LEAF🍃: Language Learners' English Essays and Feedback Corpus

Shabnam Behzad
Georgetown University
shabnam@cs.georgetown.edu

Omid Kashefi
Educational Testing Service (ETS)
okashefi@ets.org

Swapna Somasundaran
Educational Testing Service (ETS)
ssomasundaran@ets.org

Abstract

This paper addresses the issue of automated feedback generation for English language learners by presenting a corpus of English essays and their corresponding feedback, called LEAF, collected from the "essayforum" website. The corpus comprises approximately 6K essay-feedback pairs, offering a diverse and valuable resource for developing personalized feedback generation systems that address the critical deficiencies within essays, spanning from rectifying grammatical errors to offering insights on argumentative aspects and organizational coherence. Using this corpus, we present and compare multiple feedback generation baselines. Our findings shed light on the challenges of providing personalized feedback and highlight the potential of the LEAF corpus in advancing automated essay evaluation.

1 Introduction

The educational technology landscape is undergoing a radical transformation, shaped by the profound influence of the Internet and AI. Assessment and evaluation are integral parts of education (Valenti et al., 2003), and the cost-effectiveness and efficiency issues in providing corrections and feedback to students are noteworthy considerations in modern educational practices (Shen et al., 2023).

Recent work have mainly focused on sentence-level grammatical feedback comment generation (Nagata, 2019; Han et al., 2019; Pilan et al., 2020; Hanawa et al., 2021; Behzad et al., 2023a,b; Coyne, 2023) and there have been extensive efforts in the field of NLP focused on essay scoring (Ke and Ng, 2019), but the task of generating effective and personalized feedback remains insufficiently explored and studied. Current automated essay scoring models offer useful overall scores; however, they lack the granularity desired by both learners and instructors seeking more detailed insights. Delivering effective feedback on student essays is essential for enhancing learning outcomes, yet it poses significant challenges and could be very time-consuming, particularly when providing personalized and informative guidance (Carless, 2006; Hattie and Timperley, 2007).

Some studies looked at the use of NLP techniques to provide feedback comments to learners. Criterion (Burstein et al., 2004) was one of the first tools that offered comments on different parts of the essay but was not open-sourced. ArgRewrite (Zhang et al., 2016; Kashefi et al., 2022) is a revision assistant for argumentative writings. Liu et al. (2017) present a feedback system that incorporates predefined questions for each predefined essay feature (e.g. Grammar, Sentence Diversity, Supporting Ideas). These questions advise students to review their essays, but they are not very specific, for example:

> The communication of your ideas needs more strong cohesive cues and devices, such as transitions and connective phrases that link ideas. Do you use them correctly?

More recently, Gong et al. (2021) and Zhang et al. (2022) proposed methods for Chinese feedback generation based on sets of keywords and features, and sets of templates respectively. Feedback Prize competition (Baffour et al., 2023) introduced tasks involving segmenting essays into sections and assigning discourse labels like lead, position, claim, and evidence. Participants also predicted the effectiveness rating of these labels. Han et al. (2023) presented EssayCoT, a Chain-of-Thought prompting strategy that uses scores predicted from their essay scoring system to generate feedback. Behzad et al. (2024) shows that while LLMs such as GPT-4 are capable of producing reasonable feedback, they still lack certain crucial elements associated with *constructive feedback* (Ende, 1983; Ovando, 1994; Omer and Abdularhim, 2017).

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 433–442
June 16-21, 2024 ©2024 Association for Computational Linguistics

Most current feedback systems cannot account for the diversity in students' writing and often appear generic, failing to be customized according to the specific content of the essay. In fact, to the best of our knowledge, there are no publicly available datasets as a starting point for this direction of research. In this paper, (i) We present **LEAF**[1]: a dataset of **l**anguage **l**earners' **E**nglish **e**ssays **a**nd **f**eedback, which is the first publicly available English essay feedback generation corpus including open-ended essay-level feedback on multiple aspects such as grammar, organization and arguments of the essay, (ii) We present a few strong baselines using Llama2-7B and retrieval-augmented prompts, sharing results on how they perform on different aspects of *constructive feedback*, and (iii) Our human evaluation shows that Llama2-7B, although smaller, performs better than Llama2-13B on this task. Our proposed retrieval-augmented approach (primed on a few examples of LEAF) improves these results further by better highlighting essay weaknesses.

2 LEAF Corpus

Essay Forum[2] is a platform that assists both native English speakers and those who learn English as a second language in enhancing their writing skills. Users can upload their essays, and educators can provide them with feedback. Our data was collected by crawling the *Writing Feedback* forum of the website. This forum is mainly used by English language learners preparing for English proficiency tests. Data from essayforum has been used in previous work but for the argument generation task (Bao et al., 2022). We used similar approaches for preprocessing but added extra steps to account for feedback posts. These steps included filtering essays based on topic (if an essay needs a reference image/figure, it is removed), and feedback posts based on length (if feedback posts were too short, we removed them). To ensure high feedback quality, we only kept feedback posts from active users (who posted at least 5 feedback posts), and used like/post ratio as another filtering step. Note that in many instances, students may find the feedback helpful, but the act of "liking" the feedback is not a widespread practice on this platform. Thus, we established a like/post rate threshold of 0.3, considering the overall distribution of all ratios.

Comparison with GPT-4. Several works have compared the feedback provided by instructors/peers with that generated automatically, discussing the advantages and disadvantages associated with each feedback source (Bai and Hu, 2017; Zhang and Hyland, 2018; Tian and Zhou, 2020). In a previous work by the same authors (Behzad et al., 2024) human feedback from essayforum.com was studied and compared with feedback from GPT-4 (OpenAI, 2023). Human evaluators were trained to evaluate feedback comments based on the following evaluation criteria:

C1: Which feedback is more **accurate** and **relevant** to the essay content?
C2: Does this feedback highlight the essay's **weaknesses**?
C3: Does this feedback highlight the **strengths** of the essay?
C4: Is the feedback **specific** and **actionable**?
C5: Which feedback is more **helpful** overall?

These criteria are synthesized from the literature on *constructive feedback* (Ende, 1983; Ovando, 1994; Omer and Abdularhim, 2017), except for the first one which we specifically added to see if there are hallucinations in LLM-generated feedback.

Multiple different prompting strategies were explored (Liu et al., 2023) including detailed task description in the prompt, few-shot (using random examples in the prompt, Brown et al., 2020), and *rubric-based prompt*. We developed the *rubric-based* prompt tailored explicitly for this task by reviewing various rubrics related to learner and argumentative writing (Kneupper, 1978; Faigley and Witte, 1981; Burstein et al., 2003; Kashefi et al., 2022), we summarized the essential points for the model's attention in the prompt. Refer to Behzad et al. (2024) for more details on the prompt design.

The study showed that the use of a rubric-based prompt with GPT-4 yielded the highest quality feedback. Additionally, we discovered that although GPT-4 feedback tends to be generally more encouraging and positive, humans excel in providing more specific and actionable comments that concentrate on the most critical issues in the essay. Motivated by these findings, we opted to integrate the feedback of educators in LEAF and GPT-4, creating a new set of feedback that encompasses high-quality elements from both sources. This approach aimed

[1] https://github.com/shabnam-b/LEAF

[2] https://essayforum.com/

Source	Feedback
Human	The prompt restatement + opinion is incorrect. It does not contain the original reference points for the topic basis. What is presented is the personal opinion of the writer, causing a failed task accuracy score. Yes, it is going to fail even as you provide a proper response to the agree or disagree question. That is because you did not provide a proper restatement or representation of the original prompt. Once there is a prompt alteration, as in your presentation, it will automatically have a failing TA score. <quotation>Because having</quotation>. You cannot start a sentence with a connecting word. It is in fact, unnecessary in your sentence presentation. It was only used as an unnecessary word filler. Avoid using unnecessary words in your sentence formation, this can have a negative GRA effect.
GPT-4	Your essay largely addresses the given topic, which is that employers should give their employees longer vacations. The arguments presented are coherent and relevant. However, there's room for improvement in several areas. 1) Spelling and Grammar: Your essay does contain some grammatical errors. For example, in the sentence, "Firstly, the first reason why letting workers to have more day off…", it should be "more days off". In the same sentence, instead of "letting workers to have", it should be "letting workers have". Also, there is unnecessary repetition of words - like "Firstly, the first..", which makes the sentence redundant and confusing. The positioning and use of prepositions need some corrections too. 2) Point of View: Your stance is clear; you agree with the statement that employers should give employees longer vacations. However, it would be more engaging if you introduced your viewpoint at the beginning of the essay itself. 3) Arguments to Support the Point of View: You presented pertinent arguments that longer holidays can enhance work performance and foster creativity. The examples provided underscore your arguments well. But, they could advantage from being more precise and detailed, explore different sectors or types of employees rather than focusing just on those working in Marketing. Adding statistics or research findings can strengthen your arguments further. 4) Be careful with your final sentence: it doesn't wrap up your ideas in a satisfying way. It would be advisable to provide a more comprehensive conclusion that briefly recaps the main points of the essay and your viewpoint. Continue practising and refining your writing. With time and revision, you'll make significant progress. Good luck!
AI-Augmented (our corpus)	The essay you wrote sufficiently addresses the prompt regarding longer vacation periods for employees. However, several areas could use further refinement and improvements. 1) Topic Response: You started the essay by sharing your personal opinion instead of restating the Given prompt. While it's great that you took a stance, it's crucial to offer a clear restatement of the topic first. Doing so can significantly enhance the clarity and depth of your response to the prompt. 2) Spelling, Grammar, and Stylistic Issues: There are several grammatical errors within your essay. For instance, "letting workers to have more day off…" should be corrected to "letting workers have more days off...". Plural and singular terms should align with their respective contexts. Avoid unnecessary repetition and starting a sentence with a connecting word like "Because". Avoid using unnecessary filler words as they could negatively affect your grammar, accuracy, and coherence. 3) Clarity of Point of View: Although your viewpoint is clear, the presentation could be improved. Consider introducing your stance in the opening paragraph for a more effective presentation of your opinion. 4) Argument Support: Your arguments for why longer holidays could improve work performance and creativity are generally good. However, more specific details, wider field examples, and addition of statistical information or research findings could provide concrete evidence to solidify your position. 5) Conclusion: Your concluding paragraph should provide a brief summary of your discussion and reaffirm your standpoint. Your current conclusion doesn't wrap up your essay in a comprehensive manner. Overall, your essay shows potential but will benefit significantly from better grammar usage, increased precision in arguments, and an improved conclusion. Keep practicing and focusing on these areas for better scores.

Table 1: Example of feedback comments written by a human, by GPT-4 rubric-based prompt, and our corpus. We revise GPT-4 feedback with comments from human feedback to have data points that include elements from both sources.

to enhance the overall quality of constructive feedback across various aspects (See Table 1).

AI-Augmented Feedback. We used GPT-4 as a tool and designed a prompt using both human feedback and previously generated feedback by GPT-4. The prompt asked for a revision of the previous feedback, by incorporating and using elements from human feedback. These AI-augmented feedback, along with human-generated feedback, is incorporated and released with our corpus. LEAF comprises 5,886 essay and feedback pairs, distributed across the train (5,300), development (280), and test (306) sets, with an average length of 292 for feedback and 342 for essays.

Corpus Availability. LEAF is accessible to researchers through the GitHub repository provided in §1. Licensing information is available in Appendix A.

3 Retrieval-augmented Feedback Generation

To demonstrate the effectiveness of LEAF in enhancing the qualitative aspects of (open-source) AI-generated feedback, we conducted experiments with two Llama 2 models (7B and 13B) (Touvron et al., 2023). Two of the authors studied different aspects of 30 Llama-generated feedback comments and findings were similar to Behzad et al. (2024): LLMs may overlook crucial aspects in their feedback such as relevance to the topic, and guidance on organizing the essay based on the writing task prompt. Furthermore, AI-generated comments may lack specificity and actionability in certain instances. As part of this paper, we present a retrieval-augmented feedback generation baseline that would use a few *relevant* LEAF examples to prime the LLMs to the task. We use the LEAF training set as a reference corpus and retrieve relevant examples to use in a few-shot setting (Lewis et al., 2020; Rubin et al., 2022; Mi-

Llama 2	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
vanilla-13B	0.850	0.329	0.061	0.162	0.029
vanilla-7B	0.858	0.384	0.075	0.174	0.063
retrieval augmented (random)-7B	**0.861**	**0.413**	**0.092**	0.180	0.085
retrieval augmented (similar)-7B	**0.861**	0.412	**0.092**	**0.183**	**0.086**

Table 2: Automatic evaluation results for essay feedback generation task. We used Llama 2 in all experiments.

Llama 2	C1 (relevance)	C2 (weakness)	C3 (strength)	C4 (specificity)	C5 (helpfulness)
vanilla-13B	2.3	2.2	**2.8**	1.8	1.9
vanilla-7B	2.3	2.4	**2.8**	2.2	2.2
retrieval augmented (random)-7B	**2.5**	2.6	2.5	**2.6**	**2.6**
retrieval augmented (similar)-7B	**2.5**	**2.7**	2.4	2.5	2.4

Table 3: Human evaluation results for essay feedback generation task. Evaluators were asked to score 1,2 or 3. Details available in Appendix B

alon et al., 2023). We compare this approach with *vanilla* Llama and when examples are *randomly* selected for the retrieval-augmented generation.

To find *relevant* examples, we look at **similarity in writing task prompt** since analyzing instances of our dataset revealed that a common issue among learners is that they are not always addressing the topic properly. Furthermore, human feedback frequently provides insights into structuring the essay in accordance with the specific task prompt. For instance, when the prompt requires expressing agreement or disagreement, or necessitates discussing multiple perspectives, guidance is offered by educators on structuring individual paragraphs to ensure a coherent essay that effectively addresses the prompt within the allotted examination timeframe. Such recommendations are typically absent in feedback generated by automated systems.

We hypothesize by priming LLMs with a few essays on related task prompts, coupled with their corresponding feedback, the generated feedback is more likely to incorporate suggestions concerning essay organization, and relevance to the topic and given task prompt.

To identify similar essays, we employed Sentence-BERT (Reimers and Gurevych, 2019). We first conducted a semantic search with a bi-encoder (paraphrase-mpnet-base-v2) and then re-rank using a cross-encoder (ms-marco-MiniLM-L-6-v2, trained on the MS Marco Passage Ranking task (Reimers and Gurevych, 2021)). Then we picked the top 3 most similar instances and arranged them in the prompt, with the most relevant example positioned last.

4 Results and Discussion

Automatic metric scores are presented in Table 2, showing marginal enhancements with the incorporation of data from our corpus in both random and similar retrieval-augmented settings. However, recognizing the importance and reliability of human evaluation in NLG applications (Celikyilmaz et al., 2020), we also conduct a human evaluation study to compare different settings.

We evaluated 20 essays with their feedback comments from our 4 baselines (80 feedback in total). Six human evaluators experienced in NLP/linguistics research (3 female, 3 male. Native or native-like English speakers) were given detailed guidelines (Appendix B) on aspects discussed in §2 and asked to rate each feedback (from 1 to 3, ordinal scale). Each sample was evaluated by 2 people and then their scores were averaged.

Human evaluation results are available in Table 3. In nearly all aspects, our retrieval-augmented baselines yield improvements compared to vanilla models. This shows that our data, even in a random few-shot setting, can improve performance on this task. As hypothesized, retrieving essays with similar topics enhances the generated feedback's ability to highlight and discuss weaknesses.

Surprisingly, in the vanilla setting, Llama 7B outperforms Llama 13B, suggesting that smaller models may still yield reasonable results for this task. Refer to Table 4 for examples of generated feedback.

Despite the improvements observed, the evaluation scores suggest significant room for further improvement, particularly for future use by students in the real world. Furthermore, another key concern is relevance and accuracy, as models often propose unnecessary additions and edits, especially in well-written essays. For instance, common suggestions, such as requesting additional evidence and specific data, may be deemed unnecessary for students striving to meet a word count. In particular, Llama-13B performed better in scenarios involving well-written essays.

Lastly, evaluators observed that generated feedback tends to inaccurately praise aspects not represented in the essay. For example, statements such as "Your essay presents a clear stance on the topic" may be given despite the arguments in the essay lacking clarity and coherence.

5 Conclusion

In this paper, we introduced **LEAF**: a dataset of **l**anguage **l**earners' **E**nglish **e**ssays **a**nd **f**eedback, representing the first publicly available English essay feedback generation corpus. Each essay is paired with both human-written feedback and feedback collaboratively generated by a human and GPT-4. We contend that the latter demonstrates higher quality, considering various crucial aspects of constructive feedback. Our experiments reveal that (i) LLMs face inherent challenges in generating personalized and constructive feedback effectively, and (ii) LEAF can serve as an evaluation benchmark and a valuable resource for training and reference to improve the performance of AI-generated feedback.

6 Limitations

One limitation of our dataset, LEAF, stems from its collection from online writing forums. Despite all of our careful preprocessing and data cleaning efforts, we acknowledge the potential presence of noise. Additionally, the feedback comments are user-provided, and these users may not necessarily be English instructors. We also acknowledge that AI-augmented feedback could be inaccurate and irrelevant in some cases, prompting us to release both human feedback and AI-augmented feedback within our dataset. We encourage researchers to use them with discretion.

Another limitation of our study is the exclusive use of English essays and feedback comments within our dataset. Future work could look into providing feedback in English learners' native languages or creating corpora for learners of other languages. Additionally, we did not study the language proficiency levels and writing abilities of the writers. Instructors might frame their feedback differently based on these criteria, but this has not been our focus when creating the dataset.

Furthermore, if a system or model incorporates this dataset at any stage of its life cycle and is intended for deployment and use with different target populations, it is crucial to carefully consider potential biases in the training data, including native language and level of English proficiency.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. Additionally, we would like to extend our thanks to Tatsuya Aoyama, Rupayan Mallick, Negar Siyari, and Devika Tiwari for their contributions to this study.

References

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246, Toronto, Canada. Association for Computational Linguistics.

Lifang Bai and Guangwei Hu. 2017. In the face of fallible awe feedback: how do students respond? *Educational Psychology*, 37(1):67–81.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundara. 2024. Assessing online writing feedback resources: Generative ai vs. good samaritans. In *LREC-COLING*.

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023a. ELQA: A corpus of metalinguistic questions and answers about English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2047, Toronto, Canada. Association for Computational Linguistics.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2023b. Sentence-level feedback generation for English language learners: Does data augmentation help? In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 53–59, Prague, Czechia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25(3):27.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

David Carless. 2006. Differing perceptions in the feedback process. *Studies in higher education*, 31(2):219–233.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Steven Coyne. 2023. Template-guided grammatical error feedback comment generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 94–104, Dubrovnik, Croatia. Association for Computational Linguistics.

Jack Ende. 1983. Feedback in clinical medical education. *The Journal Of The American Medical Association (JAMA)*, 250(6):777–781.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, 32(4):400–414.

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. IFlyEA: A Chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248, Online. Association for Computational Linguistics.

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*.

Wen-Bin Han, Jhih-Jie Chen, Chingyu Yang, and Jason Chang. 2019. Level-up: Learning to improve proficiency level of essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 207–212, Florence, Italy. Association for Computational Linguistics.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. Argrewrite v. 2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, pages 1–35.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Charles W Kneupper. 1978. Teaching argument: An introduction to the toulmin model. *College Composition and Communication*, 29(3):237–241.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. 2017. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4):502–513.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of* *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ahmad Omer and Mohhamed Abdularhim. 2017. The criteria of constructive feedback: The feedback that counts. *Journal of Health Specialties*, 5(1):45–45.

OpenAI. 2023. Gpt-4 technical report.

Martha N. Ovando. 1994. Constructive feedback: A key to successful teaching and learning. *International Journal of Educational Management*, 8(6):19–22.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 332–339, Marseille, France. European Language Resources Association.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 605–611, Online. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Xinshu Shen, Hongyi Wu, Xiaopeng Bai, Yuanbin Wu, Aimin Zhou, Shaoguang Mao, Tao Ge, and Yan Xia. 2023. Overview of CCL23-eval task 8: Chinese essay fluency evaluation (CEFE) task. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 282–292, Harbin, China. Chinese Information Processing Society of China.

Lili Tian and Yu Zhou. 2020. Learner engagement with automated feedback, peer feedback and teacher feedback in an online efl writing context. *System*.

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*.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 37–41, San Diego, California. Association for Computational Linguistics.

Zhe Victor Zhang and Ken Hyland. 2018. Student engagement with teacher and automated feedback on l2 writing. *Assessing Writing*.

Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022. Automatic comment generation for Chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223, Abu Dhabi, UAE. Association for Computational Linguistics.

A Data License

Essayforum.com has the following terms of use:

> Once you have made a post on the EssayForum.com website, you consent to us dealing with it (within and outside of the forums) as we see fit. By submitting your post, you agree to grant us a royalty-free, perpetual, non-exclusive, unrestricted, worldwide license to use, reproduce, modify, adapt, translate, enhance, transmit, distribute, publicly perform, display, or sublicense any such content in any medium (now in existence or hereinafter developed) and for any purpose, including commercial purposes, and to authorize others to do so. Unless you complete the thread deletion procedure (which is free only if no other members posted in your thread and requires a fee if other users answered your question), you cannot withdraw or retract the post. You cannot seek payments from us in relation to this licence at any time now or in the future. If you do not want your post to be used in this or any other way that we may find appropriate you should not post it on EssayForum.com.

Nevertheless, we contacted them, asking if their data could be shared and used for research purposes. They responded: "You can use the data as long as you reference it to our site."

B Annotation Guidelines

We will share our complete guidelines since we believe human evaluation is very important for this task. Here, we share our scoring instructions which were part of the instruction sent to evaluators:

A) Does feedback contain irrelevant content or incorrect suggestions?
   1. Yes, mainly incorrect/irrelevant information
   2. There are some incorrect/irrelevant information, but also some correct ones
   3. Feedback is 100% correct and relevant.

B) Does this feedback highlight the strengths of the essay?
   1. No positive comments in the feedback
   2. One of the strengths is mentioned
   3. More than one strength is mentioned

C) Does this feedback highlight the essay's weaknesses?
   1. No relevant weakness is mentioned, or only low-priority/not major weaknesses are mentioned
   2. 1-2 weaknesses are discussed, including major ones
   3. More than 2 weaknesses are discussed, including major ones

D) Is this feedback specific and actionable?
   1. Not at all, it's very general
   2. Specific on some points, but not everything is discussed in detail
   3. Specific and actionable on most points

E) How helpful is this feedback?
   1. Not helpful
   2. Somewhat helpful
   3. Helpful

C Generated Feedback Examples

In Table 4, we provide examples of feedback comments from different resources for the following essay:

> **Some people believe that everyone has a right to have access to university education and that the government should make it free for all students no matter what financial background they have. To what extent do you agree or disagree ?**
>
> It is widely believed that all people who are entitled to university education should have provision of free-of-charge education for all, whether rich or poor. From my perspective, I totally disagree with this statement.
>
> On the one hand, there are various compelling reasons why students should be allowed to pursue college for free. It could be explained that it would help reduce the huge tuition from their children learning. As a result, their costs of living were less heavy. Besides, the poor and difficult children would have a chance to go to school more and that would create a more educated workforce.
>
> On the other hand, if the free-college policy was carried out, the national economy would have to suffer varying pressure. the educational quality could be low. That is because the teacher would lose the effort to teach with the little salary, which is not enough to pay for their living. Furthermore, the material facilities for equipping every school wastes a lot that causes budget deficits and the nation would go down day- by-day in an economic recession.
>
> To conclude, although tuition waivers seem to be a good idea for people , I believe that the government should not effectuate abolishing tuition fees.

Theses examples show that there is still a lot of room for improvement in the task of personalized feedback generation on students' essays.

Source	Feedback
Llama 2:	
vanilla-13B	* The essay presents a clear and concise thesis statement that sets the tone for the rest of the essay. * The writer provides some valid points on both sides of the argument, but could have provided more supporting evidence for their position. * The essay could benefit from more specific examples and details to support the writer's claims. * The writer's use of transitions and cohesive devices could be improved to make the essay flow better. * The conclusion could be more concise and summarize the main points made in the essay. * The essay could benefit from more attention to sentence-level grammar and mechanics. Overall, the essay presents a good effort to discuss the topic, but could benefit from more development and refinement of the ideas presented.
vanilla-7B	This essay presents a clear and well-structured argument on the topic of free university education. The writer provides both pros and cons of the issue and offers a well-reasoned conclusion. However, there are a few areas that could be improved upon: * The introduction could be stronger. Instead of simply stating the topic, the writer could provide a brief overview of the issue and explain why it is important. * Some of the sentences are quite long and could be broken up for easier reading. For example, the first sentence of the second paragraph is quite long and could be broken up into two or three shorter sentences. * The writer could provide more specific examples to support their arguments. For instance, they could provide data on the economic impact of free university education or discuss the potential consequences of reducing funding for universities. * The conclusion could be more concise and to the point. Instead of summarizing the main points, the writer could provide a clear and definitive conclusion that summarizes their overall position on the issue. * Overall, this is a good effort, and with some revisions, the essay could be even stronger.
retrieval augmented	Your essay presents a clear and well-structured argument against the idea of providing free university education to all students, regardless of their financial background. You have effectively acknowledged the potential benefits of free education, while also highlighting the potential drawbacks. However, there are a few areas where you could improve your essay: 1. Use of Evidence: While you have provided some valid points against free education, you could have strengthened your argument by incorporating empirical evidence or statistics to support your claims. This would have made your argument more convincing and persuasive. 2. Clarity of Arguments: At times, your argumentation could be clearer and more direct. For instance, you could have explicitly stated how the free-college policy would lead to a decrease in the quality of education or how it would result in budget deficits for the national economy. 3. Conclusion: Your conclusion effectively summarizes your main points, but it could be more impactful if you had provided a more definitive answer to the question. Instead of simply stating that free education is not a good idea, you could have provided a clear explanation of why you hold this position. Overall, your essay demonstrates a good understanding of the topic and a clear perspective on the issue. With some refinements, your argument could be even stronger.
Our Corpus:	
human	<quotation>It is widely believed</quotation> This is an incorrect restatement. There is a world of difference between "some people" and "it is widely believed". The difference being that the former indicates a measure of a group of people, while the latter, refers to a majority of supporters for the given idea. That is not what the original prompt is referring to so you will be scored down for that exaggeration. Now, you are also missing a summarized supporting presentation for your opinion. The thesis statement should have followed with the 2 reasons that you would be discussing in the body of paragraphs. By using the incorrect discussion format in the reasoning section, you will be scored down in terms of word count and valid discussion presentations. The examiner will not consider the comparative discussion you presented in the essay and will deduct that from the existing word count. The essay will then fall under the requirement and be scored as an incompletely developed reasoing essay. It is highly likely that you will fail the test based on these reasons.
AI-Augmented	1. Relevance to the topic: Your essay does touch upon the concept of free university education, which is on point with the given topic. However, your interpretation might not fully align with the original prompt. Specifically, be careful with assertions like "it is widely believed," which could exaggerate the perspective the prompt asks you to consider. The prompt refers to "some people", not a majority. 2. Spelling, Grammar, and Stylistic Issues: While your command of English is commendable, there are some grammar and style concerns. For instance, "children learning" should be "children's learning". Wordings like "the poor and difficult children" and "the nation would go down day-by-day" could be phrased better for clarity and precision. Remember to maintain consistency in punctuation; for example, there should be no space before a comma. 3. Author's POV: Your stance on the issue was initially clear when you mentioned you 'disagree' with the statement that university education should be free. However, your arguments seemed to discuss both sides of the issue leading to a probable dilution of your stance. After stating your position, your subsequent arguments should aim to support and reaffirm this stance. 4. Arguments: While your arguments for both sides of the debate are solid, your essay could be more impactful with better structuring. After your initial stance, the primary focus should be on providing evidence and reasons that back your opinion. Offering deeper insights into why the cons outweigh the pros could strengthen your argument. You could also consider discussing alternative methods to make university education more affordable, reinforcing your disagreement with free education. 5. Thesis Statement: Develop a concise thesis statement that not only states your position, but also outlines the main supporting points that you will be discussing in the body paragraphs of your essay. Remember, writing is a skill that develops with practice. With attention to these areas, your essays will significantly improve. Keep practicing and refining your essays.

Table 4: Example of model-generated feedback and the gold feedback from our corpus