# WHOM SHOULD A STUDENT TRUST? UNSUPERVISED RELIABILITY ESTIMATION FOR MULTI-TEACHER DISTILLATION

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

## ABSTRACT

Multi-teacher distillation provides diverse supervision for training language models. However, teacher-generated responses vary in quality and suitability for the student, making it difficult to determine which teachers to trust and which responses to learn from. Human annotations and external LLM judges can guide selection, but add cost and may introduce evaluator-specific biases, motivating unsupervised approaches. Existing student-based methods select responses using token probabilities to measure familiarity or loss gradients to characterize their effects on optimization. However, familiarity does not directly indicate learning benefit, and computing parameter gradients requires additional backward passes. Therefore, we propose Student-informed Unsupervised Reliability Estimation (SURE), which jointly estimates teacher reliability and response-selection probabilities without external quality judgments. In particular, SURE measures coherence among student gradients from different teachers' responses to the same prompt without backpropagating through the frozen student. It then models this evidence over a teacher-prompt bipartite graph to iteratively refine teacher reliability and response-selection probabilities, integrating within-prompt coherence with cross-prompt teacher reliability. Experiments across pretrained students from different model families show consistent improvements over other unsupervised multi-teacher distillation baselines. Further analyses show higher judged quality of the selected responses and support the value of estimating teacher reliability across prompts for curating effective multi-teacher supervision.

## 1 INTRODUCTION

As large language models (LLMs) grow in scale and capability, their generated responses have become a useful source of supervision for training other models (Xu et al., 2024; Honovich et al., 2023; Hsieh et al., 2023; Wang et al., 2023), reducing the need for human annotators to write training responses. A common approach to use this supervision is distillation, in which a student model learns from teacher-generated responses to improve its reasoning and instruction-following abilities (Gu et al., 2024; Chen et al., 2024b; Jiang et al., 2023). In multi-teacher distillation, multiple teachers generate candidate responses to each prompt, creating a pool with varied reasoning strategies and writing styles (Jin et al., 2026). However, diversity alone does not guarantee reliable supervision. Teacher-generated responses vary in factual accuracy, instruction adherence, and suitability for the student (Li et al., 2024a), making response selection essential to effective distillation. Selecting among these responses raises a broader question: *Which teachers consistently provide useful supervision across prompts?* Estimating reliability from a teacher's responses across prompts would allow us to use that teacher-level signal alongside prompt-specific evidence when selecting individual responses.

External evaluation can guide response selection through human annotations, task-specific verifiers, or LLM judges (Zhang et al., 2025; Lee et al., 2024a). Human annotations provide direct quality assessments but are costly to collect at scale. Task-specific verifiers check responses against explicit criteria, such as reference answers or executable tests, making them useful when correctness can be established automatically (Wang et al., 2024b; Chen et al., 2023). Open-ended instructions, however, often lack such checks and require broader assessments of relevance and instruction adherence (Ye



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., 2024; Kim et al., 2024). LLM judges address this need by evaluating responses in context, but scoring large response pools adds inference expense and may introduce evaluator-specific biases (Wang et al., 2024a; Chen et al., 2024a). These costs and dependencies motivate *unsupervised multi-teacher distillation*: estimating teacher reliability and selecting responses using the response pool and the student itself, without external quality labels or judge scores (Agarwal et al., 2025; Lee et al., 2024b).

To select responses without external quality judgments, existing student-based methods obtain signals from the student's prediction probabilities or gradients. Probability-based methods score responses by their likelihood under the student, favoring familiar supervision with only a forward pass (Ren et al., 2025; Li et al., 2025). However, likelihood indicates how familiar a response is to the student, not necessarily how much the student would benefit from learning it. Gradient-based methods examine how teacher-generated supervision affects the student's optimization, but computing parameter gradients requires additional backward passes and increases selection costs (Kwon et al., 2024; Pan et al., 2024; Zhao et al., 2025). Moreover, a response's gradient alone does not establish whether it provides reliable supervision. Assessing responses individually also overlooks information about the teachers that produced them (Wadhwa et al., 2025; Yang et al., 2025). A teacher's responses across multiple prompts can reveal patterns of reliability that are difficult to identify from a single response. Using these patterns along with prompt-specific assessments can reduce dependence on noisy individual scores (Han et al., 2025; Zhang et al., 2024). This motivates an efficient method that estimates teacher reliability across prompts and response-selection probabilities within each prompt.

To that end, we propose SURE, Student-informed Unsupervised Reliability Estimation, a framework that uses student-derived gradient coherence among candidate responses to estimate teacher reliability for unsupervised multi-teacher distillation. SURE computes response gradients with a frozen student's final hidden states without backpropagating. From these gradients, we extract cross-teacher supervision signals to obtain cleaner within-prompt coherence, and construct a gradient coherence matrix for each prompt to quantify how strongly each candidate's learning signal is supported by other responses to the same prompt. SURE then iteratively estimates teacher reliability through a bipartite graph linking teachers to prompts using per-prompt coherence matrices, allowing teacher reliability estimated across prompts along with the prompt-specific gradient coherence to inform response-selection probabilities. Therefore, the final distillation set is selected based on student-derived learning signals for individual responses and teacher reliability inferred from the broader response pool. We summarize our contributions as follows:

- We design an unsupervised gradient-coherence measure based on extracted cross-teacher supervision signals to quantify how strongly each candidate's learning signal is supported by other responses within the same prompt, guiding response selection based on each candidate's potential contribution to student learning.
- We introduce a teacher-prompt bipartite graph that aggregates per-prompt gradient coherence to estimate teacher reliability across prompts, and jointly refines reliability and response-selection probabilities using both cross-prompt teacher evidence and within-prompt supervision support.
- We conduct extensive experiments across student models and downstream benchmarks, showing that SURE improves distillation performance on different tasks, while further analyses confirm higher-quality selected responses and the contribution of estimated teacher reliability for unsupervised multi-teacher distillation.

## 2 Related Work

**Multi-teacher distillation.** Multi-teacher distillation trains a student on responses generated by multiple frozen teacher models, allowing the student to learn diverse reasoning strategies and response styles (Cui et al., 2024; Panigrahi et al., 2026). As candidate responses can vary in quality, prior work commonly curates teacher-generated supervision through external evaluation (Chen et al., 2024c; Liu et al., 2024b; Xia et al., 2024; Wu et al., 2023; Yuan et al., 2023). For tasks with verifiable outcomes, rule-based verifiers can identify responses that satisfy explicit correctness criteria (Xin et al., 2025; Lightman et al., 2024; Ni et al., 2023; Le et al., 2022). For open-ended

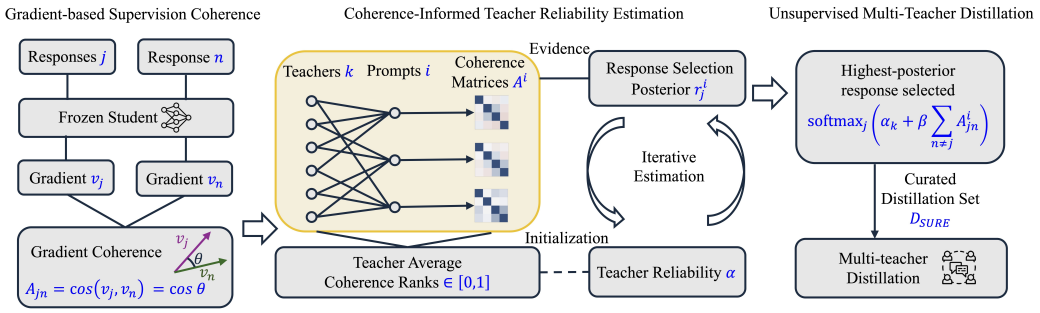



Figure 1: SURE utilizes within-prompt student-gradient coherence with cross-prompt teacher reliability estimated over a teacher-prompt bipartite graph to refine response-selection probabilities and curate the final unsupervised distillation set.

instructions, model-based evaluators provide broader assessments of response quality. For example, UltraFeedback uses GPT-4 to evaluate responses on dimensions such as helpfulness (Cui et al., 2024). While effective, these approaches rely on external quality assessments that may introduce evaluator-specific biases and do not directly consider how suitable the supervision is for the target student (Raina et al., 2024; Zheng et al., 2023). This motivates unsupervised student-informed curation methods that select among multi-teacher responses without external quality labels or judge scores for multi-teacher distillation (Zhou & Ai, 2024; Liu et al., 2024a; Li et al., 2024b).

**Unsupervised multi-teacher distillation set curation.** Unsupervised student-informed methods use the student's prediction probabilities or gradients to identify suitable teacher-generated response (Wang et al., 2025; Xu et al., 2025; Wei et al., 2025; Tang et al., 2025; Li et al., 2024a). Probability-based methods require only forward passes: GRAPE selects responses with the highest student likelihood (Zhang et al., 2026), while RSR balances student alignment and informativeness using token likelihoods and ranks (Yang et al., 2026). However, these signals mainly characterize the student's current predictions. Gradient-based methods provide a more direct learning-related signal: G-Vendi measures diversity across the response pool using model-induced gradients (Jung et al., 2026), while GRACE measures teacher-student compatibility from student parameter gradients on teacher responses (Panigrahi et al., 2026). These methods require additional backward passes, and derive their curation signals from only response-level gradient evidence without modeling whether their source teachers consistently provide reliable supervision across prompts (Sun et al., 2025; Dai et al., 2025; San Joaquin et al., 2024). In contrast, SURE obtains gradient-informed learning signals without backpropagating through the student and aggregates prompt-specific coherence evidence across the response pool to estimate teacher reliability, considering local response support with cross-prompt teacher information for unsupervised multi-teacher response selection.

## 3 Methodology

In unsupervised curation for multi-teacher distillation, candidate responses vary in quality and suitability for the student, making it unclear which teachers to trust and which responses to learn from. Given a prompt $q$, let $\{y_j\}_{j=1}^K$ denote the candidate responses from $K$ teachers, where $T_j$ identifies the teacher that generated $y_j$. Our goal is to select, without external quality labels, the response that best supports student learning while estimating which teachers provide reliable supervision across prompts. To address this task, as shown in Figure 1, SURE consists of two components that use student-derived signals to compare candidate responses and estimate teacher reliability:

- **Gradient-Based Supervision Coherence Modeling:** SURE extracts cross-teacher supervision signals to obtain cleaner within-prompt coherence to quantify how strongly each candidate's learning signal is supported by other responses.
- **Coherence-Informed Teacher Reliability Estimation:** SURE aggregates per-prompt coherence through a teacher-prompt bipartite graph to iteratively estimate teacher reliability and response-selection probabilities, combining cross-prompt teacher reliability with prompt-specific coherence to determine which responses to select.

### 3.1 GRADIENT-BASED SUPERVISION COHERENCE MODELING

To compare how candidate responses could support student learning during distillation, we measure gradient coherence among responses within each prompt using student-derived gradients. Before computing coherence, we extract cross-teacher supervision signals from the normalized response gradients to obtain cleaner within-prompt coherence estimates. This allows within-prompt coherence to more directly reflect how strongly each candidate's learning signal is supported by other responses to the same prompt.

**Student-derived gradients for characterizing learning signals.** For each candidate response $y_j$, we use the frozen student to predict each token given the prompt and the preceding response tokens. Let $h_{jt}$ denote the final hidden state passed to the output layer to predict token $y_{jt}$. The output layer maps $h_{jt}$ to vocabulary logits, which are converted by softmax into the next-token distribution $p_{jt}$. Taking the teacher-provided token as the target, we compute the cross-entropy loss $\ell_{jt}$ and its gradient with respect to $h_{jt}$:

$$g_{jt} = \nabla_{h_{jt}} \ell_{jt} = W^\top \big(p_{jt} - e_{y_{jt}}\big), \tag{1}$$

where $W$ is the student's output-layer weight matrix and $e_{y_{jt}}$ is the one-hot vector for the teacher-provided token. The hidden-state gradient captures the local learning signal induced by the teacher-provided token under the student's current prediction. It can be computed directly from the student's predicted probabilities and output weights without backpropagating through the student. To summarize this gradient evidence for each response, we average its token-level gradients to obtain $\bar{g}_j$ and normalize it as $u_j = \bar{g}_j / \|\bar{g}_j\|_2$. This produces a unit-length representation of the response-level learning signal, allowing subsequent coherence comparisons to reflect alignment between learning signals rather than differences in gradient magnitude.

**Extracting cross-teacher supervision signals.** Although $u_j$ captures the learning signal induced by an individual prompt-response pair, part of this signal may be systematically shared across responses generated by the same teacher. To obtain cleaner supervision signals for cross-teacher comparison, for each teacher $k$, we estimate $m_k$ as the global average of the student-derived normalized gradients induced by its responses across prompts, capturing the gradient component consistently of that teacher. We then subtract this teacher-average component from each corresponding response gradient and normalize the resulting vector:

$$m_k = \widehat{\mathbb{E}}\big[u_j \mid T_j = k\big], \qquad v_j = \frac{u_j - m_{T_j}}{\|u_j - m_{T_j}\|_2}, \tag{2}$$

The resulting cross-teacher supervision signal $v_j$ captures the response-specific learning signal after subtracting the gradient pattern shared across responses from the same teacher, providing a cleaner basis for comparing supervision across teachers within a prompt.

A response's gradient alone does not establish whether it provides reliable supervision. Therefore, we compare the extracted cross-teacher supervision signals of candidate responses to the same prompt and use their coherence as evidence of support from other responses. For responses $y_j$ and $y_n$, we measure pairwise gradient coherence using cosine similarity, defined as $A_{jn} = \cos(v_j, v_n)$ for $j \neq n$. Higher similarity indicates stronger coherence between the cross-teacher supervision signals of the two responses under the student. Together, these pairwise comparisons form a per-prompt coherence matrix that provides prompt-specific evidence for the subsequent estimation of teacher reliability and response-selection probabilities.

### 3.2 COHERENCE-INFORMED TEACHER RELIABILITY ESTIMATION

Within-prompt gradient coherence provides prompt-specific evidence for comparing candidate responses. However, coherence within a single prompt remains local evidence: it does not reveal whether a teacher consistently provides reliable supervision across prompts, and may also be affected by differences among valid responses. To utilize evidence beyond individual prompts, we aggregate coherence information across prompts to estimate each teacher's reliability. SURE refines these teacher reliability estimates and response-selection probabilities through a bipartite graph linking teachers and prompts. This allows response selection to consider both teacher reliability inferred across prompts with the gradient coherence of candidate responses within each prompt.

**Coherence-informed bipartite modeling of teacher reliability.** We assign each teacher $k$ a learnable reliability score $\alpha_k$ shared across prompts. Let $z \in \{1, \ldots, K\}$ denote the index of the response to select. Since the preferred response is not labeled, we treat $z$ as a latent variable and infer it using both teacher reliability across prompts and gradient coherence among candidate responses within each prompt. Teacher reliability captures how consistently a teacher provides supported supervision across the response pool, while gradient coherence provides prompt-specific support for individual responses. We define the prior probability of selecting response $y_j$ using a softmax over the reliability scores of the corresponding teachers:

$$\pi_j = P(z = j \mid T_1, \ldots, T_K) = \frac{\exp(\alpha_{T_j})}{\sum_{n=1}^{K} \exp(\alpha_{T_n})}. \tag{3}$$

We then model how within-prompt gradient coherence supports response selection. Let $A$ denote the per-prompt matrix of pairwise gradient coherence scores. If $y_j$ is the selected response, we model coherence scores involving $y_j$ using a Gaussian distribution with mean $\mu_s$, while scores among the remaining responses are modeled with mean $\mu_b$. We use a shared variance $\sigma^2$ and constrain $\mu_s \geq \mu_b$, so that stronger coherence between $y_j$ and other candidate responses provides greater support for selecting $y_j$. The likelihood of the observed coherence scores conditioned on selecting $y_j$ is:

$$p(A \mid z = j) = \prod_{n \neq j} \mathcal{N}(A_{jn}; \mu_s, \sigma^2) \prod_{\substack{a<b \\ a \neq j,\ b \neq j}} \mathcal{N}(A_{ab}; \mu_b, \sigma^2). \tag{4}$$

We use this Gaussian likelihood as an approximation for estimating selection probabilities, without requiring the observed cosine similarities to follow an exact Gaussian distribution.

To aggregate prompt-specific coherence evidence across the response pool, as shown in Figure 1, we represent teachers and prompts as a bipartite graph. Each teacher node is connected to the prompts for which it provides a candidate response, and each prompt is associated with its per-prompt gradient coherence matrix. This structure links a teacher's responses across prompts and allows prompt-specific coherence evidence to inform how reliably that teacher provides supported supervision.

**Iterative teacher reliability estimation.** We iteratively refine teacher reliability and response-selection probabilities over the bipartite graph. The two estimates inform each other throughout this process: *gradient coherence provides response-specific evidence of how strongly each candidate is supported by the other responses, while teacher reliability provides cross-prompt evidence about how consistently the teacher provides reliable supervision.* These two sources of evidence jointly determine response-selection probabilities, which are then accumulated across prompts to update the reliability of the corresponding teachers.

We initialize teacher reliability by ranking teachers within each prompt according to the average gradient coherence of their responses with the other candidates. We scale the ranks to $[0, 1]$, assigning higher ranks to responses with greater coherence, then average each teacher's ranks across prompts and standardize these averages to obtain the initial reliability scores. To estimate teacher reliability across prompts, let $A^i$ denote the gradient coherence matrix for prompt $i$, $T_j^i$ the teacher of response $y_j^i$, and $\pi_j^i$ its selection prior from equation 3. Keeping $A^i$ fixed, we estimate the teacher reliability scores $\alpha$, the Gaussian means $\mu_s, \mu_b$, and the shared variance $\sigma^2$ by maximizing:

$$J(\alpha, \mu_s, \mu_b, \sigma^2) = \sum_i \log \left[ \sum_{j=1}^{K} \pi_j^i \, p(A^i \mid z^i = j) \right] - \frac{\lambda}{2} \|\alpha\|_2^2, \tag{5}$$

where $\alpha$ collects the reliability scores of all teachers. For each prompt, the objective considers every candidate as a possible selected response and weights its coherence likelihood by the corresponding selection prior. Maximizing this objective estimates teacher reliability and the Gaussian parameters from coherence evidence across prompts, while the regularization term with coefficient $\lambda$ limits overly large reliability scores. At each iteration, we update the posterior probability $r_j^i$ of selecting response $y_j^i$ by combining its teacher-reliability prior with its within-prompt coherence evidence. Under the shared-variance likelihood, this probability is expressed as:

$$r_j^i = \frac{\pi_j^i \, p(A^i \mid z^i = j)}{\sum_{n=1}^{K} \pi_n^i \, p(A^i \mid z^i = n)} = \mathrm{softmax}_j \left( \alpha_{T_j^i} + \beta \sum_{n \neq j} A_{jn}^i \right), \qquad \beta = \frac{\mu_s - \mu_b}{\sigma^2} \geq 0, \tag{6}$$

where $\beta$ is determined by the estimated Gaussian means and variance. Thus, each response is scored using both the reliability of its teacher estimated across prompts and its gradient coherence with the other candidate responses within the current prompt.

We then use these posterior response-selection probabilities across prompts to update teacher reliability. When optimizing equation 5, we keep $r_j^i$ fixed and update $\alpha$ so that teachers whose responses receive greater posterior support across prompts are assigned higher reliability scores:

$$\alpha^{\text{new}} = \arg\max_{\alpha} \left[ \sum_i \sum_{j=1}^{K} r_j^i \log \pi_j^i - \frac{\lambda}{2} \|\alpha\|_2^2 \right]. \tag{7}$$

This update favors teachers whose responses consistently receive stronger posterior selection support across prompts, allowing teacher reliability to be estimated from evidence accumulated across the response pool rather than from any single prompt. Meanwhile, we update the Gaussian parameters using the same posterior probabilities. To estimate $\mu_s$, we weight each response pair by the probability that one of its responses is selected; to estimate $\mu_b$, we use the probability that neither response is selected. We enforce $\mu_s \geq \mu_b$ and estimate the shared variance from the corresponding weighted squared differences around these means. These updates repeat until the regularized objective converges or the iteration limit is reached.

After estimation, we use the final teacher reliability scores and Gaussian parameters to compute the posterior selection probabilities in equation 6, and retain the highest-probability response for each prompt to construct the student's fine-tuning set $D_{\text{SURE}}$. Therefore, the final distillation set curation considers teacher reliability estimated across prompts with gradient coherence among candidate responses within each prompt, without requiring external quality labels or judge scores.

# 4 Experiments

## 4.1 Experimental Settings

**Training settings.** To evaluate SURE for unsupervised multi-teacher distillation, we use UltraFeedback (Cui et al., 2024) as the candidate response pool. After deduplication and removal of incomplete response groups, the resulting pool contains 63,748 prompts, each with four candidate responses generated by a pool of 17 teacher models. For each selection method, we retain one response per prompt and full fine-tune Qwen3-4B-Base, Qwen3-8B-Base, Llama-3.1-8B, and Llama-3.2-3B students on the resulting curated set. We train for one epoch with a learning rate of $10^{-5}$ and a batch size of 256. Training settings are kept identical across selection methods for each student to isolate the effect of unsupervised data curation on multi-teacher distillation. See Appendix A for detailed training settings.

**Evaluation benchmarks.** We evaluate the fine-tuned students across three categories: (i) *instruction following* with AlpacaEval 2 (Dubois et al., 2023; 2024) and IFEval (Zhou et al., 2023). AlpacaEval 2 evaluates response quality against reference responses and reports both raw and length-controlled (LC) win rates to account for response-length effects. IFEval evaluates how well models follow verifiable instructions and reports strict and loose prompt-level accuracy; (ii) *multi-turn response quality* with MT-Bench (Bai et al., 2024), which evaluates responses over two-turn conversations and reports the average score across both turns; and (iii) *knowledge-intensive reasoning* with MMLU (Hendrycks et al., 2020) and MMLU-Pro (Wang et al., 2024c), which evaluate performance across diverse domains using multiple-choice accuracy. See Appendix A for evaluation settings.

**Baselines.** We compare SURE against unsupervised response-selection baselines applied to the same UltraFeedback candidate pool: (i) *Uniform Sampling* (Xia et al., 2025) uniformly selects one response per prompt using a fixed seed; (ii) *Max Token Length* (Zhao et al., 2024) selects the response with the greatest number of tokens under the student's tokenizer; (iii) *GRAPE* (Zhang et al., 2026) selects the response with the lowest student perplexity, favoring responses that better match the student's distribution; and (iv) *RSR* (Yang et al., 2026) selects the response with the lowest Rank-Surprisal Ratio, which combines token ranks and surprisal under the student to balance alignment with informativeness. All methods curate for the same prompts to ensure equal training-set sizes and isolate the effect of unsupervised response selection on distillation.

Table 1: Downstream student performance after full fine-tuning. The best result for each student is bolded. Abs. Δ denotes SURE's improvement over each baseline in the seven-metric average, with MT-Bench scaled to 0-100 before averaging.

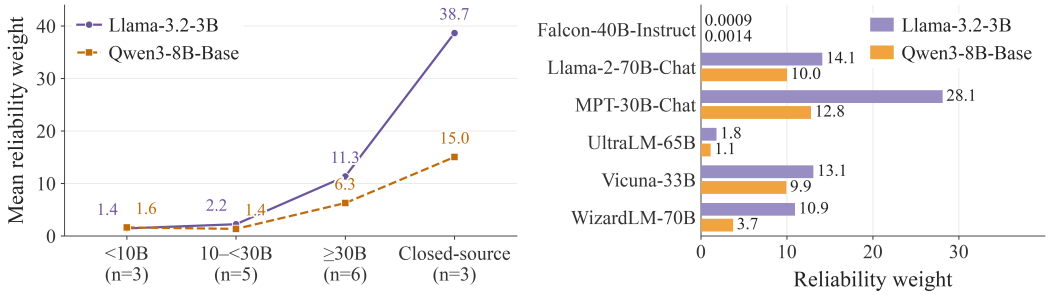
| Method | AlpacaEval 2 | | IFEval | | MT-Bench | MMLU | MMLU-Pro | Abs. Δ |
|---|---|---|---|---|---|---|---|---|
| | Raw | LC | Strict | Loose | | | | |
| *Qwen3-4B-Base* | | | | | | | | |
| Uniform Sampling | 4.59 | 5.59 | 30.31 | 32.90 | 5.519 | 71.41 | 36.99 | 6.30 |
| Max Token Length | 7.81 | 8.40 | 29.39 | 34.75 | 6.069 | 72.23 | 38.94 | 4.12 |
| GRAPE | 5.76 | 6.98 | 32.16 | 35.86 | 5.594 | 72.28 | 39.17 | 4.70 |
| RSR | 4.86 | 6.10 | 30.31 | 33.83 | 5.694 | 71.56 | 37.77 | 5.67 |
| SURE (Ours) | **8.32** | **10.52** | **40.11** | **45.47** | **6.125** | **72.95** | **42.45** | |
| *Qwen3-8B-Base* | | | | | | | | |
| Uniform Sampling | 5.62 | 7.44 | 37.52 | 41.22 | 5.500 | 75.34 | 41.12 | 6.65 |
| Max Token Length | 11.78 | 12.74 | 35.67 | 39.56 | 6.006 | 74.57 | 38.74 | 5.24 |
| GRAPE | 6.44 | 8.77 | 37.71 | 42.88 | 5.869 | 74.96 | 41.44 | 5.56 |
| RSR | 6.35 | 8.36 | 33.27 | 37.71 | 5.719 | 74.82 | 40.25 | 7.41 |
| SURE (Ours) | **11.99** | **16.21** | **45.47** | **49.91** | **6.744** | **75.71** | **43.09** | |
| *Llama-3.2-3B* | | | | | | | | |
| Uniform Sampling | 1.37 | 1.18 | 18.11 | 19.59 | 3.850 | 54.79 | 24.72 | 3.10 |
| Max Token Length | 1.96 | 1.59 | 16.27 | 20.70 | 4.188 | 55.50 | 24.91 | 2.45 |
| GRAPE | 1.59 | 1.44 | 19.22 | 22.00 | 4.037 | 54.74 | 24.53 | 2.29 |
| RSR | 1.46 | 1.34 | 15.53 | 19.22 | 4.169 | 54.17 | **26.03** | 2.93 |
| SURE (Ours) | **3.13** | **3.09** | **22.55** | **26.80** | **4.300** | **55.77** | 25.59 | |
| *Llama-3.1-8B* | | | | | | | | |
| Uniform Sampling | 2.29 | 2.48 | 19.59 | 23.29 | 4.319 | 59.45 | 29.09 | 4.86 |
| Max Token Length | 3.44 | 3.12 | 21.26 | 24.95 | **4.900** | 62.78 | 31.18 | 2.52 |
| GRAPE | 3.16 | 3.27 | 24.03 | 28.84 | 4.581 | 61.88 | 30.14 | 2.32 |
| RSR | 2.77 | 2.96 | 18.48 | 21.81 | 4.419 | 61.45 | 33.16 | 4.08 |
| SURE (Ours) | **5.18** | **5.24** | **26.80** | **30.50** | 4.825 | **63.35** | **34.06** | |

### 4.2 Main Results

**Superior distillation performance across downstream benchmarks.** As shown in Table 1, SURE achieves the best performance on most student-benchmark settings, with particularly strong gains on instruction-following tasks, which is the primary focus of UltraFeedback distillation. On AlpacaEval 2 LC win rate and IFEval Loose prompt accuracy, SURE reaches 10.52% and 45.47% on Qwen3-4B, improving over the strongest baseline by 2.12% and 9.61%. On Qwen3-8B, SURE reaches 16.21% and 49.91%, with gains of 3.47% and 7.03% over the strongest baseline. Similar improvements are observed for both Llama students. These results show that considering prompt-specific gradient coherence with teacher reliability across prompts improves unsupervised distillation without external judge labels, and produces supervision that better suits the student.

**Robust effectiveness across model families and scales.** SURE improves the seven-metric average over every baseline for all student models, with absolute gains of 4.12%-6.30% on Qwen3-4B, 5.24%-7.41% on Qwen3-8B, 2.29%-3.10% on Llama-3.2-3B, and 2.32%-4.86% on Llama-3.1-8B. These results show that SURE remains effective across both Qwen and Llama families and across student sizes from 3B to 8B, indicating that student-derived gradient coherence and cross-prompt teacher reliability estimation are able to generalize across different model architectures and scales. Moreover, SURE shows larger gains for the 8B student within both model families, which shows that the benefit of SURE is maintained and can become more significant as student capacity increases.

**Cross-prompt teacher reliability beyond prompt-specific signals.** Compared with baselines, SURE shows clear improvements over methods that rely only on prompt-specific student-derived signals. On Qwen3-4B, SURE achieves 42.45% on MMLU-Pro, compared with 39.17% for GRAPE and 37.77% for RSR; on Llama-3.1-8B, SURE reaches 63.35% on MMLU, compared with 61.88% for GRAPE and 61.45% for RSR. GRAPE and RSR derive their signals from token-level statistics of individual responses. In contrast, SURE aggregates per-prompt gradient coherence across prompts to estimate teacher reliability. These improvements suggest that prompt-specific response evidence




(a) Mean reliability by teacher group.

(b) Reliability of the $\geq$ 30B teachers.

Figure 2: Estimated teacher reliability weights for Llama-3.2-3B and Qwen3-8B-Base students.

alone can overlook useful information about how reliably each teacher provides supervision across the response pool. Therefore, cross-prompt teacher reliability provides more comprehensive evidence for selecting effective supervision in unsupervised multi-teacher distillation.

### 4.3 Analysis on Estimated Teacher Reliability

To better examine how estimated teacher reliability varies with teacher scale, we group the teachers by parameter size. The grouped results in Figure 2a show that SURE assigns higher reliability on average to stronger teacher groups, but the reliability differences cannot be explained by teacher scale alone. For Llama-3.2-3B, the mean reliability weight increases from 1.4 for teachers below 10B parameters, to 11.3 for teachers $\geq$ 30B and 38.7 for closed-source teachers; Qwen3-8B also shows a similar overall increase. However, substantial variation appears even among similarly large teachers. Within the $\geq$ 30B group in Figure 2b, MPT-30B-Chat receives the highest reliability for both students with weights of 28.1 and 12.8, while larger teacher models such as Llama-2-70B-Chat receive lower weights and Falcon-40B-Instruct receives almost no weight. This shows that SURE does not simply favor larger teachers, but instead estimates which teachers consistently provide supervision that is well supported across prompts. The reliability pattern also depends on the student: for example, the gap between MPT-30B-Chat and Llama-2-70B-Chat is much larger for Llama-3.2-3B student than for Qwen3-8B student. This supports the motivation of SURE that useful supervision depends not only on teacher quality, but also on its suitability for the student.

### 4.4 Pairwise Analysis of Teacher Preferences in Response Selection

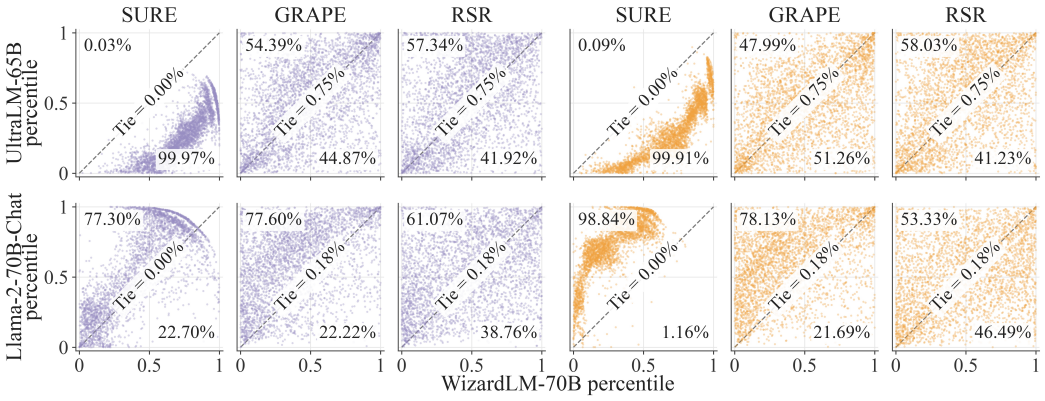



Figure 3: Pairwise teacher preferences for Llama-3.2-3B (left) and Qwen3-8B-Base (right). Response-selection scores are converted to percentile ranks for comparison. Points above and below the diagonal indicate prompts favoring the y- and x-axis teachers, and diagonal points indicate ties.

To compare pairwise teacher preferences across methods, we convert response scores to percentile ranks, where the diagonal separates preferences for the two teachers and diagonal points indicate ties. In Figure 3, SURE shows much clearer preferences than GRAPE and RSR: it favors WizardLM-

Table 2: Ablation study of SURE. Best results are bolded. Abs. Δ denotes SURE's improvement over each baseline in the four-metric average, with MT-Bench scaled to 0-100 before averaging.

| Model | Method | AE2 Raw | AE2 LC | MT-Bench | MMLU | Abs. Δ |
|---|---|---|---|---|---|---|
| Qwen3-8B-Base | SURE (Ours) | **11.99** | **16.21** | **6.744** | **75.71** | |
| | w/o Cross-Teacher Extraction | 10.94 | 13.89 | 6.569 | 75.37 | 1.37 |
| | w/o Reliability Estimation | 5.94 | 7.41 | 5.900 | 75.09 | 5.98 |
| Llama-3.2-3B | SURE (Ours) | **3.13** | **3.09** | **4.300** | **55.77** | |
| | w/o Cross-Teacher Extraction | 2.71 | 2.64 | 4.287 | 55.06 | 0.43 |
| | w/o Reliability Estimation | 1.73 | 1.58 | 3.925 | 54.95 | 1.87 |

70B over UltraLM-65B on 99.97% of shared prompts for Llama-3.2-3B and 99.91% for Qwen3-8B, and favors Llama-2-70B-Chat over WizardLM-70B on 77.30% and 98.84% for the two students. However, reliability does not determine selection alone: for Llama-3.2-3B, WizardLM-70B is still preferred over Llama-2-70B-Chat on 22.70% of prompts despite its lower reliability weight in Figure 2b, showing that prompt-specific gradient coherence remains influential. GRAPE and RSR also produce ties on 0.75% and 0.18% of prompts, because identical responses receive the same prompt-specific scores regardless of the teacher, which explains the unchanged tie rates across students. By contrast, SURE complements local coherence with teacher reliability estimated across prompts, and provides additional evidence when prompt-specific signals cannot distinguish the candidates.

### 4.5 External Quality Analysis of Selected Responses

To further assess the curated distillation sets, we use UltraFeedback's provided rubric-based GPT-4 "Overall" ratings as external labels for quality analysis. As shown in Figure 4, SURE selects responses with higher mean ratings than GRAPE and RSR. For Llama-3.2-3B, SURE reaches 7.221 compared with 6.677 for GRAPE and 6.541 for RSR; for Qwen3-8B, SURE reaches 7.117 compared with 6.816 and 6.507. These results show that SURE's unsupervised selection based on prompt-specific gradient coherence and cross-prompt teacher reliability also identifies responses with higher externally assessed quality, which provides additional empirical support beyond downstream fine-tuning performance.

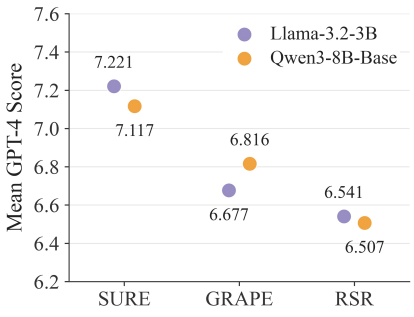



Figure 4: Mean UltraFeedback GPT-4 "Overall" ratings of selected responses.

### 4.6 Ablation Analysis of SURE

Table 2 shows that both cross-teacher supervision signal extraction and teacher reliability estimation contribute to SURE's downstream performance. Removing cross-teacher extraction decreases the average by 1.37% on Qwen3-8B and 0.43% on Llama-3.2-3B, indicating that the extracted signals provide cleaner within-prompt coherence for comparing candidate learning signals. Removing reliability estimation causes larger drops of 5.98% and 1.87%. This supports the importance of aggregating coherence evidence across prompts to estimate teacher reliability rather than relying only on within-prompt response comparisons. The full SURE achieves the best result on every reported metric, confirming that coherence from cross-teacher supervision signals and cross-prompt teacher reliability provide complementary signals for unsupervised distillation.

## 5 Conclusion

We presented SURE, an unsupervised framework for multi-teacher distillation that uses both gradient-based supervision coherence within each prompt and teacher reliability estimated across prompts. Across student models from different families and scales, SURE improves downstream distillation performance over existing selection baselines on multiple tasks. Further analyses show that SURE selects responses with higher externally assessed quality, while the estimated teacher reliability captures teacher-specific differences that vary with the target student and beyond teacher scale alone. These highlight cross-prompt teacher reliability as a key source of evidence for selecting effective supervision beyond prompt-specific student signals without external judgment labels.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods, polishing manuscript writing and proof writing for clarity and readability, and finding related work. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret experimental results. We have reviewed all AI-assisted implementations and writing. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our experiments throughout the paper and supplementary materials. The datasets, student models, training setup, evaluation benchmarks, and compared baselines are described in Section 4.1. Detailed training configurations, inference settings, benchmark evaluation procedures, and LLM-judge settings for benchmarks such as AlpacaEval 2 and MT-Bench are provided in Appendix A. The complete codebase and all experiment configurations are included in the supplementary material. Upon publication, we will make the code and configurations publicly available under the MIT License.

## REFERENCES

Ishika Agarwal, Krishnateja Killamsetty, Lucian Popa, and Marina Danilevsky. Delift: Data efficient language model instruction fine-tuning. In *International Conference on Learning Representations*, volume 2025, pp. 100488–100509, 2025.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ktrw68Cmu9c.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024a.

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 *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=ffLblkoCw8.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*, volume 2024, pp. 34767–34797, 2024c.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Qiang Yue, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BOorDpKHiJ.

Qirun Dai, Dylan Zhang, Jiaqi W Ma, and Hao Peng. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities. In *EMNLP (Findings)*, pp. 7079–7102, 2025.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4hturzLcKX.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Tao Han, Huaixuan Shi, Xinyi Ding, Xi-Ao Ma, Huamao Gu, and Yili Fang. Mixture of experts based multi-task supervise learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14256–14264, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428, 2023.

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, 2023.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3134–3154, 2023.

Ruihan Jin, Pengpeng Shao, Zhengqi Wen, Jinyang Wu, Mingkuan Feng, Shuoyang, Chu Yuan Zhang, and Jianhua Tao. Exploring knowledge purification in multi-teacher knowledge distillation for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7pvJoB4aKO.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in llm reasoning. *Advances in Neural Information Processing Systems*, 38:90649–90685, 2026.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 21921–21942, 2024.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=uydQ2W41KO.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. Aligning large language models by on-policy self-judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11442–11459, 2024b.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16189–16211, 2024a.

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 *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, 2024b.

Zhuang Li, Yuncheng Hua, Thuy Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. Scar: Data selection via style consistency-aware response ranking for efficient instruction-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12790, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. *Advances in Neural Information Processing Systems*, 37:97800–97825, 2024a.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 22353–22373, 2024b.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, pp. 26106–26128. PMLR, 2023.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15395–15406, 2024.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good graces: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, volume 2026, pp. 54779–54822, 2026.

Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7499–7517, 2024.

Xuan Ren, Qi Chen, and Lingqiao Liu. Efficiently selecting response generation strategies for synthetic data construction by self-aligned perplexity. In *EMNLP (Findings)*, pp. 11584–11605, 2025.

Ayrton San Joaquin, Bin Wang, Zhengyuan Liu, Nicholas Asher, Brian Lim, Philippe Muller, and Nancy Chen. In2core: Leveraging influence functions for coreset selection in instruction finetuning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10324–10335, 2024.

Zexu Sun, Yiju Guo, Yankai Lin, Xu Chen, Qi Qi, Xing Tang, Ji-Rong Wen, et al. Uncertainty and influence aware reward model refinement for reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2025, pp. 102240–102261, 2025.

Zinan Tang, Xin Gao, Qizhi Pei, Zhuoshi Pan, Mengzhang Cai, Jiang Wu, Conghui He, and Lijun Wu. Middo: Model-informed dynamic data optimization for enhanced llm fine-tuning via closed-loop learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6882–6902, 2025.

Somin Wadhwa, Chantal Shaib, Silvio Amir, and Byron C Wallace. Who taught you that? tracing teachers in model distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3307–3315, 2025.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Nice data selection for instruction tuning in llms with non-differentiable evaluation metric. In *Forty-second International Conference on Machine Learning*, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024c.

Jingyu Wei, Bo Liu, Tianjiao Wan, Baoyun Peng, Xingkong Ma, and Mengmeng Guo. Ji2s: Joint influence-aware instruction data selection for efficient fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 514–527, 2025.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In *EMNLP (Findings)*, pp. 2698–2711, 2025.

Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *International Conference on Learning Representations*, volume 2025, pp. 72274–72303, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *International Conference on Learning Representations*, volume 2024, pp. 30745–30766, 2024.

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, 2025.

Chuanguang Yang, Xinqiang Yu, Han Yang, Zhulin An, Chengqing Yu, Libo Huang, and Yongjun Xu. Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9148–9156, 2025.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, et al. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42123–42150, 2026.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*, volume 2024, pp. 55361–55414, 2024.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950, 2023.

Dylan Zhang, Qirun Dai, and Hao Peng. The best instruction-tuning data are those that fit. *Advances in Neural Information Processing Systems*, 38:141172–141208, 2026.

Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, and Shiming Ge. Coupled confusion correction: Learning from crowds with sparse annotations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16732–16740, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, volume 2025, pp. 12476–12505, 2025.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=0AZAjkXhit`.

Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, et al. Beyond similarity: A gradient-based graph method for instruction tuning data selection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24391–24404, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yuhang Zhou and Wei Ai. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 265–282, 2024.

# A IMPLEMENTATION

## A.1 TRAINING HYPERPARAMETERS.

As shown in Table 3, we use the same data construction and SFT training configurations for SURE and all baselines to ensure a fair comparison. We evaluate four student models: Qwen3-4B-Base, Qwen3-8B-Base, Llama-3.2-3B, and Llama-3.1-8B. Each method selects one response for each of the 63,748 UltraFeedback prompts, and all students are full-parameter fine-tuned for one epoch in BF16 with a learning rate of $1 \times 10^{-5}$, cosine scheduling with a 0.03 warmup ratio, an effective batch size of 256, and a maximum sequence length of 8,192 tokens. All experiments are conducted on NVIDIA H200 GPUs using DeepSpeed ZeRO-2 with seed 42. The same seed is also used for the Uniform Sampling baseline during response selection.

Table 3: Data construction and SFT hyperparameters for SURE and baselines.

| **Hyperparameter** | **Value** |
|---|---|
| *Data Construction* | |
| Source Dataset | UltraFeedback |
| Retained Prompts | 63,748 |
| Candidate Responses per Prompt | 4 |
| Selected Responses per Prompt | 1 |
| *SFT Training Configuration* | |
| Fine-tuning Method | Full-parameter SFT |
| Precision | BF16 |
| Training Examples | 63,748 |
| Learning Rate | $1 \times 10^{-5}$ |
| LR Scheduler | Cosine |
| Warmup Ratio | 0.03 |
| Optimizer | Fused AdamW |
| Adam $(\beta_1, \beta_2)$ | $(0.9, 0.999)$ |
| Adam $\epsilon$ | $10^{-8}$ |
| Weight Decay | 0.0 |
| Maximum Gradient Norm | 1.0 |
| Epochs | 1 |
| Effective Batch Size | 256 |
| Maximum Sequence Length | 8,192 tokens |
| Sequence Packing | Disabled |
| *Infrastructure* | |
| Distributed Training | DeepSpeed ZeRO-2 |
| Seed | 42 |

### A.2 INFERENCE HYPERPARAMETERS.

All benchmarks follow the default evaluation settings of their official implementations. We use vLLM 0.11.0 for inference, with a maximum context length of 16,384 tokens and random seed 42. No additional system prompt is used for any evaluation. AlpacaEval 2 uses GPT-4o to perform log-probability-weighted pairwise judging against GPT-4-1106-preview reference responses, while MT-Bench uses GPT-4o for single-response scoring with the official task-specific rubrics. Following the official MT-Bench generation settings, we use category-dependent temperatures: 0.7 for writing and roleplay, 0.1 for STEM and humanities, and 0.0 for extraction, math, coding, and reasoning tasks. IFEval is evaluated with the official constraint checks and reports strict and loose prompt-level accuracy. MMLU follows the official five-shot multiple-choice evaluation, selecting the answer token with the maximum log-likelihood among the candidate choices without generating reasoning. MMLU-Pro uses five-shot greedy generation and exact-answer evaluation with the official answer-extraction procedure. Detailed benchmark-specific inference and judge configurations are reported in Tables 4 and 5.

Table 4: Inference and judging settings for LLM-judge benchmarks.

| **Hyperparameter** | **AlpacaEval 2** | **MT-Bench** |
|---|---|---|
| *Inference Configuration* | | |
| Number of Questions | 805 | 80 |
| Few-shot Examples | 0 | 0 |
| Decoding | Sampling | Category-dependent |
| Temperature | 0.7 | 0.0 / 0.1 / 0.7 |
| Top-$p$ | 1.0 | 1.0 |
| Top-$k$ | -1 | -1 |
| Maximum New Tokens | 2,048 | 1,024 per turn |
| *LLM Judge Configuration* | | |
| LLM Judge Model | GPT-4o | GPT-4o |
| Judge Temperature | 0.0 | 0.0 |
| Judge Maximum New Tokens | 1 | 2,048 |
| Reference Model | GPT-4-1106-preview | GPT-4 |

Table 5: Inference and judging settings for rule-based benchmarks.

| **Hyperparameter** | **IFEval** | **MMLU** | **MMLU-Pro** |
|---|---|---|---|
| *Inference Configuration* | | | |
| Number of Questions | 541 | 14,042 | 12,032 |
| Few-shot Examples | 0 | 5 | 5 |
| Decoding | Sampling | Maximum answer-token log-likelihood | Greedy |
| Temperature | 0.7 | N/A | 0.0 |
| Top-$p$ | 1.0 | N/A | 1.0 |
| Top-$k$ | -1 | N/A | -1 |
| Maximum New Tokens | 4,096 | 1 | 2,048 |

### A.3 TEACHER COMPOSITION OF THE ULTRAFEEDBACK RESPONSE POOL

After deduplication and removal of incomplete response groups, the UltraFeedback pool contains 63,748 prompts with four candidate responses per prompt, including 254,992 responses from 17 teacher models. Table 6 reports the number of responses contributed by each teacher. For analysis, we group open-source teachers by parameter scale into $<$ 10B, 10-30B, and $\geq$ 30B groups, while closed-source teachers are treated as a separate group as their parameter counts are not publicly available. These same groups are used in Figure 2 to examine how SURE's estimated teacher reliability varies across teacher scales while retaining teacher-level differences within each group.

Table 6: Teacher composition of the cleaned UltraFeedback response pool before selection, grouped by model scale for the teacher-reliability analysis in Figure 2.

| Group | Teacher | Responses |
|---|---|---|
| $< 10$B | alpaca-7b | 16,513 |
| | llama-2-7b-chat | 16,730 |
| | wizardlm-7b | 16,549 |
| $10 - 30$B | llama-2-13b-chat | 16,585 |
| | pythia-12b | 832 |
| | starchat | 13,921 |
| | ultralm-13b | 16,525 |
| | wizardlm-13b | 16,747 |
| $\geq 30$B | falcon-40b-instruct | 16,458 |
| | llama-2-70b-chat | 16,534 |
| | mpt-30b-chat | 16,510 |
| | ultralm-65b | 16,610 |
| | vicuna-33b | 16,699 |
| | wizardlm-70b | 16,476 |
| Closed-source | bard | 8,326 |
| | gpt-3.5-turbo | 16,516 |
| | gpt-4 | 16,461 |
| **Total** | | **254,992** |

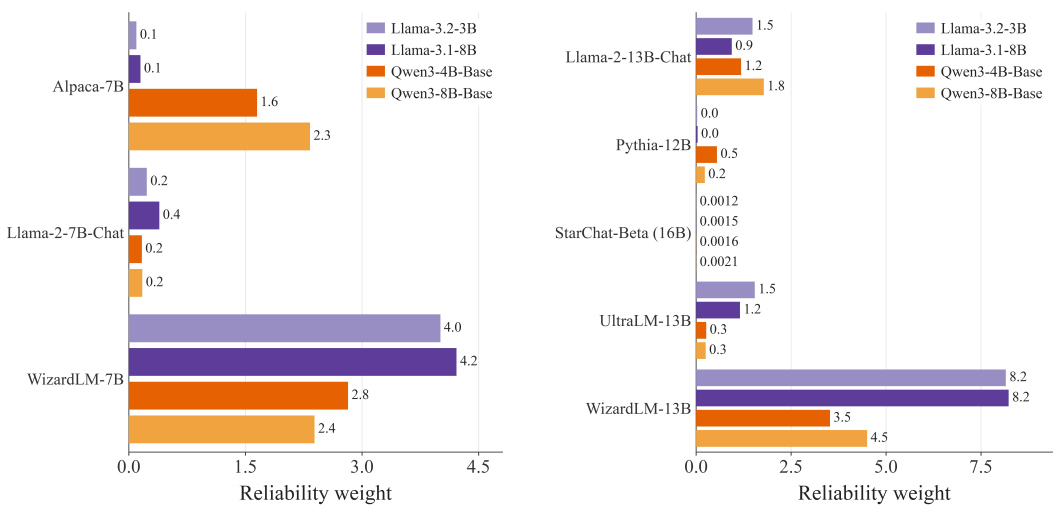
# B Additional Results

## B.1 Comparative Runtime Analysis

Table 7: End-to-end runtime for unsupervised multi-teacher distillation set curation on UltraFeedback response pool.

| Student | Runtime (min) | | |
|---|---|---|---|
| | **SURE** | **GRAPE** | **RSR** |
| Qwen3-4B-Base | 14.12 | 12.40 | 12.84 |
| Qwen3-8B-Base | 20.75 | 17.95 | 18.32 |
| Llama-3.2-3B | 10.74 | 8.90 | 9.23 |
| Llama-3.1-8B | 18.20 | 16.15 | 16.49 |

As shown in Table 7, curation runtime increases with student model size, mainly due to the higher cost of forward inference for larger students. SURE computes hidden-state gradients directly from the student's forward-pass outputs without backpropagating through the model, keeping its gradient-informed distillation set curation computationally efficient. Compared with GRAPE and RSR, SURE introduces only about 1.3-2.8 additional minutes of runtime across the four students, primarily from its iterative teacher reliability estimation. This modest overhead enables SURE to incorporate cross-prompt teacher reliability while adding only a small runtime cost, contributing to substantially stronger downstream student performance.




(a) Reliability of the <10B teachers

(b) Reliability of the 10 - 30B teachers.

(c) Reliability of the ≥30B teachers.

(d) Reliability of the closed-source teachers.

Figure 5: Estimated teacher reliability weights for all students.

### B.2 Group-wise Analysis of Estimated Teacher Reliability

Figure 5 provides the teacher-level reliability weights within each group used in the main analysis. Broad reliability patterns are similar across students, with several teachers receiving consistently higher weights within their groups, but substantial student-dependent differences remain. Within the $<$ 10B group, the Qwen students assign noticeably higher reliability to Alpaca-7B than the Llama students, while WizardLM-7B receives the highest weight for all four students. Similar variation appears among closed-source teachers: the Llama students assign high reliability to both GPT-3.5-Turbo and GPT-4, while the Qwen students show different relative preferences between these teachers. Within the same parameter groups, larger teachers are also not consistently assigned higher reliability. For example, MPT-30B-Chat receives higher weights than several larger 65B-70B teachers across students. These results further show that the estimated reliability weights do not simply follow teacher size. Instead, student-derived gradient coherence captures how well each teacher's responses align with the learning signals of a particular student, so teachers providing more suitable supervision can receive higher reliability even when they are smaller teacher models.

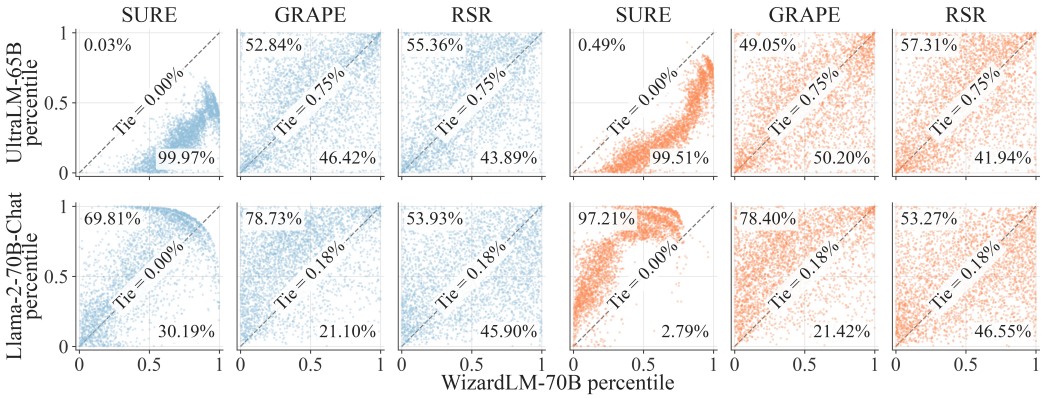



Figure 6: Pairwise teacher preferences for Llama-3.1-8B (left) and Qwen3-4B-Base (right). Response-selection scores are converted to percentile ranks for comparison. Points above and below the diagonal indicate prompts favoring the y- and x-axis teachers, and diagonal points indicate ties.

### B.3 Additional Pairwise Analysis of Teacher Preferences

Figure 6 shows that the pairwise preference patterns also remain for the remaining Llama-3.1-8B and Qwen3-4B-Base students. SURE strongly favors WizardLM-70B over UltraLM-65B on 99.97% and 99.51% of shared prompts, while GRAPE and RSR produce much more balanced preferences. For Llama-2-70B-Chat versus WizardLM-70B, SURE favors Llama-2-70B-Chat on 69.81% of prompts for Llama-3.1-8B and 97.21% for Qwen3-4B-Base, showing that the strength of teacher preference varies with the student. At the same time, WizardLM-70B remains preferred on 30.19% and 2.79% of prompts, indicating that teacher reliability does not override prompt-specific gradient coherence. The same 0.75% and 0.18% baseline tie rates also appear here because GRAPE and RSR assign the same score when paired teachers produce identical responses. These results further illustrate how SURE considers student-dependent cross-prompt teacher reliability with local coherence for response selection.

# C CASE STUDY

Table 8: Candidate teacher responses for the race-position case study of Ultrafeedback.

| |
|---|
| **Prompt:** You're running a race and pass the person in 2nd place. What place are you in now? |
| **Teacher A (Bard):** If I pass the person in second place, I will be in first place. |
| **Teacher B (Falcon-40B-Instruct):** You are now in first place after passing the person in second place. |
| **Teacher C (WizardLM-13B):** If you've passed the person in second place during a race, that means you have moved into their position, and thus, you are now at second place yourself. |
| **Teacher D (WizardLM-7B):** As an AI language model, I do not have a physical presence to run a race nor do I have a concept of time. However, in a simple scenario, if you pass the person in 2nd place while running a race, you are now in 1st place. |

Table 9: Comparison of response-selection signals for the race-position case study. Bold values indicate each method's selected response.

| **Method** | **Measurement** | **Teacher A Response** | **Teacher B Response** | **Teacher C Response** | **Teacher D Response** |
|---|---|---|---|---|---|
| SURE | Mean gradient coherence | 0.2202 | 0.2509 | 0.1513 | $-0.1110$ |
| | Estimated teacher reliability | 0.1478 | 0.0011 | 8.2288 | 4.2166 |
| | Posterior probability ↑ | 0.0197 | 0.0002 | **0.8309** | 0.1492 |
| GRAPE | Avg. NLL ↓ | 1.3732 | 1.4620 | 1.9729 | **1.1601** |
| RSR | Rank-surprisal ratio ↓ | 2.9585 | **2.2963** | 3.0873 | 2.6163 |
| GPT-4 | Overall rating ↑ | 5 | 4 | **9** | 3 |
| **Answer Correctness** | | Incorrect | Incorrect | **Correct** | Incorrect |

For the race-position case in Table 8 and 9 for the Llama-3.1-8B student, Teacher C provides the only correct response and receives the highest GPT-4 "Overall" rating of 9, but its mean gradient coherence of 0.1513 is lower than Teacher A at 0.2202 and Teacher B at 0.2509. However, Teacher C comes from WizardLM-13B, which receives the highest estimated reliability weight of 8.2288, and SURE assigns its response the highest posterior probability of 0.8309. The interaction between the two signals is also evident for the other candidates. For example, Teacher B has the highest gradient coherence but a reliability weight of only 0.0011, resulting in a posterior probability of 0.0002, while Teacher D has the second-highest reliability weight of 4.2166 but negative coherence of $-0.1110$, reducing its posterior probability to 0.1492. GRAPE and RSR instead select Teachers D and B, both of which give incorrect answers. This case demonstrates the benefit of SURE considering within-prompt gradient coherence together with teacher reliability estimated across prompts. Gradient coherence provides evidence about the current response, and accumulated cross-prompt evidence helps identify teachers that more consistently provide suitable supervision. Their combination prevents either signal from dominating the selection and enables SURE to select the correct and most suitable response to the student even when it does not have the highest within-prompt coherence.