Improving Generative AI Student Feedback: Direct Preference Optimization with Teachers in the Loop

Juliette Woodrow
Stanford University
jwoodrow@stanford.edu

Sanmi Koyejo
Stanford University
sanmi@cs.stanford.edu

Chris Piech
Stanford University
piech@cs.stanford.edu

ABSTRACT

High-quality feedback requires understanding of a student's work, insights into what concepts would help them improve, and language that matches the preferences of the specific teaching team. While Large Language Models (LLMs) can generate coherent feedback, adapting these responses to align with specific teacher preferences remains an open challenge. We present a method for aligning LLM-generated feedback with teacher preferences using Direct Preference Optimization (DPO). We integrate preference data collection into the grading process. This creates a self-improving pipeline keeping the teacher-in-the-loop to ensure feedback quality and maintain teacher autonomy. To evaluate effectiveness, we conducted a blind controlled study where expert evaluators compared feedback from multiple models on anonymized student submissions. Evaluators consistently preferred feedback from our DPO model over GPT-4o. We deployed the system in two offerings of a large university course, with nearly 300 students and over 10 teaching assistants per term, demonstrating its feasibility in real classroom settings. We share strategies for automated performance monitoring using critic models. We explore methods for examining fairness across protected demographics.

Keywords

Alignment, Direct Preference Optimization, Automated Feedback, Human-in-the-loop, Preference-based fine-tuning, Generative AI

1. INTRODUCTION

Effective feedback is a cornerstone of student learning in any course, yet crafting high-quality feedback is one of education's most complex tasks. Research in education theory defines feedback as a two-stage process: "noticing" errors and misconceptions in student work and "responding" with clear feedback that guides the student on how to refine their thinking [7]. This work focuses on automating the responding stage. Crafting high-quality responses to students requires teachers to draw on multiple forms of expertise simultaneously: deep subject knowledge, pedagogical experience, and course-specific context. For instance, consider a teacher responding to a student's work on event independence in a probability course. When giving feedback, the teacher combines her understanding of probability theory, her experience teaching probability, and course-specific elements like key terminology and teaching style. Large Language Models (LLMs) offer promising capabilities for generating detailed feedback at scale, potentially helping to reduce this burden on teachers [1, 34, 17]. However, a critical challenge remains: while LLMs can generate coherent feedback, it is not clear how to systematically adapt their outputs to specific course contexts and align with individual teaching preferences. Thus, we pose *The Feedback Alignment Challenge*: how can we automatically adapt LLM feedback generation to match specific teacher preferences?

We introduce a method for refining LLM-generated feedback using Direct Preference Optimization (DPO) [31] to better align with specific teacher preferences. During grading, our system presents teaching assistants (TAs) with two AI-generated feedback suggestions for each student response. The TA can choose one, modify it, or write their own. Importantly, this system is only relating to the feedback text the student sees and not the grade the student receives. This "teacher-in-the-loop" design ensures quality feedback while maintaining teacher autonomy. These preferences are used to fine-tune the model between assignments and the AI-generated feedback evolves over time to better match teacher expectations, leading to a "self improving" system. In a blind controlled study, we find that our method produces feedback that is significantly preferred over feedback generated by GPT-4o. We show, through a state of the art critic model for evaluating feedback alignment [32], that our model outperforms GPT-4o in four out of five metrics. Additionally, we present our own critic model for evaluating feedback quality across a broader range of student work, introducing new metrics for evaluating feedback alignment. We deployed this system in two offerings of a large university course, each with nearly 300 students and over 10 TAs to demonstrate its feasibility in real classroom settings.

Main Contributions

1. We present a method for iteratively refining automated feedback suggestions by incorporating preference data collected during the grading process. Our approach leverages Direct Preference Optimization to better align

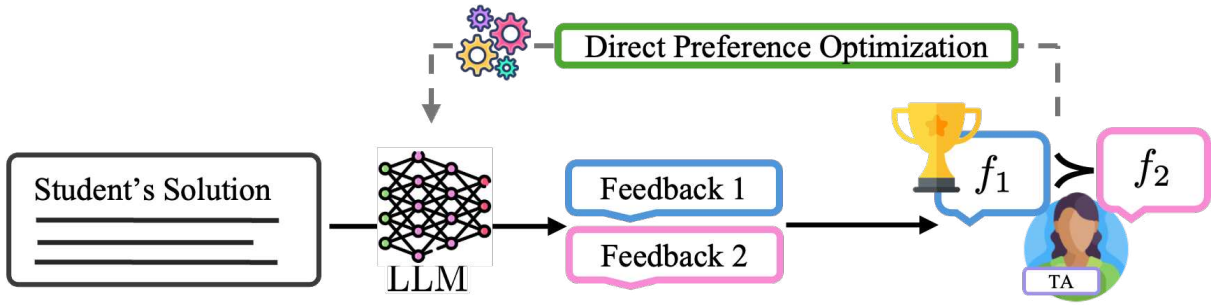

Figure 1: For each student's response, the LLM generates two feedback options. During grading, a teaching assistant (TA) selects the preferred feedback or writes her own. After the assignment is graded, the collected preferences are used to fine-tune the LLM via Direct Preference Optimization (DPO). The updated model is then used to generate feedback on the next homework assignment.

feedback suggestions with teacher expectations.

2. We present results from a blind controlled study demonstrating that our method produces feedback that is significantly preferred over feedback generated by GPT-4o.
3. Using a state-of-the-art critic model for evaluating feedback alignment, we demonstrate that our model outperforms GPT-4o in four out of five metrics.
4. We present a new critic model for assessing model performance and feedback alignment.
5. To demonstrate the feasibility of integrating this method into classroom settings, we deployed our system in a large university course, where TAs used suggestions for feedback generated by the DPO model.
6. We explore how to measure fairness of LLM generated feedback with an initial exploration of gender that extends to other protected demographics.
7. Our code is available at `https://juliettewoodrow.github.io/dpo_feedback/`.

1.1 The Feedback Alignment Challenge

The *Feedback Alignment Challenge* is to generate student feedback that aligns with course-specific preferences. This includes adapting feedback to match terminology, grading rigor, and explanation style used by the course staff. Evaluating feedback alignment requires both teacher input and automated methods. The most reliable method of verifying feedback alignment is for teachers to perform a blind evaluation to judge how well the feedback matches their preferences, but this is time-intensive. To scale, automated methods are also needed. Scarlatos et al. introduce a method for evaluating whether LLM-generated feedback aligns with instructor preferences when student solutions are known to be incorrect [32]. Expanding the evaluation to consider feedback on any student submission (correct or incorrect) requires an additional critic model that assesses feedback on two more dimensions: assertiveness and helpfulness.

1.2 Background and Related Work

Feedback Theory. Feedback is a powerful way to help students improve while learning [15]. It consists of two key stages: Noticing and Responding [7]. Brown's theory provides a generative explanation for how mistakes arise, suggesting that errors are not random but emerge from underlying misconceptions or incomplete mental models [6]. When students recognize an error—either on their own or through feedback—they must not only acknowledge it but also engage in a repair process, actively working to correct their understanding. This process helps address the underlying cognitive issue, reducing the likelihood of similar mistakes in the future. This theory has been utilized in intelligent tutoring systems, emphasizing that feedback should not merely list errors but actively guide students through the repair process [37]. Effective feedback must first notice the student's mistake and then respond in a way that facilitates meaningful revision—simply identifying errors is insufficient; feedback must help students "repair" and "feed forward" into future learning [8].

AI Generated Feedback. Recent advances in large language models (LLMs) have transformed the landscape of automated feedback generation in education [30, 38, 25, 5, 24, 30, 35, 12, 29, 19, 10]. While these systems demonstrate impressive capabilities in producing fluent and coherent feedback [18], they face several pedagogical challenges. Key limitations include lack of specificity, inability to fully replicate nuanced human insights, and limited understanding of educational context [16, 11, 34]. Previous research has explored various approaches to evaluating feedback generated by LLMs, including assessing the quality of feedback itself and investigating whether LLMs can effectively evaluate their own responses [17, 20, 1]. One prominent direction in this area is to use an LLM as a judge, or critic, to evaluate AI-generated feedback [21, 39].

Preference Learning. Collecting preference data has emerged as a promising approach for improving AI systems in educational contexts [26, 32]. Recent work demonstrated the potential of using GPT-4 [27] to generate synthetic preference data for training open source models to provide feedback on student work in math [32]. Our work extends these approaches by collecting preference data from teaching assistants during actual course deployment, providing a more authentic signal for model improvement. To address the limitations of automated systems, researchers have increasingly advocated for human-in-the-loop approaches in educational AI [23]. These systems aim to combine the efficiency of AI-generated feedback with human oversight to ensure accuracy and pedagogical appropriateness. While existing work has established the importance of human oversight [16], our approach goes further by implementing a systematic preference-based learning framework that continuously improves feedback quality through direct interaction with course staff.

2. METHODOLOGY

2.1 Direct Preference Optimization

In preference optimization, a language model improves its responses by learning from human feedback. This is done through comparative feedback, where different responses to the same input are evaluated against each other. Given two possible outputs for the same input, a human annotator selects the preferred one, forming a preference pair (y_w, y_l), where y_w represents the preferred (winning) response, and y_l represents the dispreferred (losing) response. A preference dataset $\mathcal{D}$ consists of tuples (x, y_w, y_l), where x is the input. Traditional preference learning methods, such as Reinforcement Learning from Human Feedback [9], involve a two-step process. First, a reward model is trained on preference data to assign a scalar score to generated responses. Then, the reward model guides reinforcement learning updates, typically using Proximal Policy Optimization [33].

Direct Preference Optimization (DPO) simplifies preference learning by eliminating the reward model entirely [31]. Instead of learning an explicit reward function, DPO directly optimizes the language model to increase the probability of generating preferred responses over dispreferred ones. The loss function minimizes the following:

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

Here, $\pi_\theta(y \mid x)$ represents the probability that the fine-tuned model assigns to response y given an input x, while $\pi_{\text{ref}}(y \mid x)$ is the probability assigned by a reference model, which is typically the original model before fine-tuning. The parameter β is a hyperparameter that controls how much the fine-tuned model deviates from the reference model during optimization. The function σ is the sigmoid function, which maps the log probability ratio to a value between 0 and 1. By optimizing this objective, the model gradually increases the likelihood of generating preferred responses over dispreferred ones, aligning with human preferences without requiring a reward model.

2.2 DPO Feedback Pipeline

As shown in Figure 1, our pipeline consists of three key stages: preference data collection, model training, and inference. During grading, TAs generate preference data. After grading, this data is used to fine-tune the model using Direct Preference Optimization (DPO). The updated model is then deployed to generate feedback suggestions for the next assignment, creating a self-improving system that adapts to instructor preferences over time. During grading, two feedback suggestions, f_1 and f_2, are shown along with each student response. TAs can select one, modify it, or write entirely new feedback. These interactions produce preference pairs (y_w, y_l), where y_w is the preferred (winning) response and y_l is the dispreferred (losing) one. If a TA selects one verbatim, it becomes y_w, while the other is labeled y_l. If the TA edits a suggestion or writes new feedback, their modified or custom response becomes y_w, generating two preference pairs with each original suggestion serving as y_l. These preference pairs are then used to fine-tune the model via DPO. Each training example includes a structured prompt with instructions, the question text, an example TA solution, the student's work, and the corresponding (y_w, y_l) pair. Inference follows a two-step approach, mirroring the feedback process: noticing and response generation. First, the noticing step identifies key observations about the student's work. This can either be a rubric, automated test cases, or another LLM. These observations are then included in a structured prompt alongside the question text, an example TA solution, and the student's work. The fine-tuned model then generates feedback based on this prompt, which is presented to TAs for review. All AI-generated feedback remains subject to human oversight before being delivered to students.

In our implementation, we fine-tuned Meta's Llama 3.1 8B Instruct [2] using Hugging Face's DPOTrainer. Each training period, corresponding to a full assignment's preference data, took approximately 7 hours on 3 A6000 GPUs. The average preference dataset size was 1,408 examples. At inference time, we specifically used GPT-4o [28] to complete the noticing step. During grading, each student response had two AI-generated feedback suggestions-one from our DPO fine-tuned Llama model and one from GPT-4o [28]. This setup served two purposes: ensuring high-quality feedback options (course staff had indicated that GPT-4o was a strong baseline) and enabling (noisy) direct model comparisons within each assignment. We developed the prompts in collaboration with course staff to ensure they were aligned with pedagogy and reflected feedback styles they were comfortable with. We open source the prompts, code setup, and further details on computational costs and infrastructure at `https://juliettewoodrow.github.io/dpo_feedback/`.

3. EXPERIMENTAL SETUP

Our study was conducted within a large university course covering the fundamentals of probability. All feedback generation and model training took place in the context of this specific course, its assignments, and the preferences of its teaching staff. The course included five main assignments, and our training process followed an iterative cycle aligned with these assignments. We define a **Generation** as a full model update cycle, beginning with data collection from grading one assignment and concluding with the deployment of a newly fine-tuned model for the next assignment. The first model used in deployment was trained via Supervised Fine-Tuning (SFT) on past course data, including student submissions and TA-written feedback, but without any pairwise preference data. This model was used to generate feedback for the first assignment, and after grading, preference data was collected and used to train a model using DPO. This marked the transition to Generation 2, where the newly fine-tuned DPO model was used to generate feedback for the second assignment. Each subsequent generation followed the same process: collecting preference data from the previous assignment, fine-tuning the model, and deploying it for the next assignment. All models after Generation 1 were fine-tuned with DPO.

3.1 Controlled Human Evaluation Study

We conducted a controlled study with 10 expert evaluators, all current or former course staff, to assess model performance. Evaluators reviewed a fixed set of student responses across three problems from course assignments, comparing feedback generated by different models under blind, randomized conditions. Each evaluator was shown 20 feedback pairs

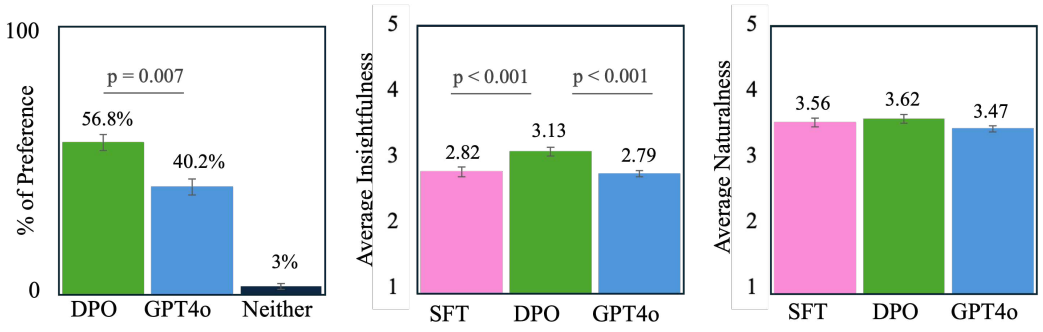

Figure 2: Results from controlled study comparing three feedback models. **Left**: Expert evaluators showed a statistically significant preference for DPO feedback over GPT-4o. **Middle**: DPO feedback was rated significantly more insightful than both GPT-4o and SFT. **Right**: No significant differences between the models in naturalness. Error bars show standard error of the mean.

Feedback Type	COR. (yC)	REV. (yR)	SUG. (yS)	POS. (yP)	Score (score)
DPO	**0.932 ± 0.004**	**0.980 ± 0.002**	0.163 ± 0.006	**0.989 ± 0.002**	**0.724 ± 0.004**
GPT-4	0.862 ± 0.006	0.973 ± 0.003	**0.229 ± 0.007**	0.905 ± 0.005	0.664 ± 0.005
Teacher	0.852 ± 0.008	0.982 ± 0.003	0.321 ± 0.011	0.934 ± 0.006	0.682 ± 0.007

Table 1: Evaluation of feedback quality using the Scarlatos et al. [32] framework across five dimensions: correctness (COR.), not giving away the answer (REV.), improvement suggestions (SUG.), positive tone (POS.), and overall score, with standard errors reported. Each metric ranges from 0 to 1, where higher values indicate better performance in that category. Teacher feedback is included as a reference point, but not highlighted in the comparison as this method focuses on evaluating LLM-generated feedback quality.

per problem, drawn from two model comparisons: DPO (Generation 4) vs. GPT-4o or SFT vs. GPT-4o. Neither the DPO model nor the SFT model had seen these specific student submissions before. For each pair, evaluators received the question text, anonymized student response, a TA-provided solution, and two feedback options. They were not informed of which models were used, how many total models were evaluated, or any other methodological details. They rated each feedback option on Insightfulness (1–5), measuring whether the feedback provided meaningful information and guidance, and Naturalness (1–5), assessing how human-like it sounded. They then selected their preferred response based on alignment with their teaching preferences or, if neither was appropriate, marked "Neither response is appropriate for students." This design ensured a direct comparison of models under identical conditions while incorporating diverse evaluator perspectives. The evaluation covered student work and feedback on three assignment problems: The first involved the Central Limit Theorem; The second focused on statistical estimation with bootstrapping; The third examined combinatorics, where students applied independence assumptions and the complement rule.

3.2 Critic Models

In addition to human evaluation, we use two automated critic models to evaluate feedback alignment. A critic model is an AI-based evaluator that assess text quality based on a rubric. Our first approach adopts the rubric-based evaluation framework developed by Scarlatos et al. [32]. This framework evaluates feedback across the following dimensions: correctness, whether the answer is revealed, presence of suggestions, diagnosing errors, encouraging tone, and overall score. Since our DPO task does not focus on error diagnosis, we exclude this dimension from our results, though we retained it in the prompt to the critic model to maintain consistency with the original framework. This method was designed specifically for incorrect student answers, but our setting encompasses feedback for both correct and incorrect student responses. To handle this broader scope, we developed a critic model tailored to our goals. This method assesses feedback on three dimensions: Assertiveness, measuring how bold or timid the feedback is; Accuracy, ensuring the feedback correctly describes the student's solution and avoids hallucinations; and Helpfulness, determining whether the feedback is useful in addressing key conceptual misunderstandings. We selected these dimensions because they apply across all types of student submissions—correct or incorrect. Furthermore, assertiveness and helpfulness were not included in the earlier framework, yet we view them as critical for high-quality feedback. Assertiveness reflects how confidently and specifically the feedback refers to the student's work (e.g., "Good job" vs. "Nice use of the complement rule"), while helpfulness captures how actionable the feedback is in promoting learning. Both critic models use GPT-4o [28] as the evaluator. We open source the prompts for our critic model at our website: `https://juliettewoodrow.github.io/dpo_feedback/`.

3.3 Deployment in a Large University Course

We deployed our feedback generation system across two course offerings: the first with 330 students and 15 TAs, and the second with 289 students and 13 TAs. Generations 1 and 2 are from the first offering and Generations 3-5 are from the second offering. Each Generation corresponds to one assignment. Each problem in an assignment was graded by a single TA, and each assignment had between 2 and 4 problems. During grading, for each student, TAs chose between feedback from GPT-4o and our model, or wrote their own. The system only related to the feedback text the students would see and not the grade they received. Unlike the controlled study, real-world deployment introduced several confounding factors. Since each problem was graded

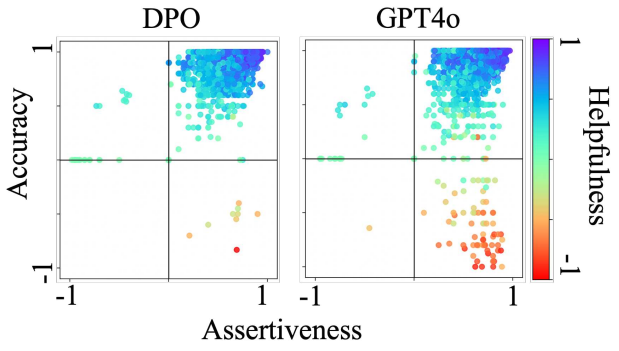

Figure 3: Each point represents a feedback instance evaluated by our critic model. DPO feedback (left) clusters more consistently in high accuracy and helpfulness.

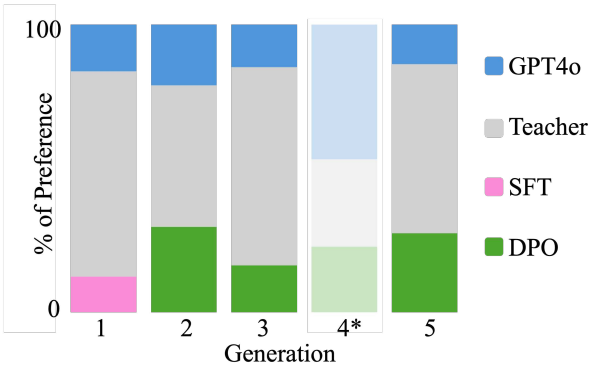

Figure 4: Percent of feedback preferences across five generations of deployment. In each generation, TAs could select feedback from available AI models or write their own (Teacher). Generation 4 (marked with *) had one-off implementation issues.

by a single TA, individual grader preferences could influence feedback selection. Additionally, assignments changed across generations, with no repeated problems, making it difficult to isolate model improvements from variations in problem difficulty or grader subjectivity.

4. RESULTS

4.1 Controlled Human Evaluation Study

The controlled evaluation found that expert evaluators preferred feedback from the DPO model 56.8% of the time, significantly more often than GPT-4o at 40.2% ($p = 0.007$). In 3.0% of cases, evaluators found both models' feedback inappropriate for students, meaning neither response met the minimum quality threshold for course feedback (Figure 2, leftmost plot). The DPO model received an average insightfulness score of 3.13, significantly higher than GPT-4o at 2.79 and the SFT model at 2.82, with p-values of 0.0004 and 0.0005, respectively. Naturalness ratings were similar across all models, with no statistically significant differences.

In the open-ended justifications, when evaluators preferred DPO over GPT-4o they explained it was due to the DPO feedback having greater specificity, problem relevance, encouragement, and actionable suggestions. In these instances, GPT-4o responses were described as generic, occasionally harsh, and less problem-specific. When evaluators preferred GPT-4o over DPO, they cited the conciseness and clarity of the GPT-4o feedback, noting that DPO was sometimes overly verbose or awkwardly phrased. When comparing SFT to GPT-4o, evaluators found SFT feedback to be more problem-specific but noted that the feedback often mirrored the TA-provided solution exactly, rather than presenting alternative approaches. We include more qualitative findings on our website: `https://juliettewoodrow.github.io/dpo_feedback/`.

4.2 Critic Models

In the Scarlatos et al. [32] evaluation framework (Table 1), the DPO-trained model performed strongly across multiple dimensions, outperforming GPT-4o in four of five metrics: correctness, not revealing answers, maintaining a positive tone, and overall score. GPT-4o performed better than the DPO model in providing improvement suggestions. Teacher-written feedback is included as a baseline for consistency with Scarlatos et al. [32] framework, serving as an aspirational standard rather than a benchmark.

Our specialized critic model (Figure 3) provided additional insights by evaluating feedback along three dimensions: assertiveness, accuracy, and helpfulness. The visualization shows that the DPO model's feedback consistently clusters in regions of high accuracy and high helpfulness, with moderate variance in assertiveness. In contrast, GPT-4o's feedback shows notably higher variance in both accuracy and helpfulness dimensions, particularly evidenced by a substantial cluster of feedback instances rated as both inaccurate and unhelpful. Quantitative analysis revealed that GPT-4o feedback was 8.73 times more likely to be classified as unhelpful and 7 times more likely to be inaccurate compared to DPO-generated feedback. Both models were found to be comparable in terms of likelihood of assertiveness.

4.3 Deployment in a Large University Course

The deployment results, presented in Figure 4, show that TAs predominantly preferred writing their own feedback, with teacher-written feedback accounting for 71.2% of responses in Generation 1, 49.2% in Generation 2, 68.9% in Generation 3, and 58.7% in Generation 5. When selecting AI-generated feedback, TAs showed varying preferences across generations. In Generation 1, where only the SFT model was available, TAs preferred GPT-4o over SFT (16.3% vs. 12.5%). After implementing DPO, TAs showed stronger preference for our model, selecting DPO feedback 29.7% of the time compared to 21.1% for GPT-4o in Generation 2, and 27.5% versus 13.8% in Generation 5. Generation 3 showed more balanced selection between DPO and GPT-4o (16.3% vs. 14.8%). Generation 4 data was impacted by two implementation issues not present in other generations: a formatting bug in our code prevented models from accessing equations in student explanations, and preference data was available for only two problems instead of the usual three to four.

5. DISCUSSION

Redundancy in Feedback. In university courses, students usually receive feedback from multiple teaching assistants on a single assignment, leading to diverse perspectives on their

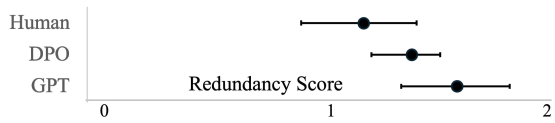

Figure 5: Redundancy scores measured by scaled pairwise similarity of feedback text (lower indicates better performance). Error bars show standard error.

work. We examined whether automated feedback preserves this diversity or generates redundant responses. Using pairwise cosine similarity of feedback embeddings, we measured redundancy for our DPO-tuned model, GPT-4o, and human graders, shown in Figure 5. Feedback similarity was scaled by question text similarity and averaged across students and assignments. While error bars overlap, indicating no statistically significant differences, we observe a trend: GPT-4o exhibits the highest redundancy, human feedback the lowest, and our DPO model falls in between. Lower redundancy suggests greater variation in feedback style, resembling the diversity seen with multiple human graders.

Monitoring Fairness. It is critically important that any method used in an assessment process – even if the model is only providing feedback – should be "fair" with respect to protected demographics, such as gender and ethnicity. How can we monitor our feedback generating system for signs that it is behaving unfairly? Prior work has built fairness monitoring systems based on probabilistic inference techniques, fairness-aware evaluation metrics, and bias mitigation strategies [4, 14, 3]. In our context, monitoring fairness is made especially challenging because we don't explicitly know the demographics of our students. Given this constraint, we approximate the probability of gender (our initial fairness inquiry) from students' first names in line with prior work [13, 36]. We computed a "parity" comparison by calculating the Pearson correlation coefficient between inferred gender probability and feedback ratings. Our analysis found no statistically significant correlation in any case. The full analysis is available at `https://juliettewoodrow.github.io/dpo_feedback/`.

While we consider this a good indication of a fair system, this analysis has limitations. Name-based gender inference can be inaccurate, excludes non-binary identities, and does not account for potential biases in human labels. More critically, parity analysis does not rule out the possibility that the LLM could be treating similar work (from different demographics) differently in a way that is counter-balanced by other factors. A counterfactual fairness evaluation would instead examine whether an individual would receive the same feedback in both the real world and a counterfactual world where only a protected demographic is changed [22]. While true counterfactual fairness requires understanding causal structures, we suggest an approximate approach that controls for student work similarity, using embedding-based similarity or performance-based grouping to compare feedback within similar work categories. Working to develop more rigorous fairness analysis and to integrate mitigation techniques into the DPO pipeline is an area of important future work.

6. LIMITATIONS

This study was conducted in two offerings of one university course, limiting generalizability. We did not evaluate how students engage with AI-generated feedback or its long-term impact on learning outcomes, revision habits, or perceptions of feedback. Future studies should assess whether AI feedback leads to meaningful improvements in student learning. Additionally, the effectiveness of DPO depends on the base model's quality. If the initial LLM (e.g., Llama 3.1 8B Instruct [2]) has biases or weaknesses, the DPO-tuned models may have those as well. Lastly, our study focused on text-based student submissions, making it unclear how well the system applies to handwritten work, diagrams, or other visual formats. Future work should explore extending AI-generated feedback to multimodal contexts where text-based analysis alone may be insufficient.

7. FUTURE WORK

Ensuring fairness in AI-generated feedback requires evaluation across all protected demographics beyond probabilistic estimates. Future work should collect demographic data and incorporate fairness-aware training strategies to mitigate biases early in the optimization process. In addition to fairness, a deeper understanding of feedback optimization is needed. Figure 3 suggests that there might be a local maximum at timid and neutral feedback (around the left side x axis) and a global maximum at assertive, accurate, and helpful feedback (top right corner). This raises key questions about how Direct Preference Optimization (DPO) optimizes the feedback space. Unlike reinforcement learning approaches that use an explicit reward model, DPO directly optimizes preference probabilities, making its update dynamics less interpretable. It is unclear whether DPO consistently pushes the model toward the true optimal feedback distribution or gets stuck in suboptimal preference plateaus. Future work should explore the topology of the preference space in feedback alignment. A better understanding of these dynamics could inform strategies to improve convergence and ensure that DPO consistently finds high-quality feedback policies.

8. CONCLUSION

In this paper, we demonstrate the potential of DPO for aligning LLM-generated feedback with teacher preferences, offering a scalable yet adaptable approach to feedback generation. By directly gathering and incorporating teacher preferences, we create a system that improves over time while respecting the unique instructional styles and pedagogical goals of each course. Rather than replacing teachers, this approach enhances their ability to provide high-quality feedback efficiently, keeping them in control of the process. With demonstrated success in a large university course and strong TA preference for DPO-generated feedback over GPT-4o in a controlled study, this work highlights how LLMs can support and adapt to specific teaching practices.

9. ACKNOWLEDGEMENTS

Juliette Woodrow and Chris Piech sincerely thank the Carina Foundation for making this research possible. Sanmi Koyejo acknowledges support by NSF 2046795, NSF 2205329, IES R305C240046, the MacArthur Foundation and Schmidt Sciences.

10. REFERENCES

[1] D. Agostini and F. Picasso. Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. *Intelligenza Artificiale*, 18(1):121–138, 2024.

[2] AI@Meta. Llama 3 model card. 2024.

[3] C. Ashurst and A. Weller. Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery.

[4] P. Awasthi, A. Beutel, M. Kleindessner, J. Morgenstern, and X. Wang. Evaluating fairness of machine learning models under uncertain and incomplete information, 2021.

[5] R. Banjade, P. Oli, M. Sajib, and V. Rus. Identifying gaps in students' explanations of code using llms. In A. Olney, I. Chounta, Z. Liu, O. Santos, and I. Bittencourt, editors, *Artificial Intelligence in Education, AIED 2024*, volume 14830 of *Lecture Notes in Computer Science*. Springer, Cham, July 2024.

[6] J. S. Brown and K. VanLehn. Repair theory: A generative theory of bugs in procedural skills. *Cognitive science*, 4(4):379–426, 1980.

[7] D. L. Butler and P. H. Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, 65(3):245–281, 1995.

[8] D. Carless. Differing perceptions in the feedback process. *Studies in higher education*, 31(2):219–233, 2006.

[9] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2023.

[10] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gasevic, and G. Chen. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT . In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325, Los Alamitos, CA, USA, July 2023. IEEE Computer Society.

[11] I. Estévez-Ayres, P. Callejo, M. Hombrados-Herrera, et al. Evaluation of llm tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education*, May 2024.

[12] H. Gabbay and A. Cohen. Combining llm-generated and test-based feedback in a mooc for programming. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 177–187, New York, NY, USA, 2024. Association for Computing Machinery.

[13] Gender API. Gender api - determines the gender of a first name, 2024. [cited 2 Jan 2024].

[14] A. Ghosh, P. Kvitca, and C. Wilson. When fair classification meets noisy protected attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 679–690. ACM, Aug. 2023.

[15] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[16] Q. Jia, J. Cui, H. Du, P. Rashid, R. Xi, R. Li, and E. Gehringer. Llm-generated feedback in real classes and beyond: Perspectives from students and instructors. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 862–867, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[17] Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer. On assessing the faithfulness of llm-generated feedback on student assignments. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 491–499, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[18] Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer. On assessing the faithfulness of llm-generated feedback on student assignments. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 491–499, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[19] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, and E. Gehringer. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 5–16, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[20] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[21] C. Koutcheme, N. Dainese, A. Hellas, S. Sarsa, J. Leinonen, S. Ashraf, and P. Denny. Evaluating language models for generating and judging programming feedback. *arXiv preprint arXiv:2407.04873*, 2024.

[22] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness, 2018.

[23] J. K. Matelsky, F. Parodi, T. Liu, R. D. Lange, and K. P. Kording. A large language model-assisted education tool to provide feedback on open-ended responses, 2023.

[24] H. McNichols, W. Feng, J. Lee, A. Scarlatos, D. Smith, S. Woodhead, and A. Lan. Automated distractor and feedback generation for math multiple-choice questions via in-context learning, 2024.

[25] R. Mok, F. Akhtar, L. Clare, C. L. Liu, L. Ross, and M. Campanelli. Using ai large language models for grading in education: A hands-on test for physics. *arXiv preprint arXiv:2411.13685*, 2024.

[26] P. Oli, R. Banjade, A. M. Olney, and V. Rus. Can llms identify gaps and misconceptions in students' code explanations? *arXiv preprint arXiv:2501.10365*, 2024.

[27] OpenAI. Gpt-4 model documentation, 2024. Accessed:

2024-02-19.

[28] OpenAI. Gpt-4o model documentation, 2024. Accessed: 2024-02-19.

[29] T. Phung, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generating high-precision feedback for programming syntax errors using large language models. In M. Feng, T. KÃ¤ser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 370–377, Bengaluru, India, July 2023. International Educational Data Mining Society.

[30] T. Phung, V.-A. Pădurean, A. Singh, C. Brooks, J. Cambronero, S. Gulwani, A. Singla, and G. Soares. Automating human tutor-style programming feedback: Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 12–23, New York, NY, USA, 2024. Association for Computing Machinery.

[31] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[32] A. Scarlatos, D. Smith, S. Woodhead, and A. Lan. Improving the validity of automatically generated feedback via reinforcement learning. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, editors, *Artificial Intelligence in Education*, pages 280–294, Cham, 2024. Springer Nature Switzerland.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

[34] H. Seo, T. Hwang, J. Jung, H. Kang, H. Namgoong, Y. Lee, and S. Jung. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, 15(2), 2025.

[35] H. Tanwar, K. Shrivastva, R. Singh, and D. Kumar. Opinebot: Class feedback reimagined using a conversational llm, 2024.

[36] A. D. VanHelene, I. Khatri, C. B. Hilton, S. Mishra, E. D. Gamsiz Uzun, and J. L. Warner. Inferring gender from first names: Comparing the accuracy of genderize, gender api, and the gender r package on authors of diverse nationality. *medRxiv*, 2024.

[37] K. VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

[38] J. Woodrow, A. Malik, and C. Piech. Ai teaches the art of elegant coding: Timely, fair, and helpful style feedback in a global course. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 1442–1448, 2024.

[39] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.