.35	79.72	62.56	80.13	75.94
gpt-oss-20b	82.91	80.16	64.57	81.34	77.25
OLMo-32B	82.14	81.68	62.47	81.93	77.06
Nemotron-30B	83.57	81.24	65.71	82.38	78.23
Phi-4-reasoning+	84.36	77.62	59.18	83.55	76.18
DeepSeek-V3.2	86.53	83.16	67.72	85.44	80.71
GLM-5	77.31	84.76	68.19	86.53	79.20
GPT-6 Astra	**90.73**	**86.48**	**71.16**	**88.59**	**84.24**

Table 10: Complete single-teacher results for student Ministral-3B. Percentages; Macro is the equal-domain average. Bold marks the best score in each column.

Teacher	Math	Science	Code	Reasoning	Macro
Base	69.78	61.93	50.37	64.82	61.73
Qwen3.5-4B	70.63	61.18	44.52	65.07	60.35
Qwen3.5-9B	72.13	64.68	54.32	68.75	64.97
Qwen3.5-27B	76.45	68.92	59.24	71.58	69.05
Gemma-E4B	69.73	67.15	45.68	65.31	61.97
Gemma-12B	76.71	69.36	68.59	72.14	71.70
Gemma-31B	81.69	74.26	63.91	78.37	74.56
Ministral-3B	68.14	60.63	51.28	64.75	61.20
Ministral-8B	77.25	68.53	60.91	73.38	70.02
Ministral-14B	80.75	71.28	63.94	76.41	73.10
Qwen3.8-27B	79.93	70.27	62.68	76.14	72.26
Llama-Scout	81.87	73.42	65.69	78.26	74.81
gpt-oss-20b	81.72	74.85	67.23	79.61	75.85
OLMo-32B	81.20	74.51	65.76	78.32	74.95
Nemotron-30B	83.92	76.46	68.15	80.63	77.29
Phi-4-reasoning+	83.27	71.93	63.48	80.71	74.85
DeepSeek-V3.2	85.89	78.31	70.67	83.24	79.53
GLM-5	86.28	79.65	71.92	84.37	80.56
GPT-6 Astra	**88.32**	**82.74**	**74.21**	**85.67**	**82.74**

D COLLABORATION RESULTS AND CALIBRATION SUMMARIES

D.1 THE NINE SHARED CANDIDATE POOLS

Table 11 gives all nine realized collaboration rows. The common-pool macro comparison is in Table 2; alternative fixed and per-domain references are in Table 6. Pair identifiers remain consistent across these comparisons.

Table 11: All nine answer-selection collaboration results. Qwen4B is the judge. Percentages.

ID	Candidates	Math	Science	Code	Reasoning	Macro
1	Qwen4B+GemmaE4B	74.80	69.20	46.80	70.80	65.40
2	Qwen4B+Ministral3B	73.20	64.80	48.40	68.40	63.70
3	GemmaE4B+Ministral3B	73.20	68.80	50.80	70.40	65.80
4	Qwen9B+Gemma12B	82.80	77.60	63.60	80.80	76.20
5	Qwen9B+Ministral14B	82.40	74.40	62.40	78.80	74.50
6	Gemma12B+Ministral14B	83.60	79.60	67.60	81.60	78.10
7	Gemma31B+Qwen3.8-27B	91.60	88.00	81.60	90.00	87.80
8	Gemma31B+OLMo32B	88.40	82.80	71.60	86.40	82.30
9	Qwen3.8-27B+OLMo32B	89.20	82.80	73.20	86.80	83.00

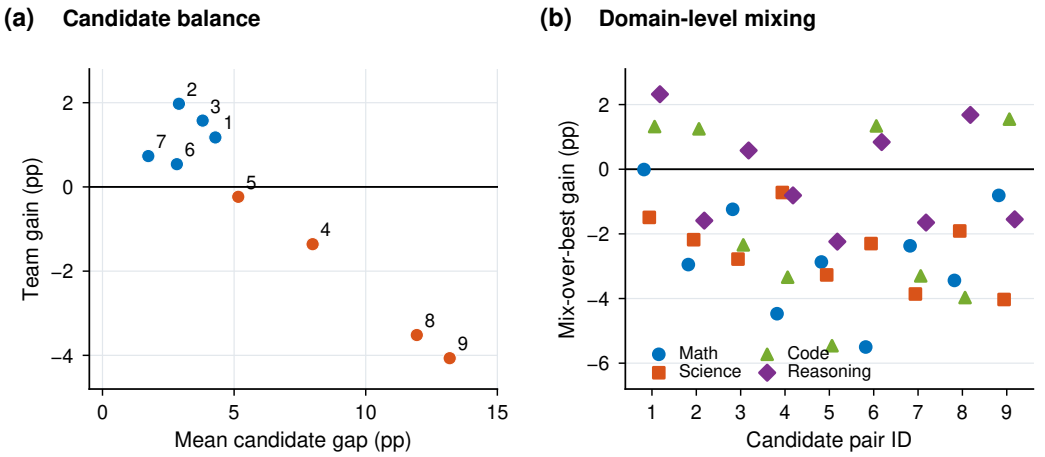


Figure 6: **Candidate balance and all domain-level mixture contrasts.** (a) The nine macro team gains against the better fixed candidate. (b) All 36 mixture gains against the better single teacher in each domain. Pair IDs follow Table 2.

D.2 EXPANDED STUDY AND CONSISTENCY SCREEN

We exclude three configurations whose paired-run candidate marginals differ from their corresponding standalone evaluations by more than 3.5 pp, and conduct the subsequent calibration analysis on the remaining 23 configurations. Table 12 reports the excluded cells for completeness.

Table 12: Flagged cells in the three configurations omitted from the expanded analysis. Values are retained from its consistency screen.

Pair	Domain	Candidate	Paired run	Reference	Difference
DeepSeek-V3.2 + GPT-6 Astra	Code	DeepSeek-V3.2	91.2	83.0	8.2
DeepSeek-V3.2 + GPT-6 Astra	Code	GPT-6 Astra	84.8	76.8	8.0
GLM-5 + GPT-6 Astra	Code	GLM-5	94.0	85.8	8.2
GLM-5 + GPT-6 Astra	Code	GPT-6 Astra	85.2	76.8	8.4
GPT-6 Astra + GPT-6 Astra	Math	GPT-6 Astra	98.0	94.2	3.8
GPT-6 Astra + GPT-6 Astra	Science	GPT-6 Astra	96.0	92.4	3.6
GPT-6 Astra + GPT-6 Astra	Code	GPT-6 Astra	92.8	76.8	16.0
GPT-6 Astra + GPT-6 Astra	Reasoning	GPT-6 Astra	96.8	92.3	4.5

Table 13: Domain summaries reported for heterogeneous pairs in the expanded collaboration study. Δ, D, H and accuracies are in percentage points/percent; η is a conditional probability. These rounded group summaries are not joint event counts for a representative pair.

Domain	Δ	D	H	Stronger reference	Collaboration	η
Math	6.4	15.6	4.6	84.5	83.2	0.677
Science	9.8	20.1	5.1	80.9	77.6	0.634
Code	12.6	25.1	6.2	69.2	65.5	0.578
Reasoning	7.6	17.4	4.9	82.5	80.7	0.654

D.3 Paired-run coverage across the expanded pool

Table 14 gives the retained paired-run summaries used to draw Figure 2a. Each row is the reported mean of four domain-level joint statistics, rounded to one decimal. The cross-scale combinations are the held-out candidate pairs in the expanded forecast study. Identical displayed model names in a self-pair refer to two generated samples, not identical answers. Minor deviations from $D = 2H + \Delta$ at the displayed precision are rounding effects.

Table 14: **Retained paired-run supply summaries.** The row identifiers belong to the expanded study, not the nine-pool comparison. All values are percentage points. Self-pairs provide a resampling reference.

ID	Group	Candidate pair	D	Δ	H
P01	Peer	Qwen4B (two samples)	9.7	0.3	4.7
P02	Peer	Qwen4B + GemmaE4B	20.7	4.5	8.1
P03	Peer	Qwen4B + Ministral3B	20.1	2.9	8.6
P04	Peer	GemmaE4B (two samples)	8.9	0.3	4.3
P05	Peer	GemmaE4B + Ministral3B	20.5	3.7	8.4
P06	Peer	Ministral3B (two samples)	9.3	0.3	4.5
P07	Mid	Qwen9B (two samples)	8.1	0.3	3.9
P08	Mid	Qwen9B + Gemma12B	20.1	7.9	6.1
P09	Mid	Qwen9B + Ministral14B	18.6	5.0	6.8
P10	Mid	Gemma12B (two samples)	6.8	0.2	3.3
P11	Mid	Gemma12B + Ministral14B	16.2	2.8	6.7
P12	Mid	Ministral14B (two samples)	7.6	0.4	3.6
P13	Strong	Gemma31B (two samples)	4.8	0.2	2.3
P14	Strong	Gemma31B + Qwen3.8-27B	10.5	1.9	4.3
P15	Strong	Gemma31B + OLMo32B	20.1	12.1	4.0
P16	Strong	Qwen3.8-27B (two samples)	4.5	0.1	2.2
P17	Strong	Qwen3.8-27B + OLMo32B	20.6	13.2	3.7
P18	Strong	OLMo32B (two samples)	7.4	0.2	3.6
P19	Frontier	DeepSeek-V3.2 (two samples)	3.7	0.1	1.8
P20	Frontier	DeepSeek-V3.2 + GLM-5	8.2	2.0	3.1
P22	Frontier	GLM-5 (two samples)	2.9	0.3	1.3
P25	Cross-scale	Qwen4B + GPT-6 Astra	31.0	28.4	1.3
P26	Cross-scale	GemmaE4B + GPT-6 Astra	27.8	24.8	1.5

D.4 EARLIER COVERAGE-ONLY RECORDS

Table 15 reports the complete lower and upper coverage bounds for the candidate pairs.

Table 15: Complete complementary-coverage bounds (%).

ID	Candidates	Math	Science	Code	Reasoning
1	Qwen4B×Qwen4B	63.60 / 74.40	58.00 / 67.20	37.20 / 47.20	61.60 / 70.40
2	Qwen4B×GemmaE4B	62.40 / 79.60	54.80 / 77.20	33.20 / 56.80	58.00 / 77.60
3	Qwen4B×Ministral3B	62.00 / 79.60	52.80 / 72.00	34.00 / 58.80	56.00 / 74.80
4	GemmaE4B×GemmaE4B	66.00 / 74.40	65.20 / 74.00	42.80 / 52.80	65.20 / 73.60
5	GemmaE4B×Ministral3B	61.60 / 78.80	54.40 / 77.20	38.40 / 60.00	56.80 / 77.20
6	Ministral3B×Ministral3B	65.60 / 74.00	57.20 / 66.80	45.60 / 55.60	60.40 / 69.60
7	Qwen9B×Qwen9B	74.80 / 81.60	64.80 / 73.20	51.20 / 60.80	71.20 / 78.80
8	Qwen9B×Gemma12B	72.40 / 87.60	63.60 / 85.60	48.80 / 74.40	69.20 / 86.80
9	Qwen9B×Ministral14B	72.00 / 86.80	62.00 / 81.20	48.40 / 72.80	68.40 / 84.40
10	Gemma12B×Gemma12B	79.20 / 85.20	76.80 / 83.20	62.80 / 71.60	78.00 / 84.00
11	Gemma12B×Ministral14B	75.20 / 88.00	68.00 / 86.00	56.40 / 75.60	72.00 / 86.80
12	Ministral14B×Ministral14B	78.00 / 84.00	70.40 / 78.00	60.40 / 69.60	74.00 / 81.60
13	Gemma31B×Gemma31B	87.20 / 91.20	84.80 / 89.20	75.60 / 82.40	86.00 / 90.00
14	Gemma31B×Qwen3.8-27B	85.60 / 94.40	81.60 / 91.60	73.20 / 87.60	84.00 / 92.80
15	Gemma31B×OLMo32B	78.80 / 92.80	70.40 / 90.80	54.00 / 84.40	76.40 / 92.00
16	Qwen3.8-27B×Qwen3.8-27B	89.60 / 92.80	83.60 / 88.40	79.20 / 85.20	87.20 / 91.20
17	Qwen3.8-27B×OLMo32B	79.20 / 94.00	70.00 / 90.00	54.80 / 86.40	76.40 / 92.40
18	OLMo32B×OLMo32B	79.20 / 85.20	70.40 / 78.00	54.40 / 64.00	76.80 / 83.20
19	DeepSeek-V3.2×DeepSeek-V3.2	92.80 / 95.20	86.00 / 90.00	80.40 / 85.60	90.40 / 93.60
20	DeepSeek-V3.2×GLM-5	92.00 / 96.80	84.40 / 93.60	78.40 / 90.80	90.00 / 96.40
21	DeepSeek-V3.2×GPT-6 Astra	93.20 / 98.80	86.80 / 97.60	80.80 / 95.20	90.80 / 98.00
22	GLM-5×GLM-5	94.00 / 96.00	88.40 / 91.20	83.60 / 88.00	92.80 / 95.20
23	GLM-5×GPT-6 Astra	94.40 / 98.80	88.00 / 97.60	83.60 / 95.60	92.40 / 98.00
24	GPT-6 Astra×GPT-6 Astra	97.60 / 98.40	95.20 / 96.80	91.60 / 94.40	96.40 / 97.60

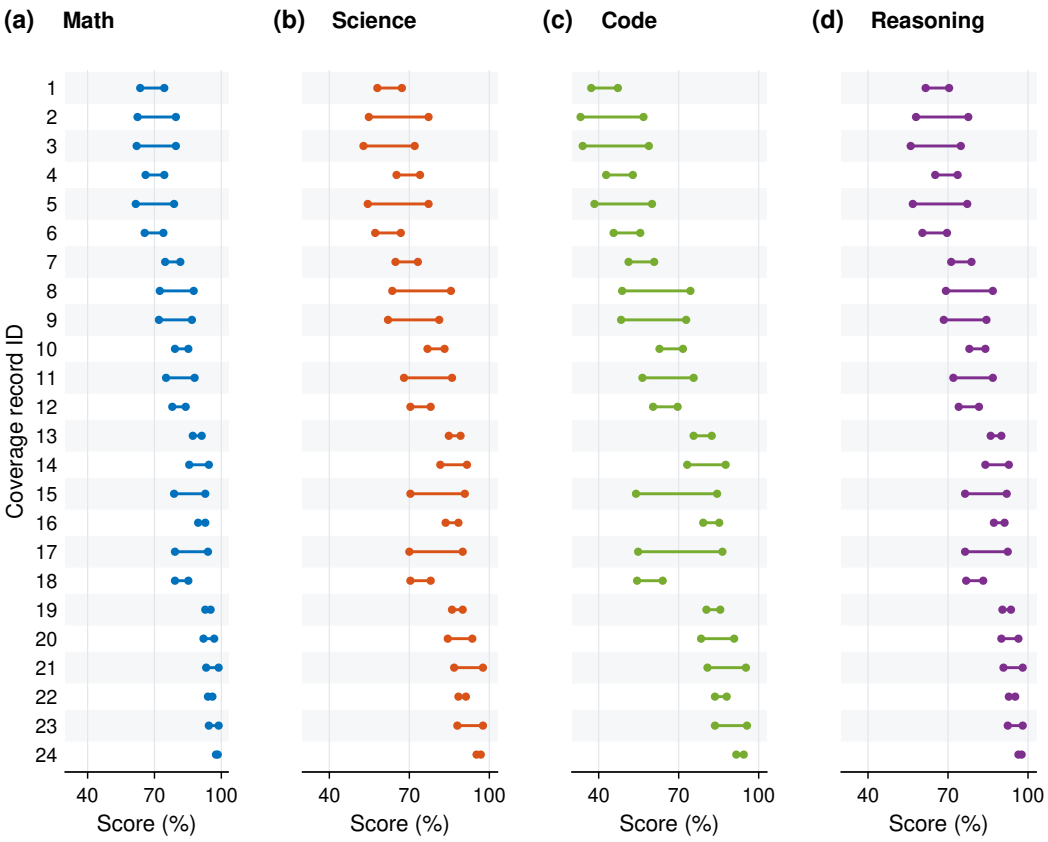


Figure 7: **Complete earlier lower/upper coverage records.** Each interval joins the two values in a row of Table 15; these are not uncertainty intervals or actual judge-selection scores. They are kept separate from standalone marginals and the expanded calibration summaries.

E Uniform Two-Teacher Mixing

Table 16 gives all four-domain results for the nine uniform mixtures. Each student sees 4,000 prompts total, with half of each domain assigned to each teacher. The primary macro comparison is with the better fixed single-teacher student, not the mean of two teachers and not a per-domain teacher oracle. Table 6 supplies those additional descriptive references.

Table 16: All nine uniform two-teacher SFT results. Qwen4B is the student. Percentages.

ID	Candidates	Math	Science	Code	Reasoning	Macro
1	Qwen4B+GemmaE4B	71.32	64.30	45.58	69.93	62.78
2	Qwen4B+Ministral3B	68.38	60.67	47.01	65.54	60.40
3	GemmaE4B+Ministral3B	69.34	63.01	43.42	68.19	60.99
4	Qwen9B+Gemma12B	73.05	69.44	50.50	71.58	66.14
5	Qwen9B+Ministral14B	76.84	65.97	51.12	72.59	66.63
6	Gemma12B+Ministral14B	74.21	67.86	57.92	75.67	68.92
7	Gemma31B+Qwen3.8-27B	78.83	70.79	56.01	76.21	70.46
8	Gemma31B+OLMo32B	77.92	72.92	55.34	79.54	71.43
9	Qwen3.8-27B+OLMo32B	80.55	70.80	58.76	76.04	71.54

For Qwen4B+GemmaE4B, the mixture scores 62.7825% before rounding, versus 62.00% for the better fixed single-teacher student. Its +0.78 pp gain is the only positive macro mix-over-best result.

Eight of nine mixtures also underperform the constituent mean. A mixture beating that mean would still not necessarily surpass the better teacher. At the domain level, 8 of 36 comparisons with the better single-teacher score are positive.

The mixture design changes both the identity and coverage of supervision: each teacher is queried on only half as many prompts as in its single-teacher run. Hence the contrast is not an isolated test of knowledge conflict or of an optimal teacher-allocation mechanism. Structured multi-agent distillation and adaptive prompt assignment answer different questions (Chen et al., 2024b; Zhang et al., 2026).

F Complete Transfer-Sensitivity Results

The following eight groups share the 1.5B student described in Section 6.2. Mistral-Nemo-12B is abbreviated as Mistral-Nemo12B in the teacher-pair table.

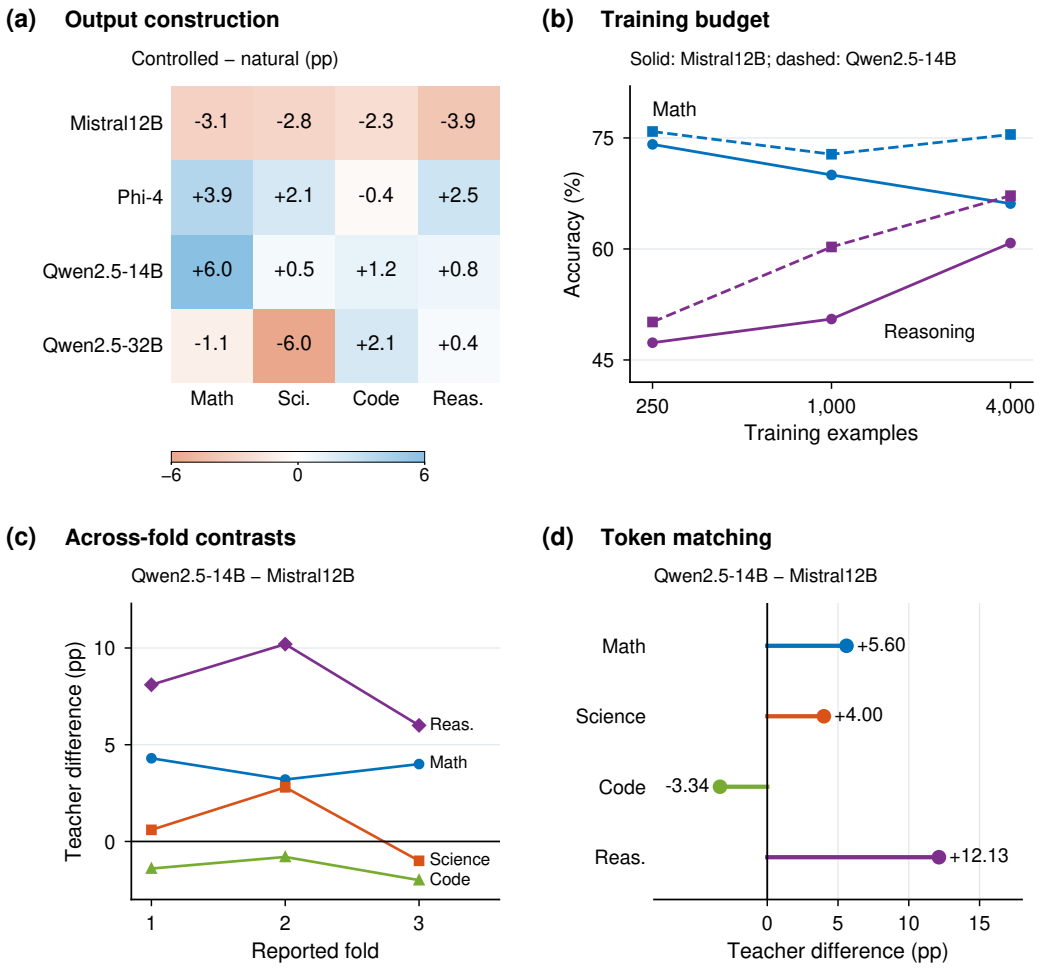


Figure 8: **Transfer depends on supervision, budget, and domain.** A separate 1.5B-student study. (a) Length/style-controlled minus natural supervision. (b) Math and Reasoning across training budgets. (c,d) Qwen2.5-14B minus Mistral-Nemo-12B supervision across folds and under 1,000-example token-matching scheme.

F.1 TEACHER IDENTITY AND NATURAL-OUTPUT SUPERVISION

Table 17: Sensitivity study 1: natural-output single-teacher supervision, 4,000 examples, learning rate 10^{-5}. Student: R1-Distill-Qwen-1.5B. Percentages.

Teacher	Math	Science	Code	Reasoning	Macro
Base	71.20	52.80	29.60	53.20	51.70
R1-Distill14B	68.27	54.93	32.27	60.53	54.00
GPT-5.5 (2026-04-24)	73.60	55.07	32.13	66.67	56.87
QwQ-32B	71.47	50.13	32.00	69.87	55.87
Mistral-Nemo-12B	69.20	54.27	32.67	64.67	55.20
Phi-4	77.47	46.53	35.87	68.40	57.07
Qwen2.5-14B	69.47	56.40	30.40	66.40	55.67
Qwen2.5-32B	76.00	55.60	34.00	71.07	59.17

F.2 LENGTH AND STYLE CONTROL

Table 18: Sensitivity study 2: natural versus length/style-controlled outputs; 4,000 examples and learning rate 10^{-5}. Percentages.

Teacher	Output	Math	Science	Code	Reasoning
Mistral-Nemo-12B	Natural	69.20	54.27	32.67	64.67
Mistral-Nemo-12B	Controlled	66.13	51.47	30.40	60.80
Phi-4	Natural	77.47	46.53	35.87	68.40
Phi-4	Controlled	81.33	48.67	35.47	70.93
Qwen2.5-14B	Natural	69.47	56.40	30.40	66.40
Qwen2.5-14B	Controlled	75.47	56.93	31.60	67.20
Qwen2.5-32B	Natural	76.00	55.60	34.00	71.07
Qwen2.5-32B	Controlled	74.93	49.60	36.13	71.47

F.3 TRAINING-EXAMPLE BUDGET

Table 19: Sensitivity study 3: training-example budget with controlled outputs and learning rate 10^{-5}. Percentages.

Teacher	Examples	Math	Science	Code	Reasoning
Mistral-Nemo-12B	250	74.13	54.67	28.80	47.33
Mistral-Nemo-12B	1,000	70.00	57.73	31.73	50.53
Mistral-Nemo-12B	4,000	66.13	51.47	30.40	60.80
Qwen2.5-14B	250	75.87	52.40	32.00	50.13
Qwen2.5-14B	1,000	72.80	54.00	30.00	60.27
Qwen2.5-14B	4,000	75.47	56.93	31.60	67.20

F.4 LEARNING RATE

Table 20: Sensitivity study 4: learning-rate comparison with 4,000 controlled-output examples. Percentages.

Teacher	Learning rate	Math	Science	Code	Reasoning
Mistral-Nemo-12B	10^{-5}	66.13	51.47	30.40	60.80
Mistral-Nemo-12B	3×10^{-6}	70.80	53.07	35.20	53.87
Qwen2.5-14B	10^{-5}	75.47	56.93	31.60	67.20
Qwen2.5-14B	3×10^{-6}	72.13	53.47	30.40	61.33

F.5 JOINTLY CORRECT FILTERING

Table 21: Sensitivity study 5: ordinary versus jointly correct supervision. Percentages.

Teacher	Scheme	Examples	Math	Science	Code	Reasoning
Mistral-Nemo-12B	Ordinary	1,000	70.00	57.73	31.73	50.53
Mistral-Nemo-12B	Jointly correct	1,029	73.20	56.67	34.27	58.00
Qwen2.5-14B	Ordinary	1,000	72.80	54.00	30.00	60.27
Qwen2.5-14B	Jointly correct	1,029	75.20	56.40	32.00	58.80

F.6 TOKEN MATCHING

The 300-example set combines the three-domain jointly correct label with token matching; the 1,000-example set is labelled four-domain exact token matching.

Table 22: Sensitivity study 6: two token-matching schemes. Percentages.

Teacher	Scheme	Examples	Math	Science	Code	Reasoning
Mistral-Nemo-12B	Joint correct + tokens	300	73.60	54.93	30.67	51.87
Qwen2.5-14B	Joint correct + tokens	300	74.27	56.00	31.33	53.33
Mistral-Nemo-12B	Four-domain tokens	1,000	69.20	55.47	31.47	50.27
Qwen2.5-14B	Four-domain tokens	1,000	74.80	59.47	28.13	62.40

F.7 DATA FOLDS

All three available folds are included for both teacher comparisons.

Table 23: Sensitivity study 7: all three folds for each of two teacher comparisons. Controlled outputs; 1,000 examples; learning rate 10^{-5}. Two runs are included for each student. Percentages.

Teacher	Fold	Math	Science	Code	Reasoning
Mistral-Nemo-12B	1	70.30	55.60	31.50	53.70
Qwen2.5-14B	1	74.60	56.20	30.10	61.80
Mistral-Nemo-12B	2	72.00	54.20	31.80	51.40
Qwen2.5-14B	2	75.20	57.00	31.00	61.60
Mistral-Nemo-12B	3	68.60	55.00	31.20	54.20
Qwen2.5-14B	3	72.60	54.00	29.20	60.20
Phi-4	1	72.60	54.00	30.40	63.40
Qwen2.5-32B	1	76.80	54.60	32.20	60.60
Phi-4	2	73.40	51.00	30.80	60.80
Qwen2.5-32B	2	76.40	54.00	34.40	61.00
Phi-4	3	71.80	55.40	31.40	56.60
Qwen2.5-32B	3	74.60	56.40	29.20	63.80

F.8 MULTI-TEACHER CONSTRUCTION AND GROUPING

Complete-case uses 3,803 examples and full-data uses 4,000.

Table 24: Sensitivity study 8: Teacher-group comparisons under complete-case and full-data constructions.

Teacher pair	Group	Construction	Math	Science	Code	Reasoning
Phi-4 + Mistral-Nemo12B	Complementary	3,803 complete	70.93	52.40	32.93	68.13
Phi-4 + Qwen2.5-14B	Similar	3,803 complete	72.80	53.33	35.33	67.47
Phi-4 + Mistral-Nemo12B	Complementary	4,000 full	72.93	52.40	35.87	68.27
Phi-4 + Qwen2.5-14B	Similar	4,000 full	71.73	51.07	30.67	68.27
Qwen2.5-32B + R1-Distill14B	Complementary	3,803 complete	70.53	53.20	35.73	68.00
Qwen2.5-32B + Q3-14B	Similar	3,803 complete	70.80	52.27	38.67	67.47
Qwen2.5-32B + R1-Distill14B	Complementary	4,000 full	70.80	54.53	35.73	68.80
Qwen2.5-32B + Q3-14B	Similar	4,000 full	70.40	53.33	37.47	69.07

G Calibration and Decision Evaluation Details

G.1 Information available at each stage

Equation 6 pools the pairs' correct judge selections divided by their total exactly-one-correct examples. This is disagreement-weighted pooling. In a conditional cross-pair forecast, the new pair's (s, D, Δ) are treated as measured inputs. Good prediction in this setting demonstrates stability of the judge calibration conditional on those inputs. The finite-probe T1 study addresses that additional source of error, scoring its chosen action on examples not in the current probe.

G.2 T1: single-candidate use versus selection

For each pair and domain, the probe estimates the candidate marginals and joint correctness, and a pair-held-out calibration yields predicted team accuracy. A positive predicted gain selects collaboration; otherwise the rule chooses the candidate preferred by the probe. The retrospective oracle is the better evaluated alternative on the scoring items. Generic-score, capability-profile, disagreement, and parameter-count rules use their respective supplied signals; their T1 thresholds are selected optimistically on the scoring data. These favorable signal baselines are not independently trained routers.

At 125 probe items per domain, Table 25 gives all reported T1 regrets and decision-accuracy values. The budget sweep uses 10, 25, 50, 100, and 125 items.

G.3 T2: supervision choice within a candidate pair

For each of the nine pairs, the evaluated alternatives are the two single-teacher students and their uniform mixture. The oracle comparison is per domain. The probe rule chooses a teacher from its pre-SFT task score rather than using its observed post-training ratio τ. Always-mix is a separate baseline. The reported regrets are 1.11 pp for the probe rule, 2.16 pp for generic-score and profile rules, 0.84 pp for parameter count, and 2.01 pp for always-mix.

The T2 population differs from T1's expanded collaboration population. Its per-domain oracle also differs from the fixed-teacher reference of Figure 1. These distinctions explain why a macro mixing contrast and a teacher-selection regret need not be equal. The T2 evaluation is intentionally pair-local and therefore measures supervision choice within the available candidate pair rather than full-pool teacher retrieval.

The held-out forecast pairs are Qwen3.5-4B with GPT-6 Astra and Gemma-E4B with GPT-6 Astra. The finite-probe results in Table 3 use the held-out-item evaluation protocol.

Table 25: **Selection regret at 125 probe items per domain.** T1 chooses single-candidate use or collaboration on held-out items. T2 chooses a teacher from a restricted pair, with the two single-teacher students and their mixture as the oracle alternatives. Lower regret is better; bold marks the best non-oracle regret in each task. T1 signal thresholds are chosen optimistically.

	T1: inference selection		T2: teacher choice
Selection rule	**Regret (pp)**	**Agreement (%)**	**Regret (pp)**
Task-local probe	**0.44**	73.3	1.11
Generic-score proxy	0.86	66.6	2.16
Capability profile	0.86	69.9	2.16
Disagreement rate	1.62	55.2	–
Parameter count	1.66	56.1	**0.84**
Always collaborate	1.57	58.3	–
Always use one candidate	1.47	35.5	–
Always mix two teachers	–	–	2.01
Per-domain oracle	0.00	100.0	0.00

H Capability Proxies and Deployment Accounting

H.1 Generic Capability Proxies

Table 26: Generic capability-proxy scores used in our analysis.

Model	**MMLU-labelled proxy**	**GPQA-labelled proxy**
Qwen3.5-4B	85.80	76.20
Qwen3.5-9B	88.10	81.70
Qwen3.5-27B	90.20	85.50
Gemma-E4B	82.25	58.60
Gemma-12B	85.99	78.80
Gemma-31B	89.82	84.30
Ministral-3B	72.00	30.00
Ministral-8B	76.10	42.94
Ministral-14B	79.40	48.04
Qwen3.8-27B	91.70	89.20
Llama-Scout	84.60	57.20
gpt-oss-20b	85.30	71.50
OLMo-32B	80.90	48.60
Nemotron-30B	88.26	75.44
Phi-4-reasoning+	85.42	68.90
DeepSeek-V3.2	89.73	82.40
GLM-5	90.21	86.00
GPT-6 Astra	95.50	96.00

Across all 18 teachers, the Spearman correlations between the two capability proxies and mean SFT macro performance across students are 0.521 and 0.445, respectively. Excluding GPT-6 Astra, the corresponding correlations are 0.431 and 0.341. We retain the strongest teacher in all primary comparisons to avoid artificially weakening the teacher-selection baseline.

Supplementary profile diagnostics. We further decompose each profile entry as

$$a_{md} = \mu + \ell_m + \delta_d + r_{md},$$

where ℓ_m captures model-level capability, δ_d captures shared task difficulty, and r_{md} represents the residual capability shape. The residual component has a standard deviation of 4.0 points and a mean interval half-width of 9.5 points, indicating that fine-grained residual structure is substantially harder to resolve than overall capability level. Profile distance has an association of 0.33 with measured

complementary coverage, while residual-shape similarity has an association of 0.51 with observed collaboration gain. These results support a narrower identifiability conclusion: marginal task scores, even when measured exactly, are insufficient to determine paired correctness, which requires joint observations.

Additional transfer-matrix summaries. After centering the 18×12 teacher-by-(student, domain) gain matrix within each column, 94.3% of its squared singular-value mass lies in the leading direction, indicating a strong shared component in teacher effectiveness across students and domains. Among the nine models evaluated in both teacher and collaboration-partner roles, teacher value and partner value have a rank correlation of -0.88. Because the two roles use different references and conversion mechanisms, we treat this correlation as descriptive evidence of role dependence rather than as a standalone universal relation. The complete transfer matrix and the explicit OLMo32B/Qwen3.8 reversal in the main text provide the primary cross-role evidence.

H.2 Accuracy, Cost, and the Deployment Horizon

Let A_{SFT} and A_{col} denote accuracies on a common probability scale, $T > 0$ the one-time cost of supervision generation and training, and c_S and c_{col} the per-query costs of the deployed SFT model and collaboration system, respectively. For a deployment horizon of $N > 0$ queries and a cost penalty $\lambda > 0$, define

$$U_{\mathrm{SFT}} = A_{\mathrm{SFT}} - \lambda \left(c_S + \frac{T}{N} \right), \qquad U_{\mathrm{col}} = A_{\mathrm{col}} - \lambda c_{\mathrm{col}}.$$

Let

$$K = A_{\mathrm{SFT}} - A_{\mathrm{col}} + \lambda(c_{\mathrm{col}} - c_S).$$

Distillation achieves at least the same utility as collaboration when

$$K \geq \frac{\lambda T}{N}.$$

For $K > 0$, this condition is equivalent to

$$N \geq \frac{\lambda T}{K}.$$

When $K \leq 0$, collaboration remains preferable for every finite positive deployment horizon under this utility. In the special case of equal accuracy with $c_{\mathrm{col}} > c_S$, the break-even condition reduces to

$$N \geq \frac{T}{c_{\mathrm{col}} - c_S}.$$

Thus, the relevant denominator is the additional per-query cost of collaboration relative to standalone deployment. When the target model also serves as the collaboration judge, its inference cost remains part of c_{col}, while the deployed student's inference cost is represented by c_S. Deployment constraints can be incorporated directly by restricting the feasible action set before applying the utility comparison.