Improving Suggestions For Student Feedback Using Direct Preference Optimization (DPO)

Juliette Woodrow
Stanford University
jwoodrow@stanford.edu

Chris Piech
Stanford University
piech@cs.stanford.edu

ABSTRACT

Delivering effective feedback to students is both critical and challenging, especially at scale. In this paper, we present our experience deploying a grading system fine-tuned with Direct Preference Optimization (DPO) to generate high-quality suggestions for feedback in a university-level computer science course. These feedback suggestions are shown to teaching assistants (TAs), who review, edit, and approve them before they are presented to students, ensuring the final feedback meets course standards. Our system is designed to iteratively collect preference data from TAs during the grading process and use this data to refine an open-source large language model. This iterative process allows the grading model to improve progressively throughout a single iteration of a course with minimal additional effort from the teaching team.

We evaluated the system across three problem sets using metrics such as grader preferences, NLP metrics (BERT Similarity and ROUGE), and feedback diversity. The results show that the DPO-tuned model was preferred over GPT4o in some cases, however, TAs mostly chose to write their own feedback instead of chosing an LLM generated response. In terms of diversity in feedback to a single student across multiple questions, the DPO-tuned moel demonstrated moderate improvements compared to GPT4o but still fell short of the natural variation seen in human-generated feedback.

Our findings highlight the potential of DPO fine-tuning for improving automated feedback systems, but they also underscore the need for larger datasets and architectural innovations to replicate human alignment and diversity. This work demonstrates a promising approach to integrating iterative preference-based fine-tuning into grading workflows and provides a foundation for future research in automated feedback for education.

Keywords

Template, LaTeX, text tagging

1. INTRODUCTION

Feedback is an essential part of the learning process, enabling students to identify and address mistakes, reinforce their understanding, and ultimately improve over time. Effective feedback provides not only an evaluation of correctness but also clear, actionable guidance to help students move forward. We conceptualize the feedback process as two distinct steps: (a) observing and interpreting what the student did correctly or incorrectly, and (b) transforming those observations into actionable, tailored feedback that can be delivered to the student. While many advancements in educational technology have focused on analyzing student work for correctness, a different challenge lies in transforming those analyses into clear, constructive, and personalized feedback. Recent advancements in large language models (LLMs) and human preference-based fine-tuning provide a promising approach to addressing this need.

In this paper, we present our experience using an open-source large language model (LLM) to address the second step of the feedback process. We developed a system that iteratively improves the model's ability to generate high-quality feedback by incorporating preference data collected from teaching assistants (TAs). During the grading of each problem set, the system gathers preference data from TAs with minimal additional effort on their part. This data is then used to fine-tune the LLM using Direct Preference Optimization (DPO). By repeating this process for each problem set, the system continuously refines the model to better align with human preferences and enhance its feedback generation capabilities over time. Notably, this approach allows the model to improve throughout the duration of a single course, adapting its feedback to better meet the needs of students and TAs as the course progresses. Each problem set contributes a new set of preference data, enabling iterative training and ongoing improvement.

Our results demonstrate both the promise and limitations of this approach. On one problem set, the fine-tuned model generated feedback that TAs preferred over GPT4o, indicating improved alignment with human grading standards. However, on subsequent assignments, TAs often chose to write their own feedback, highlighting areas where the model still falls short. Additionally, we observed that as the model learns to align more closely with TA preferences, it tends to lose diversity in its feedback responses. This lack of variation contrasts with human grading, where different graders naturally bring unique perspectives and styles to their feed-

back.

1.1 Research Questions

This work aims to explore the following research questions:

RQ1: Does fine-tuning a language model using Direct Preference Optimization (DPO) improve the alignment of generated feedback with human grading preferences?

RQ2: Does the feedback maintain diversity across multiple feedback responses, mirroring the variation typically observed in human feedback?

These questions address both the quality and variability of feedback, which are critical for creating an effective grading assistant that supports student learning while preserving the nuances of human grading.

1.2 Main Contributions

This work presents the following main contributions:

1. We present a novel, self-sustaining method for incorporating preference data collected during the grading process to iteratively refine feedback generation. To the best of our knowledge, we are the first to apply Direct Preference Optimization (DPO) to the task of generating high-quality suggestions for feedback.
2. We present our experience and results deploying this system in a real-world university course, demonstrating its feasibility and potential in improving automated feedback generation.
3. We explore future directions to address challenges such as maintaining diversity in feedback and replicating the richness of human-generated responses.

1.3 Background and Related Work

This section reviews relevant research in automated feedback generation, direct preference optimization (DPO), and the challenges of implementing AI-assisted grading systems in educational settings.

Automated Feedback Generation. Automated feedback generation has long been a central challenge in education, aiming to provide students with timely, personalized insights while minimizing the workload for instructors. Recent advancements in large language models (LLMs) have sparked growing interest in leveraging these tools for student feedback [6, 8, 5, 4]. Previous studies have shown that LLM-based feedback can be highly fluent, coherent, and human-like [3]. However, concerns about the faithfulness and accuracy of generated content persist [3, 2].

Direct Preference Optimization (DPO). Direct Preference Optimization is a reinforcement learning method that simplifies the process of aligning language models with human preferences [7]. Instead of explicity learning a scoring function for data, DPO uses the preference choices themselves to steer the model toward producing responses that align with human preferences. This method has shown promise in various natural language processing tasks, including sentiment control, summarization, and dialogue generation [9].

To the best of our knowledge, we are the first to apply DPO to the task of delivering feedback in a course.

Challenges in AI-Assisted Grading. While LLM-based systems show promise for automating aspects of the grading process, several challenges remain:

1. **Hallucination:** LLMs are prone to generating content that is unfaithful to the input document, which can be particularly problematic in educational settings [3].
2. **Mathematical accuracy:** AI grading has been found to be prone to mathematical errors, especially in STEM fields like physics [5].
3. **Maintaining diversity:** There is a unique challenge in maintaining the diversity of feedback that multiple human graders would provide while ensuring accuracy and alignment with grading standards.
4. **Specificity and comprehensiveness:** Feedback generated by AI systems may lack project-specific details or fail to cover all aspects of a student's work comprehensively [2].

Human-in-the-Loop Approaches. To address these challenges, many researchers advocate for human-in-the-loop approaches that combine the efficiency of AI-generated feedback with human oversight [4, 2]. These approaches aim to: ensure the accuracy and appropriateness of feedback maintain human oversight in the grading process, and continuously improve the AI system through human input and preferences.

Our method aligns with this approach, using DPO to fine-tune an LLM based on human grader preferences while maintaining full human control over grading decisions.

2. METHODOLOGY

2.1 Model Training

The model training process consisted of two major components: creating a preference dataset and fine-tuning models using supervised learning and Direct Preference Optimization (DPO). We then repeated this step multiple times throughout the course, specifically each time a problem set was graded.

Preference Dataset Collection. The dataset was collected from a core computer science course at an R1 university with over 300 students. The course grading is overseen by a team of expert teaching assistants (TAs), ensuring high-quality, consistent feedback.

For the preference dataset, feedback was collected during the grading of Problem Sets 3, 4, and 5. The process involved the following steps:

- Initial Response Generation: For each question, the language model (either GPT-4o or a fine-tuned variant) generated multiple feedback responses. In some cases, two distinct responses were generated for each student submission.
- TA Review: TAs reviewed these responses side-by-side. They selected their preferred response, wrote their own feedback if neither response was satisfactory, or edited an existing response to align with grading standards.
- Preference Pair Creation: Based on TA Actions:
  - If one response was selected, it was marked as preferred, and the other response was marked as unpreferred, forming a preference pair.
  - If the TA provided their own feedback, it was treated as the preferred response over both generated responses, creating two preference pairs.

Model Training. The training pipeline included the following steps:

- Supervised Fine-Tuning (SFT): The base Llama 3.1 8b instruct model [1] was first fine-tuned using historical grading data. This was done with regular supervised fine-tuning (and not using the preference dataset). This resulted in what we refer to as the SFT-Llama model.
- Preference-Based Fine-Tuning with DPO: Using the collected preference dataset, Direct Preference Optimization was applied to the SFT-Llama model. This approach fine-tuned the model to increase the likelihood of generating preferred responses as determined by the TAs. DPO iteratively updated the model parameters to better align the model's output with human preferences while penalizing unpreferred responses.

The final model, DPO-Llama, was then evaluated on subsequent problem sets to measure its alignment with TA preferences and overall feedback quality.

Computational Details. All training was conducted using three Nvidia A6000 GPUs, and inference was performed using a single Nvidia A6000 GPU.

2.2 Inference Pipeline

The inference pipeline addresses the two distinct steps in feedback generation: first, noticing what is correct or incorrect in a student's work, and second, crafting this information into clear, actionable feedback. In this work, the noticing step was handled by GPT4o, which received the student's work and a task-specific prompt to produce a concise summary indicating what the student did correctly or incorrectly. This summary, along with the student's work, the question details, and an additional prompt, was then passed to the fine-tuned DPO-Llama model. The DPO-Llama model was solely responsible for transforming this information into feedback for the student.

Although GPT4o was used for the noticing step in this work, the pipeline is designed to be flexible. Any system capable of indicating correctness—such as a human TA filling out a rubric, automated unit tests, or other evaluative tools—could serve as the input for the second step. This flexibility allows our model and technique to be integrated into any type of grading system. Though we have only evaluated it in a computer science setting, it is designed to be adapted to a wide range of course content and grading systems, making it a versatile tool for improving feedback generation across diverse educational contexts.

2.3 Evaluation

The evaluation of our models was conducted across three problem sets: Problem Set 4, Problem Set 5, and Problem Set 6, using data from a core computer science course at an R1 university with over 300 students. All evaluations involved real teaching assistants (TAs) from the course, ensuring that the results reflected authentic grading practices. Importantly, no feedback generated by the models was ever shown to students without explicit oversight and validation by a TA. Every piece of feedback provided to students was either directly approved, edited, or entirely rewritten by the TAs, guaranteeing fairness, accuracy, and alignment with course standards.

During the grading of each problem set, TAs were presented with the student's work and feedback suggestions generated by two models: GPT4o and our fine-tuned Llama model. In Problem Set 4, TAs were presented with feedback from SFT-Llama and in Problem Set 5 and Problem Set 6 TAs were presented with feedback from DPO-Llama. For each feedback instance, the TA reviewed the two options provided and selected one, or wrote their own feedback if neither suggestion was satisfactory.

The performance of the models was assessed using a combination of metrics to evaluate alignment with human preferences and feedback diversity. Preference accuracy, defined as the percentage of times a model's feedback was preferred over an alternative, was the primary evaluation metric. This metric provided a clear indication of how well the models aligned with TA preferences. In addition to preference accuracy, natural language processing (NLP) metrics were used to further analyze the feedback. Specifically, BERT similarity measured semantic similarity, capturing how well the meaning of the model's feedback aligned with that of the TA's feedback. ROUGE similarity, on the other hand, measured n-gram overlap, providing a structural comparison of the feedback texts. Together, these metrics offered a comprehensive view of model performance and alignment with human preferences.

To evaluate the diversity of feedback generated by the models, we introduced a variance metric designed to address the second research question. For this analysis, we calculated the pairwise similarity between all feedback generated for a single student across all questions in a problem set. These similarity scores were averaged for each student on a per-problem-set basis, resulting in a metric that captured the degree of uniformity or variation within the feedback provided to a single student. To account for the inherent similarity of the question texts themselves, we subtracted the

average similarity of the question texts from each student's feedback similarity score. Finally, these adjusted scores were averaged across all students within a problem set. This approach allowed us to isolate the diversity of the feedback itself, independent of difference in question text, and to evaluate whether the models could generate sufficiently varied feedback across different questions for the same student.

3. RESULTS

We share the results of our experience deploying the open source model for providing suggestions for feedback to TAs in a real course.

3.1 Grader Preferences

Figure 1 summarizes grader preferences across PSET4, PSET5, and PSET6. In PSET5, when TAs selected between model-generated feedback and GPT4o feedback without writing their own, our fine-tuned DPO-Llama model was preferred over GPT4o, indicating that the DPO-tuned model was more aligned with TA preferences in this setting. However, for PSET4 and PSET6, the TAs consistently chose to write their own feedback instead of selecting model-generated responses. This makes it challenging to draw strong conclusions, as the preferences are highly dependent on the individual TA who graded the problem. Since only one TA graded each problem, the results likely reflect grading style variations among the TAs themselves. These limitations emphasize the need to collect more data to better understand the trends and ensure consistency. Despite this, the results from PSET5 suggest that, at least in some cases, DPO-Llama can perform well in aligning with TA preferences.

3.2 Comparison to Human-Written Feedback

Table 1 compares model-generated feedback to human-written feedback using NLP metrics such as BERT similarity and ROUGE-1. These metrics help quantify how closely the model-generated feedback resembled what a human might write. While GPT4o generally outperformed DPO-Llama in both metrics across most problems, DPO-Llama demonstrated improvements in a few cases, such as BERT similarity for PSET5 Problem 3 and PSET6 Problem 2. However, these instances of better performance were not widespread, suggesting that the amount of fine-tuning data available—limited to only two problem sets—may not have been sufficient for DPO-Llama to consistently outperform GPT4o. These results highlight that while DPO fine-tuning shows promise, achieving parity or superiority with GPT4o may require more data and further iterations.

3.3 Variation in Feedback Across Questions

Figure 2 examines how feedback varied across questions for a single student within a problem set, normalized by the similarity of the question texts. Humans naturally varied their feedback more, with the lowest normalized similarity scores, likely because human graders approach each question differently and, in many cases, different graders evaluate each question. GPT4o, by contrast, consistently produced highly similar feedback across questions, indicating less variation. DPO-Llama demonstrated reduced similarity compared to GPT4o on PSET4 and PSET6, suggesting more varied feedback, but its similarity scores on PSET5 were closer to those of GPT4o. Neither model, however, approached the level of diversity seen in human feedback, emphasizing the challenge of replicating human-like richness in automated feedback systems.

Table 1: Comparison of BERT Similarity and ROUGE-1 Scores for GPT and LLaMA Outputs Across Problem Sets

PSet	Problem	Model	BERT Similarity	ROUGE-1
Pset 4	Problem 1	**GPT4o**	**0.734**	**0.593**
		LLaMA	0.536	0.352
	Problem 2	GPT4o	0.109	0.027
		LLaMA	**0.128**	**0.030**
	Problem 3	**GPT4o**	**0.525**	**0.238**
		LLaMA	0.512	0.224
Pset 5	Problem 1	**GPT4o**	**0.525**	**0.238**
		LLaMA	0.512	0.224
	Problem 2	**GPT4o**	**0.397**	**0.295**
		LLaMA	0.382	0.278
	Problem 3	GPT4o	0.444	0.276
		LLaMA	**0.459**	**0.277**
Pset 6	Problem 1	**GPT4o**	**0.231**	**0.206**
		LLaMA	0.183	0.078
	Problem 2	GPT4o	0.344	0.177
		LLaMA	**0.351**	**0.189**

3.4 Summary and Future Directions

This study reflects our experience deploying a DPO-tuned model for feedback in a small-scale setting. While the initial results show promise, they also highlight the need for further investigation and larger-scale applications. In PSET5, DPO-Llama demonstrated its potential by being preferred over GPT4o when TAs did not write their own feedback, and in certain cases, it achieved competitive NLP metrics. However, its performance was not consistent across all problem sets and problems, suggesting that the amount of preference data collected was insufficient to fully optimize the model.

4. DISCUSSION

The results of this study provide preliminary answers to our two primary research questions: whether fine-tuning a language model using Direct Preference Optimization (DPO) improves alignment with human grading preferences, and whether the feedback maintains diversity across multiple responses, mirroring human variation.

RQ1: Does fine-tuning a language model using Direct Preference Optimization (DPO) improve the alignment of generated feedback with human grading preferences?

The results indicate that DPO fine-tuning has the potential to improve alignment with human preferences, but the evidence is mixed due to the limited scope of our experience. On PSET5, the DPO-Llama model was preferred over GPT4o when TAs selected between model-generated feedback without writing their own, suggesting that DPO fine-tuning can enhance the model's ability to align with TA grading preferences. However, for PSET4 and PSET6, TAs opted to write their own feedback instead of selecting either model's responses, highlighting the dependency on the individual TA's grading style and preferences. Since only one TA graded each problem, the results are heavily influenced by individual tendencies, making it challenging to generalize

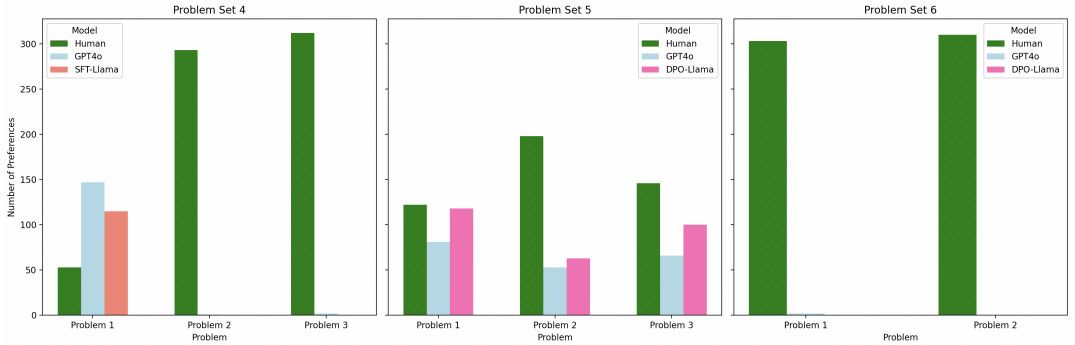

Figure 1: Human grader preference for feedback on multiple questions on each problem set. Humans mostly chose to write their own feedback instead of picking a model generated response.

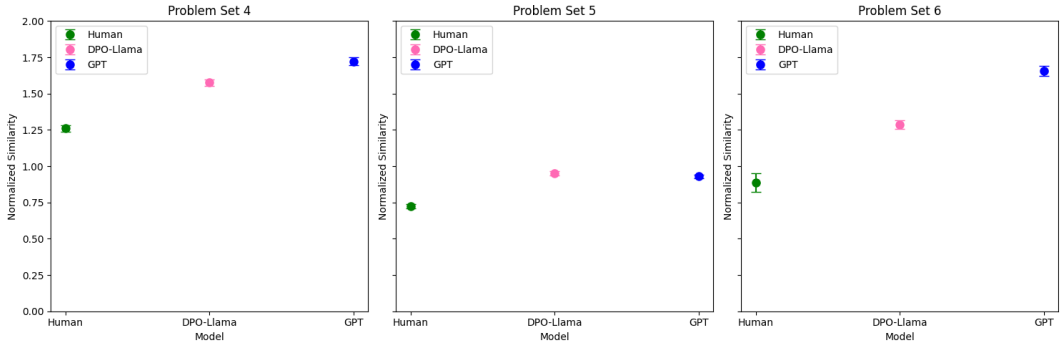

Figure 2: Similarity in feedback generated for each student across three questions within a problem set, averaged over all students. Error bars represent the standard error of the mean. Higher similarity indicates feedback for each question was more uniform, while lower similarity reflects more diverse feedback for each student. Lower similarity is preferred, as it suggests better and more tailored feedback for students.

the findings. Nevertheless, the preference for DPO-Llama on PSET5 is a promising sign that DPO fine-tuning could be effective, especially when sufficient preference data is available. Future work will need to focus on scaling this approach across more problems and courses to better understand the relationship between DPO fine-tuning and alignment with human preferences.

RQ2: Does the feedback maintain diversity across multiple feedback responses, mirroring the variation typically observed in human feedback?

Our analysis of feedback diversity, measured by normalized similarity scores across questions for the same student, shows that human feedback exhibits the greatest variation, as expected. Human graders naturally vary their responses across questions, often due to differences in individual grading styles or the involvement of multiple graders. GPT4o, in contrast, consistently produced highly similar feedback, indicating limited variation. The DPO-Llama model showed less similarity (and thus more variation) compared to GPT4o on PSET4 and PSET6, but this improvement was not observed on PSET5. This suggests that while DPO fine-tuning can introduce greater diversity into model-generated feedback, its performance is inconsistent and remains far from human-level variation. The relatively low diversity across all models highlights a key limitation in the current approach and points to the need for further architectural and methodological adjustments to better capture the richness of human feedback.

In summary, our findings suggest that DPO fine-tuning can improve alignment with human grading preferences under certain conditions and shows potential to enhance feedback diversity, albeit inconsistently. However, the results underscore the need for larger datasets and additional iterations of fine-tuning to fully realize these improvements. These insights lay the groundwork for future research aimed at refining automated feedback systems to better replicate the alignment and variation of human feedback.

5. FUTURE WORK

This work represents a first step in deploying DPO-tuned models for feedback generation in an educational setting, and several directions for future research have emerged. One immediate priority is to continue running this process and collecting additional data across more problem sets and courses. With more data, we aim to investigate whether there is a lower bound on the number of iterations (or size of preference dataset) required for the model to reach a meaningful threshold, such as achieving performance competitive with a closed-source LLM like GPT4o. This would provide valu-

able insights into the scalability and practicality of using DPO in real-world educational contexts.

Another key direction involves addressing the observed lack of variance in model-generated feedback across questions for the same student. Developing architectures or training methodologies that encourage diversity in responses is essential to better replicate human feedback patterns. One promising approach is to introduce a regularization term during training that explicitly penalizes overly similar responses within the same problem set for a single student.

Additionally, exploring more sophisticated frameworks, such as actor-critic models, could further enhance the feedback generation process. Actor-critic models could be used to better evaluate and refine the quality of the feedback, potentially balancing alignment with human preferences and diversity in responses. This could allow for more nuanced improvements in both the content and variation of the feedback provided by the model.

In summary, future work should focus on expanding the preference dataset, refining the model to better capture human-like variance, and experimenting with advanced training techniques. These efforts will be instrumental in further advancing the field of automated feedback generation and moving toward models that are not only aligned with human preferences but also capable of delivering varied and pedagogically effective feedback at scale.

6. LIMITATIONS

This study has several limitations that should be acknowledged. First, the number of graders involved in the evaluation was small (only one grader per quesion on the problem set). With only one grader, we were not able to calculate an inter-rater reliability score. As a result, the findings are likely influenced by individual grader preferences and grading styles, limiting the generalizability of our results. Additionally, the project was initiated in the middle of the course, which restricted our ability to collect long-term metrics and observe trends over an extended period. This constraint also limited the amount of preference data available for fine-tuning the model.

We only compared the performance of our fine-tuned models to a single language model, GPT4o, which may not represent the full spectrum of high-performing models. Furthermore, we only experimented with fine-tuning one open-source model, Meta's Llama 3.1 8B instruct model, which may not be the optimal choice for this task. Exploring other models or fine-tuning techniques could yield different results.

These limitations emphasize the need for future studies with more graders, improved experimental design to assess inter-rater reliability, and broader comparisons across multiple models and datasets to gain a more comprehensive understanding of the effectiveness of DPO-tuned models in educational settings.

7. CONCLUSION

This paper explored the deployment of a DPO-trained grading system designed to iteratively improve feedback alignment with human preferences through seamless integration into existing grading workflows. While the results demonstrate potential for enhancing automated feedback, our findings highlight the need for larger datasets and targeted architectural improvements to achieve consistent alignment and diversity comparable to human feedback. These insights provide a promising starting point for advancing automated feedback systems in education.

8. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] W. Xiao, Z. Wang, L. Gan, S. Zhao, W. He, L. A. Tuan, L. Chen, H. Jiang, Z. Zhao, and F. Wu. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications, 2024.