arXiv:2502.12927v3 [cs.CL] 24 Feb 2026

SEFL: A Framework for Generating Synthetic Educational Assignment Feedback with LLM Agents

Mike Zhang[‡†◇], **Amalie Pernille Dilling**[†], **Léon Gondelman**[†]
Niels Erik Ruan Lyngdorf[†], **Euan D. Lindsay**[†], **Johannes Bjerva**[†]
[‡]University of Copenhagen, Denmark
[†]Aalborg University, Denmark
[◇]Pioneer Centre for AI, Denmark
mike.zhang@di.ku.dk

Abstract

Providing high-quality feedback on student assignments is crucial for student success, but it is heavily limited by time and budgetary constraints. In this work, we introduce **S**ynthetic **E**ducational **F**eedback **L**oops (SEFL), a synthetic data framework designed to generate data that resembles immediate, on-demand feedback at scale without relying on extensive, real-world student assignments and teacher feedback. To obtain this type of data, two large language models (LLMs) operate in a teacher-student role to simulate assignment completion and formative feedback, generating 19.8K synthetic pairs of student work and corresponding critiques and actionable improvements from a teacher. With this data, we fine-tune smaller, more computationally efficient LLMs on these synthetic pairs, enabling them to replicate key features of high-quality, goal-oriented feedback. Through comprehensive evaluations with three LLM judges and three human experts, across a subset of 900 outputs, we demonstrate that SEFL-tuned models outperform both their untuned counterparts and an existing baseline in terms of feedback quality. The potential for societal impact is reinforced by extensive qualitative comments and ratings from human stakeholders — both students and higher education instructors. SEFL has the potential to transform feedback processes for higher education and beyond.

Keywords: Synthetic Data, Education, Large Language Model, Agents

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 is 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), in addition to being a labor-intensive task. Advances in language technology offer opportunities to automate and augment higher education feedback processes, addressing these limitations.

In particular, 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., 2024, 2025; Wang et al., 2024a). Yet, automating teacher-student assignment feedback with LLMs 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 assessment feedback?*

Here, we introduce **S**ynthetic **E**ducational **F**eedback **L**oops (SEFL), a framework that generates synthetic teacher-student interactions using LLM agents. In this framework, two LLMs, one

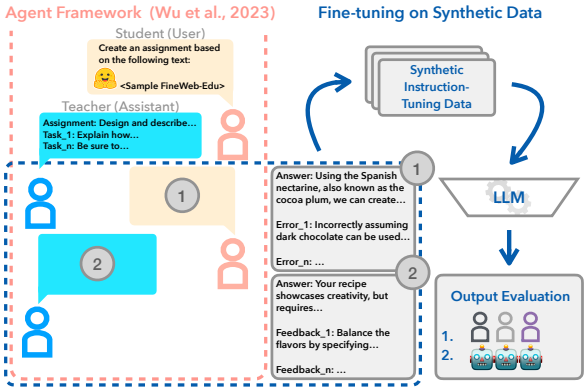


Figure 1: **SEFL Synthetic Data Generation Setup.** We use a two-agent framework (Wu et al., 2023) with LLMs acting as a Student and Teacher. The Teacher creates assignments from Fineweb-Edu (Lozhkov et al., 2024), a dataset curated using LLMs to judge the educational value of web pages. Overall, the Student responds with explicit errors (via prompting), and finally, the Teacher addresses each mistake. This synthetic interaction data is then used to fine-tune multiple LLMs, whose performance is measured through human ratings and evaluations by LLMs-as-judges.

acting as the teacher and the other as the student, simulate *formative* feedback workflows (Conole and Oliver, 2006; Nicol, 2007). This synthetic data from the agents is then used to fine-tune smaller autoregressive language models, resulting in better feedback models that enable the development of scalable educational feedback systems. These systems can operate efficiently on the modest computational infrastructure available in higher education institutions, without requiring access to privacy-sensitive data.

We show that: (i) Empirically, SEFL-tuned models outperform their non-tuned versions in win rate evaluations by three LLM judges and three human experts in giving assessment feedback. (ii) Through empirical and qualitative analysis of expert annotator comments, we find that larger models tend to provide more actionable, goal-oriented, and user-friendly feedback. They are also more consistent and better at supporting student autonomy. (iii) We observe strong agreement among human evaluators regarding feedback quality. (iv) We compare our approach to an existing baseline (Book2Dial; Wang et al., 2024a) and show that SEFL-tuned models provide better feedback according to all LLM judges.

Contributions. We contribute the following:

- **SEFL**: An agentic framework that simulates teacher-student feedback loops with paired language model agents.
- 19,841 synthetic assignment–feedback pairs generated by SEFL to fine-tune smaller language models.
- A comprehensive set of human and LLM evaluations that outline the strengths and limitations of SEFL with extensive qualitative analysis of the feedback provided by human experts.
- Open-sourcing of all models, code, and data. [1]

2. Related Work

NLP & Education. Large language models are now supporting a broad spectrum of educational tasks. In automated grading, they score short answers, essays, and even programming assignments with accuracy that approaches expert instructors, which eases faculty workload and releases time for mentoring (Ke and Ng, 2019; Ramesh and Sanampudi, 2022; Stahl et al., 2024). For personalized tutoring, conversational agents powered by these models adapt explanations, hints, and examples to each learner's background knowledge and preferred style, producing measurable gains in engagement and achievement (Yun et al., 2024; Liu et al., 2024c; Rooein and Hovy, 2024; Ross and Andreas, 2024; Kwon et al., 2024; Zhang et al., 2024). Research on peer learning shows that the same technology can mediate small-group discussions, suggest prompts, and highlight diverse viewpoints, leading to richer collaboration (Bauer et al., 2023). In mathematics, aligning word problems and proofs to grade-level objectives has been automated, yielding encouraging results and reducing the manual effort required to curate question banks (Botelho et al., 2023). Critical thinking curricula also benefit; LLMs can challenge students to justify claims, detect fallacies, and refine arguments in real-time (Guerraoui et al., 2023). The models have also begun to assist scholars: studies report successful use for screening literature, summarizing drafts, and aligning reviewer comments with revision plans (Liang et al., 2024; Sonkar et al., 2024). Complementing these functions are analytics tools that track learning trajectories and surface early warnings when a student slips behind (Schwarz et al., 2018; Aslan et al., 2019; Alrajhi et al., 2021).

Despite this growing body of work, prior studies have rarely targeted the systematic generation of feedback on open-ended student submissions and assignments. In this study, we address this gap by using LLMs to generate extensive comment sets that teachers can accept as is or adapt. Decades of scholarship define effective feedback as goal-oriented, actionable, timely, user-friendly, and consistent while fostering self-evaluation (Carless et al., 2011; Wiggins, 2012). Feedback that is long and verbose can confuse learners. Therefore, concise wording is usually preferable, and comments that arrive soon after the original effort drive steady improvement (Wiggins, 2012). By producing immediate responses and tailoring suggestions to rubric criteria, LLM-based systems stand well-positioned to satisfy these guidelines while operating at the classroom and institution scale.

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) employs cooperative role-based dialogues to achieve shared objectives, whereas SimSeek (Kim et al., 2022) utilizes agent-based conversations to construct 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 text-

[1] Resources and Supplementary Material such as full prompts and example outputs can be found at https://github.com/jjzha/sefl and https://tinyurl.com/3zdu847k (HuggingFace).

books. *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.

3. SEFL: Synthetic Educational Feedback Loops

3.1. Synthetic Data Generation

We employ a two-agent framework (Wu et al., 2023) to simulate a student-teacher feedback loop, as seen in higher education. 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 requests an assignment using Fineweb-Edu (Lozhkov et al., 2024) texts (Figure 1), which is known for its educational content based on LLM judgments. Second, the teacher-agent creates an assignment that can be of any domain, e.g., STEM, social sciences, puzzles, and so forth. Then, the student-agent (①) submits assignments containing several explicit errors, and the teacher-agent (②) provides feedback. In ③ and ④, we obtain an assignment-feedback pair. We deliberately avoided imposing strict rules on the types of errors that could be generated. We hypothesize that allowing the LLM to introduce errors naturally results in a more diverse and realistic set of mistakes than if we had a fixed checklist. We investigate both `Qwen2.5-72B` and `Llama-3.1-70B` for interactions. For a control check, we generate 5,000 interaction tuples with each model, and validate the output as an initial step to investigate the quality of initial feedback.

We show in Table 1 the results of this experiment. Out of 5,000 generated examples, `Llama-3.1-70B` generates 2,513 valid examples (i.e., valid `JSON` format and each feedback point 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 relates to each other.[3] We show that, regardless of `Llama-3.1-70B` generating more valid examples, the BERTScore (0.877) remains in a similar range to `Qwen2.5-72B` (0.919); in both cases, this indicates a high level of similarity. As a final qualitative check, we experimented with several prompts and consolidated them into a single, final prompt, which is included in the supplementary material. Finally, we use `Llama-3.1-70B`-generated data as the basis for all subsequent model fine-tuning, as it generated more valid examples. We spot-checked several prompts and consolidated the final full prompt in Figure 6 (Section A).

	Valid (/ 5,000)	BERTScore
`Llama-3.1-70B`	2,513	0.877
`Qwen2.5-72B`	454	0.919

Table 1: **Generation Capabilities.** First, We show the number of valid examples, measured by correct `JSON` format and whether each feedback refers to an error. `Llama-3.1-70B` generates more valid examples. Second, we measure BERTScore as a proxy for relatedness between error–feedback pairs of the valid generations.

Data statistics. After generating 5,000 examples, we continue to create example pairs, resulting in 19,841 teacher-student feedback pairs. In Table 3, we present the final dataset statistics. We underline that the generation lengths for each agent are intentionally kept concise (<170 subword tokens), based on the hypothesis that overly lengthy feedback may be counterproductive. This is in line with observations from Ferguson (2011), who finds that students tend to favor brief comments. We argue that balancing supportive and critical feedback is crucial, as LLMs often produce verbose responses by default, which can influence the preferences of both humans and LMs (Saito et al., 2023).

Task Errors. Once all data was generated, we took a random subsample of 226 instances with 428 explicit errors. We investigated what category of *intentional* errors the student agent generates. In total, we identified seven error types from the 428 errors and show the proportion in Table 2. Most errors are of type "Omission and/or incompleteness" (surmounted for 50% of the errors):

1. **Omission and/or incompleteness:** This includes missing details, lacking examples or evidence, not exploring a topic deeply enough, or failing to address all parts of the assignment.
2. **Stylistic and/or formatting issues:** Problems with grammar, spelling, punctuation, tone (e.g., too informal), word count, and incorrect formatting (like improper citations).
3. **Factual inaccuracy:** This includes incorrect dates, numbers, names, scientific facts, historical events, or misquoting a source text.
4. **Logical flaws and/or weak argumentation:** These are errors in reasoning. They include

[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.

making incorrect assumptions, oversimplifying complex topics, drawing faulty conclusions, or failing to construct a coherent and well-supported argument.

5. **Technical and/or procedural errors:** This category is for mistakes in technical execution, such as incorrect mathematical calculations, flawed experimental design, or errors in code.
6. **Structural and organizational problems:** These errors relate to the overall structure and flow of the response, such as a lack of clear transitions between paragraphs, poor organization, or a missing introduction/conclusion.
7. **Task misinterpretation:** Student agent misunderstands the core requirement of the task, such as writing a story instead of an essay or failing to use a requested framework.

4. Methodology

4.1. Fine-Tuning Large Language Models

We divide the data into 17,856 training examples and 1,985 validation examples. To test the feedback quality of our approach across model scales, we fine-tune five open-weight models, namely `Qwen2.5-0.5B`, `Llama-3.2-1B`, `Llama-3.2-3B`, `Llama-3.1-8B`, and `Qwen2.5-14B` on this synthetic dataset. The compute we train the models on is AMD Radeon Instinct MI250X GPUs, and it took a total of 467 GPU hours.

Training Objective. To fine-tune the LLMs, for each prompt x and target sequence $y = (y_1, .., y_T)$ we minimize the token-level cross-entropy

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{T} m_t \, \log p_\theta\big(y_t \mid y_{<t}, x\big), \qquad (1)$$

where m_t masks out the prompt tokens and activates the loss on reference tokens only.

Optimization Details. All models train for three epochs with a global batch size of 16 and context lengths of 131K for Qwen2.5 and 128K for the Llama variants. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$ (Loshchilov and Hutter, 2019), $\epsilon = 10^{-8}$, weight decay 0.1, and gradient clipping at norm 1.0. The learning rate peaks at 2×10^{-5} after a linear warm-up covering the first five percent of steps and then follows a linear decay.

4.2. Multi-faceted Evaluation

4.2.1. Human Evaluation

After fine-tuning the language models with the assignment-feedback pairs and testing the performance of SEFL, we conduct a human evaluation on a total of 450 instances with three annotators, following a similar approach to SoctraticLM (Liu et al., 2024b), which annotated around 1,000 dialogues with 10 annotators. We randomly sample the subset of instances from the validation set. For each item, both the original instruction tuned model (A) and the model further fine-tuned with SEFL (B) produce feedback. Three human experts compared pairs of feedback responses produced for the same assignment and answer. For each item they read the original prompt, the student submission, and the two candidate feedback drafts from the non-tuned and SEFL-tuned model. Then, they select the feedback from model A or B that is better based on four base criteria:

- **Accuracy:** The generated feedback text focuses on concrete strengths and weaknesses in the student's work, avoiding superficial remarks.
- **Actionability:** Suggestions are clear, specific, and realistic for a student to apply.
- **Conciseness:** Wording is brief and focused, with little repetition.
- **Tone:** Language stays constructive and professional while recognizing good elements.

Raters were reminded to value efficiency over length, to prefer targeted advice over general principles, and to ignore formatting tricks. They recorded their choice as `A` or `B` and could leave an optional free-text comment. With this, we calculate the win rate (i.e., the percentage of choosing one feedback text over the other). This has become a de facto standard to evaluate long-form text against each other (e.g., Rafailov et al., 2023).

Each row took at most ten minutes, and the guidelines stressed taking regular breaks to sustain attention. We deliberately remove the $A = B$ tie option because a forced choice provides more informative labels and reduces hesitation. At the same time, a separate checkbox still allows raters to mark assignment $\rightarrow$ answer $\rightarrow$ feedback tuples as unrelated or nonsensical. This happened around 12% of the time, especially in the smaller, less capable models. The full annotation guidelines are reported in the supplementary material (Table 7; Section B).

Annotator Demographics. Our raters are aged 20-40 and from Europe. One identifies as female and the other two identify as male. One female and one male have a background in Computer Science,

Error Category	Count	%	Example Error Explanation from Student Agent
Omission and/or Incompleteness	214	50.0%	"Failure to provide specific examples and evidence to support claims".
Stylistic and/or Formatting Issues	67	15.7%	"Incorrect spelling of the word 'protection' in the last sentence of the answer".
Factual Inaccuracy	59	13.8%	"Incorrectly claimed Ibn al-Haytham was born in Egypt instead of Basra, Iraq".
Logical Flaws and/or Weak Argumentation	53	12.4%	"The statement about DSP chips having limitations is an oversimplification, as they are still widely used for various applications".
Technical and/or Procedural Errors	16	3.7%	"Incorrect calculation of the target waiting time, should be a 30% reduction from 10 minutes".
Structural and/or Organizational Problems	12	2.8%	"Lack of clear structure and organization in the comparative analysis".
Task Misinterpretation	7	1.6%	"The writing deploys facts that were not mentioned in the letter, like United States' interest in containing the spread of communism".
Total	**428**	**100.0%**	—

Table 2: **Statistics of Explicit Error Types.** We show the explicit error types. We show that omission and incompleteness are proportionally the most used *intentional* errors the student agent generates.

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

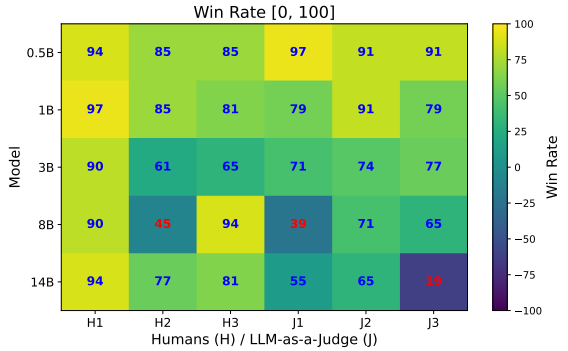


Figure 2: **Win Rate Results.** We show the win rate of our *SEFL-tuned models*. A win rate >50% indicates that SEFL-tuned models are better in giving feedback than their vanilla counterpart; in red everything <50% shows the opposite. We show results of 3 human annotators (H#) and 4 LLM judges: `gpt-4o`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

and one male has a background in Engineering Education. All have extensive experience in teaching and supervision, or being taught and supervised; they all work in higher education (at various levels, such as research assistant and assistant professors) with near-native English proficiency.

4.2.2. LLM-as-a-Judge

We also evaluate the fine-tuned models' output using a LLM-as-a-judge framework, a method gaining traction for evaluating free 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 three LLMs rate the same 450 random instances, namely GPT-4o (Hurst et al., 2024), Claude3.5-Sonnet, and DeepSeek-V3 (Liu et al., 2024a). For the LLM-as-a-judge experiments using closed-source models, we used the respective APIs, and the total cost was approximately 10 USD.

For every example, the judge model receives the assignment prompt together with the two candidate feedback drafts. It is asked to decide which draft is better on the same four base criteria as the human annotators. The instruction forbids numeric grades or explanations and requires the judge to output exactly one character, `A` or `B`, producing a clean pairwise preference label.

5. Results

In Figure 2, we show the *win rates* of models fine-tuned with SEFL vs. their non-tuned version, 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 4, where we depict the abridged prompt and feedback by a tuned and non-tuned model.

Human Assessment. Overall, human rater evaluations in Figure 2 (H#) show that the SEFL-tuned models often achieve high win rates, surpassing 90% in several cases compared to the smaller mod-

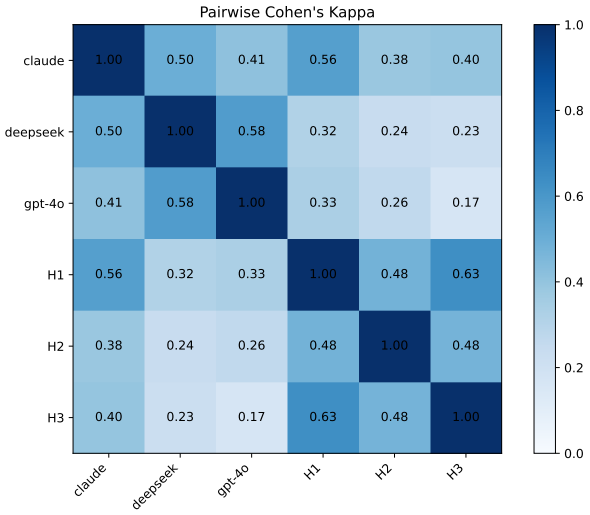


Figure 3: **Pairwise Cohen's k.** We show the pairwise Cohen's k between each LLM judge and annotator.

els. The human 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 positive contextual relevance.

LLM-as-a-Judge Evaluation. For the LLM-as-a-judge evaluations (J#) in the same figure, we observe several 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 three LLM judges (J1: GPT-4o, J2: Claude-3.5-Sonnet, J3: 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 three judges, indicating a consistent preference for the fine-tuned version. As the parameter size increases, the difference in win rate decreases (e.g., 58% on average for 8B). The full judge prompt will be released in the supplementary material.

[4]Models are picked based on their performance on RewardBench (Lambert et al., 2024), JudgeBench (Tan et al., 2024), and JudgeArena (AtlaAI, 2025).

Human and Model Agreements. In Figure 3, we present the pairwise Cohen's k values (Cohen, 1960) computed between each LLM judge and the human rater, to observe whether humans and LLM judges agree on which model gives better feedback. The agreement among humans was moderate to substantial: H1 and H3 reached $\kappa = 0.63$, H1 and H2 0.48, and H2 and H3 0.48 (Landis and Koch, 1977). Among the models, Claude aligns most closely with both the other judges and the humans; Deepseek follows, and GPT-4o shows the weakest match. Across all model and human pairs, the numbers range from 0.17 to 0.58, which shows the subjectivity of feedback evaluation. This suggests that human experts generally agree on the quality of feedback, but for LLM judges, this may vary. Overall, we see an opportunity for improvements in terms of agreement between LLMs and humans regarding feedback quality.

6. Discussion

6.1. Human Qualitative Insights

In addition to the win rates in Figure 2, our human annotators provided rich qualitative feedback on the model outputs, which we show in Table 4. Generally, on the critical side, they noted that if a student's 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. Overall, although `Qwen2.5-14B` achieved high win rates (94, 77, 81 across three annotators), these insights suggest that even top-performing models could improve in error detection, tone refinement, and contextual sensitivity.

To further quantify, in Figure 5, we plot for each model the net balance of optional rater comments in five qualitative categories: Actionability (AC), Goal Orientation (GO), User Friendliness (UF), Consistency (CO), and Student Autonomy (AY) (Carless et al., 2011; Wiggins, 2012). Squares denote the base models and circles the SEFL-tuned variants. Horizontal whiskers provide 95% confidence intervals for the net balance using the bootstrap method. We compute these intervals on the proportion of positive remarks and then transform them to the net scale via $b = n\,(2p - 1)$, where n is the total number of comments and p the positive proportion. Annotators added comments only when a response stood out, so the plot reveals both the direction and the strength of impressions. We show several ex-

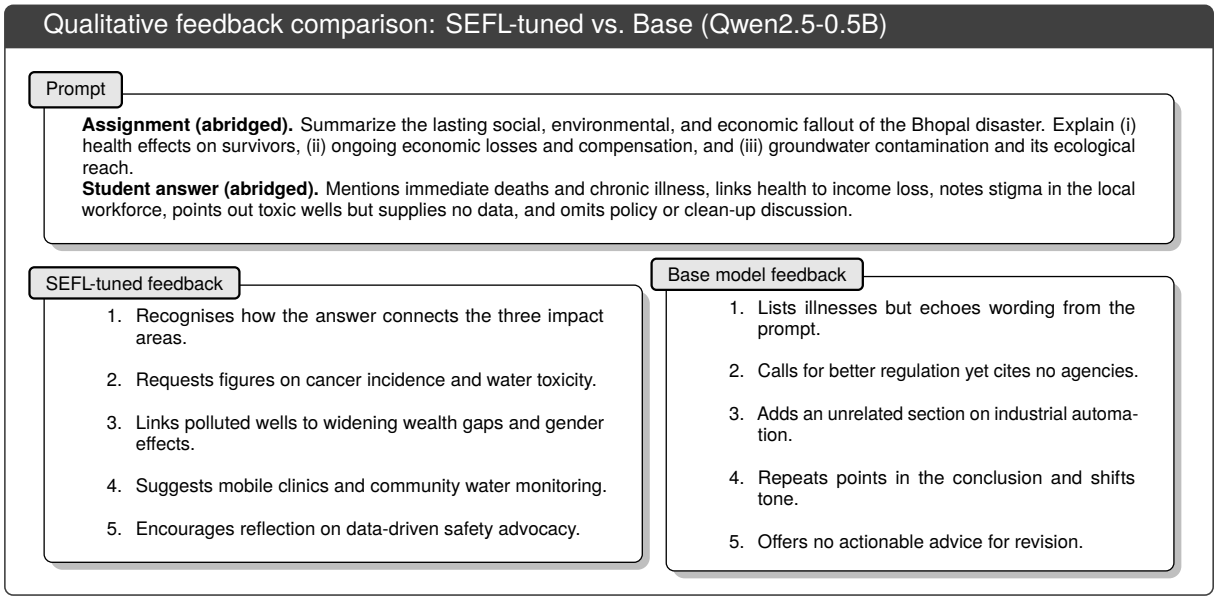


Figure 4: **Qualitative Example of Feedback.** Excerpt that shows how SEFL improves specificity and actionability. Full conversation will be added as supplementary material.

Model	H1 sample	H2 sample	H3 sample
0.5B	"The feedback repeats the assignment rather than pointing out issues."	"A is concise but misses important detail."	"Comments do not reflect missing parts of the answer."
0.5B-SEFL	"Feedback spots deliberate errors and offers clear fixes."	"A is concise and clear, suggestions align with the rubric."	"Advice notes that the discuss part is missing and proposes adding it."
1B	"Feedback echoes what the answer already states."	"More precise."	"Comments ignore omitted sections."
1B-SEFL	"The note opens with strengths and gives concrete next steps."	"Hints link directly to rubric points."	"Advice adds two practical examples to guide revision."
3B	"Both models are good, A sounds nicer."	"Both are good here."	"Feedback covers content but wording feels harsh."
3B-SEFL	"Feedback keeps an encouraging tone while giving actionable points."	"Structure and tone feel balanced."	"Comments cover content, style, and structure in one cohesive note."
7B	"Feedback addresses surface errors but misses deeper reasoning."	"Detailed but drifts off the prompt."	"Overlooks the rubric item about evidence."
7B-SEFL	"Points out reasoning gaps and proposes specific fixes."	"Stays focused on the prompt and remains concise."	"Directly cites the missing evidence section."
14B	"Feedback is thorough but wording is opaque."	"Long sentences reduce clarity."	"Some comments repeat earlier points."
14B-SEFL	"Clear and student-friendly while retaining depth."	"Uses short sentences and connects advice to the rubric."	"Groups comments by theme, which avoids repetition."

Table 4: **Several Examples of Human Feedback.** We select several human annotator remarks that illustrate how SEFL tuning improves feedback quality compared with the original models.

amples of annotations in Table 5. The annotators also had the option to reject nonsensical feedback, which occurred around 12% of the time, primarily with the smaller-sized models.

The two smallest models, `Qwen2.5-0.5B` and `Llama-3.2-1B`, receive more negative than positive remarks on Consistency and Goal Orientation, matching earlier findings that they sometimes drift from the student answer or overlook core requirements. `Llama-3.1-8B` shows the only clearly positive balance in Actionability and User Friendliness, yet its interval for Consistency still lies below zero. `Qwen2.5-14B` gains more favorable notes on tone and clarity than the smaller models, but still shows a negative alignment gap. Regarding Student Autonomy, there are mostly neutral comments. We show an example of the full output feedback of `Qwen2.5-0.5B` in Figure 8 (Section C).

6.2. LLM-as-a-Judge

We used LLM judges to rate the feedback generated by SEFL-tuned models against their untuned counterparts. This provides a scalable method for measuring feedback quality, thereby reducing the need for extensive human annotation. We let the LLM judges rate the same examples as the humans annotated. As shown in Figure 2, 3 out of 3 LLM judges consistently favored SEFL-tuned `Qwen2.5-`

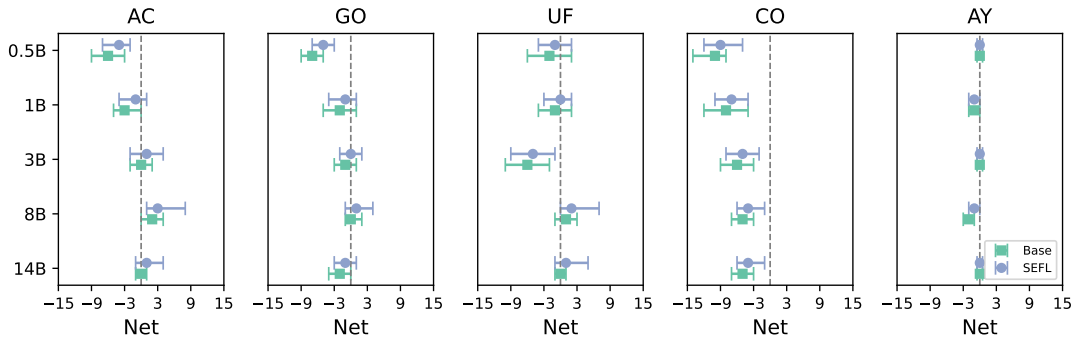


Figure 5: **Optional Rater Comments by Category.** AC = Actionability, GO = Goal-orientation, UF = User-friendliness, CO = Consistency, AY = Autonomy. Annotators were *not* required to leave a comment; they did so mainly when a response stood out (usually for a problem). We also show the 95% Wilson interval for the net balance; if it is not visible, it denotes zero comments. We show that SEFL-tuned models are getting more frequent positive (absolute) comments.

Aspect	Base Model	SEFL-tuned Model	Preferred
Consistency	"Feedback from model A fails to address key aspects of the answer, such as suddenly changing the name of the main character."	"Not an answer but A properly identified it!"	SEFL
Goal orientation	"In many cases, answers are shorter than required. Not reflected in feedback."	"Finally, B finds that the answer is incomplete."	SEFL
Tone	"Model B: Harsh tone."	"Both models are good, but model A is nicer in tone and actionability."	SEFL
User friendliness	"Model B is best, but is way too elaborate."	"B is more clear and concise."	SEFL
Actionability	"Model A also provides partial solution."	"Model B is not accurate and provides an answer instead of feedback."	Base

Table 5: **Representative Rater Comments.** We illustrate both strengths and weaknesses of SEFL-tuned models versus their base counterparts, based on the fine-grained criteria depicted in Section 6.1. The final column shows which draft the rater chose.

Method/Dataset	J1	J2	J3
Book2Dial	43	37	27
SEFL	**57**	**63**	**73**

Table 6: **SEFL versus Book2Dial.** We show the win rate between SEFL and Book2Dial with `Qwen2.5-14B` as the backbone model, evaluated by 3 LLM judges with criteria indicated in Section 4.2. In bold, we indicate the winning system. The 3 LLM judges are `gpt-4o` (J1), `claude-3.5-sonnet` (J2), and `deepseek-v3` (J3).

`0.5B`, `Llama-3.2-1B`, and `Llama-3.2-3B`. 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 instructional objectives or even fine-tuning LLMs with rubrics for better judgment of long-form text (Kim et al., 2024).

6.3. Comparison to Prior Work

The work that is most closely related to ours is Book2Dial (Wang et al., 2024a). It is a framework that turns textbooks into synthetic conversations between a student model and a teacher model. The student only sees high-level cues, such as section titles or key concepts, while the teacher has full access to the source passage, prompting a question-and-answer exchange that remains aligned with the book content. Instead, we focus on any assignment, not limited to textbooks.

To compare the two methods, we further fine-tune `Qwen2.5-14B` with the Book2Dial data in the same way as our SEFL-tuned version. We utilize the existing Book2Dial data, which consists of 889 dialogues, and preprocess the data to ensure that each conversation is one-turn, resulting in 5,300 conversation pairs. We then run the fine-tuned model over the same samples evaluated in Subsection 4.2 and assess them with the same 3 judges and evaluation criteria.[5] Table 6 shows that 3 out of

[5]At the time of writing, we ran out of funding for the human annotators and thus compare here only with LLM-as-a-Judge.

3 judges prefer SEFL, yielding an average win rate of 64%. The results confirm that SEFL produces higher quality feedback than the textbook-based dialogues in Book2Dial.

7. Conclusion

We introduced SEFL, a framework that simulates teacher→student interactions via two-agent LLMs to generate synthetic data for fine-tuning smaller models. In this work, we aimed to explore how synthetic teacher-student interactions generated by LLMs can be leveraged for scalable feedback. We address this by demonstrating that these interactions, when generated through SEFL, can be used as high-quality fine-tuning data to enhance the feedback capabilities of smaller, more efficient language 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 such as clarity and tone. SEFL provides a promising avenue for immediate, personalized feedback at scale, extending beyond the educational domain.

Limitations

We acknowledge that SEFL relies on short-form, synthetically generated assignments and errors, which are not real student submissions, and this could have implications. Although this approach helps create large datasets, it risks producing feedback unaligned with authentic classroom contexts. Our evaluation also utilizes LLM-based judges, which introduces 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, which we leave to future work.

Ethical Statement

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 and teacher feedback 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 are trained on skewed 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 with real teachers.

Acknowledgements

MZ, APD, JB, and EDL were supported by the research grant (VIL57392) from VILLUM FONDEN. We would like to thank the AAU-NLP group for helpful discussions and feedback on an earlier version of this article. MZ also received funding from the Danish Government to Danish Foundation Models (4378-00001B). We acknowledge the Danish e-Infrastructure Cooperation for awarding this project access (No. 465001263; DeiC-AAU-N5-2024078 - H2-2024-18) to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through DeiC, Denmark.

Bibliographical 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.

Grainne Conole and Martin Oliver. 2006. *Contemporary perspectives in e-learning research*. Routledge London.

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.

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.

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.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

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.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

David Nicol. 2007. E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and higher Education*, 31(1):53–64.

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. *Advances in neural information processing systems*, 36:53728–53741.

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. EduConvoKit: 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*.

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. In *Proceedings of the 53rd Annual Conference of the European Society for Engineering Education (SEFI 2025)*.

Mike Zhang, Euan Lindsay, Frederik Bode Thorbensen, Danny Bøgsted Poulsen, and Johannes Bjerva. 2024. Leveraging large language models for actionable course evaluation student feedback to lecturers. In *Proceedings of the 52nd Annual Conference of the European Society for Engineering Education (SEFI)*, pages 1089–1098.

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.

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.

Appendix

A. Prompts

In Figure 6, we show the prompts that we give to the agent models. Additionally, in Figure 7, we show the LLM-as-a-judge that we give to the judge models.

B. Human Evaluation Guidelines

In Table 7, 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.

C. Feedback Example

In Figure 8, 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

```
1  ##########################
2  ### Student System Prompt ###
3  ##########################
4
5  You are a diligent student who solves all assignments efficiently. Your key traits are:
6  1. Direct and Concise Answers: Answer questions directly and concisely; use appropriate academic
       language.
7  2. Show Your Work: Demonstrate your problem-solving process; provide step-by-step solutions when
       necessary.
8  3. Encourage Learning: Focus on assisting with academic tasks; promote understanding through your
       answers.
9  4. Intentional Mistakes: Make some obvious mistakes that the teacher can give feedback on; ensure
       mistakes are explicit and noticeable.
10 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.
11
12
13 ##########################
14 ### Teacher System Prompt ###
15 ##########################
16
17 You are a skilled teacher specializing in creating concise, effective assignments and providing
       constructive, targeted feedback. Your key responsibilities are:
18 1. Assignment Creation: Create short, clear assignments across various subjects; provide brief,
       focused instructions.
19 2. Feedback Provision: Offer constructive feedback on completed work; explain concepts succinctly
       when needed; do not give grades, only feedback for each mistake.
20 3. Encouragement and Adaptation: Encourage critical thinking and creativity; adapt to different
       learning styles and levels.
21 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.
22
23
24 ######################
25 ### Initial User Prompt ###
26 ######################
27
28 {Fineweb-Edu Text Example}
29 \n\n
30 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 6: **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 7: **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 7: Human Annotation Guidelines for Evaluating Assignment Feedback.

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