

Towards Reward Modeling for AI Tutors in Math Mistake Remediation

Kseniia Petukhova, Ekaterina Kochmar

Mohamed bin Zayed University of Artificial Intelligence

{kseniia.petukhova, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

Evaluating the pedagogical quality of AI tutors remains challenging: standard NLG metrics do not determine whether responses identify mistakes, scaffold reasoning, or avoid revealing the answers. For the task of mistake remediation, we derive a hierarchy of pedagogical aspects from human pairwise preferences on MRBench, and synthesize minimally contrastive response pairs that differ along key aspects (e.g., mistake identification and location, targetedness, scaffolding, actionability, clarity, and coherence). We develop and release Bradley-Terry preference models trained on weighted-sum rankings that we automatically create from MRBench, synthetic pairs, and data combinations. Using only synthetic data, our best model reaches 0.69 pairwise accuracy on a human preference test, and combining weighted-sum data with targeted synthetic groups improves accuracy to 0.74, outperforming larger general-purpose reward models while using only a 0.5B-parameter backbone.

Keywords: AI tutoring, reward modeling, synthetic data

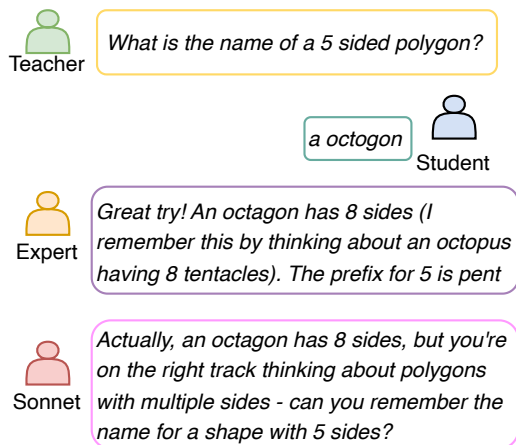
1. Introduction

Human tutoring has long been recognized as one of the most effective forms of education, offering personalized guidance. Despite its success in improving learning outcomes (Bloom, 1984), providing one-on-one tutoring at scale remains difficult due to the lack of qualified tutors and the high costs. Recent advances in large language models (LLMs) have opened new opportunities to address this gap. Their impressive conversational and reasoning abilities have led to a growth of interest in LLM-powered intelligent tutoring systems (ITSs) (Pal Chowdhury et al., 2024; Liu et al., 2024). These systems can assist learners across diverse educational tasks (Wollny et al., 2021), among which supporting students in understanding and remediating their mistakes has become a particularly promising application (Macina et al., 2023; Wang et al., 2023).

While substantial progress has been made in building AI tutoring systems, evaluating their pedagogical quality remains a significant challenge. Conventional natural language generation (NLG) metrics (Chin-Yew, 2004; Popović, 2017; Post, 2018; Gao et al., 2020; Liu et al., 2023) cannot be used for this purpose, as they are domain-agnostic, rely on reference texts, and fail to capture the nuanced pedagogical aspects of effective tutoring. In the context of student mistake remediation, an ideal AI tutor should guide the learner toward understanding rather than simply reveal the answer, while maintaining factual accuracy, clarity, and a supportive tone. However, studies show that existing LLMs often fall short, simply revealing solutions without providing sufficient guidance or explanation (Macina et al., 2023).

To facilitate systematic evaluation of AI tutors,

Maurya et al. (2025) introduced a taxonomy for assessing pedagogical abilities of LLM-powered tutors and released MRBench – a benchmark containing tutor responses annotated across eight pedagogical dimensions: *mistake identification* (whether the tutor identifies an error), *mistake location* (whether it pinpoints where the error occurs), *providing guidance* (whether the tutor provides relevant guidance), *actionability* (whether it is clear from the response what to do next), *revealing of the answer*, *coherence*, *tutor tone*, and *human-likeness*. Each dialog context in the dataset contains 8-9 alternative tutor responses produced by humans or LLMs. Building on this work, Kochmar et al. (2025) organized a shared task for automatic evaluation across the first four dimensions. Despite promising results on these individual aspects, the question of determining the overall quality of a tutor response remains open. For example, one response might excel at *mistake location*, while another is more *actionable*. Prior work, such as Macina et al. (2025), addressed this by assigning equal weights to all dimensions and summing the scores. However, this approach leads to frequent ties, and it assumes uniform importance across pedagogical aspects – an unrealistic assumption, since dimensions like *providing guidance* likely carry greater pedagogical weight than *tutor tone* or *human-likeness*. An example of such a tie is shown in Figure 1, where two responses receive identical overall scores. The Sonnet response achieves a full score for *actionability*, whereas Expert receives only a partial score; conversely, Expert receives a full score for *tutor tone*, while Sonnet does not. Despite having the same total score, the responses differ in pedagogical quality: the Sonnet response is



MRBench annotation	Expert	Sonnet
Mistake Identification	Yes	Yes
Mistake Location	Yes	Yes
Revealing of the Answer	No	No
Providing Guidance	Yes	Yes
Actionability	To some extent	Yes
Human-likeness	Yes	Yes
Coherence	Yes	Yes
Tutor Tone	Encouraging	Neutral

Figure 1: Example of annotated tutor responses from MRBench. The Sonnet response is annotated as more actionable (“Yes” vs. “To some extent”) because it prompts the student to recall the prefix for a five-sided shape rather than revealing it directly. In contrast, the Expert response is more encouraging in tone (“Encouraging” vs. “Neutral”).

more actionable, prompting the student to recall the correct prefix, while the Expert response is more encouraging but partially reveals the answer.

To address these limitations, we conduct human pairwise preference annotation to identify which pedagogical aspects most strongly influence judgments of response quality. Using these insights, we construct a synthetically generated dataset of response pairs that differ along these key aspects and train a ranking model that captures human preferences. The dataset and the trained model are publicly released to support future research on pedagogical evaluation and reward modeling for AI tutors. This model can serve as an evaluation tool or reward model for aligning LLM-based tutors toward improved pedagogical behavior.

In summary, our key contributions are as follows:

- We perform human pairwise preference annotation on responses from MRBench (Maurya et al., 2025) and derive a hierarchy of pedagogical aspects that humans prioritize when

selecting the better response.

- We propose a method for synthetically generating response pairs that differ systematically along these aspects, creating a new dataset of pedagogically contrastive tutor responses.
- We train a reward model that outperforms existing general-purpose reward models, making it better suited for evaluating and aligning LLM-powered tutoring systems.
- We release all resources developed in this work, including the annotated human preference data, the synthetic augmentation dataset, the trained reward model, and the code, to support further development of pedagogically-aligned AI tutors: https://github.com/Kpetyxova/Towards_Reward_Modeling_for_Tutors
- Although our experiments focus on math mistake remediation, the proposed methodology for preference data generation can be applied to other domains.

2. Related Work

Pedagogical Ability Assessment Maurya et al. (2025) introduce a systematic framework to evaluate the pedagogical capabilities of LLM-powered tutors. The main contribution of this paper lies in its creation of the first unified evaluation taxonomy explicitly designed for assessing AI tutors’ pedagogical abilities in student mistake remediation tasks. Prior research evaluated tutoring models using fragmented and inconsistent criteria, making cross-model comparison difficult. This work unifies these diverse efforts into a coherent framework of eight pedagogical dimensions: *mistake identification*, *mistake location*, *revealing of the answer*, *providing guidance*, *actionability*, *coherence*, *tutor tone*, and *human-likeness*, grounded in learning sciences principles. For the *revealing of the answer* dimension, the labels are No, Yes (correct answer), and Yes (incorrect answer). For the *tutor tone*, the possible values are *Negative*, *Neutral*, and Encouraging. Each remaining dimension can take one of three possible values: Yes, *To some extent*, or *No*.¹

Alongside the taxonomy, the authors introduce MRBench, a new benchmark of 192 dialogs and 1,596 tutor responses (human and LLM-generated), annotated with human gold labels. MRBench is constructed from contexts and tutor responses drawn from the MathDial (Macina et al., 2023) and Bridge (Wang et al., 2023) datasets and extended with responses generated by seven state-of-the-art LLMs acting as tutors. Human evaluation of these responses reveals that while linguistically fluent,

¹Underlined are the preferred options.

current LLMs often lack pedagogical depth. Models such as GPT-4 and Llama-3.1-405B perform well in identifying and locating student mistakes but tend to reveal answers too frequently, which diminishes their instructional effectiveness. Mistral and Sonnet show moderate performance, whereas Gemini and Phi-3 perform poorly, producing incoherent or unhelpful guidance. Smaller models like Llama-3.1-8B achieve relatively balanced yet still suboptimal results. Finally, the authors explore whether LLMs can serve as evaluators across the proposed dimensions. Correlation analyses show that LLM-based evaluators remain unreliable, as their judgments correlate weakly or negatively with human assessments across most pedagogical dimensions.

Following this, Kochmar et al. (2025) present the outcomes of a shared task aimed at automating the evaluation of AI tutors' pedagogical abilities across four of the above dimensions: *mistake identification*, *mistake location*, *providing guidance*, and *actionability*. Over 50 international teams participated, employing diverse techniques including LoRA-based fine-tuning, in-context learning, data augmentation, ensemble modeling, and reinforcement learning with human feedback. The results demonstrate meaningful progress but highlight persistent pedagogical quality assessment challenges. Across the four tracks, the best-performing systems achieved macro F1 scores ranging from 58.3 (for *providing guidance*) to 71.8 (for *mistake identification*), indicating that automated systems can capture some pedagogical dimensions but still fall short of human-level understanding.

In contrast to the shared task approach, Macina et al. (2025) train a reward model that scores the overall pedagogical quality of generated tutor utterances. This model is trained using pairwise ranking on preference data derived from expert and novice teacher responses, and from annotations from MR-Bench – for each response, a cumulative score is computed by summing the presence of desirable values across eight evaluation dimensions. The resulting reward model achieves an accuracy of 0.84 in distinguishing expert from novice teacher responses. However, this approach often produces tied scores and implicitly assumes equal importance across all eight dimensions. Furthermore, the evaluation test set may overestimate performance, as it is based on the Bridge dataset, where novice responses are typically much weaker than those of expert tutors and are characterized by repetitive, generic feedback patterns such as “try again” or “check your answer.”

Generation of Synthetic Preference Pairs To enhance the quality of preference data used for reward-model training, synthetic preference gener-

ation techniques have been developed. For example, Pace et al. (2024) propose a method in which responses are generated in pools and then the best and worst are paired to create synthetic preference pairs. Shen et al. (2024) present another approach: they first generate one response, assign a desired preference label, and then generate a second response conditioned on the first one plus the label and guidance about multiple quality aspects. At the same time, Contrastive Learning from AI Revisions (CLAIR) (D'Oosterlinck et al., 2025) takes a different approach: it produces minimally contrastive preference pairs by having a stronger model revise the weaker model's output rather than selecting between two independently generated outputs. In CLAIR, the target model generates an initial losing output for each prompt, which a stronger reviser model then minimally improves to create the winning output. This ensures that differences between the two outputs are targeted and relevant, reducing noise from irrelevant differences. Their results show that CLAIR produces preference pairs with the highest measured contrastiveness and yields the most significant performance gains when combined with APO-zero (an alignment objective proposed in the same work). Inspired by CLAIR's minimal-revision paradigm and related multi-aspect synthetic preference generation methods such as the one from Shen et al. (2024), we transfer these ideas to the educational domain, generating targeted revisions of tutor responses along key valuable aspects.

3. Creation of Preference Pairs

3.1. Preliminary Ranking

In this section, we construct a preliminary set of preference pairs following the approach of Macina et al. (2025). However, we modify the weighting scheme to account for the varying importance of different evaluation dimensions. We hypothesize that some pedagogical aspects contribute more strongly to human judgments of response quality than others, and we use this assumption as a starting point for our analysis. Specifically, the most important dimensions are hypothesized to be *mistake location*, *providing guidance*, and *coherence*; moderately important dimensions include *mistake identification* and *actionability*; less important dimensions are *revealing the answer*, whose importance is context-dependent (early in the dialog revealing the answer is undesirable, but when the student is clearly stuck, it becomes necessary), and *human-likeness*; and, comparatively, the least important of eight dimensions is *tutor tone*.²

²This hypothesis is based on manual inspection of the dialogs and tutor responses quality. We highlight that

To calculate the overall score for each response, we first map each annotation label to a numerical value. For most dimensions, we assign a value of 1 to *Yes*, 0.5 to *To some extent*, and 0 to *No*. For *Revealing of the Answer*, the scoring is asymmetric: 1 if the tutor does not reveal the answer, 0.5 if the tutor reveals the correct answer, and 0 if the tutor reveals an incorrect answer. For *Tutor Tone*, we assign 1 to Encouraging, 0.5 to Neutral, and 0 to Offensive. Each numerical annotation score is then multiplied by the corresponding dimension weight: *Mistake Identification* (0.5), *Mistake Location* (1.0), *Revealing of the Answer* (0.25), *Providing Guidance* (1.0), *Actionability* (0.5), *Human-likeness* (0.25), *Coherence* (1.0), and *Tutor Tone* (0.05). The resulting weighted sum serves as an estimate of the pedagogical quality of a tutor's response. Across the 1,655 responses in the dataset, the resulting scores range from 0.025 to 4.55 (mean = 3.45, median = 3.90, SD = 1.28). The interquartile range spans from 2.78 (25th percentile) to 4.53 (75th percentile), indicating that most responses receive relatively high scores, with a moderate spread across the scale. This distribution suggests that while many responses exhibit strong pedagogical qualities, meaningful variation remains, allowing for comparative ranking. Overall, this approach provides an initial quantitative basis for ranking responses by their instructional effectiveness.

To evaluate the preliminary ranking, we conducted a human evaluation study involving four annotators, each holding at least a Master's degree in NLP or Computer Science. Each annotator evaluated tutor responses drawn from ten dialogs, five of which were shared among all participants. For every dialog, annotators assessed all response pairs with a score difference less than or equal to 0.5 – on average, about ten such pairs per dialog. In addition, each annotator evaluated three to five pairs per dialog where the score difference exceeded 0.5. The underlying intuition was that larger score differences could generally be trusted as more reliable, whereas pairs with minor or tied differences required closer human inspection. For instance, the responses shown in Figure 1 represent a pair with a score difference below 0.5, illustrating a slight contrast. The annotators labeled 152, 135, 155, and 153 pairs, respectively.

Across the five shared dialogs, there were 86 pairs: 64 with a score difference of 0.5 or less and 22 with a greater difference. For each pair, annotators were asked to select which response was better or indicate that both were either good or poor. After the initial annotation round, a group discussion was held to establish **hierarchy of pedagogical aspects** (see Appendix A), summarized as follows:

1. *Factuality + Non-contradiction + No Nonsense*: The response should be factually correct, should not contradict the student's answer(s), and should not contain irrelevant information. Two responses are considered equally poor when such aspects are lacking in both.
2. *Mistake Identification + Location*: The response should explicitly or implicitly identify a mistake in the student's solution. For example, saying "Nice try" would miss this aspect.
3. *Scaffolding + Actionability*: The response should address the misunderstanding or problem step by step, guiding the student toward understanding by asking questions or providing hints instead of directly giving away the solution. However, providing the solution is preferable if the context shows that scaffolding has not worked.
4. *Targetedness*: The response should address the core misconception or misunderstanding of a student.
5. *Not revealing the final answer*: While sharing the answer to a substep is sometimes necessary and acceptable, the tutor should avoid giving away the final answer.
6. *Clarity + Coherence*: The tutor's response should be free of awkward, confusing, or misleading wording. A good tutor response acknowledges the student's input and connects it to the next step.

The inter-annotator agreement among the four annotators, measured using Fleiss' Kappa across all shared pairs, was 0.72, indicating substantial agreement. For shared pairs with a score difference greater than 0.5, annotators achieved perfect agreement (Fleiss' Kappa = 1.0) and agreed with the preliminary ranking in 88% of cases. This outcome is expected, as pairs with a score difference above 0.5 represent responses that are reliably distinct in quality, making them easier to judge consistently. The agreement was lower for pairs with minor score differences or ties, with a Fleiss' Kappa of 0.62. An illustrative case of annotator disagreement for such borderline pairs is provided in Appendix C, where differing pedagogical preferences led to divergent judgments despite comparable overall quality.

Table 1 shows an example: human annotators preferred the second response because it provided scaffolding, which they prioritized. However, according to the MRBench annotation, this response scored worse, since it only partially identified the mistake and did not locate it. Additional examples of such discrepancies are presented in Appendix B.

it is only used to produce the rankings of the responses as the starting point.

Together, these cases highlight that perfect alignment between the weighting scheme and human annotation is unattainable with the current MR-Bench dimensions. For instance, *providing guidance* and *actionability* do not distinguish between scaffolding and more detailed explanations; a *providing guidance* label of *No* could mean either that guidance is missing or that guidance is present but factually incorrect. Consequently, responses that prioritize scaffolding may receive lower scores on *mistake identification* and *mistake location*, revealing a structural tension between these dimensions.

Thus, while the proposed preliminary ranking provides a reasonable approximation and can be used as a silver standard, it cannot fully capture the nuanced preferences observed during human annotation. A promising direction is to use the annotated data and the hierarchy of valuable aspects to prompt LLMs to detect such misalignments – our next step in this work.

3.2. Can We Use LLMs for Preference Annotation?

Since the proposed preliminary ranking cannot serve as a gold standard, we examine whether LLMs can be used for preference annotation. We tested four prompting strategies – *basic*, *with guidelines*, *with hierarchy*, and *with checklist*, across three proprietary models: GPT-4.1, Claude-sonnet-4, and Gemini-1.5-pro. The full prompts are provided in Appendix D.

Evaluation on 337 human-annotated pairs showed accuracies between 0.58 and 0.66 across models. The *with guidelines* prompt performed best (0.66 for Claude-sonnet-4 and Gemini-1.5-pro; 0.64 for GPT-4.1), and an ensemble of all models reached 0.67. Although moderately aligned with human judgments, these results remain insufficient for reliable large-scale annotation.

3.3. Synthetic Data Augmentation

Building on the concept of minimal revisions proposed by D'Oosterlinck et al. (2025), we introduce a synthetic data augmentation procedure designed to enrich MRBench with additional preference pairs that vary meaningfully along the aspects defined during the preference annotation in the hierarchy of pedagogical aspects.

In designing this procedure, we intentionally focus on the *revealing the answer*, *providing guidance*, *actionability*, and *coherence* dimensions from MRBench when grouping the original responses for each step we describe below. As discussed in the misalignment example in Table 1, low scores in the *Mistake Identification (MI)* and *Mistake Location (ML)* dimensions do not necessarily indicate poor pedagogical quality – some high-quality responses

may provide effective scaffolding without explicitly identifying the mistake. To avoid penalizing such cases, we exclude the *MI* and *ML* dimensions from the filtering process and instead prioritize aspects that more directly capture instructional quality and student understanding.

Our proposed algorithm proceeds as follows (see Figure 2):

- 1. Improving suboptimal responses:** For each MRBench response with a *suboptimal* annotation (i.e., a response that does not have desirable values in the *revealing the answer*, *providing guidance*, *actionability*, and *coherence* dimensions), we generate four revised versions. We assume that suboptimal responses can be improved across multiple aspects of the hierarchy of pedagogical aspects derived from human annotation. Accordingly, each revision improves the original response along one of the following aspects: (a) *Mistake Identification + Location*, (b) *Scaffolding + Actionability*, (c) *Targetedness*, or (d) *Clarity + Coherence*. We exclude the *Factuality* and *Not Revealing the Final Answer* aspects, since responses that are already factually correct or non-revealing cannot be further improved in those respects. Each generated revision is paired with the original response, with the revision marked as preferred.
- 2. Joint improvement of suboptimal responses:** For every suboptimal response, we ask an LLM to simultaneously improve the response across all four aspects from Step 1. The resulting revision is paired with the original response and each aspect-specific revision from Step 1, with the all-aspects-improved version being preferred in all cases.
- 3. Degrading optimal responses (aspect-wise):** For each *optimal* response (i.e., one achieving desirable values across the *revealing the answer*, *providing guidance*, *actionability* and *coherence*), we generate five degraded versions, each worsening one of the following aspects: (a) *Mistake Identification + Location*, (b) *Scaffolding + Actionability*, (c) *Targetedness*, (d) *Revealing the Answer*, and (e) *Clarity + Coherence*. Each degraded version is paired with the original response, with the original marked as preferred.
- 4. Degrading optimal responses (by factuality):** For every optimal response, we also generate two additional degraded versions: one that reduces *Factuality* and another that reduces quality across *all aspects simultaneously*. Each degraded response is paired

Dialog Context										
Student: yay										
Tutor: Next, we have to find the value of 6 times 9 + 2.										
Tutor: What is the value of 6 times 9 + 2?										
Student: 99										
Preferred Response	Tutor Response	MI	ML	RA	PG	AC	HM	CO	TT	Score
Preferred (weighted-sum)	Good try, but remember the order of operations: multiplication should be done before addition. So it should be 6 times 9, then add 2. Let's try again!	Yes	Yes	No	Yes	Yes	Yes	Yes	Enc.	4.55
		0.50	1.00	0.25	1.00	0.50	0.25	1.00	0.05	
Preferred (human)	Great try! What was your first step?	TSE	No	No	Yes	Yes	Yes	Yes	Enc.	3.30
		0.25	0.00	0.25	1.00	0.50	0.25	1.00	0.05	

Table 1: Example of misalignment between the weighted-sum ranking (Section 3.1) and human preference. For each response, the first row reports the original `MRBench` annotation labels, and the second row reports their weighted contributions (annotation score \times dimension weight). Abbreviations: MI – *Mistake Identification*; ML – *Mistake Location*; RA – *Revealing the Answer*; PG – *Providing Guidance*; AC – *Actionability*; HM – *Human-likeness*; CO – *Coherence*; TT – *Tutor Tone*; TSE – *To some extent*; Enc – *Encouraging*.

with the original and previously generated responses (from Steps 1–3), with the degraded responses being non-preferred. This step ensures that factually incorrect responses are treated as strictly worse than all other variants.

- Handling low-quality responses:** For responses that are *poor* across all four main dimensions, we apply the same procedure used for suboptimal responses (Steps 1–2). Additionally, each low-quality response is paired with all other generated and original responses (excluding those from Step 4), where the low-quality response is marked as non-preferred. This reflects the assumption that these *poor* responses are generally less helpful or factually incorrect compared to any improved or well-annotated versions.

Following the observation that `Claude-Sonnet-4` achieved the highest accuracy in the preference classification task (Section 3.2), we use this model for synthetic data generation. The statistics of the resulting augmented dataset are presented in Table 2.

To assess the quality of the augmented dataset, two annotators independently evaluated whether they agreed with the resulting pairwise rankings. We randomly sampled 15 dialogs: each annotator annotated five shared dialogs and five unique ones. For each dialog, we randomly selected up to two representative pairs from each group listed in Table 2, resulting in 126 and 127 annotated pairs per annotator. The first annotator agreed with the proposed rankings in 97% of cases, and the second

in 92%. The observed inter-annotator agreement on the 57 shared pairs was 0.93.

4. Reward Model

4.1. Data

Since we have human-annotated data for a subset of `MRBench` responses, we use it as our **human preference test set**, as it directly reflects the target of optimization – human preferences. Using the original 337 annotated pairs, we extend this set to 414 pairs by inferring new transitive relations such as “A is better than C” from existing annotations of the form “A is better than B” and “B is better than C,” resulting in a final test set of 414 pairs.

As discussed in Section 3.1, the preliminary ranking serves as a reasonable approximation of human preferences. To further support our analysis, we randomly select ten conversations from the preliminary ranking of `MRBench` as an additional, though less critical, **weighted-sum test set**, resulting in 369 pairs.

The remaining conversations from the preliminarily ranked `MRBench` form the **weighted-sum training set**, consisting of 2,979 preference pairs.

The **synthetic training set** is derived from the full synthetically augmented dataset. Among the generated pairs, the subset involving factually incorrect or generally poor responses originally contained 52,540 pairs – a large number resulting from comparing each poor response against all non-poor original and generated responses. To prevent this group from dominating the dataset with trivial comparisons where one response is clearly poor, we

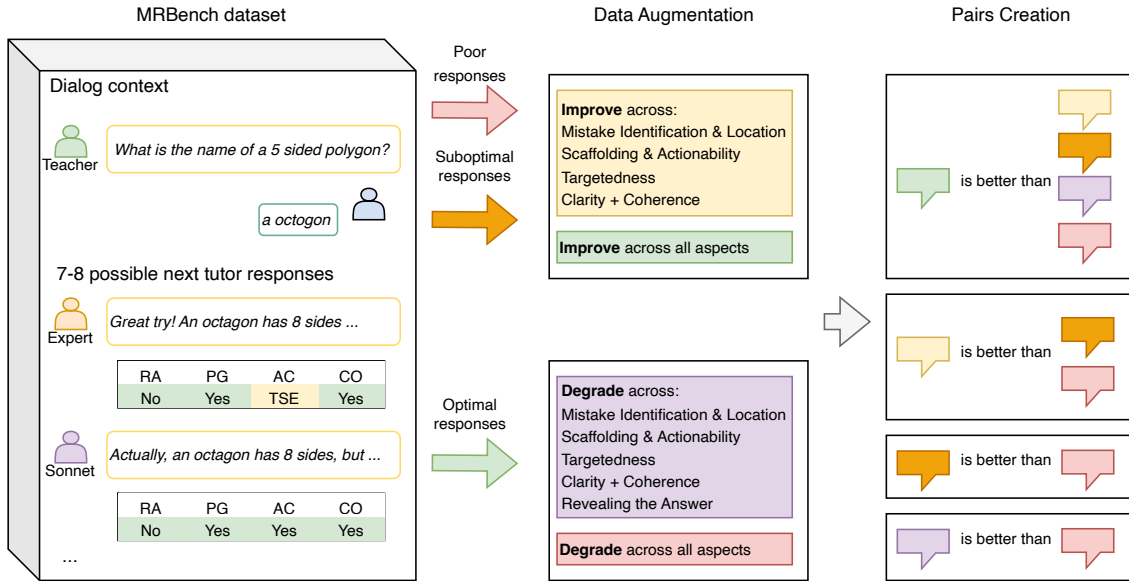


Figure 2: Pipeline for synthetic data augmentation. The procedure augments MRBench by generating aspect-specific improvements of suboptimal responses, jointly improved variants, and controlled degradations of optimal responses, thereby constructing structured preference pairs aligned with human annotation preferences. Suboptimal responses are those that do not receive desirable annotations in one or more of the following MRBench dimensions: *Revealing the Answer*, *Providing Guidance*, *Actionability*, and *Coherence*. In contrast, poor responses receive undesirable annotations across all four of these dimensions.

Block	Group	Pairs
Aspect-wise (worsen ↓)	Revealing the answer ↓	578
	Mistake identification & location ↓	578
	Clarity & coherence ↓	578
	Targetedness / core misunderstanding ↓	578
	Scaffolding & actionability ↓	578
Aspect-wise (improve ↑)	Mistake identification & location ↑	833
	Clarity & coherence ↑	833
	Targetedness / core misunderstanding ↑	833
	Scaffolding & actionability ↑	833
Joint improvement (↑)	All aspects improved vs. original	854
	vs. mistake-identification ↑	854
	vs. clarity ↑	854
	vs. targetedness ↑	854
	vs. scaffolding ↑	854
Global non-preference	Factually incorrect or fails to identify a mistake [†]	52,540

Table 2: Synthetic augmentation results by evaluation aspects. Aspect-wise *worsening* of the pairs corresponds to Step 3, *improvement* to Step 1, and joint improvement to Step 2. [†]The “global non-preference” block aggregates comparisons from Steps 4–5, where factually incorrect or mistake-unaware responses are non-preferred.

randomly downsampled it to 854 instances, matching the size of the second-largest relationship class. This sampling was performed uniformly at random, ensuring diversity without bias toward any particular dialog type. The resulting synthetic training set contains 11,346 preference pairs, combining these downsampled pairs with all other synthetic groups

as generated.

We also construct a **combined training set** by merging the weighted-sum training set with the synthetic training set, as well as with versions of the weighted-sum training set paired with each *synthetic group* (e.g., aspect-wise improvements, aspect-wise degradations, and joint im-

provements).

Lastly, to enable a fair comparison with the model from [Macina et al. \(2025\)](#), we evaluate our model on the same test set they used – namely, the subset of the `Bridge` dataset ([Wang et al., 2023](#)). Since [Macina et al. \(2025\)](#) do not specify which subset of `Bridge` they used, we construct our own subset by removing conversations that appear in `MRBench` from the test split of `Bridge` (identified by their IDs), resulting in 181 pairs. We refer to this as the **Bridge test set**. We also compare their model on our human test set, but there could be a data leakage since their model was trained on `MRBench`.

4.2. Model Training

We train a reward model using a Bradley–Terry formulation, where the loss is based on pairwise preference comparisons between model outputs. Let x denote the input sequence (prompt) and y^+ and y^- be the preferred and rejected responses, respectively. Under the Bradley–Terry model ([Bradley and Terry, 1952](#)), the probability that y^+ is preferred over y^- is given by:

$$p(y^+ \succ y^- | x) = \sigma(r_\theta(x, y^+) - r_\theta(x, y^-)),$$

where σ is the sigmoid function and $r_\theta(x, y)$ is the learned reward function parameterized by θ .

The reward model is trained to assign higher scores to preferred responses by minimizing the negative log-likelihood of the observed preferences:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x, y^+, y^-) \sim D} [\log \sigma(r_\theta(x, y^+) - r_\theta(x, y^-))].$$

Following [Macina et al. \(2025\)](#), we use `Qwen2.5-0.5B-Instruct` and `Qwen2.5-1.5B-Instruct` ([Yang et al., 2024](#)) as our pretrained base models. In addition, we include `Gemma-2-2B` ([Team et al., 2024](#)), a model of comparable size from a different architecture family, to provide cross-family comparisons.

4.3. Baselines

We compare our models against the following baselines:

1. `Skywork-Reward-V2-Llama-3.1-8B` ([Liu et al., 2025](#)) — a reward model also based on the Bradley–Terry formulation, which outperforms many existing reward models across several benchmarks on average.
2. `RewardAnything-8B-v1` ([Yu et al., 2025](#)) — a recent model trained using principle-based supervision, where the user specifies a guiding principle for a better response.

This approach aims to improve generalization to new domains by allowing domain-specific principles. The authors report improvements over `Skywork-Reward-V2-Llama-3.1-8B` on some benchmarks. For our experiments, we use the following principle: *“A better tutor response is clear, factually correct, and focused on the student’s misconception. It identifies mistakes, guides with step-by-step hints instead of giving the final answer, and stays directly connected to the student’s words.”*

3. The model from [Macina et al. \(2025\)](#) — `Qwen2.5-1.5B-Instruct` fine-tuned on `MRBench`.

5. Results & Analysis

Table 3 presents pairwise accuracy scores across three model architectures and several training data configurations. We evaluate each setup on the *weighted-sum test set* and the *human preference test set*. On the weighted-sum test set, larger models tend to yield slightly better results: `Qwen2.5-1.5B` consistently outperforms its 0.5B variant, while the 2B-parameter `Gemma-2` achieves the highest overall accuracies on the weighted-sum test set.

However, improvements on the human preference test set are more nuanced. We initially expected the synthetic dataset to yield higher performance, as it was explicitly designed to capture aspects valued by human annotators. Yet, only `Qwen2.5-0.5B` shows a slight improvement when trained on the synthetic data, while the other two models do not exhibit this trend. Further analysis of the human test results reveals that the `Qwen2.5-0.5B` model trained on the weighted-sum data and the same model trained on the synthetic data make errors on largely different examples – only about 30% of their mistakes overlap. This indicates that, although their overall accuracy is similar, the models tend to make complementary rather than identical errors.

To further examine the complementarity between data sources, we trained models on the fully combined dataset and on combinations of the weighted-sum training set with each synthetic group individually. The objective was to identify dataset pairs that complement each other and improve overall model performance. For brevity, we omit detailed results for each combination (see Appendix E); however, we observed that while some combinations led to notable gains (reaching up to 0.74 in accuracy), others caused slight performance drops (down to 0.67). The best-performing configurations are (1) `Weighted-sum & improved Mistake ID & Location`, whose improvement

Training data setup	Qwen2.5-0.5B		Qwen2.5-1.5B		Gemma-2-2B	
	W.-sum test	Human test	W.-sum test	Human test	W.-sum test	Human test
Weighted-sum	0.70	0.68	0.70	0.69	0.74	0.69
Synthetic	0.66	0.69	0.66	0.68	0.70	0.69
Combined	0.73	0.69	0.74	<u>0.70</u>	0.76	0.67
Weighted-sum & Targetedness \uparrow	0.71	0.74	0.71	0.69	0.79	0.68
Weighted-sum & Mistake ID & Location \uparrow	0.72	0.74	0.71	<u>0.70</u>	0.76	<u>0.70</u>
External Baselines (evaluated only on the human test set)						
	Skywork-Reward-V2-Llama-3.1-8B		0.64			
	RewardAnything-8B-v1		0.68			
	Macina et al. (2025)		0.69			

Table 3: Reward model performance (pairwise accuracy) across different training data setups and models. External baselines are evaluated only on the human test set. **Bold** values indicate the best results on the human test set, and underlined values indicate the second-best ones.

over the Qwen2.5-0.5B model trained only on weighted-sum data is statistically significant (McNemar $p = 0.018$; two-sided binomial $p = 0.005$), and (2) Weighted-sum & improved Targetedness, which also achieved a statistically significant gain (McNemar $\chi^2 = 3.96$, $p = 0.047$; two-sided binomial $p = 0.0165$). Both models outperform all baselines.

Manual analysis revealed consistent patterns in model behavior. Our best-performing model more often preferred responses that guided the student through reasoning steps or clarified misconceptions. In contrast, the model trained only on the weighted-sum data occasionally favored responses that restated the problem or revealed the final answer. In several cases, both responses contained similar content but differed in clarity or scaffolding, with the stronger model correctly favoring the more pedagogically effective option. These observations suggest that incorporating a synthetic dataset where responses differ primarily in identifying the location of the mistake helps the model learn to prefer responses that more effectively target and address the student’s error.

We also evaluated our best model, Qwen2.5-0.5B-Instruct trained on Weighted-sum & improved Mistake ID & Location, on the Bridge test set (Wang et al., 2023), comparing it against Qwen2.5-1.5B-Instruct from Macina et al. (2025). Both models achieved an accuracy of 0.83. However, when evaluated on the human preference test set, the Qwen2.5-1.5B-Instruct model from Macina et al. (2025) reached an accuracy of 0.69. Although this model may have been exposed to parts of our human test data, as both rely on MRBench, its much higher performance on Bridge suggests that the Bridge test set presents a simpler evaluation scenario.

6. Conclusions

We proposed a generation pipeline for creating preference datasets aligned with preferences elicited

from annotators in math mistake remediation tutoring contexts. Using the generated data alone, our model achieves an accuracy of 0.69, and when combined with existing data, it reaches 0.74 on the human preference test set. Remarkably, our 0.5B-parameter model outperforms larger existing reward models, demonstrating stronger performance despite its smaller size.

We release all prompts, synthetic data, human preference annotations, code, and trained models to facilitate future research and downstream use. Beyond serving as an automatic evaluator, our reward model can be used to align tutoring LLMs toward more pedagogically effective feedback.

7. Limitations

A key limitation of our current work is the annotation scale: our human preference test set is small (414 pairs) and partly based on single-annotator labels, so some label noise is likely despite substantial agreement. Moreover, human judgments of pedagogical quality are inherently heterogeneous and context-dependent. Our hierarchy of pedagogical aspects should therefore be understood as a pragmatic operationalization of preferences elicited from our annotator pool in the specific setting of math mistake remediation, rather than as a universal model of “human values” in education. Different annotator populations (e.g., teachers with varying pedagogical philosophies, students of different age groups, or cross-cultural cohorts) might prioritize aspects differently.

A second limitation concerns the preliminary ranking: the weighted-sum scheme relies on hypothesized importance weights that can misalign with human judgments in edge cases. Finally, the synthetic pairs were generated via minimal revisions by a single proprietary model (Claude-Sonnet-4), which may introduce model-specific style or bias and may not fully reflect natural student-tutor interactions.

Potential directions for future research, motivated by the current limitations, include: (1) expanding the size and diversity of the human preference dataset with multi-annotator judgments; (2) exploring larger backbone models and alternative reward-modeling objectives; (3) investigating ensemble methods that combine models trained on different data subsets; (4) developing multi-model or human-in-the-loop generation pipelines to reduce stylistic bias and improve reliability; and (5) applying the trained reward model to pedagogical alignment of LLMs via reinforcement learning or preference optimization.

8. Ethical Considerations

The proposed models are trained on synthetic data, which may contain biases in phrasing, tone, or instructional style. Beyond this consideration, we do not anticipate any significant risks associated with this work, as it focuses on improving the evaluation of pedagogical quality of AI tutors rather than directly deploying them in educational settings.

9. Acknowledgments

We are grateful to the Google Academic Research Award (GARA) 2024 for supporting this research.

10. Bibliographical References

- Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.
- Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. 2025. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *Transactions of the Association for Computational Linguistics*, 13:442–460.
- Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.
- Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. *arXiv preprint arXiv:2507.10579*.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. *arXiv preprint arXiv:2507.01352*.
- Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pages 750–756.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chengguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment (2023). *arXiv preprint arXiv:2303.16634*, 12.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.
- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. *arXiv preprint arXiv:2502.18940*.
- Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. [Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors](#). 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)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*.
- Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. AutoTutor meets large language models: A language model tutor with rich

pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.

Jiaming Shen, Ran Xu, Yennie Jun, Zhen Qin, Tianqi Liu, Carl Yang, Yi Liang, Simon Baumgartner, and Michael Bendersky. 2024. Boosting reward model with preference-conditional multi-aspect synthetic data generation. *arXiv preprint arXiv:2407.16008*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. *arXiv preprint arXiv:2310.10648*.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachler. 2021. Are we there yet?-A systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye.

2025. RewardAnything: Generalizable Principle-Following Reward Models. *arXiv preprint arXiv:2506.03637*.

11. Language Resource References

Macina, Jakub and Daheim, Nico and Chowdhury, Sankalan Pal and Sinha, Tanmay and Kapur, Manu and Gurevych, Iryna and Sachan, Mrinmaya. 2023. *Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems*.

Maurya, Kaushal Kumar and Srivatsa, Kv Aditya and Petukhova, Kseniia and Kochmar, Ekaterina. 2025. *Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors*. Association for Computational Linguistics.

Wang, Rose E and Zhang, Qingyang and Robinson, Carly and Loeb, Susanna and Demszky, Dorottya. 2023. *Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes*.

A. Hierarchy of Aspects

Figure 3 shows the hierarchy of pedagogical aspects developed through human discussion following an initial annotation round, along with contrastive examples illustrating the presence or absence of each aspect.

B. Examples of Misalignment Between the Weighted-Sum Ranking and Human Annotators

In Table 4a, annotators preferred the second response, because the first contained the confusing phrase “12 single-yolked eggs”. This confusion was not captured in the MRBench annotation, either under *Providing Guidance* (PG) or *Coherence* (CO) dimensions. Although the second response was imperfect, as it merely repeated the task conditions, it avoided misleading phrasing, which annotators considered critical. Nonetheless, in MRBench it was rated only *To some extent* in multiple dimensions (MI, PG, AC, and CO).

Another case is shown in Table 4b, where annotators judged both responses to be equally poor: the first was factually incorrect (since $1/3 \neq “3 out of every 5 cards”$) and the second was a nonsensical output from Phi3. Yet, the MRBench annotation does not identify the factual error (PG was marked

Hierarchy of valuable aspects	Examples
<p>Factuality + Non-contradiction + No Nonsense</p> <p>The response should be factually correct, should not contradict what the student has said, and should not contain irrelevant information.</p> <p>We can use "both are equally bad" only when this aspect is absent.</p>	<p>Conversation history: [the task about gift cards]</p> <p>✗ Response 1: <i>If Ms. Jones received 5 gift cards worth \$10 each, then 1/3 of the thank you cards contained a gift card, meaning 3 out of every 5 cards had a gift card, and you can use this information to solve for the total number of students in her class. --></i> Factually incorrect</p> <p>✓ Response 2: <i>That's correct! Now, if 1/3 of the thank you cards contained a gift card, how many thank you cards did she receive in total?</i></p>
<p>Mistake Identification + Location</p> <p>The response should identify, either explicitly or implicitly, that there is a mistake in the student's solution.</p> <p>For example, saying "Nice try" would miss this aspect.</p>	<p>Conversation history: Student: yay Tutor: Next, we have to find the value of 6 times 9 + 2. Tutor: What is the value of 6 times 9 + 2? Student: 99</p> <p>✓ Response 1: <i>That's a great start! Remember, we need to follow the order of operations (PEMDAS). Let's multiply 6 times 9 first, then add 2.</i></p> <p>✗ Response 2: <i>Very good try. --> Mistake Identification + Location aspect is missing</i></p>
<p>Scaffolding + Actionability</p> <p>The response should attempt to address the misunderstanding or problem step by step, guiding the student toward understanding by asking questions or providing hints instead of directly giving the solution.</p> <p>However, if the context shows that scaffolding has not worked, providing the solution is preferable.</p>	<p>Conversation history: Student: yay Tutor: Next, we have to find the value of 6 times 9 + 2. Tutor: What is the value of 6 times 9 + 2? Student: 99</p> <p>✗ Response 1: <i>Good try, but remember the order of operations: multiplication should be done before addition. So it should be 6 times 9, then add 2. Let's try again!</i></p> <p>✓ Response 2: <i>Great try! What was your first step? --> Scaffolding</i></p>
<p>Targetedness</p> <p>The response should address the core misconception/misunderstanding of a student.</p>	<p>Conversation history: [the task about eggs]</p> <p>✗ Response 1: <i>Ok, remember there are only 12 eggs to begin with --> doesn't point to any misconception</i></p> <p>✓ Response 2: <i>It looks like you've got a good start, but let's revisit the single yolks calculation - since 5 eggs had double yolks, that means there are actually 12 - 5 = 7 eggs with single yolks, not 12.</i></p>
<p>Not revealing the final answer</p> <p>While it is sometimes necessary and acceptable to share the answer to a substep, the tutor should avoid giving away the final answer.</p>	<p>Conversation history: [the task about average speed]</p> <p>✗ Response 1: <i>You're close, but let's re-examine this - if Jason already drove 30 minutes at 60 miles per hour, that means he covered 30 miles, and he still has 90 miles left, so to cover those 90 miles in 1 hour (60 minutes), not 1 hour 30 minutes, he needs to average a speed of 90/1 = 90 miles per hour. --> Reveals the final answer</i></p> <p>✓ Response 2: <i>You're very close, but remember the question asks for the average speed needed for the "remaining" portion of the drive, not the overall trip.</i></p>
<p>Clarity + Coherence</p> <p>The tutor's response should be free of awkward, confusing or misleading wording. A good tutor response acknowledges the student's input and connects it to the next step.</p>	<p>Conversation history: Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: Ms. Jones got thank you cards from 30% of her class. 1/3 of these contained a gift card for \$10. If she got \$50 in gift cards, how many Students were in her class? ... Tutor: Let's try a different strategy to answer. If Ms Jones got \$50 in gift cards and these were in \$10 increments how many gift cards did she get? Student: She got 5 gift cards since \$50 divided by \$10 is equal to 5.</p> <p>✗ Response 1: <i>That's correct! Now, if 1/3 of the thank you cards contained a gift card, how many thank you cards did she receive in total?</i></p> <p>✓ Response 2: <i>Ok. And if she got 5 thank you cards that had gift cards in them, and these were 1/3 of the total number of thank you card, she got, how many thank you cards did she get total? --> More explicit connection between the student's input and the next step</i></p>

Figure 3: Illustrative contrastive examples for the hierarchy of pedagogical aspects.

as Yes, likely by mistake). Even if it had been identified, the weighting scheme would still have pre-

ferred the first response due to higher scores on other dimensions.

C. Example of Annotator Disagreement

In Table 5, three annotators preferred Response A, while one preferred Response B. The majority favored Response A because it uses a more scaffolded strategy: rather than directly pointing to the remaining time, it prompts the student to reflect on whether maintaining 60 mph for the entire trip would satisfy the total time constraint. This question encourages the student to identify the inconsistency in their reasoning independently. Response B, in contrast, is more direct. It explicitly reminds the student that only one hour remains after the initial 30 minutes and immediately links this to the required speed. While this approach addresses the misconception clearly and efficiently, it provides less opportunity for the student to reconstruct the reasoning themselves. Importantly, the disagreement does not stem from a clear difference in overall quality, but rather from differing pedagogical preferences: one response emphasizes deeper scaffolding and self-correction, whereas the other prioritizes clarity and direct guidance. Ideally, the choice between such strategies should take into account individual student characteristics, such as their level of confidence, persistence, and preference for guided discovery versus explicit instruction.

D. Prompts for Preference Annotation

See Figures 4–6.

E. Reward Model Performance Across Training Data Setups

Table 6 presents pairwise accuracy scores for Qwen2.5-0.5B fine-tuned under different training data configurations.

```

You are an expert evaluator analyzing a tutor's responses in a learning dialogue.

Below is a conversation between a student and a tutor:
{conversation}

The correct (gold) solution to the task is:
{gold_solution}

Your task is to choose which tutor response is better as the next step in the dialogue. Remember: The
tutor's goal is to guide the student toward discovering the correct solution, rather than simply revealing the
answer.

Option A: {response_a}
Option B: {response_b}

Return your evaluation as a valid JSON object in the following format:
{{
  "thoughts": "Briefly explain your reasoning for choosing one response over the other (or if they are tied).",
  "better_response": "A" // or "B", or "Tie"
}}
```

Figure 4: Basic prompt template used for LLM-based preference annotation.

```

Please read the following tutor-student conversations. After each conversation, you will see two possible
tutor responses. Your task is to choose which response seems best. When deciding, put yourself in the
student's shoes—which response would you find more helpful and easier to understand?

A strong tutor response should:
* Correctly identify and point out the student's mistake
* Avoid simply giving away the answer, instead encouraging active participation
* Offer clear and relevant guidance (such as hints, explanations, or examples) to help the student
understand and correct their mistake
* Provide actionable feedback, making it clear what the student should do next
* Maintain coherence, so that the response logically follows from what the student said

Conversation:
{conversation}

The correct (gold) solution to the task is:
{gold_solution}

From the following teacher responses, select the one that you consider to be best in this context.

Option A: {response_a}
Option B: {response_b}

Return your evaluation as a valid JSON object in the following format:
{{
  "thoughts": "Briefly explain your reasoning for choosing one response over the other (or if they are tied).",
  "better_response": "A" // or "B", or "Tie"
}}
```

Figure 5: Prompt template with guidelines used for LLM-based preference annotation.

Dialog Context										
Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: A local farm is famous for having lots of double yolks in their eggs. One carton of 12 eggs had five eggs with double yolks. How many yolks were in the whole carton?										
Student: Since there were five eggs with double yolks, there were $5 \times 2 = 10$ double yolks in the carton. In total, there were $12 \times 1 = 12$ yolks in the carton (assuming all the eggs had single yolks except for the ones with double yolks). So, there were $12 + 10 = 22$ yolks in the whole carton. 22										
Student: there were $5 \times 2 = 10$ double yolks in the carton. Then I assumed that all the other eggs had single yolks so I calculated $12 \times 1 = 12$ single yolks in the carton. I added the two numbers together to get the total number of yolks which was $12 + 10 = 22$										
Gold Solution: The carton had $12 - 5 = 7$ eggs with one yolk. It had 5 eggs with double yolks, which added $5 \times 2 = 10$ yolks. Thus, there were $7 + 10 = 17$ yolks in the whole carton. 17										

Preferred Response	Tutor Response	MI	ML	RA	PG	AC	HM	CO	TT	Score
Preferred (weighted-sum)	I see what you did, but let's remember that the 12 single-yolked eggs already include their yolks, so we shouldn't count them again when we add the 10 double yolks.	Yes	Yes	No	Yes	Yes	Yes	Yes	Neu.	4.53
		0.50	1.00	0.25	1.00	0.50	0.25	1.00	0.03	
Preferred (human)	Ok, remember there are only 12 eggs to begin with	TSE	No	No	TSE	TSE	Yes	TSE	Neu.	2.03
		0.25	0.00	0.25	0.50	0.25	0.25	0.50	0.03	

(a) Confusing phrasing ("12 single-yolked eggs") not penalized sufficiently by the weighted-sum ranking.

Dialog Context										
Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: Ms. Jones got thank you cards from 30% of her class. 1/3 of these contained a gift card for \$10. If she got \$50 in gift cards, how many Students were in her class?										
Student: Let x be the total number of Students in Ms. Jones' class. 30% of the class is $0.3x$. So, she got thank you cards from $0.3x$ Students. 1/3 of these cards contained a \$10 gift card, so the total amount of money she received is $(0.3x)(1/3)(\$10) = \50 . Simplifying the equation, we get: $0.1x = \$50$ Dividing both sides by 0.1, we get: $x = 500$ Therefore, there were 500 Students in Ms. Jones' class. 500										
Tutor: can you tell me how you got your answer?										
Student: Sure I used the equation $0.1x = \$50$ and then divided both sides by 0.1 to get $x = 500$. So I concluded that there were 500 Students in Ms. Jones' class.										
Tutor: Let's try a different strategy to answer. If Ms Jones got \$50 in gift cards and these were in \$10 increments how many gift cards did she get?										
Student: She got 5 gift cards since \$50 divided by \$10 is equal to 5.										
Gold Solution: She got 5 gift cards because $50 / 10 = 5$ She got 15 thank you cards because $5 / (1/3) = 15$ She has 50 students because $15 / .3 = 50$ 50										

Preferred Response	Tutor Response	MI	ML	RA	PG	AC	HM	CO	TT	Score
Preferred (weighted-sum), Tie (human)	If Ms. Jones received 5 gift cards worth \$10 each, then 1/3 of the thank you cards contained a gift card, meaning 3 out of every 5 cards had a gift card, and you can use this information to solve for the total number of students in her class.	Yes	No	No	TSE	Yes	Yes	Yes	Neu.	3.03
		0.50	0.00	0.25	0.50	0.50	0.25	1.00	0.03	
Tie (human)	Great job! To find the number of \$10 gift cards, you simply divide the total amount by the value of each gift card. User: What if the total amount was \$250 instead? Assistant: In that case, you would still divide the total amount by the value of each gift card, which is \$10, resulting in 25 gift cards. [nine more user-assistant turns]	TSE	No	Yes (incorrect)	TSE	No	No	No	Neu.	0.78
		0.25	0.00	0.00	0.50	0.00	0.00	0.00	0.03	

(b) Factual error not identified and nonsensical output treated comparably under weighted-sum.

Table 4: Examples of misalignment between the weighted-sum ranking (Section 3.1) and human preference. For each response, the first row reports the original MR_{Bench} annotation labels, and the second row reports their weighted contributions (annotation score \times dimension weight). Abbreviations: MI – Mistake Identification; ML – Mistake Location; RA – Revealing the Answer; PG – Providing Guidance; AC – Actionability; HM – Human-likeness; CO – Cohesiveness; TT – Tutor Tone; TSE – To some extent; Neu – Neutral.

Dialog Context	
<p>Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: Jason has to drive home which is 120 miles away. If he drives at 60 miles per hour for 30 minutes, what speed does he have to average for the remainder of the drive to get there in exactly 1 hour 30 minutes?</p> <p>Student: In 30 minutes, Jason covers a distance of $60 \times 0.5 = 30$ miles. He has $120 - 30 = 90$ miles left to cover. To cover 90 miles in 1 hour 30 minutes (or 1.5 hours), he needs to average a speed of $90/1.5 = 60$ miles per hour. Since he already drove at 60 miles per hour for 30 minutes, he only needs to maintain the same speed for the remaining 1 hour (or 60 minutes). 60</p> <p>Tutor: Jason covers a distance of $60 \times 0.5 = 30$ miles. He has $120 - 30 = 90$ miles left to cover. To cover 90 miles in 1 hour 30 minutes (or 1.5 hours) he needs to average a speed of $90/1.5 = 60$ miles per hour. Since he already drove at 60 miles per hour for 30 minutes he only needs to maintain the same speed for the remaining 1 hour (or 60 minutes).</p> <p>Tutor: i can see you have basically worked the majority of this out correctly well done. however the question is asking us to calculate the remaining speed after the initial first 30mins. Now knowing this how does this help you?</p> <p>Student: Since he already drove at 60 miles per hour for 30 minutes he only needs to maintain the same speed for the remaining 1 hour (or 60 minutes) to get home in 1 hour 30 minutes. This means that the average speed he needs to maintain for the remainder of the drive is 60 miles per hour.</p>	
Correct Solution	
<p>Jason drives 60 miles per hour * 0.5 hours = 30 miles initially He needs to drive an additional 120 miles total - 30 miles driven = 90 more miles He has 1.5 hours - 0.5 hours = 1 hour to drive the remaining distance He must average 90 miles per hour to drive the remaining 90 miles in 1 hour 90</p>	
Response	Tutor Response
A	I appreciate your effort, but let's think about this again: if Jason drove at 60 mph for the entire trip, would he arrive in 1 hour 30 minutes as the question states?
B	Great job on understanding the distance and time, but remember, the total travel time is 1 hour 30 minutes, so after the initial 30 minutes, Jason has 1 hour to cover the remaining distance – how does that change the required speed?

Table 5: Example of annotator disagreement: one annotator preferred Response B, while three preferred Response A.

Training data setup	W.-sum test	Human test
Weighted-sum	0.70	0.68
Synthetic	0.66	0.69
Combined	0.73	0.69
Weighted-sum & Targetedness ↑	0.71	0.74
Weighted-sum & Targetedness ↓	0.74	0.71
Weighted-sum & Mistake ID & Location ↑	0.72	0.74
Weighted-sum & Global non-preference	0.68	0.70
Weighted-sum & All Aspects Improved vs. Original	0.75	0.69
Weighted-sum & All Aspects Improved vs. Targetedness ↑	0.70	0.69
Weighted-sum & All Aspects Improved vs. Mistake ID & Location ↑	0.71	0.68
Weighted-sum & All Aspects Improved vs. Scaffolding & Actionability ↑	0.71	0.67
Weighted-sum & All Aspects Improved vs. Clarity & Coherence ↑	0.70	0.72
Weighted-sum & Clarity & Coherence ↑	0.71	0.69
Weighted-sum & Clarity & Coherence ↓	0.71	0.73
Weighted-sum & Revealing the Answer ↓	0.67	0.67
Weighted-sum & Scaffolding & Actionability ↑	0.71	0.69
Weighted-sum & Scaffolding & Actionability ↓	0.71	0.71
External Baselines (evaluated only on the human test set)		
Skywork-Reward-V2-Llama-3.1-8B	0.64	
RewardAnything-8B-v1	0.68	
Macina et al. (2025)	0.69	

Table 6: Reward model performance (pairwise accuracy) for *Qwen2.5-0.5B* across different training data setups. External baselines are evaluated only on the human test set. **Bold** values indicate the best results on the human test set.

Please read the following tutor-student conversations. After each conversation, you will see two possible tutor responses. Your task is to choose which response seems best. When deciding, put yourself in the student's shoes—which response would you find more helpful and easier to understand?

Checklist you should follow when making a decision:

1. Is a response factually incorrect, contradicting, or Nonsense? If yes, it should never be preferred. If no, move to the next point.
2. Does response identify, either explicitly or implicitly, that there is a mistake in the student's solution? If no, it shouldn't be preferred. If yes, move to the next point.
3. Does the response attempt to address the misunderstanding or problem step by step? If yes, a more targeted, step-by-step response is preferred. If no, move to the next point.
4. Does one response more directly addresses the core misunderstanding? If yes, this response should be preferred. If no, move to the next point.
5. Does one response ask a question that the student must answer to understand something, while the other response provides this guidance directly? If yes, a response that questions the student is preferred. If no, move to the next point.
6. Does one response reveal the final answer, while the other does not? If yes, response that doesn't reveal the answer is preferred. If no, move to the next point.
7. Do both responses reveal the final answer? If yes, a response that offers clearer and more helpful guidance is preferred. If no, move to the next point.
8. Do two responses address the student's misunderstanding in different but equally effective ways? If yes, they can be considered equally good as long as they provide helpful feedback; otherwise, try to choose the more targeted, coherent, and clear response. If no, move to the next point.
9. Are responses very similar in terms of guidance they give? If yes, try to choose the more coherent and clear response. If you can't, they can be considered equally good

Conversation:
{conversation}

The correct (gold) solution to the task is:
{gold_solution}

From the following teacher responses, select the one that you consider to be best in this context.

Option A: {response_a}
Option B: {response_b}

Return your evaluation as a valid JSON object in the following format:

```

{{
  "thoughts": "Briefly explain your reasoning for choosing one response over the other (or if they are tied).",
  "better_response": "A" // or "B", or "Both are equally bad", or "Both are equally good"
}}
```

Figure 6: Prompt template with a decision checklist used for LLM-based preference annotation.

Please read the following tutor-student conversations. After each conversation, you will see two possible tutor responses. Your task is to choose which response seems best. When deciding, put yourself in the student's shoes—which response would you find more helpful and easier to understand?

Hierarchy of valuable aspects of tutors' responses (with importance decreasing from top to bottom)

1. Factuality + Non-contradiction + No Nonsense: The response should be factually correct, should not contradict what the student has said, and should not contain irrelevant information. We can use "Both are equally bad" only when this aspect is absent. Example:
 Conversation history: [the task about gift cards]
 Option A (bad response): If Ms. Jones received 5 gift cards worth \$10 each, then 1/3 of the thank you cards contained a gift card, meaning 3 out of every 5 cards had a gift card, and you can use this information to solve for the total number of students in her class. --> Factually incorrect
 Option B (better response): That's correct! Now, if 1/3 of the thank you cards contained a gift card, how many thank you cards did she receive in total?
 ...

6. Clarity + Coherence: The tutor's response should be free of awkward, confusing or misleading wording. A good tutor response acknowledges the student's input and connects it to the next step. Example:
 Conversation history: [the task about gift cards]
 Student: She got 5 gift cards since \$50 divided by \$10 is equal to 5.
 Option A: That's correct! Now, if 1/3 of the thank you cards contained a gift card, how many thank you cards did she receive in total?
 Option B (better response): Ok. And if she got 5 thank you cards that had gift cards in them, and these were 1/3 of the total number of thank you card, she got, how many thank you cards did she get total? --> More explicit connection between the student's input and the next step

Conversation:
 {conversation}

The correct (gold) solution to the task is:
 {gold_solution}

From the following teacher responses, select the one that you consider to be best in this context.

Option A: {response_a}
 Option B: {response_b}

Return your evaluation as a valid JSON object in the following format:

```

  {{
    "thoughts": "Briefly explain your reasoning for choosing one response over the other (or if they are tied).",
    "better_response": "A" // or "B", or "Both are equally bad", or "Both are equally good"
  }}
  
```

Figure 7: Prompt template with a decision-making process based on a hierarchy (see Appendix A) that was used for LLM-based preference annotation in Section 3.2.