Closing the Loop: Learning to Generate Writing Feedback via Language Model Simulated Student Revisions

Inderjeet Nair$^{\alpha}$, Jiaye Tan$^{\alpha}$, Xiaotian Su$^{\beta}$, Anne Gere$^{\alpha}$, Xu Wang$^{\alpha}$, Lu Wang$^{\alpha}$
$^{\alpha}$University of Michigan, $^{\beta}$ETH Zürich
inair@umich.edu

Abstract

Providing feedback is widely recognized as crucial for refining students' writing skills. Recent advances in language models (LMs) have made it possible to automatically generate feedback that is actionable and well-aligned with human-specified attributes. However, it remains unclear whether the feedback generated by these models is truly effective in enhancing the quality of student revisions. Moreover, prompting LMs with a precise set of instructions to generate feedback is nontrivial due to the lack of consensus regarding the specific attributes that can lead to improved revising performance. To address these challenges, we propose PROF that PROduces Feedback via learning from LM simulated student revisions. PROF aims to iteratively optimize the feedback generator by directly maximizing the effectiveness of students' overall revising performance as simulated by LMs. Focusing on an economic essay assignment, we empirically test the efficacy of PROF and observe that our approach not only surpasses a variety of baseline methods in effectiveness of improving students' writing but also demonstrates enhanced pedagogical values, even though it was not explicitly trained for this aspect.

1 Introduction

Writing high-quality essays often requires subject-specific and customized feedback from peers and experts, followed by multiple rounds of revisions (Fitzgerald and Markham, 1987; Hayes et al., 1987; MacArthur et al., 1991; Afrin and Litman, 2023). As students incorporate feedback into their writing, they not only improve the current piece but also advance the general writing skills, learn to critically self-assess their work (MacArthur, 2007), and gain a deeper understanding of the subject matter (Bangert-Drowns et al., 2004).

Recent advances in language models (LMs) (Hoffmann et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023; Jiang et al., 2024) make it possible to develop automatic feedback generation systems to provide concrete and actionable comments in a timely manner (Chamoun et al., 2024; D'Arcy et al., 2024), compared to the time-consuming process performed by humans. However, careful prompt engineering is necessary to incorporate precise instructions, ensuring that the generated feedback effectively guides students in improving the quality of their writing. More importantly, providing such detailed instructions is not a trivial task since there is still no general consensus about what attributes the feedback must entail to effectively contribute to students' learning outcomes (Nelson and Schunn, 2009). For example, Bitchener et al. (2005) show that including explanations in feedback can only improve the writing quality of specific revisions, and sometimes (e.g., if too lengthy) can negatively affect tenth-graders' overall writing performance (Tseng and Tsai, 2007).

To this end, our goal is to build an automatic feedback generation system that can be *directly optimized to maximize students' writing revision performance*, to avoid the complexity of explicitly specifying the criterion for effective feedback. However, involving actual students at every stage of the system-building process is impractical due to the time required and the potential negative impact on participants from an immature system (Latifi et al., 2021). To address this challenge, we first develop an **LM-based student simulator** that emulates the process of applying feedback to revise initial content, inspired by the recent efforts to simulate human processes (Park et al., 2023; Shanahan et al., 2023; Xu and Zhang, 2023; Lu and Wang, 2024). In our empirical evaluations of LM simulators, we discovered that by varying the temperature used in autoregressive decoding, we can effectively simulate a diverse array of behaviors to support a comprehensive testing of the feedback generator.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16636–16657
November 12-16, 2024 ©2024 Association for Computational Linguistics

We then propose a feedback generation model, **PROF**[1], that PROduces Feedback via learning from LM-simulated student revisions. Concretely, we use the LM student simulator to iteratively generate preference relations involving desirable and undesirable feedback. We then apply the Direct Preference Optimization (DPO) objective (Rafailov et al., 2023; Xu et al., 2023; Yuan et al., 2024; Pang et al., 2024) along with the preference relations to update the feedback generator. Importantly, the iterative process aims to enhance the effectiveness of the generated feedback, resulting in better implementation performance according to the student simulator.

To evaluate the feedback generated by PROF, we conduct a study on an essay assignment from an introductory economics course offered at a university in United States. In our experimental analyses, we compared the performance of our model with that of few-shot prompted `gpt-3.5`/`gpt-4` models along with other nontrivial comparisons. Our approach not only achieves a similar level of pedagogical alignment, but also outperforms these enterprise LMs in terms of implementation performance of essay revising. Notably, our model is significantly smaller in size (8 billion parameters), making it more efficient and can be easily adapted to other writing assignments. Furthermore, in our empirical experiments involving the student simulators, we observed that the feedback generated by our model aligns well with the actual human revisions, demonstrating its faithfulness in implementation.

To summarize, our work makes the following major contributions:

- We propose PROF, a method that trains feedback generation models by eliminating the need for manually defining the desired feedback attributes or relying on any large-scale annotated dataset of high-quality feedback.
- We propose a method for automatically evaluating the effectiveness of the generated feedback in terms of student implementation performance using LM-based student simulators.
- By tuning the temperature involved in autoregressive decoding, we can generate a wide-range of behaviour from the student simulator allowing us to subsequently develop a feedback generation system that caters to diverse writing traits. This versatility can be harnessed to customize feedback generators and optimize implementation performance for distinct behaviors.

2 Data Description

We collected data from the essay assignments submitted by students enrolled in the Economics 101 course at the University of Michigan, Ann Arbor, United States. This assignment explores a scenario in which "*an increase in the minimum wage in San Francisco could lead to burgeoning adoption of automation*". To discourage this outcome, two policies are proposed: a) a tax on automation and b) a ban on automation. The students are instructed to craft a persuasive letter explaining the economic consequences of a minimum wage increase. Furthermore, they are tasked with presenting arguments against one of the aforementioned policies by using the tools and principles taught in the course. Refer Appendix A.1 to view the assignment prompt.

Thereafter, each student essay is reviewed by 3 peers to obtain a set of 3 feedback which identifies areas for improvement. This assignment uses scripted peer feedback (Latifi et al., 2021; Noroozi et al., 2016) wherein the peer reviewers are expected to provide feedback along a series of prompts. Refer Appendix A.2 to view the questions in the scripted peer feedback. Finally, the author revises the original essay based on the received peer reviews.

In total, we collected 363 datapoints, each comprising the **initial writing**, **three peer feedback**, and the **revised essay**. Among these, we utilized 291 essay-revision pairs along with the 873 (291×3) feedback for our train split, while the remaining 72 datapoints were used for testing. As demonstrated later, our approach yields enhanced performance despite the limited number of datapoints and the absence of expert annotated feedback. This assignment includes a detailed rubric for evaluating and grading student-written essays. Our approach utilizes this rubric to create an essay evaluation prompt for assessing revisions produced by LM student simulators. This prompt can be found in Appendix A.3.

---

[1] Our code and data are available at `https://github.com/launchnlp/PROF`.

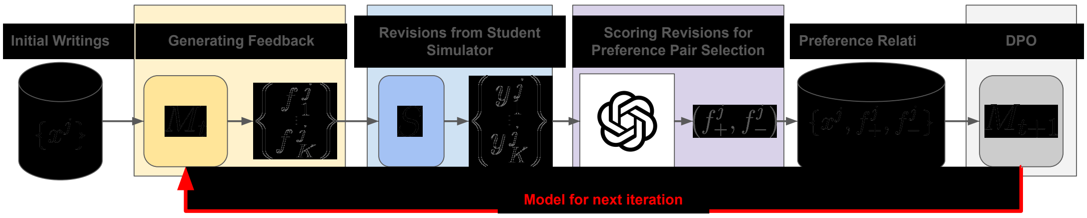

Initial Writings
Generating Feedback
Revisions from Student Simulator
Scoring Revisions for Preference Pair Selection
Preference Relati
DPO
Model for next iteration

Figure 1: **PROF Pipeline**: The depicted figure illustrates the iterative optimization algorithm used in our approach. At each iteration t, the feedback generator M_t generates multiple feedback samples, which are then evaluated for their effectiveness using the student simulator and then `gpt-4` as a judge. These evaluations are used to establish preference relations over feedback using the quality of the corresponding revised essays , which are subsequently used to update the parameters of M_t via DPO (Rafailov et al., 2023), resulting in the updated version M_{t+1}.

3 PROF: Learning to Generate Feedback with Simulated Student Revisions

To directly optimize the feedback generation on student revising performance, we have two LMs that function as **feedback generator** and **student simulator** respectively, as illustrated in Figure 1. In §3.1, we first describe the training of student simulators using the data from §2, to emulate how students integrate feedback into revisions. In §3.2 and §3.3, we present how the feedback generator is initialized and iteratively optimized by the proposed PROF method based on desirable and undesirable feedback, as measured by simulated revisions' quality.

3.1 Student Simulation

We represent the dataset of the assignment submissions as $\mathcal{D}$, whose j^{th} element can be represented as a tuple $(x^j, \{f_i^j\}_{i=1}^3, y^j)$. Here, $\{f_i^j\}_{i=1}^3$ represents the set of feedback applied by the student to revise the initial writing x^j into y^j.

To simulate the behavior of implementing one feedback in place of 3 feedback simultaneously, we instruct `gpt-3.5`[2] to combine the 3 feedback into a holistic feedback f_c^j. Please refer Appendix A.4 to view the exact prompt.

Thereafter, we fine-tune two LMs, `llama3-8b`[3] and `gpt-3.5`, of different scales to implement the feedback, i.e., generating y^j given x^j and f_c^j. We represent the trained simulator as S and use it to sample revisions during feedback generator training for initial writing x and feedback f, i.e., $y \sim S(\cdot|x, f)$.

3.2 Feedback Generator: Initialization

We initialize our feedback generator using `llama3-8b` and train it specifically for the task of generating peer feedback. To create paired data for fine-tuning, we use $\mathcal{D}$ where the jth assignment submission consists of three paired data points $\{x^j, f_i^j\}_{i=1}^3$ using the individual peer feedback. We represent the feedback generator by M from which the feedback f can be sampled for an essay x as $f \sim M(\cdot|x)$. After this initialization process, we continue training using the PROF method to optimize student revision performance

3.3 Feedback Generator: Optimization

Our approach assumes access to two functions for the iterative optimization of the feedback generator: (a) student simulator as described in §3.1 and (b) automatic essay scoring system, where we use `gpt-4` via few shot prompting owing to their strong capabilities in critically assessing the quality of natural language outputs (Zheng et al., 2023; Li et al., 2023; Zheng et al., 2024a). The rubrics employed to assess the quality of the essay are identical to the course rubrics. Refer to Appendix A.3 to view the complete prompt.

Let the feedback generator at the start of tth iteration be represented by M_t. Our objective would be to use M_t in creating desirable feedback f_+^j and undesirable feedback f_-^j for each initial essay x^j in $\mathcal{D}$. After constructing such kind of preference relationship for each datapoint, we use Direct Preference Optimization (DPO) loss (Rafailov et al., 2023) to train a new model M_{t+1} for the next iteration. The following objective describes the relation between M_{t+1} and M_t using the DPO loss:

$$M_{t+1} = \arg\min_{M_\theta} \left[\sum_{j=1}^{|\mathcal{D}|} \mathcal{L}_t(f_+^j, f_-^j, x) \right] \quad (1)$$

---

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

[3] https://llama.meta.com/llama3

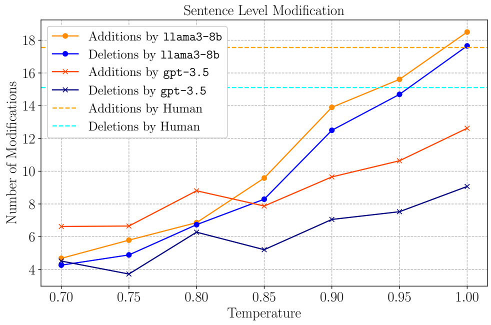


Figure 2: Temperature and sentence-level modifications

$$\mathcal{L}_t(f_+^j, f_-^j, x^j) = -\log \sigma \left(\beta \frac{M_\theta(f_+^j|x^j)}{M_t(f_+^j|x^j)} - \beta \frac{M_\theta(f_-^j|x^j)}{M_t(f_-^j|x^j)} \right) \quad (2)$$

The feedback generator M_θ is initialized with the parameters of M_t and after optimizing Eq. 1, it becomes M_{t+1}. The loss in Eq. 2 enables the model to effectively discern high quality feedback from the low quality ones by amplifying the difference in likelihood between the two, relative to the likelihood estimates of M_t.

To generate the preference pair (f_+^j, f_-^j) for each initial writing x at iteration t, we apply the following steps in sequence:

1. Sampling K different feedback from M_t for each datapoint x^j as: $\{f_k^j\}_{k=1}^K \sim M(\cdot|x^j)$.
2. The trained student simulator S is used to generate the revised version for each of the generated feedback in $\{f_k^j\}_{k=1}^K$.
3. Finally, the reward for each feedback is estimated by employing `gpt-4` as an evaluator that assesses the quality of the revised essay for each feedback. The feedback associated with the best revision and the worst revision is chosen as the preference pair.

It is important to note that our algorithm relies on minimal supervision from `gpt-4`, which is solely used for evaluating the quality of the generated essays.

4 Analysis of Student Simulators

In this section, we analyze the alignment between the properties of revision in student simulators and actual students, based on peer written feedback. Concretely, we examine the impact of the temperature parameter on the number of modifications (§4.1), revision performance (§4.2), and faithfulness to feedback (§4.3), and compare it with the real students' revision process.

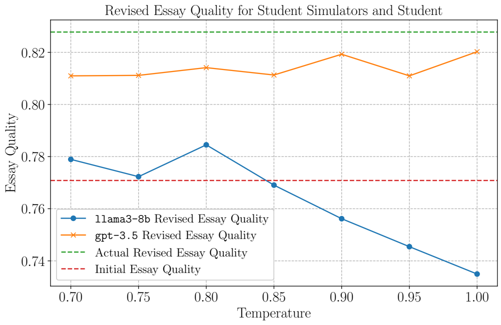


Figure 3: The quality of the revised essay by student simulators vs. actual students.

4.1 Temperature and Revisions

We first analyze the variation in the number of lexical modifications between two student simulators (initialized with `llama3-8b` and `gpt-3.5` respectively) using different temperature settings and compare these results to the actual revision process of real students. The Ratcliff/Obershelp algorithm (Black, 2004) provides us with a list of edit operations (additions and deletions) required to transform one sequence into another. We categorize contiguous additions / deletions that involve less than a sentence as word-level modifications, with the number of words involved being counted. For modifications that span an entire sentence or more, we categorize them as sentence-level modifications, with the number of sentences involved being counted.

From the plots shown in Figures 2 and 6, we see that the number of elements (words/sentences) added or deleted by the student simulators increases as the temperature is increased. This observation aligns with expectations, as higher temperature settings introduce more randomness during decoding, leading to increased number of alterations (Renze and Guven, 2024). Furthermore, the `gpt-3.5`-based student simulator tends to be more conservative than the `llama3-8b`-based counterpart, with a greater inclination towards preserving the original content.

4.2 Temperature and Revision Quality

In this analysis, we compute the quality of the revised essays based on peer feedback from both student simulators and actual students using `gpt-4` as the judge. To validate the use of `gpt-4` in evaluating essay quality, we compared its inferred scores with the scores assigned by teaching instructors of the Economics course to actual students' final re-

APPROACH	STUDENT SIMULATORS						REAL STUDENTS
	`llama3-8b`			`gpt-35-turbo`			
	0.7	0.85	1.0	0.7	0.85	1.0	
# of Faithful Rev. (F)	1.1	2.4	3.6	1.8	1.6	1.9	4.5
# of Unfaithful Rev. (U)	1.4	1.3	4.1	0.5	0.8	2.0	1.4
$\gamma = \log\left(\frac{F}{U}\right)$	-0.1	**0.3**	-0.1	**0.6**	0.3	-0.1	0.5

Table 1: The average number of faithful and unfaithful revisions from student simulators operating at different temperatures vs. actual student revisions. γ assigns higher value to faithful revisions that adhere to the feedback without making additional content changes.

vised essays. The mean squared error was 0.082 after normalizing the scores between 0 and 1, suggesting that the inferred scores are reliable and closely align with the expert-assigned scores.

Based on the findings presented in Figure 3, it is evident that actual students exhibit a higher level of effectiveness in incorporating feedback compared to the student simulators. This is a desirable outcome as it creates a more challenging environment for our feedback generator during training and testing with student simulators. Among the student simulators, the model based on `gpt-3.5` demonstrates superior implementation performance, with a slight improvement in the quality of revised essays as the temperature increases and reasonable alignment with the quality of revision from real students. *This makes* `gpt-3.5` *a suitable approach for automatically assessing feedback effectiveness.* On the other hand, the `llama3-8b` based student simulator demonstrates a modest enhancement in quality at lower temperatures but experiences a decline beyond a temperature of 0.8.

4.3 Temperature and Revision Faithfulness

Next, we assess how faithful the student simulators are in implementing the feedback, in comparison to the revisions made by real students. For this purpose, we broke down the feedback into a list of distinct recommended changes, and then examined how many of these changes were incorporated in the revised writing, as compared to the initial version. We categorized the changes as either **faithful** (i.e., they adhered to the provided feedback) or **unfaithful** (i.e., they went beyond the scope of the feedback and made additional changes). For 10 samples, the number of these instances at 3 temperature values for the student simulators was manually annotated by a fluent English speaker. This resulted in the analysis of $(3 + 3 + 1) \times 10 = 70$ revised essays, with 3 revisions from each of the student simulators and one actual revision.

Comparing the revisions generated by student simulators with those made by humans, we find that the simulators produce more unfaithful revisions and fewer faithful revisions, as shown in Table 1. While previous experiments indicate that higher temperatures result in more content-level modifications from the simulators, they do not always align with the provided feedback. To quantify the faithfulness of modifications, we define γ as the logarithm of the ratio between the number of faithful and unfaithful modifications. These results suggest that there is still room for improvement in the faithfulness of simulated revisions compared to actual human revisions.

5 Feedback Generator Setups

For our iterative optimization approach, **we use `llama3-8b` based student simulator for training and emphasize on `gpt-3.5` based one for testing**. We do this for the following three reasons: **(1)** `gpt-3.5` based student simulator better aligns with actual student revision, as demonstrated in §4.2 and thus provides a more realistic testing environment to measure implementation performance. **(2)** Using `gpt-3.5` based student simulator is prohibitively expensive when used in conjunction with an iterative optimization approach. Using `llama3-8b` as a student simulator makes our research more accessible due to lower training cost and easy access to the open-source models. **(3)** Training and testing on the same student simulator would not provide conclusive evidence of the effectiveness of our approach, as it might learn to exploit one type of student simulator while performing poorly on others. For completeness, however, we also include the effectiveness of the generated feedback using the `llama3-8b` based student simulator.

Refer Appendix B for more details.

6 Results for Feedback Generation

We consider the following types of baselines: (1) **`gpt-3.5 / gpt-4`**: Using enterprise LLMs as a few-shot feedback generator by sampling in-context examples from peer-written feedback from the train-split. (2) **`sft-from-human`**: Fine-tuning `llama3-8b` on peer review feedback.

Our method variants are named as "**PROF,** $\tau = x$" where the `llama3-8b` based feedback generator is initialized with `sft-from-human` and trained using the iterative optimization framework

APPROACH	PEDAGOGICAL EVALUATION RGQ	EAL	DM	MSSC	AVG.
`gpt-3.5`	70.6	**80.0**	78.6	**60.0**	72.3
`gpt-4`	71.4	**80.0**	**79.2**	59.4	**72.5**
`sft-from-human`	65.8	67.8	65.6	53.3	63.1
PROF, $\tau = 0.7$					
Iteration 1	66.7	77.2	76.4	57.2	69.4
Iteration 2	68.6	79.2	77.8	59.2	71.2
Iteration 3	70.6	79.4	78.3	59.7	72.0
PROF, $\tau = 0.85$					
Iteration 1	70.8	78.3	75.0	58.3	70.6
Iteration 2	70.6	79.2	77.8	58.6	71.6
Iteration 3	**71.9**	79.2	**79.2**	59.2	72.4
PROF, $\tau = 1.0$					
Iteration 1	62.2	62.5	61.4	53.0	59.8
Iteration 2	70.8	73.6	69.4	57.2	67.8
Iteration 3	70.8	76.9	75.8	58.0	70.4

Table 2: We evaluate the intrinsic quality of the generated feedback in terms of pedagogical alignment. **Green** and Blue represents best and second-best performance respectively.

described in §3 along with the `llama3-8b` based student simulator executed at temperature x.

6.1 Intrinsic Evaluation

To intrinsically evaluate the quality of the feedback generated from different approaches, we employ LM as the judge (Chevalier et al., 2024; Ke et al., 2023) and evaluate along the following **four major pedagogical dimensions** (Jurenka et al., 2024):

- **Respects Guided Questions (RGQ)**: Given that the assignment uses scripted feedback, the generated responses are expected to follow a template with 6 prompts, each accompanied by a targeted feedback. Here, we assess how well each feedback component aligns with its respective prompt. To view the prompt template, refer to Appendix A.2.

- **Encourages Active Learning (EAL)**: Measures how well the feedback guides the students to make improvements on their own without explicitly revealing the concrete changes.

- **Deepens Metacognition (DM)**: Determine the effectiveness of the feedback in identifying and addressing student errors and misconceptions within the essay.

- **Motivates and Stimulates Student Curiosity (MSSC)**: Assess how well the feedback maintains a positive and encouraging tone that fosters curiosity and motivation.

For the score generated by `gpt-4` in relation to a sample feedback, please see Appendix D.2. The idea of using a critic LLM to automatically assess the quality of feedback based on pedagogical aspects was inspired by Jurenka et al. (2024), who demonstrated a strong correlation between the generated scores and those provided by humans. In our study, we utilized `gpt-4` to automatically assign pedagogical scores to the feedback.

Results from `gpt-4` based pedagogical evaluation. We used `gpt-4` to assign scores ranging from 0 to 5 for each metric, representing lowest to highest quality. The average quality for each metric was then calculated, normalized between 0 and 100, and presented in Table 2. *Our approach significantly improves performance for* `sft-from-human` *and achieves comparable results to enterprise LLMs, despite having significantly fewer parameters.* This demonstrates the effectiveness of our approach without requiring high-quality feedback.

Among our models trained with student simulators at different temperatures, we observe the most significant improvements in the model trained at a temperature of 0.85. We believe that at lower temperatures, the student simulators only incorporate a limited number of feedback elements, as discussed in §4.3. This limitation prevents the appropriate selection of desirable and undesirable feedback at each stage of preference learning. Conversely, at higher temperatures, the student simulator demonstrates a higher degree of unfaithfulness, leading to poor implementation performance, as demonstrated in §4.3 and §4.2, respectively. By using `sft-from-human` as the foundation for our feedback generator across various temperature settings, we achieve a notably improved model in terms of pedagogical alignment.

Validation of `gpt-4`'s pedagogical evaluation. To validate the pedagogical evaluation using `gpt-4`, we used 21 pairs of essays and peer reviews and had two proficient English annotators assess the peer reviews across various pedagogical dimensions by providing a score between 0 and 5 for each metric. We then calculated the Pearson correlation between the average normalized scores assigned by humans and those inferred by `gpt-4`.

APPROACH	STUDENT SIMULATOR 0.7	0.85	1.0	AVG.
`gpt-3.5`	76.3	77.1	76.9	76.8
`gpt-4`	76.6	77.0	77.4	77.0
`sft-from-human`	78.9	78.8	80.3	79.4
PROF, $\tau = 0.7$				
Iteration 1	**80.1**	79.9	79.7	79.9
Iteration 2	77.1	79.4	80.0	78.9
Iteration 3	79.0	77.5	**80.9**	79.1
PROF, $\tau = 0.85$				
Iteration 1	79.3	**80.0**	80.8	**80.0**
Iteration 2	79.0	74.8	78.7	77.5
Iteration 3	79.1	79.5	79.5	79.4
PROF, $\tau = 1.0$				
Iteration 1	79.2	77.2	77.3	77.9
Iteration 2	79.7	**80.0**	78.2	79.3
Iteration 3	78.0	78.9	76.6	77.8

Table 3: Extrinsic evaluation using `gpt-3.5` based student simulator. **Green** and Blue represents best and second-best performance respectively. Each experiment was repeated for 5 different seeds to mitigate the impact of randomness.

We noted a moderate-to-high correlation between the generated and averaged annotated scores for dimensions such as "respects guided questions", "encourages active learning", and "deepens metacognition", with respective values of 0.31, 0.40, and 0.71. However, the dimension "motivates and stimulates student curiosity" demonstrated a lower correlation of 0.20, likely due to the subjective nature of assessing positivity and the encouraging tone of feedback. In terms of average pedagogical score, we observed a correlation of 0.40. These measures of correlation justify the validity of pedagogical evaluation from `gpt-4`.

6.2 Extrinsic Evaluation

Next, we use trained student simulators to gauge the efficacy of feedback generated by different systems. Although we discuss results by using `gpt-3.5` based student simulator, we also include the results computed using `llama3-8b` based student simulator in Table 9 of Appendix D.1. `gpt-3.5`-based simulator aligns more closely with actual human performance and was not employed in training our models, making the effectiveness evaluation more reliable and trustworthy.

For each approach, we use greedy decoding to generate feedback. The student simulators are executed at 3 different temperatures: 0.7, 0.85, and 1.0, with 5 different seeds to mitigate the impact of randomness. Finally, we compute a score based on the course rubric prompt (refer to Appendix A.3), averaging the scores across different aspects and normalizing it between 0 and 100.

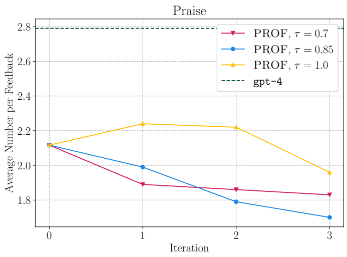
Table 3 demonstrates that our approaches consistently outperform enterprise LLMs like `gpt-3.5`/`gpt-4`. Interestingly, the results also indicate a lack of correlation between extrinsic and intrinsic evaluations. While `sft-from-human` exhibits better extrinsic performance, its intrinsic performance is lower compared to the few-shot approaches. One possible explanation for this discrepancy is that `sft-from-human` provides explicit feedback, which may negatively impact the "Encourages Active Learning" metric (Table 2), but contribute to a more effective revision process. In most cases, one round of iterative optimization leads to the best extrinsic performance. However, it is important to note that these findings are specific to the experiments conducted on a domain-specific course and may not necessarily apply to broader datasets or different contexts.

6.3 Additional Analyses

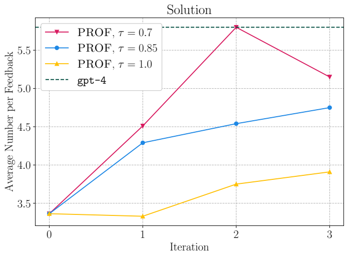
RQ1: How does our training algorithm influence different feedback categories with the number of refinement iterations? In this analysis, we break down the feedback into distinct components and classify them into one of 3 categories: **praise**, **solution**, or **problem**. Our primary objective is to verify whether PROF effectively adjusts the frequency of these elements in a manner that is consistent with the learning sciences research on maximizing feedback effectiveness (Lizzio and Wilson, 2008; Nelson and Schunn, 2009; Cho and MacArthur, 2010).

If a segment solely describes an issue, it is labeled as a **problem**. If segment contains both problem and solution, it is still categorized as a **solution**. To view how our algorithm influences the number of **praise** elements, refer to Figure 4a. We notice that the average number of **praise** elements decreases with more optimization steps, which is corroborated by previous research indicating that praise has minimal impact on student performance (Kluger and DeNisi, 1996; Ferris, 1997).

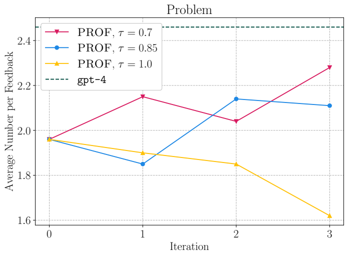
In general, the average number of **solution** and **problem** elements in the feedback (refer Figure 4b and 4c respectively) increases with the number of iterations which is known to impact implementation performance as supported by many prior works (Hayes et al., 1987; Matsumura et al., 2002; Bitchener et al., 2005; Sugita, 2006).



(a) Number of **Praise** segments

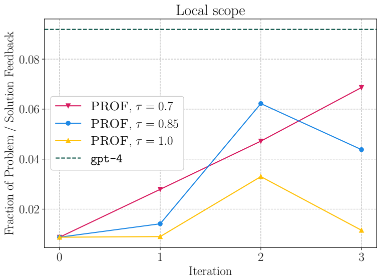


(b) Number of **Solution** segments

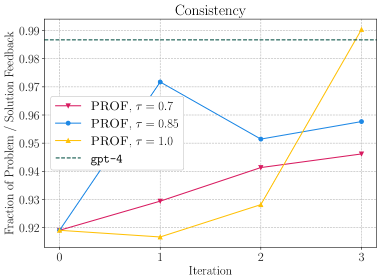


(c) Number of **Problem** segments

Figure 4: Evolution of segments



(a) Fraction of elements associated with local scope.



(b) Fraction of elements that are logically consistent.

Figure 5: Progression of problem/solution elements with refinement iterations.

Among our models trained at different temperatures, PROF, $\tau = 0.7$ significantly increases the average number of **solution** elements compared to the model trained at a temperature of 1.0 as shown in Figure 4b. When training with the student simulator at a high temperature (1.0), the feedback generator is guided to minimally increase the number of solution segments. This is intuitive because the student simulator can introduce numerous changes that may lower the quality of the essay at higher temperatures. If there are too many solution elements, the model may make excessive changes that could degrade the essay's quality. Conversely, with the student simulator being very conservative at a temperature of 0.7, our optimization algorithm responds by more aggressively increasing the number of solution components.

While **solution** provides explicit guidance on improving the initial writing, the lack of this guidance from the **problem** elements coupled with the ineffectiveness of our `llama3-8b`-based student simulator at higher temperatures (1.0) make the feedback generator produce fewer **problem** elements as the number of the training iterations increases (refer Figure 4c).

As the **problem** and **solution** feedback elements are influential in improving feedback effectiveness, we further sub-categorize these elements and analyze how they evolve with the number of refinement iterations.

RQ2: How does the faction of problem or solution feedback segments associated with a local Scope vary with the number of refinement iterations? The feedback scope is broadly categorized into two classes: (a) **local scope**, which focuses on specific words, phrases, sentences, or paragraphs and is associated with narrow aspects such as surface features; (b) **global scope**, involves considering multiple parts or the entirety of the writing. Both local-scope and global-scope feedback have been observed to result in improved essay quality after implementation (Olson and Raffeld, 1987; Lin et al., 2001; Miller, 2003). To ensure effective feedback for enhancing essay quality, it is important to include an appropriate proportion of both locally and globally scoped problem and solution segments, rather than focusing solely on one type.

Figure 5a illustrates that the fraction of locally-scoped feedback instances generally increases as the number of iterations grows, with the most significant increase observed for $\tau = 0.7$. This is desirable as initially the generated feedback has very few instances of locally-scoped elements and PROF rectifies this by increasing it appropriately. When the student simulator operates at a low temperature and makes minimal edits, the optimization algorithm encourages the feedback generator to prioritize generating feedback that focuses on local scope. This is because feedback associated with local scope is more likely to be addressed through localized changes, which align with the minimal edits made by the student simulator at low temperatures.

RQ3: Is the consistency of Problem / Solution segments improved with the number of refinement iterations? The **consistency** of a problem/solution is determined by two aspects: *intrinsic correctness* and *consistency with respect to the content*. Intrinsic correctness refers to the validity and absence of any logical fallacies in the feedback segment. Consistency with respect to the initial content refers to whether the identified problem is indeed an issue in the original content and whether the solution maintains the original stance of the essay without altering it.

Based on Figure 5b, our training approach demonstrates an improvement in the consistency of the problem/solution segments as the number of refinement iterations increases. Notably, the consistency of PROF, $\tau = 1.0$ shows the highest performance after 3 iterations. We attribute this observation to the complexity of the training environment, which influences the consistency of the generated feedback. In the case of the student simulator executed with a temperature of 1.0, which yields lower implementation performance and creates a more challenging environment for the feedback generator, the training algorithm guides the feedback generator to produce feedback with better consistency to achieve optimal implementation performance.

7 Related Works

7.1 Automatic Feedback Generation Systems

NLP systems have been developed to automatically provide formative feedback to improve students' writing (Liu et al., 2016; Zhang et al., 2019; Klebanov and Madnani, 2020). One significant challenge faced by these approaches is the creation of high-quality feedback datasets, which requires considerable time and effort. In contrast, our approach starts with peer-annotated reviews that may not initially be of high quality. However, through an iterative preference learning process, we steer our feedback generator towards producing better quality responses. Considering that many previous works have focused on collecting peer review datasets (Kang et al., 2018; Lin et al., 2023; Dycke et al., 2023), our approach can leverage these datasets for better initialization.

While Language Models (LMs) can bypass the need for high-quality annotated feedback through few-shot prompting, they are computationally intensive and expensive (Han et al., 2023; Chamoun et al., 2024; D'Arcy et al., 2024). In contrast, our feedback generator offers a cost-effective solution using smaller LMs with fewer parameters and without relying on high-quality supervision.

7.2 Iterative Preference Optimization

In recent times, a novel paradigm has emerged, which entails the iterative application of offline reinforcement learning techniques (Rafailov et al., 2023). In this approach, model generates preference relations per iteration, which are used to construct potentially more informative relations than those observed so far. This updates the model parameters, resulting in a better aligned model. Examples of such approaches include Iterative DPO (Xu et al., 2023; Xiong et al., 2023; Gulcehre et al., 2023), Self-Rewarding LMs (Yuan et al., 2024), and SPIN (Chen et al., 2024). Previous approaches build preference relations by intrinsically evaluating the quality of the generation. However, exploring the utility of generation to establish preference relations has not been investigated before. To our knowledge, we introduce a method that constructs preference relations through extrinsic assessments, quantifying the utility of generated samples.

8 Conclusion

In this paper, we present an optimization technique called PROF that focuses on maximizing students writing revision performance through LM simulation. We conducted extensive analysis to showcase the alignment between our student simulators and actual revisions, while also demonstrating the flexibility of adjusting temperature to elicit diverse behaviors from the models. Through experiments, we show that our trained models not only exhibit better effectiveness but are more pedagogically aligned.

Acknowledgements

This work is supported in part through National Science Foundation under grant 2302564. We are grateful for the resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. Finally, we would like to extend our gratitude to Prof. Mitchell Dudley and Zhen Qian for their extensive contributions in designing the assignment, collecting student data and guiding us in comprehending the data.

Limitations

We acknowledge the following limitations with our work:

- As we consider a narrow domain for our experimental analyses, we concede that some of the conclusions and findings may not be directly extensible to other domains.
- In our experimental setup, we employed LLMs to automatically assess the effectiveness of the generated feedback. As our analysis indicated a gap between some aspects of the simulated revision and the actual revision, we advise practitioners to approach the interpretation of the results with caution.

Ethical Statement

As we rely on actual student data to train the LMs on simulating student implementation process, we affirm that the Institutional Review Board (IRB) approval was obtained prior to collecting and using student data, ensuring compliance with ethical standards for human subject research. Moreover, while our research simulates student implementation process, it should not be considered as a perfect representation of the actual student behavior. We strongly advise against using these simulations as a substitute for real student implementation procedures in educational settings. Instead, they should be viewed as a complementary tool to enhance understanding of student learning processes and to improve feedback mechanisms.

References

Tazin Afrin and Diane Litman. 2023. Predicting desirable revisions of evidence and reasoning in argumentative writing. In Findings of the Association for Computational Linguistics: EACL 2023, pages 2550–2561, Dubrovnik, Croatia. Association for Computational Linguistics.

Robert L Bangert-Drowns, Marlene M Hurley, and Barbara Wilkinson. 2004. The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. Review of educational research, 74(1):29–58.

John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on esl student writing. Journal of second language writing, 14(3):191–205.

Paul E Black. 2004. Ratcliff/obershelp pattern recognition. Dictionary of algorithms and data structures, 17.

Eric Chamoun, Michael Schlichktrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. Preprint, arXiv:2405.20477.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. arXiv preprint arXiv:2401.01335.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, et al. 2024. Language models as science tutors. arXiv preprint arXiv:2402.11111.

Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. Learning and instruction, 20(4):328–338.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. arXiv preprint arXiv:2401.04259.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Dana R Ferris. 1997. The influence of teacher commentary on student revision. Tesol Quarterly, 31(2):315–339.

Jill Fitzgerald and Lynda R Markham. 1987. Teaching children about revision in writing. Cognition and instruction, 4(1):3–24.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. arXiv preprint arXiv:2308.08998.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2023. Fabric: Automated scoring and feedback generation for essays. arXiv preprint arXiv:2310.05191.

John R Hayes, Linda Flower, Karen A Schriver, James Stratman, Linda Carey, et al. 1987. Cognitive processes in revision. Advances in applied psycholinguistics, 2:176–240.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. arXiv preprint arXiv:2407.12687.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. arXiv preprint arXiv:2311.18702.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing–50 years and counting. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages 7796–7810.

Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological bulletin, 119(2):254.

Saeed Latifi, Omid Noroozi, Javad Hatami, and Harm JA Biemans. 2021. How does online peer feedback improve argumentative essay writing and learning? Innovations in Education and Teaching International, 58(2):195–206.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. arXiv preprint arXiv:2310.05470.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Moprd: A multidisciplinary open peer review dataset. Neural Computing and Applications, 35(34):24191–24206.

Sunny SJ Lin, Eric Zhi-Feng Liu, and Shyan-Ming Yuan. 2001. Web-based peer assessment: feedback for students with various thinking-styles. Journal of computer assisted Learning, 17(4):420–432.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. 2016. Automated essay feedback generation and its impact on revision. IEEE Transactions on Learning Technologies, 10(4):502–513.

Alf Lizzio and Keithia Wilson. 2008. Feedback on assessment: Students' perceptions of quality and effectiveness. Assessment & evaluation in higher education, 33(3):263–275.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. pages 1–12.

Charles A MacArthur. 2007. Best practices in teaching evaluation and revision. Best practices in writing instruction, pages 141–162.

Charles A MacArthur, Steve Graham, and Shirley Schwartz. 1991. Knowledge of revision and revising behavior among students with learning disabilities. Learning Disability Quarterly, 14(1):61–73.

Lindsay Clare Matsumura, G Genevieve Patthey-Chavez, Rosa Valdés, and Helen Garnier. 2002. Teacher feedback, writing assignment quality, and third-grade students' revision in lower-and higher-achieving urban schools. The Elementary School Journal, 103(1):3–25.

Peter J Miller. 2003. The effect of scoring criteria specificity on peer and self-assessment. Assessment & Evaluation in Higher Education, 28(4):383–394.

Melissa M Nelson and Christian D Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. Instructional science, 37:375–401.

Omid Noroozi, Harm Biemans, and Martin Mulder. 2016. Relations between scripted online peer feedback processes and quality of written argumentative essay. The Internet and Higher Education, 31:20–31.

Mary W Olson and Paul Raffeld. 1987. The effects of written comments on the quality of student compositions and the learning of content. Reading Psychology: An International Quarterly, 8(4):273–293.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. arXiv preprint arXiv:2404.19733.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST '23, New York, NY, USA. Association for Computing Machinery.

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

Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. arXiv preprint arXiv:2402.05201.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. Nature, 623(7987):493–498.

Yoshihito Sugita. 2006. The impact of teachers' comment types on students' revision. ELT journal, 60(1):34–41.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Sheng-Chau Tseng and Chin-Chung Tsai. 2007. Online peer assessment and the role of the peer feedback: A study of high school computer course. Computers & Education, 49(4):1161–1174.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. 2023. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. arXiv preprint arXiv:2312.11456.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. arXiv preprint arXiv:2312.16682.

Songlin Xu and Xinyu Zhang. 2023. Leveraging generative artificial intelligence to simulate student learning behavior. arXiv preprint arXiv:2310.19206.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. arXiv preprint arXiv:2401.10020.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 9619–9625.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. arXiv preprint arXiv:2309.11998.

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

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372.

A Dataset Prompts

In this section, we attach all the prompts associated with the dataset described in §2.

A.1 Assignment Prompt

The prompt in Table 4 was shown to the students to get the initial writing:

A.2 Peer Review Prompt

The prompt in Table 5 was shown to the students while reviewing the initial writing:

Table 4: Assignment prompt shown to the students to get the initial writing.

SITUATION

In San Francisco, the fear of losing jobs to automation after an increase to the minimum wage has motivated two similar policy proposals aimed at discouraging or banning automation. As a student of economics and someone who will soon enter the job market, you find this issue interesting and relevant. [For simplicity, assume these policies are only occurring in or are only proposed for the San Francisco area. Also assume the ban is for automation in general, not just delivery robots.]

PROMPT

Write a letter to the San Francisco Board of Supervisors regarding the two policy proposals under consideration: a tax on automation (Ms. Kim's proposal) or a ban on automation (Mr. Yee's proposal). Your letter should: Briefly explain the economic impact of a minimum wage increase (moving from nonbinding to binding) in the labor market, and its extended effect on the automation market as well as the market for a good which may be produced using labor, automation or some combination of the two; Identify one of the two proposed policies and construct an argument, based on the economics you've learned in class, for why you oppose the policy. You, personally, may oppose both policy proposals, but your paper should focus on only one policy, given the word count limit. Your argument of opposition should not be based on your support for the other proposed policy. While your letter is a normative economic assessment, the majority of the letter should consist of positive economic analysis. [While you may have strong opinions on this subject based on moral or ethical reasoning, the purpose of this assignment is to see your ability to use the economic tools you've learned to analyze the situation.] Explain the economic impact of this policy proposal on these same markets, highlighting the economic reasoning for opposing the policy; Start your analysis assuming the minimum wage increase already occurred. Be persuasive.

ITEMS TO KEEP IN MIND:

The Board of Supervisors likely has some knowledge of economics. Your explanations may assume prior general knowledge consistent with our coverage of Supply, Demand, Consumer Surplus, Producer Surplus, and Efficiency. The supervisors understand the definitions of these terms, but not necessarily how they interact specifically with this problem. For example, the Board of Supervisors do not immediately understand how a price change in the labor market affects related markets. The provided article must be cited. External references are not required but also must be cited if used. Either APA format or MLA format is acceptable. Since you are writing to the Board of Supervisors, you should take care to carefully edit and proofread your letter. Your letter should be between 400-500 words (this includes your first draft) and should follow the conventions of a professional letter, including a To and From section as well as a professional and courteous tone. Please sign your letter "A Concerned Citizen" since anonymity is needed for peer review. Please include a word count at the end of your assignment. [The word count may exclude the "To" and "From" lines, as well as the parenthetical citations within your paper. Note quotations within your paper still count. Don't forget that both the minimum and maximum limits to the word count are important on all drafts.]

Table 5: Prompt shown to the students to review an initial writing.

In providing feedback, your task is to help your peers identify areas that need the most attention. To guide you through the process of effectively providing feedback according to the essay rubric, you will be given a series of prompts that correspond to the essay rubric criteria.

PEER REVIEW GUIDELINES

- Highlight the pieces of texts that let you directly address the feedback prompts in your online responses.
- In your online responses, focus on larger issues (higher order concerns) of content and argument rather than lower order concerns like grammar and spelling.
- Be very specific in your responses, referring to your peer's actual language, mentioning terms and concepts that are either present or missing, and following the directions in the feedback prompts.
- Use respectful language whether you are suggesting improvements to or praising your peer.

PEER REVIEW FEEDBACK PROMPTS:

Understanding:

- Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.
- How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?

Critical Thinking

- Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?
- How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.
- Are all outside sources properly cited?

Response Alignment with Audience

- The letter should be understandable to a person with a basic but not sophisticated understanding of economic principles. In this context, which parts were difficult to understand? Which parts were easy to understand?

A.3 Prompt for Grading Essays

The prompt in Table 6 includes the rubric to grade an essay. On a high level, it elucidates the criteria for assigning a particular point for each of the question / instruction in the peer review prompt from Appendix A.2. The detailed rubric is given as follows:

A.4 Prompt for Combining Feedback

This prompt used for combining the 3 peer reviews into a single review that captures all the comments from the individual feedback is shown below:

> Your task is to skillfully merge feedback from multiple reviewers into one unified and coherent narrative. Maintain the original language style and tone of each review to accurately represent their feedback. In the following examples, you will be provided with input from several reviewers, and your job will be to craft a single piece of feedback that honors the format and intent of the individual contributions, without explicitly mentioning reviewer numbers.

Using the above prompt on 3 peer review shown in Table 7 results in the combined feedback shown in Table 8.

B Implementation Details: Hyperparameter Settings

For training the student simulator and initializing the feedback generator using `llama-8b-instruct`, we use LoRA (Hu et al., 2021) with a batch size of 8 and cosine annealing with warm restarts (Loshchilov and Hutter, 2016) for linear rate scheduling. The model undergoes 20 epochs of training, with an initial learning rate of 10^{-4}.

Our feedback generator is iteratively refined for 3 iterations. At each stage of sampling preference relations, we sample $K = 5$ different feedback for each datapoint. For the DPO optimization, we set $\beta = 0.1$ and train the model under this framework for 5 epoches with an initial learning rate of 10^{-5} and a batch size of 8.

For each of the above described training processes, we use `LLaMA-Factory` (Zheng et al., 2024b).

C Student Simulator: Additional Analysis

C.1 Temperature and Word-level Modifications

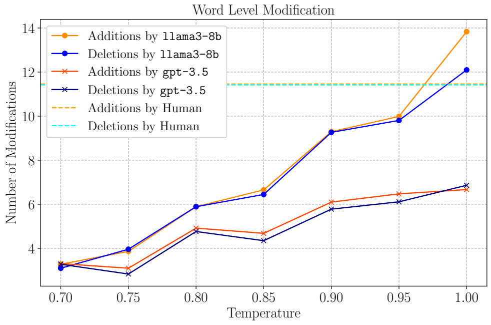


Figure 6: Temperature and Word-level modifications

The variation of the number of modified by the student simulators with temperature is shown in Figure 6.

C.2 Qualitative Analysis of Revisions

Upon analyzing the individual data points and corresponding revisions from both simulators, we have observed that while `llama3-8b` generates a higher number of revisions, these changes do not necessarily enhance the quality of the initial writing. Specifically, we have noticed that it often deletes crucial statements that are essential for the overall argumentation of the essay, while also introducing statements that may negatively impact clarity and persuasiveness. We present a qualitative example in this section to illustrate undesirable and desirable feedback implementation from the `llama3-8b` and `gpt-3.5` based student simulator respectively.

As shown in the Figure 7, the revised version from `gpt-3.5` based student simulator retains the more detailed and comprehensive explanation of how the minimum wage increase affects the labor market (*"Increasing the minimum wage ... known as a price floor."*). It explicitly mentions the concept of a price floor and explains how it leads to higher unemployment which in turn results in an increase to the quantity of the workers supplied at a lower quantity of the workers demanded (*"Therefore, with the cost of workers now increased to a level ... ultimately leading to greater unemployment. ... in quantity of workers demanded."*). Overall, the improved version provides a clearer and more nuanced understanding of the economic implications of the minimum wage increase on the

Table 6: This prompt provides the rubric for grading the essay.

- UNDERSTANDING
    - **Concepts & Accuracy**:
        - Missing elements: Missing several central economic terms and/or correct definitions. (1, 2, 3 Points)
        - Meets expectations: Almost all economic concepts central to the policies and markets are identified and/or correctly defined. (4 Points)
        - Exceeds Expectations: All economic concepts central to the policies and markets are identified and correctly defined in a way that exceeds expectations for the course. (5 Points)
    - **Linking Concepts**:
        - Missing elements: Several connections between relevant concepts and markets are missing or incorrect. (1, 2, 3 Points)
        - Meets expectations: Mostly correct connections between relevant concepts and markets in a way that demonstrates an understanding consistent for the course. (4 Points)
        - Exceeds Expectations: Building upon their definitions, the writer correctly connects the relevant concepts and markets to one another demonstrating an understanding that is sophisticated for the course. (5 Points)
    - **Conciseness**:
        - Missing elements: Over word count (>510 words), with a large number of sentences and/or words that are not directly related to the prompt. (1, 2, 3 Points)
        - Meets expectations: Within the word count (up to 510 words), with places in which descriptions are wordy, suggesting a lack of understanding.
        - Exceeds Expectations: Able to answer question succinctly in less than 500 words.
- CRITICAL THINKING
    - **Interpreting Sources**:
        - Missing elements: Interpretation suggests a limited understanding of the economics within the source(s); Writing suggests source(s) may not have been read; No citations. (1, 2, 3 Points)
        - Meets expectations: Interprets the economics within the article consistent with expectations for the course while predominantly quoting the source(s); An attempt to cite source(s) is made. (4 Points)
        - Exceeds Expectations: Accurately interprets and articulates the economics within the source(s) in a sophisticated manner while predominantly summarizing source(s); citations properly formatted. (5 Points)
    - **Analysis of Case Study**:
        - Missing elements: Opposition to a policy is missing; Or the argument does not acknowledge the economic impact of the policy. (1, 2, 3 Points)
        - Meets expectations: Argument effectively debunks the proposed solution in a manner consistent with the level of the class. Full and thorough articulation of each market interaction is not present. (4 Points)
        - Exceeds Expectations: Insightful articulation of the issues facing one of the proposed solutions. All market interactions are explored, coming to a conclusion indicating that the proposed solution is not economical. Accurately interweaves each economic concept present in the proposal into their articulation of the downsides. (5 Points)
- RESPONSE ALIGNMENT WITH AUDIENCE
    - Missing elements: Explanations do not align with the expected audience; Recommendations are inconsistent with the target audience, but carry reasonable economic analysis. Ex: Recommending government action when the audience is producers. (1, 2, 3 Points)
    - Meets expectations: Explanations generally align with audience needs but tend to be too advanced or too simple for the specific audience; Format mostly correct. (4 Points)
    - Exceeds Expectations: Assumes format described in prompt and explains concepts in a way that consistently meets audience needs. (5 Points)

Table 7: Example of 3 peer reviews. The results for combining this into a single peer review is shown in Table 8.

REVIEW 1:

Understanding 1: Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.: The author did not include anything about non-binding and what it has to do with minimum wage. Other than that, the author did not miss anything and did not include anything extra.

Understanding 2: How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?: The author can discuss substitutes relevant to automation to better connect their topic. Additionally, the author can talk more about the labor market which will help connect the job loss from automation.

Critical Thinking 1: Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?: I liked how the author talked about the graph in explaning these concepts, but I believe they could bring in real-world examples to help strengthen their argument and give the reader better understanding. Additionally, the author can talk a little bit more about the long term effects of automation.

Critical Thinking 2: How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.: Like I said before, the author can incorporate more effects of automation which would help influence the reader to take their side rather than oppose it.

Critical Thinking 3: Are all outside sources properly cited?: No, it is not cited correctly. There are punctuation and capitalization errors. **Response Alignment with Audience:** The letter should be understandable to a person with a basic but not sophisticated understanding of economic principles. In this context, which parts were difficult to understand? Which parts were easy to understand?: The author should explain the concept of deadweight loss better. However, everything else in the paper is explained in the proper manner and is easy to understand.

REVIEW 2:

Understanding 1: Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.: There are no missing terms or important details.

Understanding 2: How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?: I think that some explanations could be more clearer and easier to understand. For example, when you say that a binding price control will ensure that the market will not reach equilibrium. You could first mention what a binding price control is and also explain how that applies to minimum wage. Another place where you could connect ideas better is when you mention that "This surplus causes deadweight loss and inefficiency", even though you mentioned that the market is below the equilibrium before. I feel like the connection between that and the inefficiency is not clear. Instead, you could explain that the market is below equilibrium, which causes inefficiency immediately after or before you say that the surplus causes deadweight loss and inefficiency.

Critical Thinking 1: Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?: One thing that could help you improve your analysis on automation is the effects of society. In your paper, you only mentioned the effects of the labor market and the employers, you never really mention the effects of these changes on society. Adding this could help improve your analysis of automation and allow you to look at more negative factors.

Critical Thinking 2: How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.: One thing that you could do to improve your argument would be going more in-depth into the markets. For example, when you can mention the effect of the goods when you want to automate or not. For example, when you have automation, you will produce more goods and will therefore drop the price. Compared to a scenario which does not automate, they make less goods, which will have an increased price compared to the one with automation. Citizens will more likely want to purchase the cheaper option, which opposes the ban on automation.

REVIEW 3:

Understanding 1: Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.: You talked about all the topics and I think you layed out your essay nicely. However, I would expand upon the influence of the policies/concepts more. I don't think it's necessary to introduce producer surplus in this essay, or at least not in the way it's currently pulled in.

Understanding 2: How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?: I think adding to your second and fourth paragraphs by diving into the specific effects of each policy would strengthen your explanations about the connections between each market. The fourth paragraph is kind of confusing and I feel like you may be extrapolating past the immediate effects.

Critical Thinking 1: Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?: I couldn't tell which demand curve shift you were describing at the start of your 4th paragraph. I think it shows your shift to talking about the automation market, but it's not super clear. I think it would add to your argument to talk more about the unemployment that would be caused by banning automation for all the workers involved in that market and those in the companies forced to close due to high labor prices.

Critical Thinking 2: How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.: I think pointing out that labor and automation are substitutes was really smart and helped illustrate the point you're making. A few effects are misattributed. For example, the deadweight loss is a result of the price floor, not the surplus. Also, instead of quoting Dr. Dudley's slides, I would rephrase those definitions in your own words to demonstrate understanding.

Critical Thinking 3: Are all outside sources properly cited?: They properly cited the article provided and another outside source with in-text citations and a correct works cited page. But, you don't need to cite Dr. Dudley's slides in this paper.

Response Alignment with Audience: The letter should be understandable to a person with a basic but not sophisticated understanding of economic principles. In this context, which parts were difficult to understand? Which parts were easy to understand?: The assignment does suggest you sign off with a concerned citizen, but your sign-off works as well, but I'd change "economic" to 'economics'. You have some grammatical errors scattered throughout the essay, fixing those may make the paper a little easier to read and more professional given the audience. But overall the order of your essay and the stance you take is well-founded. Good job!

Table 8: The result of combining the 3 peer reviews shown in Table 7 using the prompt shown in A.4.

OUTPUT FROM `gpt-3.5`
Understanding 1: Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.: The author did not include anything about non-binding and what it has to do with minimum wage. Additionally, while you talked about all the topics and laid out your essay nicely, expanding on the influence of the policies/concepts more would be beneficial. There are no missing terms or important details overall, but it's not necessary to introduce producer surplus in this essay, or at least not in the way it's currently pulled in.
Understanding 2: How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?: The author can discuss substitutes relevant to automation to better connect their topic. Additionally, the author can talk more about the labor market to help connect the job loss from automation. Some explanations could be clearer and easier to understand. For example, when you say that a binding price control will ensure that the market will not reach equilibrium, you could first mention what a binding price control is and explain how that applies to minimum wage. Strengthening the second and fourth paragraphs by diving into the specific effects of each policy would also help. Clarifying that the market being below equilibrium causes inefficiency would make the connection between surplus, deadweight loss, and inefficiency clearer.
Critical Thinking 1: Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?: I liked how the author talked about the graph in explaining these concepts, but incorporating real-world examples would help strengthen the argument. One thing that could help improve your analysis on automation is discussing the effects on society, not just the labor market and employers. Additionally, discussing more about the long-term effects of automation and the unemployment caused by banning automation would be beneficial. Clarifying which demand curve shift you are describing in the fourth paragraph would also add clarity.
Critical Thinking 2: How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.: The author can incorporate more effects of automation to help influence the reader to take their side rather than oppose it. Going more in-depth into the markets, such as mentioning the effects on goods prices with and without automation, would strengthen the argument. Pointing out that labor and automation are substitutes was effective, but correctly attributing deadweight loss to the price floor rather than the surplus, and rephrasing definitions from class materials in your own words, would improve the application of economic principles.
Critical Thinking 3: Are all outside sources properly cited?: No, it is not cited correctly. There are punctuation and capitalization errors. You don't need to cite Dr. Dudley's slides in this paper.
Response Alignment with Audience: The letter should be understandable to a person with a basic but not sophisticated understanding of economic principles. In this context, which parts were difficult to understand? Which parts were easy to understand?: The author should explain the concept of deadweight loss better. However, everything else in the paper is explained in a proper manner and is easy to understand. There are some typos in your paper, such as misspelling "substitute". While the order of your essay and your stance are well-founded, fixing grammatical errors and using professional language will make the paper easier to read and more aligned with the audience. The assignment suggests signing off as "a concerned citizen," but your sign-off works as well. Good job!

APPROACH	STUDENT SIMULATORS		
	`llama3-8b`		
	0.7	0.85	1.0
`gpt-3.5`	75.3	75.7	72.4
`gpt-4`	**78.3**	76.4	75.2
`sft-from-human`	73.7	74.5	73.4
PROF, $\tau = 0.7$			
Iteration 1	76.9	76.4	**75.5**
Iteration 2	76.3	76.8	75.2
Iteration 3	76.4	77.1	74.7
PROF, $\tau = 0.85$			
Iteration 1	76.6	76.5	72.0
Iteration 2	76.5	76.2	73.6
Iteration 3	77.5	**77.8**	72.0
PROF, $\tau = 1.0$			
Iteration 1	73.5	76.4	75.1
Iteration 2	76.3	76.3	68.4
Iteration 3	75.7	75.1	75.3

Table 9: Extrinsic evaluation of the generated feedback in terms of simulated revision performance. **Green** and Blue represents best and second-best performance respectively. Each experiment was repeated for 5 different seeds to mitigate the impact of randomness.

labor market and this improves the persuasiveness of the argument.

On the other hand, the changes implemented by `llama3-8b` results in the deletion of an important line that explains why employers would prefer automation market (*"The market of goods produced by these employers ... more goods produced and greater revenue."*). While it makes the paragraph more technical by explaining how the supply curve and the demand curve is impacted using the concept of price floor(*"The supply curve will ... demand curve remains the same."*), it is unclear and not immediately deducible from the previous statements. This negatively impacts the clarity and persuasiveness of the argument.

D Analyses of Feedback Generation: Additional Details

D.1 Extrinsic Evaluation using `llama3-8b`

In this section, we attach the effectiveness of the generated feedback using `llama-3b` in Table 9. On a high-level, we notice that our approaches perform better than the few-shot approaches involving enterprise LLMs and `sft-from-human`. This indicates that for behavior traits associated with poor revision performance, our approaches are more effective.

D.2 Example of Scoring a Feedback along Pedagogical Dimensions

In this section, we attach the output from `gpt-4` which contains the analysis of a feedback along pedgagogical aspects in Table 10.

E Annotation Guidelines: Student Simulator Faithfulness

The detailed guidelines for conducting student simulator faithfulness analysis is provided in Table 11.

Table 10: Example of a feedback evaluated along pedagogical aspects by `gpt-4`.

FEEDBACK:

Understanding 1: Based on your class discussion and course readings, identify any important concepts that are missing. Identify any unnecessary concepts in use.: As I was reading through this letter, I noticed that many fundamental concepts were missing. For example, the author forgot to mention that an increase in the minimum wage, set by a binding price floor, results in a surplus of labor, as quantity supplied is greater than quantity demanded. Also, if the price of labor goes up in tandem with an imposition of taxes on automation, then the supply of goods and services produced using labor and automation will decrease, causing equilibrium price and quantity to increase and decrease respectively. Additionally, the author failed to mention how consumer, producer, and total surplus would be affected by the tax, which is a crucial component of this prompt.

Understanding 2: How can the author connect concepts in a more useful manner? For example, using your knowledge from class, how could the author improve their explanation of interactions between the various markets affected by these policies?: To reiterate, I felt that while the author does a good job at talking about the markets individually, they do not talk about how they are all interconnected. The author can explain that the change in minimum wage would not only cause a surplus in labor, but an increase in the demand for automation, because labor and automation are most possibly substitutes in the production of particular goods and services. When the ban is imposed on automation, the dynamics between these markets become much more complex. For example, in the market for goods and services, the supply of these goods and services now decreases, reducing equilibrium quantity while simultaneously increasing the equilibrium price. Therefore, not only would suppliers pay, in the form of added input costs, but so would consumers, who are now forced to pay a higher price to receive the same goods that were cheaper before the tax on automation.

Critical Thinking 1: Based on your class discussion and course readings, how could the author improve their analysis of the minimum wage increase and the automation policy (ban or tax, depending)?: The author gave a very brief description of how an increase in the minimum wage affects markets, such as the labor market, by stating that it "triggers various economic repercussions." beyond that, they did not, but should explain how it creates a surplus of labor in the labor market, as laborers are enticed by a higher wage, while employers are deterred, because they see it as a potential rise in input costs, which will significantly affect profits. They can also improve their analysis of the tax on automation by explaining how it may affect both producers and consumers of the goods and services produced by labor and automation. Hence, both producers and consumers of these goods and services may experience a decrease in producer and consumer surplus.

Critical Thinking 2: How well does the author apply economic principles to justify his/her position? Suggest one (or two) additional ways the author could apply economics to their argument in order to make this letter more persuasive.: In my opinion, I feel that there was a lot more the author could have done to effectively justify his/her position. It appeared to me that the author's primary argument against taxation was that producers will still continue to use automation regardless of a tax, which is not necessarily true. To advance their argument, the author could have spoken from an efficiency standpoint, explaining that total welfare is prevented from being maximized as a result of the tax on automation; this would occur, because of either a decrease in consumer or producer surplus or both. Also, they could have argued that automation acts as a lifeline for producers who need a cheaper and more efficient alternative to labor, and adding the tax on automation would force many of these producers to exit the market and stop producing.

Critical Thinking 3: Are all outside sources properly cited?: From my understanding, the writer used the Saltsman article on the wage floor, but did not accurately cite it. They could resolve this by citing the article on a work cited page.

Response Alignment with Audience: The letter should be understandable to a person with a basic but not sophisticated understanding of economic principles. In this context, which parts were difficult to understand? Which parts were easy to understand?: In all honesty, while the content discussed in the letter was pretty comprehensible, I found the general argument against the tax on automation to be confusing. For instance, the author stated that, "If we only put a tax on automation, automation will still replace the labor force." However, I believe that this truly depends on the magnitude of the tax on automation; if producers can generate more profit using human labor than automation, then they will use labor, and not automation like the author implied." On a more positive note, a part that I thought was very well explained was the potential solutions to the rising minimum wage, such as investing in education.

SCORE FROM `gpt-4`:

Respects Guided Question: The feedback partially aligns with the guided questions by providing a detailed critique of missing economic concepts and their interconnections, as well as suggestions to strengthen the essay's arguments. However, it falls short in directly addressing some specific guided questions, like the precise connection between the minimum wage increase and automation policies. The feedback touches on these topics but doesn't fully explore their interrelation as prompted. (2 Points)

Encourages Active Learning: The feedback encourages the author to think more deeply about the economic implications and connections between concepts like labor market surplus and the effects of automation tax. Suggestions for linking these concepts and exploring their impact on both producers and consumers promote a more active engagement with the material. However, the feedback sometimes outlines specific deficiencies and corrections rather than prompting the author to identify these gaps independently. (3 Points)

Deepens Metacognition: The feedback effectively identifies key areas where the student's understanding and application of economic principles are lacking, such as the impact of minimum wage increases on labor surplus and the overall economic system. It calls attention to the need for a deeper analysis of how these factors interact, thus helping the student recognize the complexities of economic policies. However, it could further enhance metacognition by asking reflective questions that lead the student to evaluate their reasoning process. (4 Points)

Motivates and Stimulates Student Curiosity: The tone of the feedback is constructive and focuses on areas for improvement, which can motivate the student to delve deeper into economic theories. However, it lacks elements that explicitly encourage curiosity, such as posing intriguing questions or suggesting novel viewpoints that could expand the student's perspective on the topic. (3 Points)

Adapts to Essay Quality: The feedback is well-tailored to the quality of the essay. It recognizes the student's basic understanding and accordingly provides detailed explanations and corrections that aim to elevate the student's analysis from a simplistic view of economic impacts to a more nuanced understanding. The feedback is specific where needed but also pushes for a higher-level synthesis of economic interactions. (4 Points)

Table 11: Annotation Guidelines for conducting student simulator faithfulness analysis

Your task is to assess the degree to which the student simulator incorporates the feedback generated. Please adhere to the outlined steps sequentially to mitigate potential errors and biases.

1. Annotate the feedback: Each essay is accompanied by three feedback responses from student peer evaluators, all pertaining to the **SAME** essay. These responses may contain repetitions either across the three feedback or within a single feedback. As you read through the feedback, mark each unique suggested revision for annotation.
    - Highlight each new suggestion (i.e., recommended revision) you encounter using a distinct color. Ensure your chosen color is as different as possible from others. Immediately following the highlighted sentence, provide a succinct summary of the suggestion in parenthesis.
    - If the suggestion (i.e., recommended revision) has already been proposed either within the same feedback or in a separate feedback out of the three peers' feedback, shade it in gray. Do **NOT** consider this as new feedback. Immediately following the shaded sentence, paste your previous summary of the suggestion (or a combination of suggestions).
2. Put together all distinct suggestions into a list. And document the number of recommended changes (i.e. suggestions) in the provided table. Example:
    - Suggestion1: discuss consumer/producer surplus
    - Suggestion2: discuss equilibrium
    - Suggestion3: discuss effect of min wage on supply/demand of automation market
3. Evaluate the revised essay by the real student, simulated student (temp = 0.70), simulated student (temp = 0.85), simulated student (temp = 1.00) respectively. Document the following in the provided table:
    (a) Faithful Revisions (number of recommended revisions that the student correctly implements in the revised essay)
    (b) Recommended Revisions Not Accurately Implemented and Unfaithful Revisions
        i. Ignored Changes: The student does not make any revisions to address the suggestion.
        ii. Misinterpreted Changes: The student attempts to address the suggestion, but the student's actual revision differs or deviates from the feedback. In this case, the intent of the revision is aligned with the provided suggestion.
        iii. Inadequate Changes: The student attempts to implement the suggestion but fails to address some explicit directive parts in the feedback such that the revision is highly insufficient to address the issue raised in the feedback or significantly impairs the original meaning of the suggestion.
        iv. Unfaithful Revisions: The student introduces a substantial revision that does not correspond to any provided suggestions (i.e. recommended changes). Contrary to misinterpreted and inadequate changes, the intent of the change does not aligns with any of the suggestions.

It is important to begin with the how this new minimum wage increase economically impacts the labor market. Increasing the minimum wage to a level that transitioned from non-binding to binding means that the new price of hiring workers is now above the market's previous equilibrium. Therefore, with the cost of workers now increased to a level that the market does not intend, employers will hire less workers because of not wanting to spend more on individual workers. Thus, a surplus of workers will now arise. Not to mention, this consequently impacts the automation market since employers will now seek automated labor as opposed to humans because it is now a cheaper solution. The market of goods produced by these employers will also be affected by an increased efficiency from the increased use of automation, resulting in more goods produced and greater revenue.

Initial Writing

It is important to begin with the how this new minimum wage increase economically impacts the labor market. Increasing the minimum wage to a level that transitioned from non-binding to binding means that the new price of hiring workers is now above the market's previous equilibrium, [redacted] Therefore, with the cost of workers now increased to a level that the market does not intend, employers will hire less workers because of not wanting to spend more on individual workers, [redacted] Not to mention, this consequently impacts the automation market since employers will now seek automated labor [redacted] The market of goods produced by these employers will also be affected by an increased efficiency from the increased use of automation, resulting in more goods produced and greater revenue.

gpt-3.5-turbo

It is important to begin with how this new minimum wage increase economically impacts the labor market. Increasing the minimum wage to a level that transitioned from non-binding to binding means that the new price of hiring workers exceeds the market's previous equilibrium. [redacted] Therefore, with the cost of workers now increased to a level that the market does not intend, employers will hire fewer workers because of not wanting to spend more on individual workers. Thus, a surplus of workers will now arise, [redacted]

llama3-8b-instruct

It is important to begin with the how this new minimum wage increase economically impacts the labor market. Increasing the minimum wage to a level that transitioned from non-binding to binding means that the new price of hiring workers is now above the market's previous equilibrium,[redacted]. Therefore, employers will hire less workers because of not wanting to spend more on individual workers. Thus, a surplus,[redacted] The automation market is impacted since employers will now seek automated labor as opposed to humans because [redacted] The market of goods produced by these employers will also be affected by an increased output efficiency from the increased use of automation, resulting in more goods produced at a lower cost, meaning a greater revenue for employers, even though it will displace many workers, which is opposite what your proposal desires to achieve.

Revised Writing

Figure 7: This figure illustrates the modifications made to a paragraph in the initial writing by both student simulators and a real student. It specifically highlights the ineffective changes made by the `llama-8b-instruct` simulator, in contrast to the improvements made by the `gpt-3.5` simulator, which ultimately degrade the overall quality of the essay. For further insights into the subpar performance of the `llama-8b-instruct` based student simulator, please refer to §C.2.