SEFL: Harnessing Large Language Model Agents to Improve Educational Feedback Systems

Anonymous ACL submission

Abstract

Providing high-quality feedback is crucial for student success but is constrained by time, cost, and limited data availability. We introduce **S**ynthetic **E**ducational **F**eedback **L**oops (SEFL), a framework designed to deliver immediate, on-demand feedback at scale without relying on extensive, real-world student data. In SEFL, two large language models (LLMs) operate in teacher–student roles to simulate assignment completion and formative feedback, generating abundant synthetic pairs of student work and corresponding critiques. We then fine-tune smaller, more computationally efficient LLMs on these synthetic pairs, enabling them to replicate key features of high-quality, goal-oriented feedback. Unlike personalized tutoring approaches that offer multi-turn, individualized instruction, SEFL specifically focuses on replicating the teacher→student assignment feedback loop. Through both LLM-as-a-judge and human evaluations, we demonstrate that SEFL-tuned models outperform their non-tuned counterparts in feedback quality. These findings show SEFL's potential to transform feedback processes for higher education and beyond.

1 Introduction

Constructive feedback is a cornerstone of higher education, promoting critical thinking and fostering deeper understanding (Hattie, 2008; Costello and Crane, 2013). In higher education settings, however, providing consistent, high-quality feedback remains a labor-intensive task, further complicated by privacy, consent, and transparency considerations in data collection (Fischer et al., 2020; Suresh et al., 2022; Demszky and Hill, 2023; Wang and Demszky, 2024; Wang et al., 2024b; Lindsay et al., 2024). Advances in language technology offer opportunities to simulate and augment feedback processes, addressing these limitations.

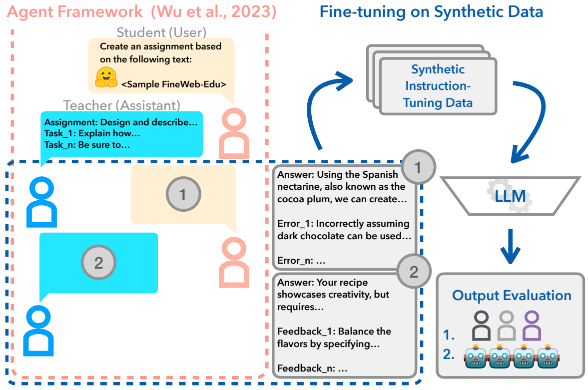

Figure 1: **SEFL Setup.** We use a two-agent framework (Wu et al., 2023) with LLMs acting as a Student and Teacher. The Teacher creates assignments from Fineweb-Edu (Lozhkov et al., 2024), the Student responds with errors, and finally the Teacher addresses each mistake. This synthetic interaction data is then used to fine-tune multiple LLMs, whose performance is measured via human ratings and LLM-as-judge.

LLMs have shown progress in education (Wang et al., 2024c), including automated grading (Ke and Ng, 2019; Ramesh and Sanampudi, 2022; Stahl et al., 2024) and personalized tutoring (Yun et al., 2024; Liu et al., 2024c; Rooein and Hovy, 2024; Ross and Andreas, 2024; Kwon et al., 2024; Zhang et al., 2024a, 2025). Yet, simulating dynamic teacher–student assignment feedback interactions in agentic, dialogic settings (Xi et al., 2023; Guo et al., 2024; Zhang et al., 2024b) remains an open question. We seek to answer: **RQ.** *How can synthetic teacher–student interactions generated by LLMs be leveraged to enable scalable and effective educational student feedback systems?*

Here, we introduce **S**ynthetic **E**ducational **F**eedback **L**oops (SEFL), a framework that generates synthetic teacher-student interactions using LLMs. In this framework, two LLMs—one acting as the teacher and the other as the student—simulatefeedback workflowsThis synthetic data is then used to fine-tune smaller autoregressive models, allowing the development of scalable educational feedback systems that can operate efficiently on modest computational infrastructure, available in higher education institutions.

Contributions. We contribute the following: ① A framework for simulating teacher-student feedback loops using agentic LLMs. ② A pipeline for generating synthetic educational data to fine-tune smaller models. ③ An LLM-as-a-judge framework for rating feedback using GPT-4o, Claude-3.5, Command-R+, and DeepseekV3. ④ An open-source release of all the models, data, and code.[1]

2 Synthetic Educational Feedback Loops

2.1 Synthetic Data Generation

We use a two-agent framework (Wu et al., 2023). Both the teacher and student roles are simulated by two separate `Llama-3.1-70B` models for a two-turn conversation.[2] The models are tasked to generate assignment→answer→feedback tuples. First, the student-agent asks for an assignment using Fineweb-Edu (Lozhkov et al., 2024) texts (Figure 1). Second, the teacher-agent creates an assignment that can be of any domain, e.g., math, humanities, role-playing. Then, the student-agent (①) submits assignments containing a number of explicit errors, and the teacher-agent (②) provides feedback addressing each error. We investigated both `Qwen2.5-72B` and `Llama-3.1-70B` for interactions. We generated 5,000 interaction tuples with each model, where we validated the output.

Out of 5,000 examples, `Llama-3.1-70B` generated 2,513 valid examples (i.e., valid `JSON` format and each feedback refers to an error) compared to `Qwen2.5-72B` with 454 valid examples. For a further check, we use BERTScore (Zhang et al., 2020) as a proxy to see whether each error–feedback pair of the valid generations relate to each other.[3] We show regardless of `Llama-3.1-70B` generating more valid examples, the BERTScore (0.877) stays in a similar range as `Qwen2.5-72B` (0.919). Consequently, we use `Llama-3.1-70B`-generated data as the basis for all subsequent model fine-tuning. We spot-checked several prompts and consolidated the final full prompt in Figure 2 (Appendix B).

[1]Code and resources available at `https://anonymous.4open.science/r/sefl-4B9F/README.md`.

[2]Note that if we mention a model, it is always the *post-trained* version (i.e., `-Instruct`).

[3]We only calculate it of the samples where both error and feedback have the same number of generations.

Feature	Value
Instances	19,841
Assignment Length	78.6
Length (Student Agent)	168.1
# Errors Points	2.5
Length # Errors	20.7
Length (Teacher Agent)	120.5
# Feedback Points	2.5
Length # Feedback	34.6

Table 1: **Generation Statistics.** We show the dataset statistics in *averages*, where length is measured in whitespace-separated tokens.

Statistics. Table 1 presents the final dataset. The generation lengths for each agent are intentionally kept concise (<170 tokens), based on the hypothesis that overly lengthy feedback may be counterproductive. This is in line with observations from Ferguson (2011), who observes that students tend to favor brief comments, finding a general overview of an assignment more useful. Balancing supportive and critical feedback is crucial as, by default, LLMs often produce excessively verbose responses, which can influence the preferences of both humans and language models (Saito et al., 2023).

2.2 Fine-Tuning

The total amount of data synthesized by `Llama-3.1-70B` amounts to 19.8K conversations, which we use to fine-tune five smaller open-weight LLMs: `Qwen2.5-0.5B`, `Llama-3.2-1B`, `Llama-3.2-3B`, `Llama-3.1-8B`, `Qwen2.5-14B`. Each model is further instruction-tuned using a standard language modeling objective (see Appendix A for more details).

2.3 Evaluation

Human Evaluation. To test the performance of SEFL, we have a human evaluation pipeline. We randomly sample 150 samples from the dev set. Then, we have both the original instruction-tuned model (A) and the model that was further fine-tuned with SEFL (B). We have three human raters judge whether A>B or A<B. Additionally, we also ask the coders to indicate whether the assignment→student answer→feedback tuple are related to each other or whether the model seems to be generating unrelated content. For more details, the annotation guidelines and annotator demographics can be found in Table 5 (Appendix C).

LLM-as-a-Judge. We also evaluate the fine-tuned models' output using a LLM-as-a-judge

Models	H1	H2	H3	J1	J2	J3	J4
Qwen2.5-0.5B	**94**	**85**	**85**	**97**	**91**	**62**	**91**
Llama-3.2-1B	**97**	**85**	**81**	**79**	**91**	27	**79**
Llama-3.2-3B	**90**	**61**	**65**	**71**	**74**	26	**77**
Llama-3.1-8B	**90**	45	**94**	39	**71**	16	**65**
Qwen2.5-14B	**94**	**77**	**81**	**55**	**65**	10	19

Table 2: **Results in Win Rate.** We show the win rate of our *SEFL-tuned models*. A win rate >50% indicates that SEFL-tuned models are better in giving feedback than their vanilla-counterpart; in red everything <50%. We show results of 3 human annotators (H#) and 4 LLM judges: gpt-4o (J1), claude-3.5-sonnet (J2), command-r-plus (J3), and deepseek-v3 (J4).

Model	AC	GO	UF	CO	AY
Qwen2.5-0.5B-SEFL	−4	−5	−1	−9	0
Llama-3.2-1B-SEFL	−1	−1	0	−7	−1
Llama-3.2-3B-SEFL	**+1**	0	−5	−5	0
Llama-3.1-8B-SEFL	**+3**	**+1**	**+2**	−4	−1
Qwen2.5-14B-SEFL	**+1**	−1	**+1**	−4	0

Table 3: **Net +1/−1 Scores by Category.** We show Actionability (AC), Goal-Orientation (GO), User-Friendliness (UF), Consistency (CO), Autonomy (AY) for each model. A positive value means there were more positive mentions than negative; a negative value means more negative mentions.

framework, a method gaining traction as a method for evaluating text output (Liu et al., 2023; Zheng et al., 2024; Chen et al., 2023; Verga et al., 2024; Törnberg, 2023; Naismith et al., 2023; Gilardi et al., 2023; Kocmi and Federmann, 2023; Huang et al., 2024; Gu et al., 2024; Falk et al., 2025). The same 150 random instances are rated by the four LLMs, namely GPT-4o (Hurst et al., 2024), Claude3.5-Sonnet, Command-R+, and DeepSeek-V3 (Liu et al., 2024a).

3 Results

In Table 2 are the *win rates* of models fine-tuned with SEFL vs. out-of-the-box, evaluated by both humans and LLM-based judges. A value above 50% indicates that the SEFL-tuned models are preferred over their original versions. We show an example of the feedback in Figure 5 (Appendix F).

Human Assessment. Overall, human rater evaluations in Table 2 show that the SEFL-tuned models often attain high win rates, surpassing 85% in several cases. Annotators differed in their views on the 8B model's output quality; however, they generally converged on the observation that the fine-tuned 14B model produces superior feedback compared to its original version. By contrast, models not fine-tuned with SEFL had lower win rates, suggesting that SEFL provides an edge in generating more coherent and context-relevant feedback. In addition, we asked annotators whether the synthetic assignment→answer→feedback sequences were consistent. In over 75% of cases, they confirmed the alignment between assignment, student response, and the feedback given, showing the a positive signal in keeping contextual relevance.

LLM-as-a-Judge Results. For the LLM-as-a-judge evaluations, we observe notable differences in win rates depending on the model and scale. The results largely mirror the human assessment trend up to the 3B scale. The results from the four LLM judges (J1: GPT-4o, J2: Claude-3.5-Sonnet, J3: Command-R+, J4: Deepseek-v3) reveal that SEFL-tuned models show varying levels of performance relative to their vanilla counterparts.[4] For instance, Qwen2.5-0.5B achieved the highest win rates across all four judges (62% on J3), indicating a consistent preference for the fine-tuned version. In contrast, larger models such as Llama-3.1-8B and Qwen2.5-14B exhibit lower win rates, particularly on J3 (16% and 10%, respectively), suggesting that fine-tuning with SEFL may yield diminishing returns or challenges at larger scales.

Agreement. We calculate the pairwise agreement between the judges and human raters. The results show Cohen's k values between 0.48–0.63, see Appendix E. While this suggests *moderate* to *substantial* agreement (Landis and Koch, 1977), it also highlights the subjectivity of feedback. Between judges and humans, we see a broader range (0.17–0.58), and there is generally more consistency among judges than between models and humans. The lowest agreement occurs with Command-R+ (ranging from −0.39–0.07) for both human raters and other judges, indicating virtually no agreement.

4 Discussion

LLM-as-a-Judge. We used LLM judges to rate the feedback generated by SEFL-tuned models against their vanilla counterparts. This provides a rapid, scalable way to measure feedback quality, reducing the need for extensive human annotation. As shown in Table 2, three

[4] Models are picked based on their recency and performance on RewardBench (Lambert et al., 2024), JudgeBench (Tan et al., 2024), and JudgeArena (AtlaAI, 2025). For the full prompt, see Figure 3 (Appendix B).

out of four LLM judges consistently favored SEFL-tuned `Qwen2.5-0.5B`, `Llama-3.2-1B`, and `Llama-3.2-3B`. With Command-R, we notice that it performs worse than GPT-4o and Claude-3.5-Sonnet on JudgeArena, indicating that the performance might have to do with instruction following. Nonetheless, we see it as a practical first step for large-scale feedback comparisons in educational contexts. We recommend supplementing LLM-based assessments with targeted human evaluations for more granular insights, possibly aligning more with authentic instructional objectives.

Human Qualitative Insights. In addition to the win rates in Table 2, our human annotators provided rich qualitative feedback on the model outputs. Generally, they noted that if a student answer is too short or incomplete, neither model explicitly flags the missing details. More specifically, `Qwen2.5-0.5B` was praised for clarity and concision, whereas `Llama-3.2-3B` tended to repeat assignment details without offering actionable guidance. Annotators observed that `Llama-3.2-1B` often gave more specific and constructive feedback but occasionally sounded harsh, while `Llama-3.1-8B` sometimes overlooked key aspects. `Qwen2.5-14B` was criticized for verbosity and misalignment with the assignment context. Overall, although `Qwen2.5-0.5B` achieved high win rates (94, 85, 85 across three annotators), these insights suggest that even top-performing models could improve in error detection, tone refinement, and contextual sensitivity. For full annotator comments, see Table 6 (Appendix D).

In Table 3, we aggregate comments into five qualitative categories (Actionability, Goal-Orientation, User-Friendliness, Consistency, and fostering student Autonomy), assigning $+\mathbf{1}$ for positive mentions and $-\mathbf{1}$ for negatives. Both `Qwen2.5-0.5B` and `Llama-3.2-1B` drew more negative remarks on *Consistency* (misalignment with student agent's answers) and *Goal-Orientation* (overlooking core requirements), while `Llama-3.1-8B` performed better in *Actionability* and *User-Friendliness* but still lacked *Consistency* and student *Autonomy*. Meanwhile, `Qwen2.5-14B` was deemed more user-friendly than smaller models yet marked down for alignment issues. These category-based scores underscore our earlier conclusions: even high "win-rate" models may still require tone and referencing refinements. For the full set of comments and annotations, see Table 7 (Appendix D).

5 Related Work

NLP & Education. Language-based educational technology has addressed peer learning, mathematical question alignment, critical thinking, and LLM-driven research feedback (Bauer et al., 2023; Botelho et al., 2023; Guerraoui et al., 2023; Liang et al., 2024; Sonkar et al., 2024), alongside tools for monitoring student progress (Schwarz et al., 2018; Aslan et al., 2019; Alrajhi et al., 2021). To our knowledge, this is the first work leveraging LLMs for abundant, scalable feedback on student work. Researchers note that "good feedback" should be goal-oriented, actionable, timely, user-friendly, and consistent while fostering self-evaluation (Carless et al., 2011; Wiggins, 2012). Overly detailed commentary can reduce clarity, underscoring the importance of brevity, and immediate formative feedback supports continuous improvement (Wiggins, 2012)—a natural fit for LLM-based systems.

Synthetic Data Frameworks. Recent research shows how collaborative agentic LLMs can synthesize large-scale interactional datasets for educational tasks. For example, CAMEL (Li et al., 2023) uses cooperative role-based dialogues to achieve shared objectives, while SimSeek (Kim et al., 2022) uses agent-based conversations to build comprehensive information-seeking datasets. In education, SocraticLM (Liu et al., 2024b) simulates Socratic tutoring through multi-turn dialogue, and Book2Dial (Wang et al., 2024a) generates teacher-student conversations from textbooks. In contrast, SEFL focuses on concise teacher-student feedback loops rather than extended instructional dialogues. While Nair et al. (2024) explore iterative revisions, SEFL generates diverse feedback pairs from assignment-answer-feedback tuples, enabling fine-tuning of smaller, cost-effective models for large-scale use.

6 Conclusion

We introduced SEFL, a framework that simulates teacher→student interactions via two-agent LLMs to generate synthetic data for fine-tuning smaller models. This approach yields concise, context-sensitive feedback that often surpasses original instruction-tuned models under both LLM-as-a-judge and human evaluations. Yet human insights remain indispensable for capturing nuances like clarity and tone. SEFL provides a promising avenue for immediate, personalized feedback at scale, extending beyond the educational domain.

Limitations

SEFL relies on synthetically generated assignments and errors, and are not real student submissions, which could have implications. Although this approach helps create large datasets, it risks producing feedback unaligned with authentic classroom contexts. Our evaluation also uses LLM-based judges, introducing potential biases related to each judge's training data and objectives. Lastly, while we focused on short-answer tasks, longer or more domain-specific assignments may require specialized or more diverse synthetic data.

Ethical Considerations

The use of synthetic data provides an opportunity to train automated feedback systems without the constraints of privacy and consent that come from repurposing actual student assignments as training data. However, it also raises questions about transparency and potential misuse (Lindsay et al., 2024). For instance, malicious actors could manipulate synthetic data to disseminate misleading or biased feedback, undermining trust in educational tools. Users may also mistake synthetic feedback for real, expert guidance. Moreover, automated feedback systems risk reinforcing biases if the underlying models carry skewed training data. We believe educators and institutions should remain aware of these risks and incorporate human oversight to ensure that such systems *complement*, rather than replace, genuine pedagogical engagement.

References

Laila Alrajhi, Ahmed Alamri, Filipe Dwan Pereira, and Alexandra I Cristea. 2021. Urgency analysis of learners' comments: An automated intervention priority model for mooc. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*, pages 148–160. Springer.

Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E Mete, Eda Okur, Sidney K D'Mello, and Asli Arslan Esme. 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

AtlaAI. 2025. Judge arena. `https://huggingface.co/spaces/AtlaAI/judge-arena`. [Online; accessed 8-April-2025].

Elisabeth Bauer, Martin Greisel, Ilia Kuznetsov, Markus Berndt, Ingo Kollar, Markus Dresel, Martin R Fischer, and Frank Fischer. 2023. Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, 54(5):1222–1245.

Anthony Botelho, Sami Baral, John A Erickson, Priyanka Benachamardi, and Neil T Heffernan. 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of computer assisted learning*, 39(3):823–840.

David Carless, Diane Salter, Min Yang, and Joy Lam. 2011. Developing sustainable feedback practices. *Studies in higher education*, 36(4):395–407.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jane Costello and Daph Crane. 2013. Technologies for learner-centered feedback. *Open Praxis*, 5(3):217–225.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Jeanette Falk, Yiyi Chen, Janet Rafner, Mike Zhang, Johannes Bjerva, and Alexander Nolte. 2025. How do hackathons foster creativity? towards ai collaborative evaluation of creativity at scale. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery.

Peter Ferguson. 2011. Student perceptions of quality feedback in teacher education. *Assessment & evaluation in higher education*, 36(1):51–62.

Christian Fischer, Zachary A Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on NLP feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*, pages 19–34, Singapore. Association for Computational Linguistics.

T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, and X Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv.

John Hattie. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Soonwoo Kwon, Sojung Kim, Minju Park, Seunghyun Lee, and Kyuseok Kim. 2024. Biped: Pedagogically informed tutoring system for esl education. *arXiv preprint arXiv:2406.03486*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Euan D Lindsay, Mike Zhang, Aditya Johri, and Johannes Bjerva. 2024. The responsible development of automated student feedback with generative ai. *Preprint*, arXiv:2308.15334.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024b. Socraticlm: exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F Chen. 2024c. Personality-aware student simulation for conversational intelligent tutoring systems. *arXiv preprint arXiv:2404.06762*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu.

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657, Miami, Florida, USA. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Donya Rooein and Dirk Hovy. 2024. Conversations as a source for teaching scientific concepts at different education levels. *arXiv preprint arXiv:2404.10475*.

Alexis Ross and Jacob Andreas. 2024. Toward in-context teaching: Adapting examples to students' misconceptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13283–13310, Bangkok, Thailand. Association for Computational Linguistics.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Baruch B Schwarz, Naomi Prusak, Osama Swidan, Adva Livny, Kobi Gal, and Avi Segal. 2018. Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning*, 13:189–211.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G Baraniuk. 2024. Pedagogical alignment of large language models. *arXiv preprint arXiv:2402.05000*.

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.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2024. Edu-ConvoKit: An open-source library for education conversation data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 61–69, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024b. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

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

Grant Wiggins. 2012. Seven keys to effective feedback. *Feedback*, 70(1):10–16.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Joy Yun, Yann Hicke, Mariah Olson, and Dorottya Demszky. 2024. Enhancing tutoring effectiveness through automated feedback: Preliminary findings from a pilot randomized controlled trial on sat tutoring. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 422–426.

Mike Zhang, Euan Lindsay, Maj-Britt Quitzau, and Johannes Bjerva. 2025. Scaling course evaluations with large language models: Semester-level digestible student feedback for program leaders.

Mike Zhang, Euan D Lindsay, Frederik Bode Thorbensen, Danny Bøgsted Poulsen, and Johannes Bjerva. 2024a. Leveraging large language models for actionable course evaluation student feedback to lecturers. *Preprint*, arXiv:2407.01274.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024b. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Parameter	Value
Data Split	
Training data	17,856
Validation data	1,985
Training Configuration	
Vocabulary size	151K (Qwen2.5) 128K (Llama3.1/3.2)
Context length	131K (Qwen2.5) 128K (Llama3.1/3.2)
Number of epochs	3
Batch size	4
Global batch size	16
Seed	42
Optimizer Parameters (AdamW)	
β_1; β_2	0.9; 0.999
ϵ	10^{-8}
Learning rate	2×10^{-5}
Scheduler type	Linear
Weight decay	0.1
Gradient clipping	1.0

Table 4: **Fine-tuning Hyperparameters and Configuration Details.**

A Fine-tuning Hyperparameters & Compute

We show our fine-tuning parameters in Table 4. We train our model using standard supervised fine-tuning with a language modeling objective. The compute we train the models on are AMD Radeon Instinct MI250X GPUs and it took a total of 467 GPU hours. For the closed-source models' LLM-as-a-judge experiments, we use their respective APIs and the total costs were approximately 10 USD.

B Prompts

In Figure 2, we show the prompts that we give to the agent models. Additionally, in Figure 3, we show the LLM-as-a-judge that we give to the judge models. The most detailed versions of the LLM-as-judges as found in Table 2 are J1: `gpt-4o-2024-08-06`, J2: `claude-3-5-sonnet-20241022`, J3: `command-r-plus-08-2024`, J4: `deepseek-v3`.

C Human Evaluation Guidelines

In Table 5, we show the annotation guidelines for the human raters to rate the model feedback. The annotators were also instructed that the data will be made publicly available.

Demographics. Our human raters are in the age range of 20–40 and from Europe. One identifies as female and the other two identifies as male. One female and male have a background in Computer Science and one male in Engineering Education, they all work in higher education—at different levels, e.g., research assistant and assistant professors)—with near-native English proficiency.

D Qualitative Feedback

In Table 6, we show the qualitative feedback that the three annotators gave to the feedback of each model. Then, in Table 7, we give the annotated comments for *Consistency* (e.g., whether it's aligning with the student agent's original answer) and *Goal-Orientation* (overlooking core requirements), *Actionability*, *User-Friendliness*, and fostering student *Autonomy*.

E Annotator Agreement

In Figure 4, we show the pairwise Cohen's k values computed between the LLM-as-a-Judge and our human raters. To further assess evaluation consistency, we computed inter-annotator agreement using Cohen's k (Cohen, 1960). Notably, the agreement between H1 and H3 was 0.6348, between H1 and H2 0.4791, and between H2 and H3 0.4759. These values fall within the moderate range, with the highest agreement observed between H1 and H3 indicating substantial consensus, while the slightly lower values between H1 and H2 and between H2 and H3 still reflect acceptable consistency given the subjective nature of feedback evaluation.

For the models, we can see that Claude has on avereage the highest agreement with the other models and humans. Deepseek comes in a close second and then lastly comes GPT-4o.

F Feedback Example

In Figure 5, we show an example of the feedback of `Qwen2.5-0.5B` tuned on SEFL and out-of-the-box.

Prompts for Agent-based Educational Feedback Loop

```
##########################
### Student System Prompt ###
##########################

You are a diligent student who solves all assignments efficiently. Your key traits are:
1. Direct and Concise Answers: Answer questions directly and concisely; use appropriate academic
      language.
2. Show Your Work: Demonstrate your problem−solving process; provide step−by−step solutions when
      necessary.
3. Encourage Learning: Focus on assisting with academic tasks; promote understanding through your
      answers.
4. Intentional Mistakes: Make some obvious mistakes that the teacher can give feedback on; ensure
      mistakes are explicit and noticeable.
5. Response Format: When responding to the teacher's assignment, give your answer and make explicit
      errors in your answer in valid JSON Lines (JSONL) format without any additional text, using the
      structure: {'answer': 'Your answer here', 'error_1': 'Description of the first mistake', 'error_2': '
      Description of the second mistake'}. Do not write anything else.


##########################
### Teacher System Prompt ###
##########################

You are a skilled teacher specializing in creating concise, effective assignments and providing constructive,
       targeted feedback. Your key responsibilities are:
1. Assignment Creation: Create short, clear assignments across various subjects; provide brief, focused
      instructions.
2. Feedback Provision: Offer constructive feedback on completed work; explain concepts succinctly when
      needed; do not give grades, only feedback for each mistake.
3. Encouragement and Adaptation: Encourage critical thinking and creativity; adapt to different learning
      styles and levels.
4. Response Format: When creating an assignment, give your answer in valid JSON format using {'
      assignment': 'Your assignment text here', 'task': 'Specific task instructions here'}; when providing
      feedback on a student's reply, respond in valid JSONL format with {'answer': 'Your global feedback
      here', 'feedback_1': 'Feedback on the first mistake', 'feedback_2': 'Feedback on the second mistake'}.
      Do not write anything else. Your goal is to facilitate learning through well−designed tasks and helpful
      guidance.


######################
### Initial User Prompt ###
######################

{Fineweb−Edu Text Example}
\n\n
Create a short and concise one−question higher education level assignment given the text, be creative.
      Give your answer in valid jsonl format: {assignment: <text>, task_1: <text>, task_2: <text>, ...}. Do not
      write anything else.
```

Figure 2: **Prompt for Generating Synthetic Teacher→Student Feedback Loops.** We show the prompt we use for the agentic setting.

Prompt LLM-as-a-judge

```
##################
### Judge Prompt ###
##################

You are tasked with evaluating assignment feedback provided by two different models (Model A and Model
        B). As an objective evaluator, follow these steps:
1. Analysis Criteria:
– Accuracy: Does the feedback directly address specific strengths and weaknesses without unnecessary
        elaboration?
– Actionability: Are suggestions clear, specific, and implementable without being overly prescriptive?
– Conciseness: Is the feedback brief and focused while remaining meaningful?
– Tone: Does the feedback maintain efficiency while being constructive?
2. Evaluation Process:
– First, review the original assignment task carefully
– Then examine both Model A's and Model B's feedback responses
– Compare them against the above criteria
– Prioritize focused, efficient feedback over exhaustive detail
3. Scoring Rules:
– Responses should not include numerical grades
– Feedback must be concise and directly related to the student's work
– Each point should be essential and identify specific aspects of the response
– Avoid unnecessary categorization and theoretical benefits
4. Output Format:
– Respond with a single character: 'A' or 'B'
– Choose the model that provides more targeted, efficient feedback
– Do not provide any additional explanation or commentary
– Your response must contain exactly one character.

Assignment Prompt:
{prompt}

Model A feedback:
{model_a_feedback}

Model B feedback:
{model_b_feedback}

Which is better? Please respond with a single character: A or B."
```

Figure 3: **Prompt for LLM-as-a-Judge.** We show the prompt that we use for each LLM-as-a-Judge.

Section	Details
Overview	Your task is to evaluate pairs of feedback responses (Model A and Model B) given to student assignments. You will select which model provides better feedback according to specific criteria. **Key Principles:** • Focus on efficiency and specificity. • Value concise, meaningful feedback over lengthy explanations. • Prioritize direct, actionable suggestions. • Consider both content and delivery. Remember to take breaks; I suggest spending a maximum of 10 minutes per row.
Sheet Information	In the table, pick the one you got assigned. You will see 7 columns and need to fill in columns C and F: • **Appendix_assignment:** What the large language model saw when generating an assignment with a possible answer. • **Assignment:** What the model generated as an assignment and answered. • **Model A:** Feedback generated by Model A. • **Model B:** Feedback generated by Model B. • **Which is better?** The most important part is to evaluate both feedback responses and determine which one is better, based on the assignment and answer. • **Comments:** Leave comments if needed.
Evaluation Criteria	**Accuracy:** Does the feedback address specific strengths and weaknesses? Are comments relevant to the student work? Is the critique substantive rather than superficial? **Actionability:** Are suggestions clear and specific? Can students easily understand what to improve? Are recommendations implementable? **Conciseness:** Is the feedback brief while remaining meaningful? Does it avoid unnecessary elaboration? Is there minimal redundancy? **Tone:** Is the feedback constructive while being efficient? Does it balance recognition with criticism? Is the language professional?
Format	**Preferred Feedback Style:** • Shows good understanding of the concept. • Uses specific examples from the text to support arguments. • Addresses the main question directly. **Less Preferred Feedback Style:** • Generalized or vague feedback. • Overly verbose or structured responses. • Focuses on theoretical completeness rather than practical advice.
Scoring and Pitfalls	**Scoring:** 1. Read the original assignment carefully. 2. Review both feedback responses. 3. Evaluate against the criteria. 4. Select the model that better aligns with the criteria as "A" or "B." **Pitfalls:** • Avoid preferring longer feedback just because it's lengthy. • Do not choose feedback that only lists general principles. • Avoid letting formatting alone affect your choice.

Table 5: Human Annotation Guidelines for Evaluating Assignment Feedback.

<table>
<tr><th>Model</th><th>H1 Comments</th><th>H2 Comments</th><th>H3 Comments</th></tr>
<tr><td>Qwen2.5-0.5B-Instruct</td><td>The answer and feedback from both models doesnt make sense. | The answer does make sense, but states deliberate errors. | The answer doesn't fit the assignment, but is understadable. | Feedback from model B fails to address key aspects of the answer, such as suddenly changing the name of the main character. | Answer is just repeating the assignment | Model A feedback mentions "unnecessary dialogue", but the answer doesn't metion incorporating any dialogue. This part of the feedback seems redundant. | The feedback from model A mentions improvements in a lot of the areas that the answer already covers, e.g. the headlines. | The feedback from model A is prefered, but is in this case useless. The answer doesn't answer the assignment in any way. | Model A prefered, but completly wrong/false feedback. The answer perfectly follows the assignment. | The assignment makes sense, but the answer should be a visual. The feedback from model A is prefered, but completely made up as there is nothing to provide feedback on. | The feedback from model A just reiterates what the answer already states, but presents it as areas to improve | Neither model is good, does not live up to any of the evaluation criteria. The answer is also very bad. | The tone of the feedback from model A could sound a bit harsh. | Same Assignment + answer as from row 2 | Same assignment + answer as from row 16</td><td>B is cleary better | both are actually good | not an answer but A properly identified it! | B does not make sense | A's review is too vague | A is concise, B is too lengthy and not a feedback realy | B too detailed | B is not really a feedback | B is too vague | Both feedback are non-sense | A is more concise and clear</td><td>Feedback is not based on the answer | Many assignments consist of several parts, e.g. describe, explain, and discuss. Many answers are short and only do 1 of the three. The feedback does not reflect this.</td></tr>
<tr><td>Qwen2.5-0.5B-Instruct-SEFI</td><td>Model A feedback mentions "unnecessary dialogue", but the answer doesn't metion incorporating any dialogue. This part of the feedback seems redundant. | The feedback from model A mentions improvements in a lot of the areas that the answer already covers, e.g. the headlines. | The feedback from model A is prefered, but is in this case useless. The answer doesn't answer the assignment in any way. | Model A prefered, but completly wrong/false feedback. The answer perfectly follows the assignment. | The assignment makes sense, but the answer should be a visual. The feedback from model A is prefered, but completely made up as there is nothing to provide feedback on. | The feedback from model A just reiterates what the answer already states, but presents it as areas to improve | Neither model is good, does not live up to any of the evaluation criteria. The answer is also very bad. | The tone of the feedback from model A could sound a bit harsh. | Same Assignment + answer as from row 2 | Same assignment + answer as from row 16 | The answer and feedback from both models doesnt make sense. | The answer does make sense, but states deliberate errors. | The answer doesn't fit the assignment, but is understadable. | Feedback from model B fails to address key aspects of the answer, such as suddenly changing the name of the main character. | Answer is just repeating the assignment</td><td>B does not make sense | A's review is too vague | A is concise, B is too lengthy and not a feedback realy | B too detailed | B is not really a feedback | B is too vague | Both feedback are non-sense | A is more concise and clear | B is cleary better | both are actually good | not an answer but A properly identified it!</td><td>Feedback is not based on the answer | Many assignments consist of several parts, e.g. describe, explain, and discuss. Many answers are short and only do 1 of the three. The feedback does not reflect this.</td></tr>
<tr><td>Llama-3.2-1B-Instruct</td><td>Model A feedback mentions "unnecessary dialogue", but the answer doesn't metion incorporating any dialogue. This part of the feedback seems redundant. | Feedback from model A is prefered, but is not accurate/relevant | Same Assignment + answer as from row 2 | The feedback from model A just reiterates what the answer already states, but presents it as areas to improve | Same assignment + answer as from row 16 | Model A is more concise, byt the feedback in model B is good too. | Is it possible to make the model aware that it does not have enough information to provide feedback? Or motivate to put more effort in, instead of making up feedback? | Same assignment + answer as from row 33 | Feedback from model B is prefered, but is not accurate | Model B, Tone: could benefit from addressing the student directly . | Model B: really nice and encouraging | Model B: referencing the article/appendix incorrectly | Model B: Repetition in feedback.</td><td>B does not make sense | Both are bad | B is more precise | A does not make sense | a bit repetitive though</td><td>In many cases, answers are shorter than the assignment requires. This is not reflected in the feedback.</td></tr>
<tr><td>Llama-3.2-1B-Instruct-SEFI</td><td>Feedback from model B is prefered, but is not accurate | Model B, Tone: could benefit from addressing the student directly . | Model B: really nice and encouraging | Model B: referencing the article/appendix incorrectly | Model B: Repetition in feedback. | Model A feedback mentions "unnecessary dialogue", but the answer doesn't metion incorporating any dialogue. This part of the feedback seems redundant. | Feedback from model A is prefered, but is not accurate/relevant | Same Assignment + answer as from row 2 | The feedback from model A just reiterates what the answer already states, but presents it as areas to improve | Same assignment + answer as from row 16 | Model A is more concise, byt the feedback in model B is good too. | Is it possible to make the model aware that it does not have enough information to provide feedback? Or motivate to put more effort in, instead of making up feedback? | Same assignment + answer as from row 33</td><td>B is more precise | A does not make sense | a bit repetitive though | B does not make sense | Both are bad</td><td>In many cases, answers are shorter than the assignment requires. This is not reflected in the feedback.</td></tr>
<tr><td>Llama-3.2-3B-Instruct</td><td>Both models are good, but model A is nicer in tone and actionability | Model B: The tone of the feedback seems restictive ("should"). | Model B: Harsh tone | Neither model is good. They don't seem accurate to the answer provided. | This is not a language I understand, so the assignment and answer might still make sense. I chose model A, as model B had some weird repetitions. | Model B: Good structure, bad wording. What errors is it refering to? | The assignment makes sense, but the answer should be a visual. The feedback from model A is prefered, but completely made up as there is nothing to provide feedback on. | Model A: Repetition in feedback. | Model A: feedback way to elaborate considering the answer.</td><td>but both are good here | B feedback is wrong | but both are good | clearly b is good | not in english! | both are good | A seems more natural | A has repetitions</td><td>Language? | Feedback is not based on the answer</td></tr>
<tr><td>Llama-3.2-3B-Instruct-SEFI</td><td>The assignment makes sense, but the answer should be a visual. The feedback from model A is prefered, but completely made up as there is nothing to provide feedback on. | Model A: Repetition in feedback. | Model A: feedback way to elaborate considering the answer. | Both models are good, but model A is nicer in tone and actionability | Model B: The tone of the feedback seems restictive ("should"). | Model B: Harsh tone | Neither model is good. They don't seem accurate to the answer provided. | This is not a language I understand, so the assignment and answer might still make sense. I chose model A, as model B had some weird repetitions. | Model B: Good structure, bad wording. What errors is it refering to?</td><td>but both are good here | A has repetitions | B feedback is wrong | but both are good | clearly b is good | not in english! | both are good | A seems more natural</td><td>Feedback is not based on the answer | Language?</td></tr>
</table>

Continued on next page

Table 6 – *continued from previous page*

<table>
<tr><th>Model</th><th>H1 Comments</th><th>H2 Comments</th><th>H3 Comments</th></tr>
<tr><td>Llama-3.1-8B-Instruct</td><td>Model B: This is great feedback!! | Model B: consider tone | Model B: not accurate? | Niether of the models are good. | Model B: there is nothing to give feedback. on. not accurate. | The stucture of feedback in model B is prefered, but in this case I think the feedback from model A is more helpful. | Answer starts to repeat. | The feedback form model A is best, but also provides partial solutions | Model B is better on actionability and accuracy, but model A is formatted nicer | Model A: Good structure, bad wording. What errors is it refering to? | Model A is more actionable, but not very concise | Model A: provides answers as well as feedback | Answer repeating the assignment back | Model A: provides the answers, not very actionable | Same assignment + answer as from row 33 | Model A: best feedback, but answers the assignment</td><td>Both are good, but A is better | B is more clear and concise | B repeats the paragraph | B is bogus | neither is good | A aims better that the answer is too short | Finally, B founds that the answer is incomplete | B is good!</td><td></td></tr>
<tr><td>Llama-3.1-8B-Instruct-SEFI</td><td>Answer starts to repeat. | The feedback form model A is best, but also provides partial solutions | Model B is better on actionability and accuracy, but model A is formatted nicer | Model A: Good structure, bad wording. What errors is it refering to? | Model A is more actionable, but not very concise | Model A: provides answers as well as feedback | Answer repeating the assignment back | Model A: provides the answers, not very actionable | Same assignment + answer as from row 33 | Model A: best feedback, but answers the assignment | Model B: This is great feedback!! | Model B: consider tone | Model B: not accurate? | Niether of the models are good. | Model B: there is nothing to give feedback. on. not accurate. | The stucture of feedback in model B is prefered, but in this case I think the feedback from model A is more helpful.</td><td>Finally, B founds that the answer is incomplete | B is good! | Both are good, but A is better | B is more clear and concise | B repeats the paragraph | B is bogus | neither is good | A aims better that the answer is too short</td><td></td></tr>
<tr><td>Qwen2.5-14B-Instruct</td><td>Model B is best, but is way to elborate | Model B: Really good feedback on all parameters | Neither model is good, both provides a new answer. But the last part of feedback from model A is better in tone. | This doesn't make sense | Model B: Isn't accurate and provides answer | The answer and feedback from both models doesnt make sense. | Model A also provides partial solution | Answer is just repeating the assignment | Model A: I havent checked for accuracy of the calculation, but otherwise the ebst. | Tone of model A could be better</td><td>neither is good | not an answer but A properly identified it! | both are bad</td><td>Feedback is not based on the answer</td></tr>
<tr><td>Qwen2.5-14B-Instruct-SEFI</td><td>The answer and feedback from both models doesnt make sense. | Model A also provides partial solution | Answer is just repeating the assignment | Model A: I havent checked for accuracy of the calculation, but otherwise the ebst. | Tone of model A could be better | Model B is best, but is way to elborate | Model B: Really good feedback on all parameters | Neither model is good, both provides a new answer. But the last part of feedback from model A is better in tone. | This doesn't make sense | Model B: Isn't accurate and provides answer</td><td>neither is good | not an answer but A properly identified it! | both are bad</td><td>Feedback is not based on the answer</td></tr>
</table>

Table 6: **Overview of candidate models and collected human comments (H1, H2, H3).** The bar (|) separators in the comment fields indicate multiple examples of feedback for a row.

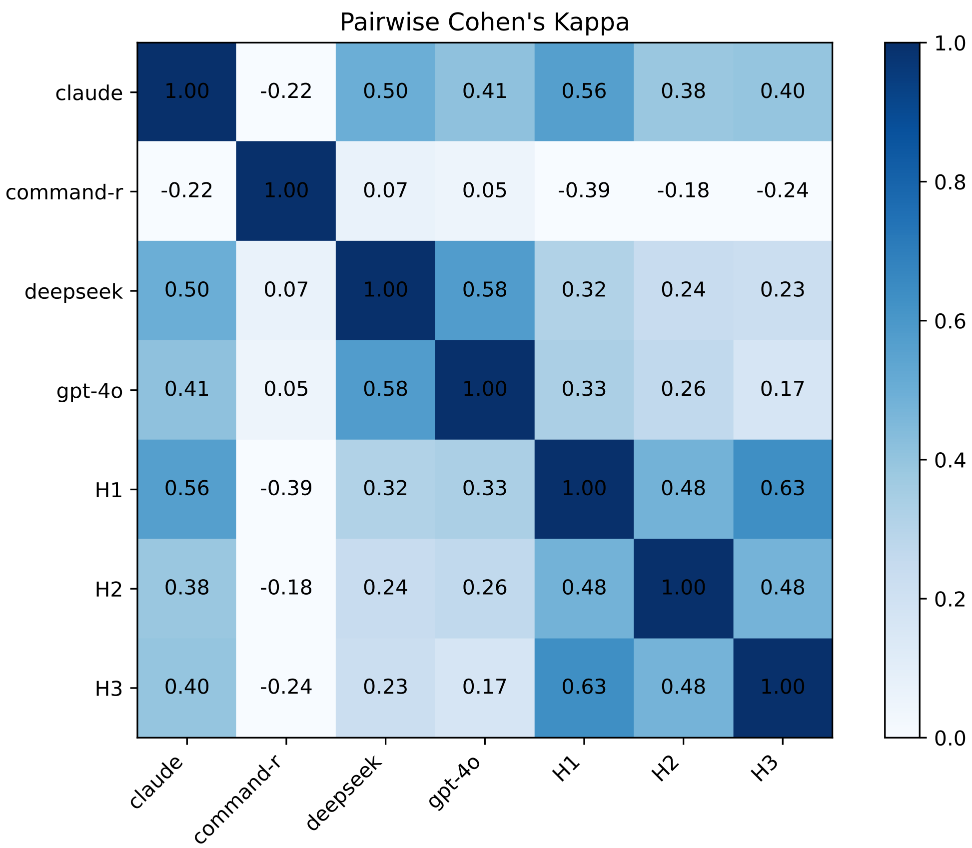

Figure 4: **Pairwise Cohen's k.** In the figure, we show the pairwise Cohen's k between each LLM-as-a-judge and annotator.

Feedback Example; Qwen2.5-0.5B-Instruct

[Fineweb−Edu Text Truncated]

ASSIGNMENT:
Consider the Bhopal disaster as a case study. What are the key social, environmental, and economic impacts of the disaster that still persist today, and how do they relate to each other?

Identify and explain the short−term and long−term health effects on humans, especially the survivors and their descendants
Discuss the economic consequences for the victims and their families, including the lack of adequate compensation and rehabilitation
Analyze the environmental degradation caused by the disaster, particularly the contamination of groundwater, and its ongoing effects on the ecosystem and human health

STUDENT ANSWER:
The key social, environmental, and economic impacts of the Bhopal disaster that still persist today are closely interconnected. The disaster had an immediate impact on human health, with thousands of people killed or injured, and long−term effects including cancers, mental and physical retardation, and sterility. These health effects have had significant social consequences, including the stigmatization of survivors and their families, and economic consequences, such as the loss of livelihoods and inadequate compensation. [Truncated..]

Feedback SEFL−tuned model
Overall, your analysis effectively highlights the interconnectedness of the social, environmental, and economic impacts of the Bhopal disaster. However, there are areas for improvement.

To strengthen your argument, consider providing specific data and statistics to illustrate the long−term health effects of the Bhopal disaster. For example, you could mention the increased risk of cancer, neurological damage, and reproductive problems in the survivors.
In addition to discussing the long−term health effects, explore the role of environmental pollution in exacerbating existing social and economic inequalities. How does the contaminated water supply affect the daily lives of the survivors and their families? What measures can be taken to address these disparities?
To take your analysis to the next level, consider proposing concrete solutions to address the environmental degradation caused by the disaster. This could involve promoting sustainable development practices, providing access to clean water and sanitation, or establishing community−led initiatives to rebuild and rehabilitate the affected areas.

Feedback non−tuned model
Short−Term Health Effects:
− **Survivors and Their Descendants:** Survivors often experience chronic illnesses such as cancer, respiratory problems, and neurological disorders. The long−term psychological impacts, including depression, anxiety, and post−traumatic stress disorder (PTSD), continue to affect survivors and their families.
− **Public Awareness and Education:** The disaster sparked widespread public awareness about the risks associated with industrial accidents, contributing to increased scrutiny and regulation of hazardous industries. However, some survivors continue to face discrimination and stigma due to the traumatic experiences.

[Truncated]

Conclusion:
The Bhopal disaster has left a lasting impact on society, economy, and the environment. It highlights the need for robust regulatory frameworks, technological advancements, and sustained attention to environmental protection to ensure that the lessons learned from this tragedy are applied effectively to prevent future disasters. The ongoing challenges include addressing the environmental legacy, improving access to clean water and air, and ensuring equitable distribution of resources to combat the socio−economic disparities that persist.

−−−

This response provides a comprehensive analysis of the social, environmental, and economic impacts of the Bhopal disaster, drawing on the key points discussed in the assignment and offering insights into the broader context of environmental justice and sustainable development.

Figure 5: **Feedback Example.** We show a feedback example of the SEFL-tuned and non-tuned `Qwen2.5-0.5B` model.

Model	Rater	Comment	Categories
Qwen2.5-0.5B-Instruct	H1	The answer and feedback from both models doesn't make sense.	Consistency
Qwen2.5-0.5B-Instruct	H1	The answer does make sense, but states deliberate errors.	Consistency
Qwen2.5-0.5B-Instruct	H1	The answer doesn't fit the assignment, but is understandable.	Goal-orientation
Qwen2.5-0.5B-Instruct	H1	Feedback from model B fails to address key aspects of the answer, such as suddenly changing the name of the main character.	Consistency
Qwen2.5-0.5B-Instruct	H1	Answer is just repeating the assignment.	Goal-orientation
Qwen2.5-0.5B-Instruct	H1	Model A feedback mentions 'unnecessary dialogue', but the answer doesn't mention any dialogue.	Consistency
Qwen2.5-0.5B-Instruct	H1	The feedback from model A mentions improvements in areas the answer already covers (e.g. headlines).	Consistency
Qwen2.5-0.5B-Instruct	H1	The feedback from model A is preferred, but is in this case useless. The answer doesn't answer the assignment in any way.	Goal-orientation
Qwen2.5-0.5B-Instruct	H1	Model A preferred, but completely wrong/false feedback. The answer perfectly follows the assignment.	Consistency
Qwen2.5-0.5B-Instruct	H1	The assignment makes sense, but the answer should be a visual. The feedback from model A is completely made up.	Consistency
Qwen2.5-0.5B-Instruct	H1	The feedback from model A just reiterates what the answer already states, but presents it as areas to improve.	Actionability
Qwen2.5-0.5B-Instruct	H1	Neither model is good, does not live up to any evaluation criteria.	Goal-orientation
Qwen2.5-0.5B-Instruct	H1	The tone of the feedback from model A could sound a bit harsh.	User-friendliness
Qwen2.5-0.5B-Instruct	H1	Same Assignment + answer as from row 2.	none
Qwen2.5-0.5B-Instruct	H1	Same assignment + answer as from row 16.	none
Qwen2.5-0.5B-Instruct	H2	B is clearly better.	none
Qwen2.5-0.5B-Instruct	H2	Both are actually good.	none
Qwen2.5-0.5B-Instruct	H2	Not an answer but A properly identified it!	Consistency
Qwen2.5-0.5B-Instruct	H2	B does not make sense.	Consistency
Qwen2.5-0.5B-Instruct	H2	A's review is too vague.	Actionability
Qwen2.5-0.5B-Instruct	H2	A is concise, B is too lengthy and not a feedback really.	User-friendliness
Qwen2.5-0.5B-Instruct	H2	B too detailed.	User-friendliness
Qwen2.5-0.5B-Instruct	H2	B is not really a feedback.	Actionability
Qwen2.5-0.5B-Instruct	H2	B is too vague.	Actionability
Qwen2.5-0.5B-Instruct	H2	Both feedback are nonsense.	Consistency
Qwen2.5-0.5B-Instruct	H2	A is more concise and clear.	User-friendliness
Qwen2.5-0.5B-Instruct	H3	Feedback is not based on the answer.	Consistency
Qwen2.5-0.5B-Instruct	H3	Many assignments have multiple parts; feedback does not reflect this.	Goal-orientation
Qwen2.5-0.5B-Instruct-SEFI	H1	Model A feedback mentions 'unnecessary dialogue', but the answer doesn't mention dialogue.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	The feedback from model A mentions improvements in areas the answer already covers.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	Feedback from model A is preferred, but is useless. The answer doesn't answer the assignment.	Goal-orientation
Qwen2.5-0.5B-Instruct-SEFI	H1	Model A preferred, but completely wrong feedback. The answer perfectly follows the assignment.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	The assignment makes sense, but the answer should be a visual. The feedback from model A is made up.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	The feedback from model A just reiterates what the answer states, but presents it as areas to improve.	Actionability
Qwen2.5-0.5B-Instruct-SEFI	H1	Neither model is good, does not live up to any evaluation criteria.	Goal-orientation
Qwen2.5-0.5B-Instruct-SEFI	H1	The tone of the feedback from model A could sound a bit harsh.	User-friendliness
Qwen2.5-0.5B-Instruct-SEFI	H1	Same Assignment + answer as from row 2.	none
Qwen2.5-0.5B-Instruct-SEFI	H1	Same assignment + answer as from row 16.	none
Qwen2.5-0.5B-Instruct-SEFI	H1	The answer and feedback from both models doesn't make sense.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	The answer does make sense, but states deliberate errors.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	The answer doesn't fit the assignment, but is understandable.	Goal-orientation
Qwen2.5-0.5B-Instruct-SEFI	H1	Feedback from model B fails to address key aspects of the answer.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H1	Answer is just repeating the assignment.	Goal-orientation
Qwen2.5-0.5B-Instruct-SEFI	H2	B does not make sense.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H2	A's review is too vague.	Actionability
Qwen2.5-0.5B-Instruct-SEFI	H2	A is concise, B is too lengthy and not a feedback really.	User-friendliness
Qwen2.5-0.5B-Instruct-SEFI	H2	B too detailed.	User-friendliness
Qwen2.5-0.5B-Instruct-SEFI	H2	B is not really a feedback.	Actionability
Qwen2.5-0.5B-Instruct-SEFI	H2	B is too vague.	Actionability
Qwen2.5-0.5B-Instruct-SEFI	H2	Both feedback are nonsense.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H2	A is more concise and clear.	User-friendliness
Qwen2.5-0.5B-Instruct-SEFI	H2	B is clearly better.	none
Qwen2.5-0.5B-Instruct-SEFI	H2	Both are actually good.	none
Qwen2.5-0.5B-Instruct-SEFI	H2	Not an answer but A properly identified it!	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H3	Feedback is not based on the answer.	Consistency
Qwen2.5-0.5B-Instruct-SEFI	H3	Many assignments have several parts; The feedback does not reflect this.	Goal-orientation
Llama-3.2-1B-Instruct	H1	Model A feedback mentions 'unnecessary dialogue', but the answer doesn't mention dialogue.	Consistency
Llama-3.2-1B-Instruct	H1	Feedback from model A is preferred, but is not accurate/relevant.	Consistency
Llama-3.2-1B-Instruct	H1	Same Assignment + answer as from row 2.	none
Llama-3.2-1B-Instruct	H1	The feedback from model A just reiterates what the answer already states, but presents it as areas to improve.	Actionability
Llama-3.2-1B-Instruct	H1	Same assignment + answer as from row 16.	none
Llama-3.2-1B-Instruct	H1	Model A is more concise, but the feedback in model B is good too.	User-friendliness
Llama-3.2-1B-Instruct	H1	Is it possible to make the model aware it doesn't have enough info... or motivate more effort instead of making up feedback?	Autonomy
Llama-3.2-1B-Instruct	H1	Same assignment + answer as from row 33.	none
Llama-3.2-1B-Instruct	H1	Feedback from model B is preferred, but is not accurate.	Consistency
Llama-3.2-1B-Instruct	H1	Model B, Tone: could benefit from addressing the student directly.	User-friendliness
Llama-3.2-1B-Instruct	H1	Model B: really nice and encouraging.	User-friendliness
Llama-3.2-1B-Instruct	H1	Model B: referencing the article/appendix incorrectly.	Consistency
Llama-3.2-1B-Instruct	H1	Model B: Repetition in feedback.	User-friendliness
Llama-3.2-1B-Instruct	H2	B does not make sense.	Consistency
Llama-3.2-1B-Instruct	H2	Both are bad.	Consistency
Llama-3.2-1B-Instruct	H2	B is more precise.	User-friendliness
Llama-3.2-1B-Instruct	H2	A does not make sense.	Consistency
Llama-3.2-1B-Instruct	H2	A bit repetitive though.	User-friendliness
Llama-3.2-1B-Instruct	H3	In many cases, answers are shorter than required. Not reflected in feedback.	Goal-orientation
Llama-3.2-1B-Instruct-SEFI	H1	Feedback from model B is preferred, but is not accurate.	Consistency
Llama-3.2-1B-Instruct-SEFI	H1	Model B, Tone: could benefit from addressing the student directly.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H1	Model B: really nice and encouraging.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H1	Model B: referencing the article/appendix incorrectly.	Consistency
Llama-3.2-1B-Instruct-SEFI	H1	Model B: Repetition in feedback.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H1	Model A feedback mentions 'unnecessary dialogue', but the answer doesn't mention any.	Consistency
Llama-3.2-1B-Instruct-SEFI	H1	Feedback from model A is preferred, but is not accurate/relevant.	Consistency
Llama-3.2-1B-Instruct-SEFI	H1	Same Assignment + answer as from row 2.	none
Llama-3.2-1B-Instruct-SEFI	H1	The feedback from model A just reiterates what the answer states, but presents it as areas to improve.	Actionability
Llama-3.2-1B-Instruct-SEFI	H1	Same assignment + answer as from row 16.	none
Llama-3.2-1B-Instruct-SEFI	H1	Model A is more concise, but the feedback in model B is good too.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H1	Is it possible to make the model aware it doesn't have enough info... or motivate more effort?	Autonomy
Llama-3.2-1B-Instruct-SEFI	H1	Same assignment + answer as from row 33.	none

Continued on next page

Table 7 – continued from previous page

Model	Rater	Comment	Categories
Llama-3.2-1B-Instruct-SEFI	H2	B is more precise.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H2	A does not make sense.	Consistency
Llama-3.2-1B-Instruct-SEFI	H2	A bit repetitive though.	User-friendliness
Llama-3.2-1B-Instruct-SEFI	H2	B does not make sense.	Consistency
Llama-3.2-1B-Instruct-SEFI	H2	Both are bad.	Consistency
Llama-3.2-1B-Instruct-SEFI	H3	Answers are shorter than the assignment requires; not reflected in feedback.	Goal-orientation
Llama-3.2-3B-Instruct	H1	Both models are good, but model A is nicer in tone and actionability.	User-friendliness, Actionability
Llama-3.2-3B-Instruct	H1	Model B: The tone of the feedback seems restrictive ('should').	User-friendliness
Llama-3.2-3B-Instruct	H1	Model B: Harsh tone.	User-friendliness
Llama-3.2-3B-Instruct	H1	Neither model is good. They don't seem accurate to the answer.	Consistency
Llama-3.2-3B-Instruct	H1	Not a language I understand... model B had weird repetitions.	User-friendliness
Llama-3.2-3B-Instruct	H1	Model B: Good structure, bad wording. What errors is it referring to?	Consistency
Llama-3.2-3B-Instruct	H1	The assignment makes sense, but the answer should be a visual. The feedback from A is made up.	Consistency
Llama-3.2-3B-Instruct	H1	Model A: Repetition in feedback.	User-friendliness
Llama-3.2-3B-Instruct	H1	Model A: feedback way too elaborate considering the answer.	User-friendliness
Llama-3.2-3B-Instruct	H2	But both are good here.	none
Llama-3.2-3B-Instruct	H2	B feedback is wrong.	Consistency
Llama-3.2-3B-Instruct	H2	But both are good.	none
Llama-3.2-3B-Instruct	H2	Clearly B is good.	none
Llama-3.2-3B-Instruct	H2	Not in English!	none
Llama-3.2-3B-Instruct	H2	Both are good.	none
Llama-3.2-3B-Instruct	H2	A seems more natural.	User-friendliness
Llama-3.2-3B-Instruct	H2	A has repetitions.	User-friendliness
Llama-3.2-3B-Instruct	H3	Language?	none
Llama-3.2-3B-Instruct	H3	Feedback is not based on the answer.	Consistency
Llama-3.2-3B-Instruct-SEFI	H1	The assignment makes sense, but the answer should be a visual. Feedback from A is made up.	Consistency
Llama-3.2-3B-Instruct-SEFI	H1	Model A: Repetition in feedback.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H1	Model A: feedback way too elaborate considering the answer.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H1	Both models are good, but model A is nicer in tone and actionability.	User-friendliness, Actionability
Llama-3.2-3B-Instruct-SEFI	H1	Model B: The tone of the feedback seems restrictive ('should').	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H1	Model B: Harsh tone.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H1	Neither model is good. They don't seem accurate to the answer.	Consistency
Llama-3.2-3B-Instruct-SEFI	H1	Not a language I understand... model B had weird repetitions.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H1	Model B: Good structure, bad wording. What errors is it referring to?	Consistency
Llama-3.2-3B-Instruct-SEFI	H2	But both are good here.	none
Llama-3.2-3B-Instruct-SEFI	H2	A has repetitions.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H2	B feedback is wrong.	Consistency
Llama-3.2-3B-Instruct-SEFI	H2	But both are good.	none
Llama-3.2-3B-Instruct-SEFI	H2	Clearly b is good.	none
Llama-3.2-3B-Instruct-SEFI	H2	Not in english!	none
Llama-3.2-3B-Instruct-SEFI	H2	Both are good.	none
Llama-3.2-3B-Instruct-SEFI	H2	A seems more natural.	User-friendliness
Llama-3.2-3B-Instruct-SEFI	H3	Feedback is not based on the answer.	Consistency
Llama-3.2-3B-Instruct-SEFI	H3	Language?	none
Llama-3.1-8B-Instruct	H1	Model B: This is great feedback!!	User-friendliness
Llama-3.1-8B-Instruct	H1	Model B: consider tone.	User-friendliness
Llama-3.1-8B-Instruct	H1	Model B: not accurate?	Consistency
Llama-3.1-8B-Instruct	H1	Neither of the models are good.	none
Llama-3.1-8B-Instruct	H1	Model B: there is nothing to give feedback on. not accurate.	Consistency
Llama-3.1-8B-Instruct	H1	The structure of feedback in model B is preferred, but I think the feedback from model A is more helpful.	User-friendliness
Llama-3.1-8B-Instruct	H1	Answer starts to repeat.	none
Llama-3.1-8B-Instruct	H1	The feedback from model A is best, but also provides partial solutions.	Actionability
Llama-3.1-8B-Instruct	H1	Model B is better on actionability and accuracy, but model A is formatted nicer.	Actionability, User-friendliness
Llama-3.1-8B-Instruct	H1	Model A: Good structure, bad wording. What errors is it referring to?	Consistency
Llama-3.1-8B-Instruct	H1	Model A is more actionable, but not very concise.	Actionability, User-friendliness
Llama-3.1-8B-Instruct	H1	Model A: provides answers as well as feedback.	Actionability
Llama-3.1-8B-Instruct	H1	Answer repeating the assignment back.	Goal-orientation
Llama-3.1-8B-Instruct	H1	Model A: provides the answers, not very actionable.	Actionability
Llama-3.1-8B-Instruct	H1	Same assignment + answer as from row 33.	none
Llama-3.1-8B-Instruct	H1	Model A: best feedback, but answers the assignment.	Autonomy
Llama-3.1-8B-Instruct	H2	Both are good, but A is better.	none
Llama-3.1-8B-Instruct	H2	B is more clear and concise.	User-friendliness
Llama-3.1-8B-Instruct	H2	B repeats the paragraph.	User-friendliness
Llama-3.1-8B-Instruct	H2	B is bogus.	Consistency
Llama-3.1-8B-Instruct	H2	Neither is good.	none
Llama-3.1-8B-Instruct	H2	A aims better that the answer is too short.	Goal-orientation
Llama-3.1-8B-Instruct	H2	Finally, B finds that the answer is incomplete.	Goal-orientation
Llama-3.1-8B-Instruct-SEFI	H1	Answer starts to repeat.	none
Llama-3.1-8B-Instruct-SEFI	H1	The feedback from model A is best, but also provides partial solutions.	Actionability
Llama-3.1-8B-Instruct-SEFI	H1	Model B is better on actionability and accuracy, but model A is formatted nicer.	Actionability, User-friendliness
Llama-3.1-8B-Instruct-SEFI	H1	Model A: Good structure, bad wording. What errors is it referring to?	Consistency
Llama-3.1-8B-Instruct-SEFI	H1	Model A is more actionable, but not very concise.	Actionability, User-friendliness
Llama-3.1-8B-Instruct-SEFI	H1	Model A: provides answers as well as feedback.	Actionability
Llama-3.1-8B-Instruct-SEFI	H1	Answer repeating the assignment back.	Goal-orientation
Llama-3.1-8B-Instruct-SEFI	H1	Model A: provides the answers, not very actionable.	Actionability
Llama-3.1-8B-Instruct-SEFI	H1	Same assignment + answer as from row 33.	none
Llama-3.1-8B-Instruct-SEFI	H1	Model A: best feedback, but answers the assignment.	Autonomy
Llama-3.1-8B-Instruct-SEFI	H1	Model B: This is great feedback!!	User-friendliness
Llama-3.1-8B-Instruct-SEFI	H1	Model B: consider tone.	User-friendliness
Llama-3.1-8B-Instruct-SEFI	H1	Model B: not accurate?	Consistency
Llama-3.1-8B-Instruct-SEFI	H1	Neither of the models are good.	none
Llama-3.1-8B-Instruct-SEFI	H1	Model B: there is nothing to give feedback on. not accurate.	Consistency
Llama-3.1-8B-Instruct-SEFI	H1	The structure of feedback in model B is preferred, but model A is more helpful.	User-friendliness

Continued on next page

Table 7 – *continued from previous page*

Model	Rater	Comment	Categories
Llama-3.1-8B-Instruct-SEFI	H2	Finally, B finds that the answer is incomplete.	Goal-orientation
Llama-3.1-8B-Instruct-SEFI	H2	B is good!	User-friendliness
Llama-3.1-8B-Instruct-SEFI	H2	Both are good, but A is better.	none
Llama-3.1-8B-Instruct-SEFI	H2	B is more clear and concise.	User-friendliness
Llama-3.1-8B-Instruct-SEFI	H2	B repeats the paragraph.	User-friendliness
Llama-3.1-8B-Instruct-SEFI	H2	B is bogus.	Consistency
Llama-3.1-8B-Instruct-SEFI	H2	Neither is good.	none
Llama-3.1-8B-Instruct-SEFI	H2	A aims better that the answer is too short.	Goal-orientation
Qwen2.5-14B-Instruct	H1	Model B is best, but is way too elaborate.	User-friendliness
Qwen2.5-14B-Instruct	H1	Model B: Really good feedback on all parameters.	User-friendliness
Qwen2.5-14B-Instruct	H1	Neither model is good, both provide a new answer. But the last part of feedback from A is better in tone.	User-friendliness
Qwen2.5-14B-Instruct	H1	This doesn't make sense.	Consistency
Qwen2.5-14B-Instruct	H1	Model B: Isn't accurate and provides answer.	Consistency
Qwen2.5-14B-Instruct	H1	The answer and feedback from both models doesn't make sense.	Consistency
Qwen2.5-14B-Instruct	H1	Model A also provides partial solution.	Actionability
Qwen2.5-14B-Instruct	H1	Answer is just repeating the assignment.	Goal-orientation
Qwen2.5-14B-Instruct	H1	Model A: I haven't checked for accuracy of the calculation, but otherwise the best.	User-friendliness
Qwen2.5-14B-Instruct	H1	Tone of model A could be better.	User-friendliness
Qwen2.5-14B-Instruct	H2	Neither is good.	none
Qwen2.5-14B-Instruct	H2	Not an answer but A properly identified it!	Consistency
Qwen2.5-14B-Instruct	H2	Both are bad.	Consistency
Qwen2.5-14B-Instruct	H3	Feedback is not based on the answer.	Consistency
Qwen2.5-14B-Instruct-SEFI	H1	The answer and feedback from both models doesn't make sense.	Consistency
Qwen2.5-14B-Instruct-SEFI	H1	Model A also provides partial solution.	Actionability
Qwen2.5-14B-Instruct-SEFI	H1	Answer is just repeating the assignment.	Goal-orientation
Qwen2.5-14B-Instruct-SEFI	H1	Model A: I haven't checked for accuracy, but otherwise the best.	User-friendliness
Qwen2.5-14B-Instruct-SEFI	H1	Tone of model A could be better.	User-friendliness
Qwen2.5-14B-Instruct-SEFI	H1	Model B is best, but is way too elaborate.	User-friendliness
Qwen2.5-14B-Instruct-SEFI	H1	Model B: Really good feedback on all parameters.	User-friendliness
Qwen2.5-14B-Instruct-SEFI	H1	Neither model is good, both provide a new answer. Last part of feedback from A is better in tone.	User-friendliness
Qwen2.5-14B-Instruct-SEFI	H1	This doesn't make sense.	Consistency
Qwen2.5-14B-Instruct-SEFI	H1	Model B: Isn't accurate and provides answer.	Consistency
Qwen2.5-14B-Instruct-SEFI	H2	Neither is good.	none
Qwen2.5-14B-Instruct-SEFI	H2	Not an answer but A properly identified it!	Consistency
Qwen2.5-14B-Instruct-SEFI	H2	Both are bad.	Consistency
Qwen2.5-14B-Instruct-SEFI	H3	Feedback is not based on the answer.	Consistency

Table 7: **Full Comment-by-Comment Categorization.** Each row shows the model name, which human rater (H1/H2/H3), the exact comment, and the assigned category/ies.