FEAT: A Preference Feedback Dataset through a Cost-Effective Auto-Generation and Labeling Framework for English AI Tutoring

Hyein Seo[1], Taewook Hwang[1], Yohan Lee[2], Sangkeun Jung[13*]
[1]Chungnam National University
[2]Electronics and Telecommunications Research Institute
[3]EurekaAI
{hyenee97,taewook5295}@gmail.com, carep@etri.re.kr, hugmanskj@gmail.com

Abstract

In English education tutoring, teacher feedback is essential for guiding students. Recently, AI-based tutoring systems have emerged to assist teachers; however, these systems require high-quality and large-scale teacher feedback data, which is both time-consuming and costly to generate manually. In this study, we propose FEAT, a cost-effective framework for generating teacher feedback, and have constructed three complementary datasets[1]: (1) DIRECT-Manual (DM), where both humans and large language models (LLMs) collaboratively generate high-quality teacher feedback, albeit at a higher cost; (2) DIRECT-Generated (DG), an LLM-only generated, cost-effective dataset with lower quality;, and (3) DIRECT-Augmented (DA), primarily based on DG with a small portion of DM added to enhance quality while maintaining cost-efficiency. Experimental results showed that incorporating a small portion of DM (5–10%) into DG leads to superior performance compared to using 100% DM alone.

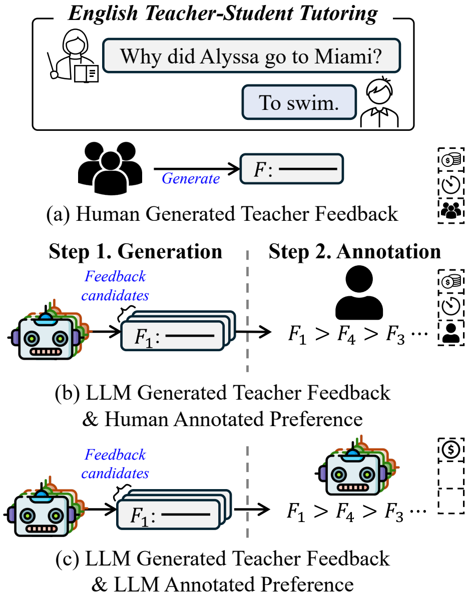


Figure 1: Teacher feedback generation and annotation process in an English tutoring system.

1 Introduction

In English education tutoring, providing appropriate teacher feedback plays a crucial role in guiding students and improving their educational outcomes (Ma et al., 2014; Fossati, 2008). Given its importance, various studies have explored automated teacher feedback generation (Meyer et al., 2024; Scarlatos et al., 2024b; Liu et al., 2020).

Figure 1 illustrates methods for generating and annotating teacher feedback for tutoring systems. As shown in (a), human-generated feedback with ranking provides high quality, but its time-consuming and costly nature makes it difficult to scale up (Chang et al., 2023).

To address this challenge, we propose **FEAT**, a cost-effective framework using large language models (LLMs) to automatically generate a large-scale teacher feedback preference dataset for tutoring AI. This enables reward- or rank-based learning, making it suitable for building human-friendly tutoring models (Ouyang et al., 2022). FEAT generates teacher feedback based on student responses, using the dialogue history between teacher and student and context as input. Moreover, we apply feedback criteria defined by Seo et al. (2025) to ensure educationally appropriate feedback.

Using FEAT, we constructed three datasets: (1) *DIRECT-Manual* (DM), which contains human and LLM-generated feedback with human-annotated rankings (high quality and high cost), (2) *DIRECT-Generated* (DG), an entirely LLM-generated and annotated preference dataset (medium quality and low cost), and (3) *DIRECT-Augmented* (DA), a hybrid dataset built on DG with a minor addition of DM (high quality and low cost).

*Corresponding author
[1]Our dataset is publicly available at https://github.com/hyenee/FEAT

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 575–589
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

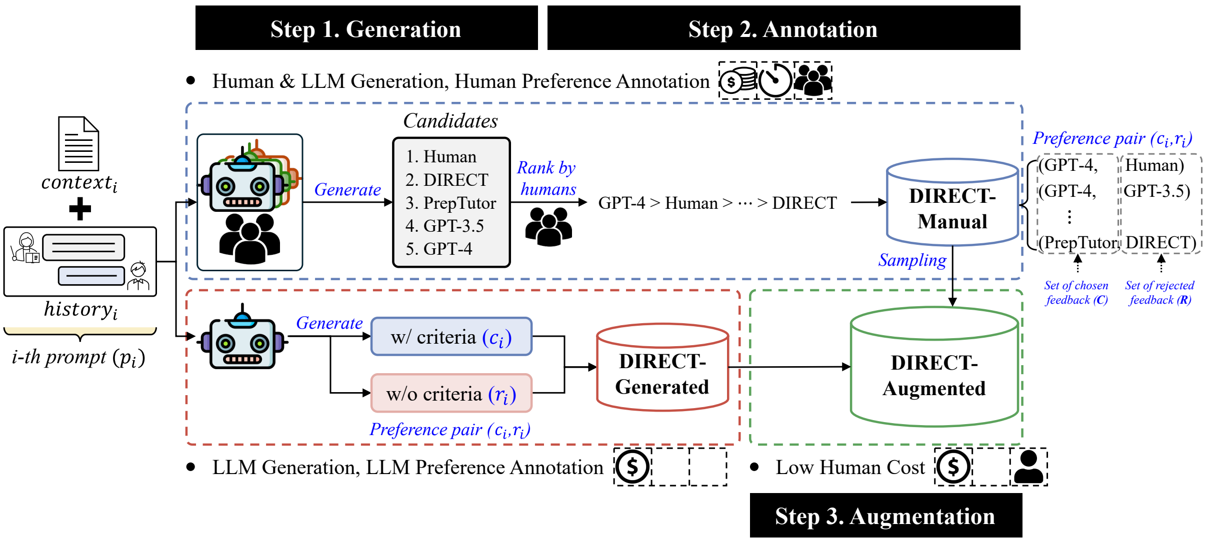


Figure 2: The architecture of the FEAT framework, illustrating the construction process of the DIRECT-Manual, DIRECT-Generated, and DIRECT-Augmented datasets. p_i, c_i, and r_i denote the i-th prompt, chosen, and rejected responses, respectively.

Our experiments showed that incorporating a small portion of DM (5–10%) into DG leads to superior performance compared to using DM alone.

Our main contributions are as follows:

- We proposed FEAT, a cost-effective framework for automated teacher feedback generation and annotation in English tutoring.
- We constructed three preference datasets: DM, DG, and DA, enabling reward- or rank-based learning.
- We confirmed that incorporating a small amount of DM into DG (DA) yielded better performance than DM alone.

2 FEAT: *F*eedback Dataset Generation Framework for *E*nglish *A*I *T*utoring

Figure 2 illustrates the construction process of our FEAT framework. FEAT applies five criteria from Seo et al. (2025)-*Correct, Revealing, Guidance, Diagnostic,* and *Encouragement*—ensuring educationally effective feedback.

2.1 DIRECT-Manual: Rank-based Preference Dataset

DM is an extended version of the DIRECT (Huang et al., 2023) dataset, simulating intelligent tutoring between teachers and students. While it ensures high quality, it relies heavily on human effort, making it time-consuming and costly.

Step 1: Feedback Generation. We collected teacher feedback data for scenarios with incorrect student answers in teacher-student dialogues from diverse sources (Human, DIRECT, PrepTutor, GPT-3.5, and GPT-4; see Appendix B), with DM previously used as private data in Liermann et al. (2024). An example from the DM is shown in Figure 3.

Step 2: Feedback Ranking via Human Annotation. Human annotators ranked the five feedback candidates using two criteria: *Correct* (specific and accurate information) and *Revealing* (avoiding direct answers). Feedback meeting both criteria received the highest rank, with Correct prioritized when only one criterion was met.

Step 3: Preference Data Construction. From ranked five feedback candidates, we created pairwise combinations; in each pair, the higher-ranked feedback is labeled ***chosen*** and the lower-ranked as ***rejected***.

2.2 DIRECT-Generated: Criteria-based Preference Dataset

DG uses LLM to automatically generate and annotate teacher feedback based on specific criteria, producing reasonably good data at a lower cost.

Step 1: Feedback Generation. Using dialogue history and context, LLM generates teacher feedback based on five criteria. We created tutoring scenarios by converting reading comprehension tasks from MCTest (Richardson et al., 2013) to generate large-scale feedback data. A sample from the MCTest is illustrated in Figure 4.

Step 2: Preference Data Construction. We generated two types of feedback: *w/ criteria* (applying

Story
Bill is ten years old.
One day his friend Tom says to him, "I am going to have a birthday party on Sunday evening. Bill, can you come to my party? " "I'll be glad to," answers Bill.
...

History
Teacher : All right. Do you know how old Bill is?
Student : He is eight years old.

Reference
Ten. (Bill is ten years old.)

Feedback Candidates
- Human: Pretty close, but he is a bit older than that.
- DIRECT: no, he is one year older than that.
- PrepTutor: can you think more? honestly, bill is now 10 years old and he still has lots of growth. how old is he?
- GPT-3.5: Hmm, that's close. But is there any information in the passage that tells us Bill's age?
- GPT-4: Take a closer look at the beginning of the passage. It mentions Bill's age there.

Ranking
GPT-4>Human>GPT-3.5>PrepTutor>Tutor

Figure 3: Sample from the DIRECT-Manual.

Story
Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house.
...

Question
Why did Alyssa go to Miami?

Answer Options
A) Swim
B) Travel
C) Visit friends
D) Laing out

Student Correct Response
The answer is visit friends.

Student Incorrect Response
The answer is swim.

Figure 4: Sample from the MCTest.

	Train	Test
DIRECT-Manual	5,025	475
DIRECT-Generated	3,996	444

Table 1: Dataset statistics.

five criteria) and *w/o criteria* (without criteria). We labeled *w/ criteria* feedback as ***chosen*** and *w/o criteria* as ***rejected***, assuming criteria-based feedback is of higher quality. Data statistics are shown in Table 1, with full details in Appendix C.

3 Teacher Feedback Ranking

To validate the preference annotations in DM, DG, and DA, we trained pairwise-based ranking models.

3.1 Ranking Models

We employed five approaches to train ranking models. Each model takes *(prompt, chosen, rejected)* as input.
Binary Classifier formulates preference learning as a binary classification task, labeling *(chosen, rejected)* pairs as 1 and *(rejected, chosen)* pairs as 0. The input sequence is depicted in Figure 11.
Reward Model (Ouyang et al., 2022) computes scalar preference scores for feedback pairs, training to assign higher scores to chosen feedback.
Direct Preference Optimization (DPO) (Rafailov et al., 2023) optimizes language model probabilities to prefer chosen feedback, using log probability differences between chosen and rejected pairs.
RankNet (Burges et al., 2005) learns score differences between feedback pairs using Binary Cross-Entropy loss. Reward Model, DPO, and RankNet share the same prompt, shown in Figure 12.
Ensemble aggregates predictions from the above four approaches through majority voting.

3.2 Scenarios for Training

We evaluated our ranking models on DM with three training configurations. The arrow (→) indicates training (left) and evaluation (right).

- **DM→DM**: Training with DM using manual annotation, serving as a performance upper bound for comparison with the other two scenarios (DG→DM and DA→DM).
- **DG→DM**: Training with DG using automatic annotation.
- **DA→DM**: Hybrid training using DG combined with a subset of DM for mixed annotation.

During training, we enhanced data diversity by including feedback from different contexts beyond the standard *(chosen, rejected)* pairs. This approach enabled the model to learn feedback comparisons across various contexts.

For evaluation, we tested the model on all possible pairs in DM. The model's pairwise predictions

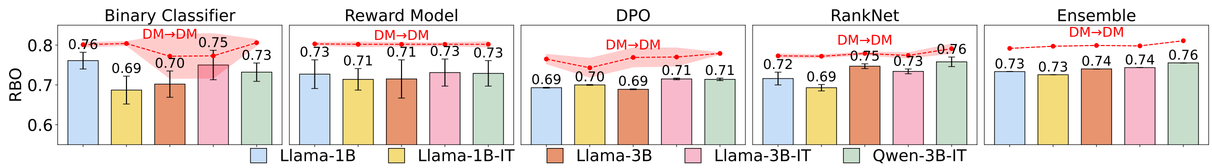
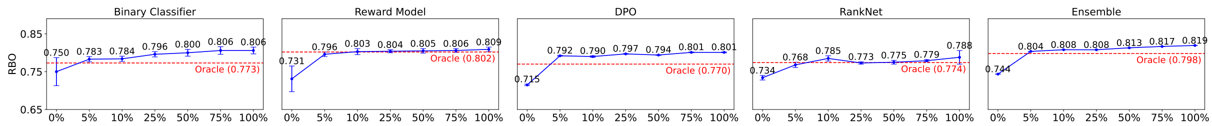
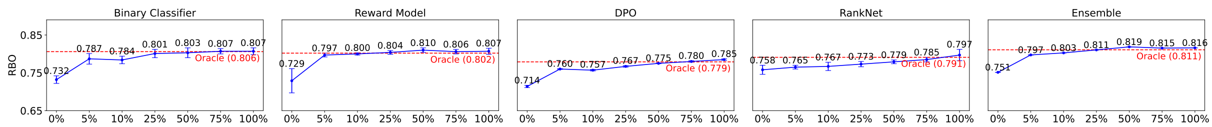


Figure 5: Ranking model performance across different approaches (with 5-seed standard deviation). Lines indicate DM→DM performance, while bars show DG→DM performance.



(a) Llama-3B-IT



(b) Qwen-3B-IT

Figure 6: Llama-3B-IT and Qwen-3B-IT performance in the DA→DM scenario. (See Appendix E for other models.)

were aggregated to create overall rankings, with accuracy scored as 1 for *chosen* > *rejected* and 0 for *chosen* < *rejected*.

4 Experiments

We designed experiments and analyzed results to address the following research questions:

- How does ranking model performance with DG compare to human-curated DM (Section 4.2)?
- How does the ratio of DM in DA affect ranking model performance (Section 4.3)?
- How does the number of criteria in DG affect ranking models performance? (Section 4.4)?

4.1 Experimental Setup

Models We trained ranking models using five open-source models: Llama-1B (Dubey et al., 2024), Llama-1B-IT, Llama-3B, Llama-3B-IT, and Qwen-3B-IT (Bai et al., 2023). Model details and hyper-parameters are provided in Appendix D.

Evaluation Metrics Rank-biased overlap (RBO) is a metric used to measure the similarity between two ranked lists. It ranges from 0 to 1, with values closer to 1 indicating higher similarity between the lists.

4.2 Comparison of Ranking Model Performance

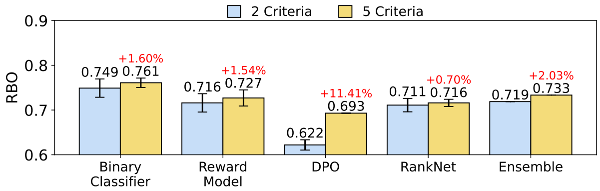
As shown in Figure 5, DM→DM performed consistently (0.77-0.80) across all sizes and approaches. While DG→DM showed lower but competitive results: the binary classifier reached 0.76 (Llama-1B), reward model 0.73 (Llama-3B-IT), DPO 0.71 (Llama-3B-IT, Qwen-3B-IT), RankNet 0.76 (Qwen-3B-IT), and Ensemble 0.76 (Qwen-3B-IT). Notably, the Ensemble maintained stable performance across different architectures, mitigating the variability seen in individual approaches.

These results indicated that teacher feedback generated by LLMs can produce rankings highly comparable to human annotator rankings, with a particularly strong trend observed in larger models.

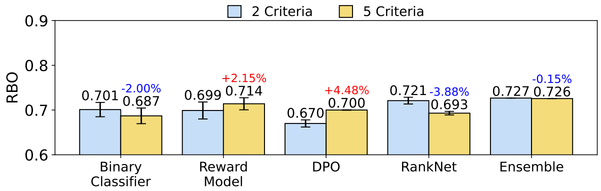
Case Study As a result of analyzing the RBO scores between the ground-truth and predicted rankings, Figure 14 demonstrates that the two rankings are nearly identical, except for the swapped positions of DIRECT and PrepTutor, resulting in an RBO score of 0.8833. In contrast, Figure 15 exhibits significantly lower agreement between the ground-truth and predicted rankings, with an RBO score of 0.4166. In DM, which is limited to five feedback candidates, the RBO score maintains a baseline similarity of at least 0.4 even when the rankings are completely different, due to the limited number of possible permutations. Additional experimental results are provided in Appendix E.

4.3 Performance Analysis by DM Ratio in the DA→DM Scenario

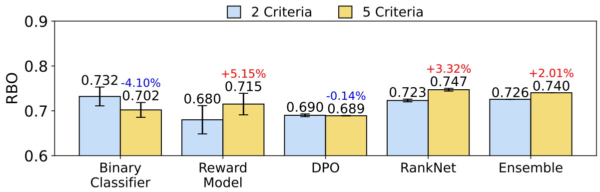
We analyzed how varying the proportion of DM (5-100%) in the DA→DM scenario affects model per-



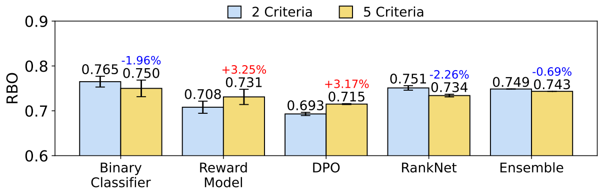
(a) Llama-1B



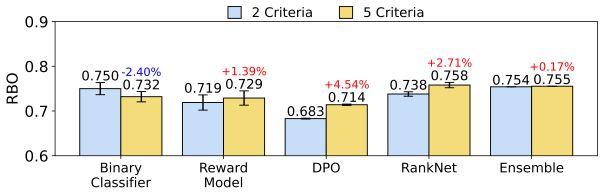
(b) Llama-1B-IT



(c) Llama-3B



(d) Llama-3B-IT



(e) Qwen-3B-IT

Figure 7: Overall performance across varying numbers of feedback criteria.

formance. Figure 6 presents the results for Llama-3B-IT and Qwen-3B-IT, the models that achieved the strongest performance under most approaches.

For Llama-3B-IT, the binary classifier, DPO, and Ensemble outperformed the DM→DM even with only 5% of human-annotated DM and DA data. Similarly, Reward Model and RankNet exceeded DM→DM performance within the 5–10% annotation range. In contrast, Qwen-3B-IT surpassed DM→DM primarily within the 50–75% or 75–100% annotation ranges. Although Qwen-3B-IT is not as efficient as Llama-3B-IT, the results suggest that high performance can be achieved with minimal human annotation costs. The overall performance of Llama-1B, Llama-1B-IT, and Llama-3B models is illustrated in Figure 13 (see Appendix E).

4.4 Performance Analysis by Number of Feedback Criteria

To investigate the impact of feedback criteria, we compared two training configurations: using all five criteria versus using only two essential criteria (Correct and Revealing). For the experiments, we generated an additional version of DG that includes only two feedback criteria and trained a ranking model in the DG→DM scenario.

Figure 7 shows that increasing the number of feedback criteria from two to five consistently improves Llama-1B across all approaches, with DPO exhibiting the largest gain (+11.41 %). For Qwen-3B-IT, every approach except the binary classifier benefits from the richer feedback, and the remaining models display improvements in selected approaches. These results suggest that incorporating richer feedback information enhances model generalization.

5 Conclusion

In this study, we proposed the **F**eedback Dataset Generation Framework for **E**nglish **AI** **T**utoring (FEAT), which utilizes LLMs to generate teacher feedback and build preference datasets for English tutoring. We evaluated ranking models on three datasets—DIRECT-Manual (DM), DIRECT-Generated (DG), and DIRECT-Augmented (DA)—constructed via FEAT.

Results showed that models based on DG performed competitively with DM-based models. Moreover, supplementing with only 5–10% human-annotated DM led to superior performance than using the full DM dataset. These findings demonstrate that high performance can be achieved with minimal human effort with our FEAT framework. In future research, we will extend our framework to broader educational scenarios.

Limitations

In this study, we explored the feasibility of LLM-based teacher feedback generation and preference dataset construction using the FEAT framework. However, the study has the following limitations:

First, while we constructed an English tutoring scenario using the MCTest dataset, further research is required to assess the generalizability of the framework across diverse educational datasets.

Second, we only conducted the ranking model experiments using 1B and 3B LLMs. Future work should explore the applicability of larger LLMs (e.g., 7B, 13B, 70B) to evaluate their impact on ranking performance.

Third, we employed a pairwise approach for ranking model training. We plan to explore preference dataset construction and training strategies applicable to pointwise and listwise ranking approaches in future research.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00004, Development of semisupervised learning language intelligence technology and Korean tutoring service for foreigners), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-0055621731482092640101), and research fund of Chungnam National University.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nischal Ashok Kumar and Andrew Lan. 2024. Improving socratic question generation using data augmentation and preference optimization. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 108–118, Mexico City, Mexico. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 89–96, New York, NY, USA. Association for Computing Machinery.

Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Nadine Chang, Francesco Ferroni, Michael J Tarr, Martial Hebert, and Deva Ramanan. 2023. Thinking like an annotator: Generation of dataset labeling instructions. *arXiv preprint arXiv:2306.14035*.

Koby Crammer and Yoram Singer. 2001. Pranking with ranking. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Ornelas, and Andrew Lan. 2024. Exploring automated distractor generation for math multiple-choice questions via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082, Mexico City, Mexico. Association for Computational Linguistics.

Davide Fossati. 2008. The role of positive feedback in intelligent tutoring systems. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 31–36.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4776–4785.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jin-Xia Huang, Yohan Lee, and Oh-Woog Kwon. 2023. Direct: Toward dialogue-based reading comprehension tutoring. *IEEE Access*, 11:8978–8987.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 133–142, New York, NY, USA. Association for Computing Machinery.

Gerd Kortemeyer. 2024. Performance of the pretrained large language model gpt-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1):47.

Charles Koutcheme, Nicola Dainese, Arto Hellas, Sami Sarsa, Juho Leinonen, Syed Ashraf, and Paul Denny. 2024. Evaluating language models for generating and judging programming feedback.

Yanyan Lan, Yadong Zhu, Jiafeng Guo, Shuzi Niu, and Xueqi Cheng. 2014. Position-aware listmle: a sequential learning process for ranking. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 449–458, Arlington, Virginia, USA. AUAI Press.

Yu-Ju Lan and Nian-Shing Chen. 2024. Teachers' agency in the era of llm and generative ai. *Educational Technology & Society*, 27(1):I–XVIII.

Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Math multiple choice question generation via human-large language model collaboration. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 941–946, Atlanta, Georgia, USA. International Educational Data Mining Society.

Ping Li, Qiang Wu, and Christopher Burges. 2007. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Wencke Liermann, Jin-Xia Huang, Yohan Lee, and Kong Joo Lee. 2024. More insightful feedback for tutoring: Enhancing generation mechanisms and automatic evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10838–10851, Miami, Florida, USA. Association for Computational Linguistics.

Haochen Liu, Zitao Liu, Zhongqin Wu, and Jiliang Tang. 2020. Personalized multimodal feedback generation in education. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1826–1840, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024a. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Jian Luo, Xuanang Chen, Ben He, and Le Sun. 2024b. PRP-graph: Pairwise ranking prompting to LLMs with graph aggregation for effective text re-ranking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5766–5776, Bangkok, Thailand. Association for Computational Linguistics.

Wenting Ma, Olusola O Adesope, John C Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, 106(4):901.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John

Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Alexander Scarlatos, Wanyong Feng, Andrew Lan, Simon Woodhead, and Digory Smith. 2024a. Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.

Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024b. Improving the validity of automatically generated feedback via reinforcement learning.

Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024c. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer.

Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. 2025. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, 15(2):671.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894.

Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. 2008. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 77–86, New York, NY, USA. Association for Computing Machinery.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. 2007. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 383–390, New York, NY, USA. Association for Computing Machinery.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022. Towards human-like educational question generation with large language models. In *Artificial Intelligence in Education*, pages 153–166, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. 2007. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 287–294, New York, NY, USA. Association for Computing Machinery.

A Related Works

A.1 Feedback in Education

In the field of education, teacher feedback plays a crucial role in enhancing students' learning experiences and achievements. In particular, immediate and appropriate feedback positively impacts students' cognitive, emotional, and motivational outcomes (Shute, 2008).

Research has been conducted on designing effective feedback strategies. Nicol and Macfarlane-Dick (2006) proposed seven principles for effective feedback, while Hattie and Timperley (2007) analyzed the impact of feedback on learning and investigated its key components. Steiss et al. (2024); Scarlatos et al. (2024c) proposed five criteria for evaluating feedback quality, designing them to help students understand clear directions for improvement and maintain motivation. Additionally, research applying feedback criteria has been conducted in fields such as programming education (Koutcheme et al., 2024).

A.2 Large Language Models in Education

Advancements in large language models (LLMs) have significantly impacted the field of education (Gan et al., 2023; Wang et al., 2024; Lan and Chen, 2024; Jeon and Lee, 2023; Dai et al., 2023). The integration of LLMs with educational technology has been applied across various domains, including automated short answer grading (Kortemeyer, 2024), automated essay scoring (Stahl et al., 2024), automated distractor generation (Feng et al., 2024; Scarlatos et al., 2024a), and automatic question generation (Luo et al., 2024a; Lee et al., 2024; Ashok Kumar and Lan, 2024; Mulla and Gharpure, 2023; Wang et al., 2022).

Research has also been conducted on automated feedback systems to provide better feedback to students (Dai et al., 2023; Meyer et al., 2024). Additionally, LLM-powered personalized feedback generation has contributed to reducing teachers' workloads and improving the efficiency of online education (Liu et al., 2020). Beyond feedback generation, LLMs have also been utilized for feedback quality assessment. Studies have proposed LLM-based feedback generation and evaluation frameworks in domains such as programming assignments (Koutcheme et al., 2024) and mathematics (Scarlatos et al., 2024c), demonstrating the potential of LLMs in educational settings.

A.3 Learning to Rank Approaches

Learning to Rank (LTR) is widely used in information retrieval and recommendation systems, aiming to learn the optimal ranking of items for a given query. LTR methodologies are generally categorized into three approaches: Pointwise, Pairwise, and Listwise.

Pointwise approaches, such as MCRank (Li et al., 2007) and PRank (Crammer and Singer, 2001), predict a relevance score for each item individually and rank them based on these scores. Pairwise approaches learn the relative preference between two items, with representative algorithms including RankNet (Burges et al., 2005), LambdaRank (Burges et al., 2006), RankSVM (Joachims, 2002), RankBoost (Freund et al., 2003), GBRank (Zheng et al., 2007), and FRank (Tsai et al., 2007). Listwise approaches consider the entire item list as a single input and optimize its order holistically, with prominent algorithms such as ListNet (Cao et al., 2007), ListMLE (Lan et al., 2014), and SoftRank (Taylor et al., 2008).

Recent LTR research has expanded to leverage LLMs for ranking tasks (Cui et al., 2023). Qin et al. (2024); Luo et al. (2024b) proposed LLM-based pairwise ranking methods, demonstrating the potential of large-scale language models in ranking optimization.

	Human	DIRECT	PrepTutor	GPT-3.5	GPT-4
Word	10.02	9.27	29.37	17.84	21.01
Token	13.98	13.25	36.60	22.61	26.58

Table 2: Average feedback word length and token length in the DIRECT-Manual dataset.

	# Data	Average Words Per:	
		Story	Question
DIRECT-Manual	5,500	193.05	11.90
MCTest	1,480	202.71	7.79

Table 3: DIRECT-Manual and MCTest dataset statistics.

B Details of DIRECT-Manual Dataset

B.1 Feedback Generation Process

The DIRECT-Manual (DM) consists of five feedback candidates, each generated through different methods. The details of these candidates are as follows:

- **Human**: Feedback written by human annotators.
- **DIRECT**: Feedback generated using GPT-2 trained on the DIRECT dataset.
- **PrepTutor**: Feedback generated using GPT-2 fine-tuned on external domain-specific feedback data.
- **GPT-3.5**: Feedback generated using GPT-3.5-turbo-0613.

Prompt for generating DIRECT-Manual

You are a proficient tutoring assistant who provides just a few clues to an user in the correct direction.
The user should understand the following passage and then answer your question.

Passage: **{passage}**

The correct answer is "{correct answer}", but the user don´t answer correctly as the following tutoring dialogues.
Generate an indirect feedback or hint to guide the user to find the answer on him/her own.

{student & teacher dialogue}

Figure 8: Prompt for generating DIRECT-Manual.

- **GPT-4**: Feedback generated using GPT-4-0613.

B.2 Feedback Ranking Process

The feedback candidates were ranked by human annotators based on two criteria:

- **Correct**: The feedback provides specific factual information based on the student's response or the given text.
- **Revealing**: The feedback guides the student toward the correct answer without explicitly stating it.

Table 2 presents the average length of feedback candidates, while Table 3 provides overall dataset statistics. Figure 8 illustrates the prompt used for feedback generation with GPT-3.5 and GPT-4 in the DM.

C Details of DIRECT-Generated Dataset

C.1 Dataset Preprocessing Process

The DIRECT-Generated (DG) dataset was constructed based on MCTest (Richardson et al., 2013), which consists of stories designed for students in grades 1–4, along with corresponding questions and four answer options.

The dataset construction involved the following preprocessing steps:

1. The question field from MCTest was used as the teacher's question.
2. The answer field was used as the student's correct response.
3. One option from the answer choices was randomly selected as the student's incorrect response.

C.2 Feedback Generation Process

We utilized LLMs to automatically generate teacher feedback. The prompt for feedback generation was designed to include the story, question, the student's incorrect response, and the correct response. Additionally, the five feedback criteria defined by Seo et al. (2025) were applied to ensure educationally effective feedback. The characteristics of each criterion are as follows:

- **Correct**: The feedback should be factually accurate and directly related to the student's response and the question.
- **Revealing**: The feedback should avoid explicitly providing the correct answer to the student.
- **Guidance**: The feedback should offer direction or hints to help the student progress towards the right answer.
- **Diagnostic**: The feedback should pinpoint and address any misconceptions or errors made by the student.
- **Encouragement**: The feedback should convey a positive and supportive tone to motivate the student.

The following LLMs were used for feedback generation:

- GPT-4o (Achiam et al., 2023)
- Claude-3[2] (Bai et al., 2022)
- Llama-3.1-70B-Instruct[3] (Dubey et al., 2024)

Figures 9 and 10 illustrate example prompts. The prompts were designed to generate teacher feedback that guides students from incorrect to correct responses. The feedback generated by all three LLMs was aggregated and then split into train and test datasets at a 9:1 ratio.

Table 4 presents examples of teacher feedback generated under different prompt strategies in the

[2] claude-3-5-sonnet-20240620

[3] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

DG dataset. Notably, when feedback criteria were not applied (w/o criteria), the generated feedback often explicitly stated the correct answer. In contrast, when feedback criteria were applied (w/ criteria), the generated feedback was more structured and pedagogically aligned.

D Implementation Details

All experiments were conducted in NVIDIA A100 (40GB VRAM) GPUs and implemented using the PyTorch. The Hugging Face (Wolf et al., 2019) was utilized for model training. All models were fine-tuned using the Low-Rank Adaptation (LoRA) (Hu et al., 2021). The versions of the models used are listed in Table 5, and detailed hyperparameter settings are provided in Table 6. The input format of the ranking model is illustrated in Figure 11 and Figure 12.

E Additional Experimental Results

Table 7 presents the overall performance across different ranking model approaches. All experiments were conducted five runs with different random seeds.

Figure 13 summarizes the results of the DIRECT-A (DA) →DIRECT-M (DM) scenario described in Section 4.3 for the Llama-1B, Llama-1B-IT, and Llama-3B. In most ranking model approaches, performance improved as the proportion of DM increased. For every model, DPO exceeds the DM→DM baseline even when trained with only 0–5% of the DM data. Notably, Llama-3B surpasses the DM→DM baseline in all methods with at most 10–25% of the DM data.

Figures 14 and 15 show examples of high and low RBO scores between ground-truth rankings and predicted rankings, respectively. When the two rankings are nearly identical, the RBO reaches 0.8333; when they diverge markedly, the score drops to 0.4166.

Prompt for generating the teacher's feedback without criteria provided

Instruction
You are an English teacher tasked with providing feedback to students.
Your goal is to provide feedback that guides the student from an incorrect answer to the correct one.
The feedback must be limited to one sentence.

Format
Respond in JSON format with the following structure:
{{
  "feedback": "Your feedback here."
}}

Story
{story}

Question
{question}

Incorrect Response
{student incorrect response}

Correct Response
{student correct response}

Figure 9: Prompt for generating the teacher's feedback with criteria provided.

Prompt for generating the teacher's feedback without criteria provided

Instruction
You are an English teacher tasked with providing feedback to students.
Your goal is to provide feedback that guides the student from an incorrect answer to the correct one.
The feedback must be limited to one sentence.

Format
Respond in JSON format with the following structure:
{{
  "feedback": "Your feedback here."
}}

Story
{story}

Question
{question}

Incorrect Response
{student incorrect response}

Correct Response
{student correct response}

Figure 10: Prompt for generating the teacher's feedback without criteria provided.

	Model Name	Generated Teacher Feedback
w/ criteria	GPT-4o	Consider the part of the story where Tom and his friends struggled the most and needed to exert extra effort to complete the task.
	Claude-3	Consider re-reading the part of the story that describes the difficulty level of fixing each item, paying special attention to which task was described as 'very hard'.
	Llama-3.1-70B	Consider re-examining the text for clues about which task Tom and his friends found particularly challenging, as indicated by the use of specific adjectives to describe the difficulty level of each task.
w/o criteria	GPT-4o	Remember, the story mentions that fixing the window was very hard, indicating it was the hardest thing for Tom and his friends to fix.
	Claude-3	While the toilet was mentioned first, the story explicitly states that fixing the window was 'very hard' and required all of them to push 'really hard', indicating it was the most challenging task.
	Llama-3.1-70B	You might want to reconsider your answer, as the passage states that fixing the window was very hard and required a lot of effort from Tom and his friends to open it.

Table 4: Examples of DIRECT-Generated for each prompt strategy.

Model Name	Version
Llama-1B	Llama-3.2-1B
Llama-1B-IT	Llama-3.2-1B-Instruct
Llama-3B	Llama-3.2-3B
Llama-3B-IT	Llama-3.2-3B-Instruct
Qwen-3B	Qwen2.5-3B-Instruct

Table 5: Model names and versions Used for training the ranking model.

Hyperparameter	Value
Training Hyperparameters	
Learning rate	5e-05
Batch size	8
Training epochs	5
Max sequence length	1,024
Random seeds	0, 42, 500, 1000, 1234
Lora Config	
Rank	16
Alpha	32
Dropout	0.05

Table 6: Hyperparameters for training the ranking model.

Input sequence for binary classifier

Select the most appropriate teacher feedback based on the context provided.

Story: **{story}**
History: **{history}**
Choose the better feedback:
1. **{chosen}**
2. **{rejected}**

Figure 11: Input data format for binary classifier.

Prompt for preference learning

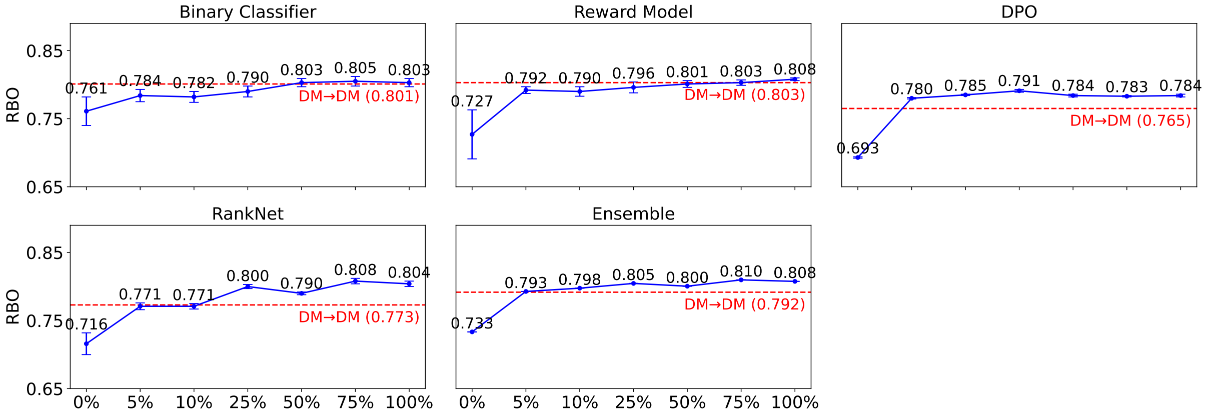
Select the most appropriate teacher feedback based on the context provided.

Story: **{story}**
History: **{history}**
Choose the better feedback.

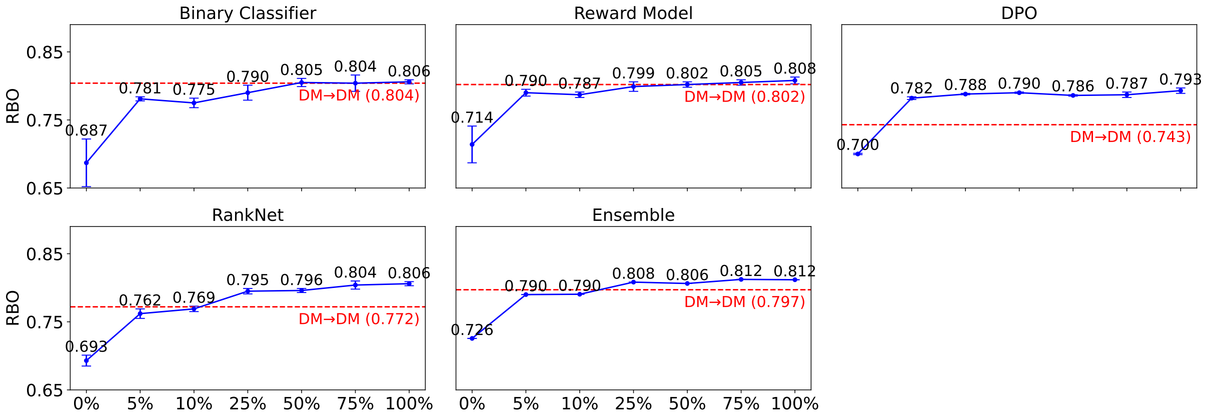
Figure 12: Prompt for reward model, DPO, and RankNet.

	Classifier		Reward Model		DPO		RankNet		Ensemble	
Model Name	DM→DM	DG→DM	DM→DM	DG→DM	DM→DM	DG→DM	DM→DM	DG→DM	DM→DM	DG→DM
Llama-1B	0.801±0.006	**0.761±0.021**	0.803±0.004	0.727±0.036	0.765±0.012	0.693±0.001	0.773±0.005	0.716±0.016	0.792	0.733
Llama-1B-IT	0.804±0.005	0.687±0.035	0.802±0.006	0.714±0.027	0.743±0.024	0.700±0.001	0.772±0.004	0.693±0.008	0.797	0.726
Llama-3B	0.772±0.056	0.702±0.033	0.802±0.005	0.715±0.048	0.769±0.024	0.689±0.001	0.779±0.004	0.747±0.006	0.799	0.740
Llama-3B-IT	0.773±0.056	0.750±0.037	0.802±0.004	**0.731±0.034**	0.770±0.023	**0.715±0.002**	0.774±0.007	0.734±0.006	0.798	0.743
Qwen-3B-IT	0.806±0.008	0.732±0.023	0.802±0.007	0.729±0.032	0.779±0.002	0.714±0.003	0.791±0.012	**0.758±0.012**	0.811	**0.755**

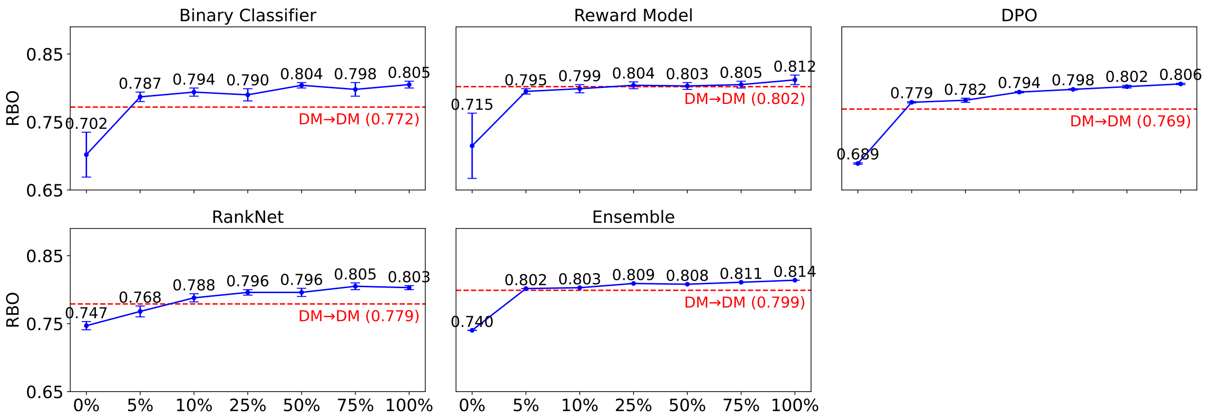
Table 7: Performance by ranking model approaches. Best results are highlighted in **bold**. The ± represents standard deviation from five results of five different seeds. IT refers to the Instruct model.



(a) Llama-1B



(b) Llama-1B-IT



(c) Llama-3B

Figure 13: Overall performance across varying DM dataset proportions in the DA→DM scenario. Error bars represent the standard deviation calculated from the results of five different seeds.

Feedback Candidates

- **Human**: It costs 20,000 to 30,000 or more dollars if you add all the expenses of a year.
- **DIRECT**: no, the cost of tuition is over us$10, 000 per year.
- **PrepTutor**: good effort! frankly, a college education is not cheap. but, students who attend college earn many years of free education. how much will a college student spend on his or her education?
- **GPT-3.5**:That's close, but it's actually higher than that. Can you think of any other expenses that students have to pay for?
- **GPT-4**: That's a good start, but remember, the $10,000 is just the average cost of tuition. There are other expenses a student needs to consider like books, housing, and personal expenses. Try to add all these up for a full picture of the annual cost.

Ground-truth Ranking: GPT-4 > GPT-3.5 > DIRECT > Human > PrepTutor
Predicted Ranking: GPT-4 > GPT-3.5 > PrepTutor > Human > DIRECT
RBO: 0.8833

Figure 14: An example with a high RBO score between the ground-truth ranking and the predicted ranking.

Feedback Candidates

- **Human**: You got it half right. There is one more answer.
- **DIRECT**: no, life today has brought new problems.
- **PrepTutor**: good job! actually, life on the planet was not that easy. people had few tools for many years. then came up with idea to create better tools. what is it?
- **GPT-3.5**: That's correct! But there was another reason mentioned in the passage. Can you think of what it might be?
- **GPT-4**: That's correct! Now, can you think about the new problems that life today has brought? Specifically, consider the various types of pollution mentioned in the passage.

Ground-truth Ranking: GPT-3.5 > PrepTutor > Human > GPT-4 > DIRECT
Predicted Ranking: GPT-4 > GPT-3.5 > DIRECT > PrepTutor > Human
RBO: 0.4166

Figure 15: An example with a low RBO score between the ground-truth ranking and the predicted ranking.