Generating Feedback for School Student Essay with Large Language Models

Anonymous ACL submission

Abstract

This paper explores using large language models (LLMs) like T5, BART, and GPT-4 Turbo to automatically generate feedback on primary and secondary school students' essays. We constructed a dataset which consists of over 740 student essays and tutor feedback across different year levels, based on that we evaluated the performance of prevalent LLMs in feedback generation. After aligning automated evaluation metrics with educational standards in helpfulness, readability, acceptance, relevance and specificity, we further conducted user studies to assess GPT-4's effectiveness in personalized feedback generation. Our findings show that GPT-4 Turbo, especially when using well-designed prompts and reasoning strategies, outperforms models like T5 and BART in providing more readable feedback. Human evaluation also supports the readability and relevance of GPT-4 Turbo's feedback but lacks helpfulness and specificity compared to real tutor feedback.

1 Introduction

Automated feedback generation in writing refers to the use of advanced algorithms, natural language processing (NLP), and machine learning techniques to evaluate and provide tailored, constructive feedback on written texts. These systems assess various aspects of writing (Song and Song, 2023;Liu et al., 2016), such as grammar (Behzad et al., 2022; Nagata, 2019), structure (Fiacco et al., 2019;Putra et al., 2021), coherence (Fabbri et al., 2021), style, and argumentation (Hartwell and Aull, 2022), and then generate personalized recommendations to help writers improve their writing skills. Automated feedback systems can quickly analyze large amounts of text and provide timely, useful feedback, helping students improve their writing skills faster. These systems are valuable in both educational and professional settings, where regular, accurate feedback can enhance learning and writing quality. While automated tools are good at identifying objective issues like grammar and clarity, they often struggle with more subjective elements such as tone, creativity, and persuasiveness. These aspects often require human intervention to provide a more nuanced assessment.

Feedback plays a key role in learning, especially in improving writing skills. Good feedback not only points out mistakes but also helps students better understand how to communicate effectively. It helps them organize their ideas, structure arguments, and express themselves more clearly. However, giving personalized feedback to many students at once is challenging, particularly in large classes. Automated feedback systems can help by reducing the workload for teachers while still giving students the guidance they need to improve their writing.

Automated feedback tools are now commonly used in online learning platforms, writing assistants, and educational apps. These tools provide immediate feedback on grammar, vocabulary, sentence structure (Fiacco et al., 2019), and even style, allowing students to quickly reflect on their work and make improvements in real-time (Grimes and Warschauer, 2010). Beyond their immediate educational benefits, automated feedback systems promote continuous learning and self-improvement. By making high-quality feedback (Palermo and Thomson, 2018) more accessible, these systems help students refine their writing through instant, data-driven insights.

Furthermore, as education and industry demand personalized learning experiences and performance improvement, automatic feedback generation serves as a cornerstone of adaptive learning environments. These systems facilitate tailored feedback that addresses the unique needs of each student or professional, thereby optimizing learning outcomes. The ongoing advancements in natural language processing and machine learning have made it possible to generate feedback not only for

basic language use but also for domain-specific tasks like technical writing, creative storytelling, or academic essays (Raiaan et al., 2024). This has broadened the applicability of feedback systems, extending their relevance across different fields and disciplines.

As the educational landscape evolves to embrace digital tools, the Feedback Generation Task represents a technological initiative aimed at enhancing the efficiency and scalability of the feedback process. By doing so, it not only supports the goals of educators but also empowers learners to take ownership of their writing development, leading to sustained growth and better academic and professional outcomes. Figure 2 shows a sample of one Year 5 student's persuasive writing alongside feedback from both a tutor and GPT using a few-shot chain-of-thought (CoT) prompt. It can be seen that GPT effectively identifies several issues in the writing, demonstrating strong comprehension and providing feedback with high readability. However, while GPT's feedback is clear and well-structured, it shows some limitations in terms of acceptance when compared to the tutor's feedback, particularly in areas that require deeper contextual understanding and nuanced suggestions.

We conducted a series of experiments using pre-trained sequence-to-sequence models, as well as GPT-4, to evaluate the performance of the generated feedback. Both automated and human evaluations were carried out to assess the models' ability to provide effective and meaningful feedback. Our main contributions are as follows:

- We benchmark state-of-the-art large language models(LLMs), including T5, BART, and GPT-4 Turbo with prompt engineering, on the task of automatic feedback generation. These models are evaluated across a range of writing samples to assess their ability to deliver meaningful, personalized feedback.

- To ensure the relevance and practicality of the feedback, we propose an evaluation framework that combines automated metrics with human judgment. By aligning these metrics with real-world educational rubrics, we ensure that the feedback generated by LLMs adheres to established educational standards.

- Using real-world datasets of student writing and teacher feedback, we rigorously assess the performance of GPT-4. Our human evaluation process, alongside automated assessments, verifies the model's ability to align with expert feedback, ensuring that the generated feedback is both pedagogically sound and applicable in real educational settings.

2 Related Work

2.1 Automatic Feedback Generation

Automatic feedback generation systems have emerged in recent years to address the need for timely, personalized feedback in educational settings. These systems have explored various techniques, including natural language processing (NLP), neural architectures, and template-based approaches, to provide formative feedback to students.

Template-Based and Rule-Based Approaches Some early systems used template-based and rule-based techniques for feedback generation. For instance, Gong (Gong et al., 2021) introduced the IFLyEA system that generates reviews using templates based on grammar correction, rhetoric recognition, and discourse coherence analysis. While these approaches are effective in specific contexts, they are limited by their inability to adapt to diverse writing styles and novel content, as they rely heavily on predefined templates and rules.

Machine Learning and Neural Architectures More advanced approaches have focused on leveraging machine learning and neural network models. For example, Fiacco (Fiacco et al., 2019) developed a hierarchical neural architecture designed to provide feedback on rhetorical structure. This method demonstrated superior performance compared to rule-based systems by analyzing both local and global writing cues. However, its reliance on specific writing structures limits its generalizability across diverse writing tasks and genres.

Natural Language Processing (NLP) and Feature Extraction Zhang et al. (Zhang et al., 2019a) introduced eRevise, a system that uses NLP techniques to extract features aligned with a predefined rubric, providing feedback on text-to-response writing. Similarly, Ke et al. (Ke et al., 2019) analyzed thesis strength in students' persuasive writing, annotating specific attributes and using language models to explain how these attributes impacted thesis quality. While these systems offer

formative feedback, their reliance on predefined rubrics and feature extraction may limit flexibility when applied to different writing types and levels.

Generation-Based Approaches Recent work has moved towards generation-based feedback models. Zhang (Zhang et al., 2022) introduced a planning-based model for generating complete comments on Chinese narrative essays, using keywords to filter out incorrect suggestions. While this model achieved significant improvements, it still faces challenges in generating highly personalized feedback across diverse topics.

2.2 Large Language Modelling in feedback generation

Pre-Trained Language Models (PLMs) The use of large pre-trained language models (PLMs) for feedback generation has gained momentum with the development of models such as BART (Lewis et al., 2019) and GPT-4. Jia et al. (Jia et al., 2022) fine-tuned BART to generate feedback for students' project reports and introduced a framework to compare the generated feedback with instructor feedback. While effective in generating coherent feedback, the model's fine-tuning process limits its adaptability to new contexts without additional retraining.

GPT-4 for Feedback Generation Ben Naismith (Naismith et al., 2023) applied GPT-4 to evaluate discourse coherence in English proficiency tests. His results showed that GPT-4 can accurately assess coherence and align with human raters. Similarly, Wei et al. (Dai et al., 2023) investigated ChatGPT's ability to provide feedback on students' project proposal reports, revealing strong alignment between ChatGPT's feedback and that of human instructors. However, these studies focus on specific writing tasks and do not extensively evaluate GPT-4's capacity across various genres and writing levels.

Many systems rely on fixed templates (Gong et al., 2021) or rules, making it difficult to adapt to different writing styles, topics, or student abilities without manual changes. Systems like eRevise (Zhang et al., 2019a) are limited to specific tasks, such as persuasive or narrative writing, which reduces their usefulness across various writing types and grade levels. Human-defined rubrics may overlook important aspects of student writing and fail to capture its full complexity. While models like GPT-4 perform well in some areas, their ability to handle a wide range of student writing hasn't been fully tested.

To address these issues, we evaluate advanced models like T5, BART, and GPT-4 Turbo for automatic feedback generation. We test these models on a diverse student writing dataset, including persuasive, narrative, and text response genres. By combining automated metrics (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019b)) with human evaluation, we ensure the feedback aligns with real-world educational standards. Our focus is on providing personalized feedback tailored to each student, moving beyond rigid templates and rubrics. Using real-world data and thorough evaluation, our approach offers a more flexible and scalable solution for generating high-quality feedback.

2.3 Automatic Evaluation and Human Marking Rubrics

In the evaluation process, the most important task is designing effective evaluation criteria. Good and effective evaluation metrics can provide a vital foundation for the evaluation, it can provide a precise and effective approach to improve the performance of the generated text and evaluate the quality of the generated text. Assessing the quality of the generated text is a significant challenge due to the probabilistic nature of AI language models. These models can produce factual errors during text generation. So we still need to cooperate with human efforts. Human evaluation (van der Lee et al., 2021;Xia et al., 2024)is essential for assessing the quality of the generated text, providing a deeper and more subjective understanding of context. Human evaluators can consider creativity, innovation, and complex tasks that automated metrics may miss. They also address ethical considerations in the text. When combined with automated metrics, human evaluation helps guide language models toward better performance.

We summarize the prominent human evaluation metrics (van der Lee et al., 2021; Jia et al., 2022 as follows: **Readability (READ)** In the realm of readability, Jia's study (Jia et al., 2022) offers a comprehensive definition that encompasses the quality of feedback concerning grammar, word choice, and coherence. This multidimensional evaluation is conducted using a five-point rating scale. The scale ranges from 0, signifying incomprehensibility, to 1, indicating a lack of fluency and co-

herence. Intermediate ratings include somewhat fluent but incoherent (2), fluent but somewhat incoherent (3), and finally, fluent and coherent (4). This nuanced approach captures the intricacies of readability assessment, providing a richer understanding of the machine-generated text's linguistic qualities. **Informativeness (INFO)** Assessing the extent to which the generated hypothesis encapsulates the essential concepts of the source text (Yuan et al., 2021). **Fluency (FLU)** Assessment of whether the text exhibits formatting issues (Yuan et al., 2021), capitalization errors, or ungrammatical sentences, which might impede readability. **Coherence (COH)** Examination of whether the text presents cohesive information of a topic, showcasing a logical progression of sentences (Dang, 2005). **Factuality (FAC)**Verifying that the hypothesis generated includes only statements that are implied by the source text. This involves covering all relevant aspects of the topic (Kryściński et al., 2019). Ensuring that all relevant aspects of the topic are covered in the text. **Semantic Coverage (COV)** Assessing the extent to which the hypothesis covers concepts from the reference texts (Nenkova and Passonneau, 2004). **Adequacy (ADE)** Verification that the output accurately conveys the same message as the input sentence (Yuan et al., 2021), without any loss, addition, or distortion. **Helpfulness** (whether the generated feedback helps students): Assessment of how helpful the generated feedback is for students, rated on a 0-4 Likert scale where 4 represents a more favorable rating. **Acceptance** Evaluation of how easy it is for students to apply the feedback to their writing, measured by user ratings on a 0-4 scale. **Specificity** Assessment of the number of suggestions, solutions, and their percentage, defining scores on a scale of 0-4. **Relevance(hallucination)** Analysis of the consistency of the generated hypothesis with the source text (Grusky et al., 2018).

3 Student Essay and Feedback Dataset Construction

The workflow 1 for this study is divided into three major stages: **Data Preprocessing, Modeling, and Evaluation**. In the Data Preprocessing stage, raw data comprising students' writing samples and corresponding tutor feedback is collected. The first task involves pairing the essays with their corresponding tutor feedback using a similarity function to ensure accurate alignment. This is followed by a human review to assess the quality of the paired data, resulting in a clean and paired dataset as output. For the Modeling stage, the cleaned dataset was used as input. In this stage, models like T5 and Bart are fine-tuned on the dataset. Additionally, prompt engineering techniques such as zero-shot (Dai et al., 2023) and few-shot learning are applied to instruct GPT-4 for generating feedback on student essays. The final stage, Evaluation, involves both automated and human evaluation. The AI-generated feedback is compared to traditional tutor feedback using automated metrics such as BLEU, ROUGE, and BERTScore. Simultaneously, human evaluation is performed by English specialists who assess both AI-generated and tutor feedback based on criteria like **helpfulness, acceptance, relevance, readability, and specificity**.

3.1 Dataset collection

Our student essay feedback dataset originated from a local education company that provides tutoring services for primary and secondary school students. The raw data is organized in a hierarchical folder structure as follows: Family ID -> Student Name -> Writing Types -> Draft Version -> Document Types -> Student Drafts and Tutor Feedback. The description of each folder level is as follows:

- Family ID: Each folder is named after the family ID, with each family potentially having more than one child.
- Student Name: Within the family folder, subfolders are named after each student.
- Writing Types: Inside each student's folder, subfolders are organized by the type of writing (e.g., persuasive, narrative, text-to-response).
- Draft Version: Further subfolders within each writing type represent the draft version (first or second draft).
- Document Types: Each draft folder contains:
- Student Drafts and Tutor Feedback: Usually in docx or pdf formats.
- Supplementary Files: Occasionally includes additional files like .png images.

This hierarchical structure allowed for organized retrieval of both student drafts and corresponding tutor feedback. However, due to the variety of formats, manual intervention, and preprocessing steps were necessary to ensure the correct data was selected and processed.

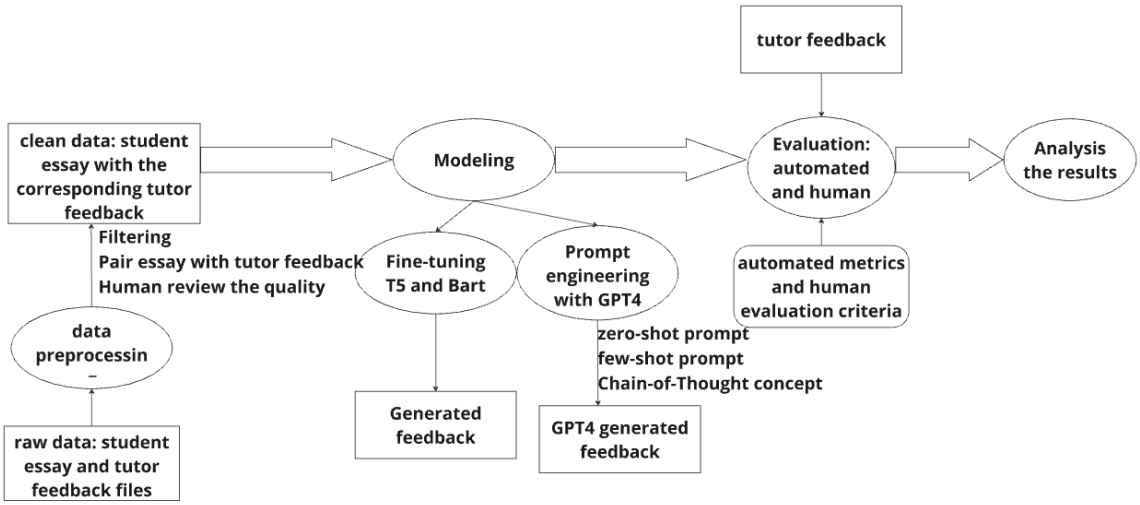


Figure 1: The whole workflow of student essay feedback generation and evaluation with large language models.

3.2 Preprocessing

Filtering for Digital Documents In our initial analysis, the writing samples cover a variety of genres, topics, writing abilities, and different year levels. The student essay formats vary from docx, pdf, to png formats. For this study, we focused on processing only the textual documents (docx and pdf) and excluded non-textual files (i.e., png files) to avoid OCR errors.

Pairing Drafts and Feedback In each student's folder, drafts and feedback were organized by writing title and draft version. However, file names were inconsistent, thus we used an eidt distance based matching algorithm to pair each draft with its corresponding tutor feedback. Specifically, We conducted the following four steps to pair student drafts with tutor feedback files:

- (i) File Name Extraction: File names were extracted without extensions.
- (ii) Similarity Calculation: A similarity score was calculated between each draft and feedback file.
- (iii) Best Match: The feedback file with the highest similarity score was selected as the best match for the draft.
- (iv) Manual Verification: After matching, a manual review was conducted to verify accuracy. Low similarity scores were checked by human evaluators to ensure correct pairing, especially in cases of poorly named files or mismatches.

Manual Review and Quality Assurance While the above automated process handled most of the matching and extraction, a manual review was necessary to ensure the integrity of the dataset. We further reviewed the file pairings and the extracted content to confirm that:

- (i) Each draft was correctly paired with its corresponding tutor feedback.
- (ii) The extracted text was complete and readable, with no missing sections due to file format issues.

This manual review guaranteed the high quality of the student essay and tutor feedback dataset.

3.3 Final Dataset Generation

After successfully matching the drafts and feedback and extracting the text, the final clean data set contains 740 student writings with their corresponding tutor feedback, spanning from Year 3 to Year 11. The dataset captures a range of writing abilities and includes 255 persuasive writing samples, 460 narrative writing samples, and 25 text responses (other writing types). Table 1 summarizes the distribution of different writing types across the year levels.

4 Large Language Models for Feedback Generation

4.1 Fine-Tuning Language Models T5 and BART

Our first task involves fine-tuning two well-established encoder-decoder language models, T5

Description	Number
Number of Persuasive writing	255
Number of Narrative writing	460
Number of Text response writing	25
Total number of essay-feedback pairs	740
Average length of essays (words)	669.13
Average length of feedback (words)	368.87
Vocabulary size of essays	17,891
Vocabulary size of feedbacks	8,442

Table 1: Statistics of Student Essay-Tutor Feedback Data set

(Raffel et al., 2020) and BART (Lewis et al., 2019), for the specific task of feedback generation on student writing. Unlike their typical use cases, such as machine translation or summarization, our models are fine-tuned to generate feedback that reflects a detailed evaluation of student compositions.

We fine-tuned the models using a dataset of 740 student writing samples, each paired with tutor feedback. Through supervised sequence-to-sequence learning, the models learned the relationships between student submissions and expert feedback. Following fine-tuning, the models were evaluated on unseen student writing samples to assess their ability to generalize and produce high-quality feedback.

4.2 Prompt Engineering with GPT-4 Turbo

In the second task, we evaluated the performance of GPT-4 Turbo, a decoder-only model, using zero-shot and few-shot prompting techniques. Unlike fine-tuning approaches, GPT-4 Turbo was used without task-specific training, relying solely on prompt engineering to generate high-quality feedback without explicit task-specific training.

We explored several prompt strategies (Stahl et al., 2024), including basic zero-shot prompts(Li et al., 2023; Dai et al., 2023), zero-shot Chain of Thought (CoT) prompts, and few-shot prompts (Jiang et al., 2023) using both steady and random samples. CoT prompting, introduced by (Wei et al., 2022), facilitates complex reasoning by asking the model to generate intermediate steps in its thought process before arriving at a final response. This allows GPT-4 Turbo to produce more nuanced and contextually aware feedback.

To assess the impact of these prompt engineering techniques, we conducted experiments on a sample of 50 argumentative essays and 50 narrative essays, ensuring the models were tested across different genres of student writing.

Zero-shot Prompt We began by providing GPT-4 Turbo with a simple, zero-shot prompt instructing it to generate feedback based on student writing. The model received no examples of student writing or feedback during this phase, relying solely on the prompt's instructions to generate feedback. This provided a baseline for comparison against more sophisticated prompting strategies.

Zero-shot Chain of Thought (CoT) Prompt In this experiment, we extended the basic zero-shot prompt by incorporating CoT reasoning (Wei et al., 2022), encouraging the model to "think step by step" when generating feedback. This technique aims to enhance the model's ability to produce more detailed and well-structured feedback.

Few-shot Prompt with Steady Samples We introduced a few-shot prompting approach (Stahl et al., 2024) where GPT-4 Turbo was provided with two consistent examples of student writing paired with corresponding feedback. By observing these examples, the model was able to learn the relationships between writing quality and feedback, which improved its ability to generate relevant feedback for new samples.

Few-shot Prompt with Random Samples In this variant of few-shot prompting, we provided the model with two random examples of student writing and feedback. This method exposed the model to a wider variety of writing styles and feedback types, potentially enhancing its ability to generalize to diverse student samples.

Few Shot Chain of Thought (CoT) Prompt with Random Samples We combined the few-shot prompt with the CoT reasoning ((Wang et al., 2022)) technique. This allowed the model not only to observe specific examples of student writing and feedback but also to engage in intermediate reasoning steps, enhancing its capability to generate detailed and personalized feedback.

5 Experiments

Setup In this section, we describe our experimental setup and the metrics used to evaluate the performance of various models in generating feedback on student writing. Our goal is to assess the efficacy of fine-tuned language models (T5 and BART) as well as GPT-4 Turbo in producing personalized, constructive feedback. We evaluate the models

Model	ROUGE-P	ROUGE-R	ROUGE-F1	ROUGE-L	BLEU-P	BERTScore
T5	0.350	0.199	0.251	0.032	0.033	-
BART	0.183	0.163	0.171	0.016	0.020	-
zero-shot	0.280	0.259	0.264	0.012	0.018	0.842
few-shot-steady	0.266	0.277	0.266	0.013	0.019	0.843
zero-shot-CoT	0.277	0.255	0.261	0.011	0.176	0.842
few-shot-dy	**0.292**	0.252	**0.266**	0.011	0.018	0.849
few-shot-dy-CoT	0.286	0.242	0.257	0.012	0.019	**0.849**

Table 2: Statistics of Automatic scores on T5, BART, and GPT-4 Turbo.

Feedback	Readability	Helpfulness	Acceptness	Specificity	Relevance
Teacher feedback	3.91	**3.71**	**3.62**	**3.58**	3.56
zero-shot prompt	3.87	3.08	3.05	3.01	2.99
few-shot steady prompt	3.92	3.64	3.59	3.53	3.52
zero-shot-CoT prompt	3.89	3.50	3.48	3.43	3.49
few-shot-dy prompt	3.92	3.59	3.56	3.55	3.57
few-shot-dy-CoT prompt	**3.94**	3.68	**3.63**	**3.59**	**3.66**

Table 3: Human evaluation on the feedback generated by GPT-4 Turbo.

using both automated metrics and human evaluation criteria to ensure a comprehensive analysis of their performance. We conducted both automated and human evaluations to assess the quality of the generated feedback.

Automated Evaluation We employed standard text generation evaluation metrics— BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019b) to evaluate the n-gram and semantic overlap between LLM generated feedback and ground-truth tutor feedback. For BLEU and ROUGE, their Precision, recall, and F1 scores were calculated separately, we also use BERTScore due to its ability to measure semantic similarity more effectively than traditional lexical-based metrics.

Human Evaluation In collaboration with an industry English specialist, we developed human evaluation criteria designed to align with standard marking rubrics (Dai et al., 2023) and effectively assess the impact of GPT-generated feedback on enhancing students' writing skills. The criteria were tailored to ensure a comprehensive evaluation of both the quality of the feedback and its usefulness in improving student compositions. To achieve this, **we focused on five key evaluation metrics: readability, helpfulness, acceptance, specificity, relevance of the feedback** in fostering improvements in student writing. These criteria provide a structured approach to comparing GPT feedback with traditional tutor feedback and gauging its potential in real-world educational contexts.

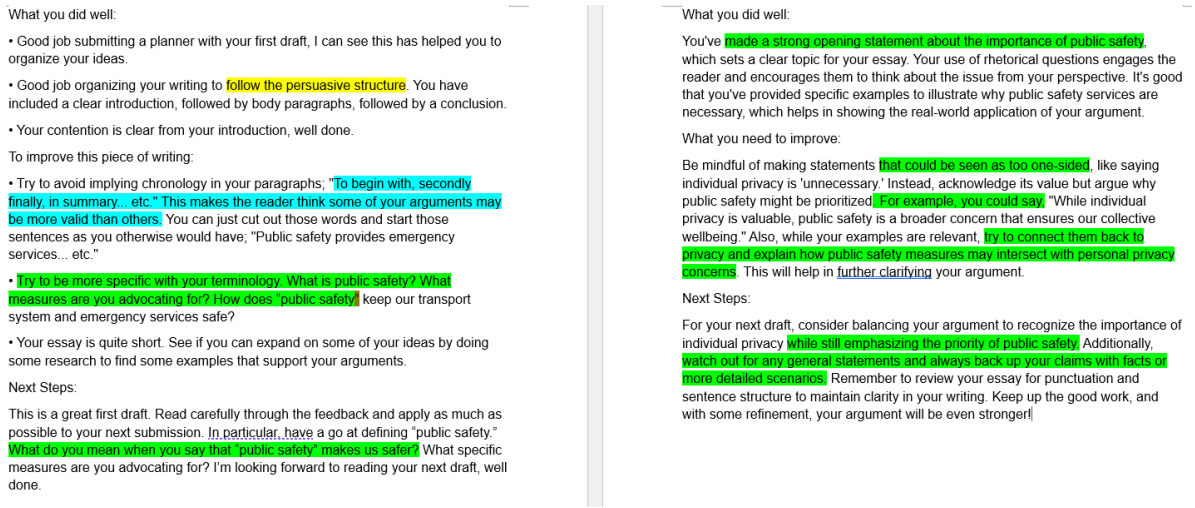
Three experienced teachers rated each feedback using a five-point scale across the defined evaluation dimensions. The average of their scores was then used to compare the quality of AI-generated feedback with that of tutor-provided feedback, allowing for a direct and consistent assessment of the two methods.

5.1 Results and analysis

Our experimental results, presented in Tables 2 and Table 3, reveal that GPT-4 Turbo outperformed T5 and BART, especially when using few-shot prompting with Chain of Thought (CoT) reasoning. This method produced the most effective feedback, particularly in terms of specificity, helpfulness, and relevance.

In the automated evaluation, BERTScore was the most reliable indicator of feedback quality, with GPT-4 Turbo scoring consistently high (0.849) in the few-shot dynamic and CoT settings. This indicates strong agreement between GPT-4 Turbo's feedback and tutor feedback. Traditional metrics like ROUGE and BLEU, which focus on word overlap, were less suited for feedback evaluation. While T5 had the highest ROUGE score (0.350), GPT-4 Turbo performed better in context-aware tasks when using few-shot CoT, achieving the best ROUGE-P score of 0.292. However, these metrics don't fully capture the richness of feedback, making BERTScore a better fit for this task.

The human evaluation showed that GPT-4 Turbo performed best with few-shot dynamic CoT prompting, scoring the highest in Readability (3.94)



What you did well:

• Good job submitting a planner with your first draft, I can see this has helped you to organize your ideas.

• Good job organizing your writing to follow the persuasive structure. You have included a clear introduction, followed by body paragraphs, followed by a conclusion.

• Your contention is clear from your introduction, well done.

To improve this piece of writing:

• Try to avoid implying chronology in your paragraphs; "To begin with, secondly finally, in summary... etc." This makes the reader think some of your arguments may be more valid than others. You can just cut out those words and start those sentences as you otherwise would have; "Public safety provides emergency services... etc."

• Try to be more specific with your terminology. What is public safety? What measures are you advocating for? How does "public safety" keep our transport system and emergency services safe?

• Your essay is quite short. See if you can expand on some of your ideas by doing some research to find some examples that support your arguments.

Next Steps:

This is a great first draft. Read carefully through the feedback and apply as much as possible to your next submission. In particular, have a go at defining "public safety." What do you mean when you say that "public safety" makes us safer? What specific measures are you advocating for? I'm looking forward to reading your next draft, well done.

What you did well:

You've made a strong opening statement about the importance of public safety, which sets a clear topic for your essay. Your use of rhetorical questions engages the reader and encourages them to think about the issue from your perspective. It's good that you've provided specific examples to illustrate why public safety services are necessary, which helps in showing the real-world application of your argument.

What you need to improve:

Be mindful of making statements that could be seen as too one-sided, like saying individual privacy is 'unnecessary.' Instead, acknowledge its value but argue why public safety might be prioritized. For example, you could say, "While individual privacy is valuable, public safety is a broader concern that ensures our collective wellbeing." Also, while your examples are relevant, try to connect them back to privacy and explain how public safety measures may intersect with personal privacy concerns. This will help in further clarifying your argument.

Next Steps:

For your next draft, consider balancing your argument to recognize the importance of individual privacy while still emphasizing the priority of public safety. Additionally, watch out for any general statements and always back up your claims with facts or more detailed scenarios. Remember to review your essay for punctuation and sentence structure to maintain clarity in your writing. Keep up the good work, and with some refinement, your argument will be even stronger!

Figure 2: Sample of Tutor and few-show CoT prompt with GPT feedback

and Relevance (3.66). It even surpassed teacher feedback in relevance, showing that GPT-4 Turbo can produce contextually aware and useful feedback.

In terms of Helpfulness and Specificity, GPT-4 Turbo's scores were very close to those of human feedback, particularly in the few-shot CoT setting. The model was rated nearly as helpful as teachers (3.68 vs. 3.71) and provided similarly specific feedback (3.59 vs. 3.58).

On the other hand, zero-shot prompting without CoT produced the weakest results, especially in Helpfulness (3.08) and Relevance (2.99), showing that structured prompting greatly improves GPT-4 Turbo's performance.

Encoder-Decoder Models Struggle T5 and BART faced challenges with generating feedback due to the complex, context-specific nature of the task. These models, designed for structured tasks like translation, were limited by their fixed input-output training approach, making it difficult to dynamically generate personalized and nuanced feedback. They also struggled with capturing long-term dependencies in student writing, which is crucial for effective feedback.

Effectiveness of Prompt Design Our experiments emphasized the importance of prompt engineering for GPT-4 Turbo. Few-shot prompts showed a significant improvement over zero-shot prompts, demonstrating that providing relevant examples can greatly enhance feedback quality. Chain of Thought (CoT) reasoning further improved the feedback by promoting deeper analysis and structure, leading to more helpful and specific responses.

Human Evaluation of Feedback Human evaluations found GPT-4 Turbo's feedback to be highly rated in terms of clarity and relevance, sometimes even surpassing teacher feedback. Few-shot CoT prompts were especially effective in providing detailed, actionable feedback, highlighting the value of both prompt variety and intermediate reasoning steps for improved performance.

6 Conclusion

This study examined the use of large language models (LLMs) like T5, BART, and GPT-4 Turbo to automatically generate feedback on student essays. While T5 and BART performed well in structured tasks, they struggled to provide detailed, personalized feedback. They had difficulty capturing the complexity and context needed for effective feedback.GPT-4 Turbo, especially when using few-shot and Chain of Thought (CoT) prompting, outperformed T5 and BART. It generated more specific, clear, and human-like feedback, with strong performance in both automated metrics (like BERTScore) and human evaluations. Teachers rated GPT-4 Turbo's feedback highly, sometimes even higher than their own.

7 Limitations

Our work has a few limitations: (i) We plan to expand the dataset to include more writing genres, and further improve prompt design to ensure the feedback is adaptable and personalized for different writing tasks and student abilities. (ii) We have not

tried in-context learning with open-sourced large language models, e.g., Llama3, the main reason is the limited amount of essays with feedback data. We plan to use data augmentation with in-context learning for better model learning. (iii) Our work only contains one round of tutor evaluation, in the real-world setting students may take the feedback and submit an improved version of their essays. Further multi-round user study can be conducted to validate the effectiveness of the generated feedback.

8 Ethics Statement

Data collection approval was received from an ethics review board. No identified student or tutor information is collected in the data collection process. All code and data used in this paper comply with the license for use.

References

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2022. Sentence-level feedback generation for english language learners: Does data augmentation help? *arXiv preprint arXiv:2212.08999*.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

James Fiacco, Elena Cotos, and Carolyn Rosé. 2019. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 310–319.

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.

Douglas Grimes and Mark Warschauer. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, 8(6).

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Kelly Hartwell and Laura Aull. 2022. Constructs of argumentative writing in assessment tools.

Qinjin Jia, Mitchell Young, Yunkai Xiao, Jialin Cui, Chengyuan Liu, Parvez Rashid, and Edward Gehringer. 2022. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. *International Educational Data Mining Society*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback ii: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3994–4004.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling chatgpt for explainable automated student answer assessment. *arXiv preprint arXiv:2305.12962*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

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.

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

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Corey Palermo and Margareta Maria Thomson. 2018. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, 54:255–270.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Parsing argumentative structure in english-as-foreign-language essays. In *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*, pages 97–109.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Cuiping Song and Yanping Song. 2023. Enhancing academic writing skills and motivation: assessing the efficacy of chatgpt in ai-assisted language learning for efl students. *Frontiers in Psychology*, 14:1260843.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019a. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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.

A Prompts examples

zero shot prompt *our task is to provide feedback for students' writing, evaluate students' writing based on these criteria: structure, coherence, the effectiveness of persuasive devices, and clarity of supporting arguments and ideas, punctuation, and cohesion. generate feedback following the template: what you did well: what you need to improve with some examples and solutions: and next steps, using the sentences that school children can understand. using at most 200 words. Student: "draft" Feedback :*

few shot prompt *Examples: 1. Original Text: "example1"Feedback : "feedback1" 2. Original Text: "example2"Feedback : "feedback2" Your task is to provide feedback for students' essays refer to the teacher's feedback in the above examples: evaluate students' writing based on these criteria: structure, coherence, effectiveness of persuasive devices, clarity of supporting arguments and ideas, punctuation, and cohesion. generate feedback following the template: what you did well: what you need to improve with some examples and solutions: and next steps, using the sentences that school children can understand. using at most 200 words. Student: "draft" Feedback :*

For the CoT prompt engineering, we only add one sentence: Let's think step-by-step.

B User case: one student's sample with both tutor feedback and GPT feedback

Draft sample *Draft """ Should children have more break time? Should children have more break time? I believe that children should have more break time. We can make friends and can play more outside games. I think you shouldn't have continuous learning time because you will get bored and if you do too much learning your eyes might hurt. Surely, you should get more break time because you can run, skip, have more fun, get more friends, play more games, and play more things. We can learn something Outside like, for example, you can learn how butterflies can fly. Also, we can make stuff and you have more time to make stuff outside. In addition, if you have a much longer break there can be many great things that can happen. For instance, you can learn or play outside with more time so you can have more fun. Also, you can have fun! Why should you have more break time? I think that you should have more break time because you can play and learn something.There are many different things that are really fun when you go outside and see nature. Certainly, you might want to have more play time and have more break time. Having a much longer break is really good if you have a longer break because you can have more fun and play more games. So it is really good to have much more break time, and there are many different things that are good if children have more break time. In conclusion one can clearly see from the above information that you should have more break time. Without a shadow of doubt in my mind, I can confidently say that children should also have less break time at the same time. Do you agree? """*

Zero-shot feedback *What you did well: You have a clear opinion on why you think lunchtime at school should be longer, which is great! You've provided several reasons to support your opinion, like wanting more time for slow eaters to play, for people to finish their games, and to have fun with friends. You've repeated your main point several times to emphasize it, showing you feel strongly about it. What you need to improve with some examples and solutions: Your writing would benefit from a clearer structure. Try to group similar ideas together into paragraphs. For example, put all your thoughts about why slow eaters need more time into one paragraph. Remember to use punctuation, like periods (.) and commas (,), to make your sentences easier to read. Like this: "If we have longer lunch times, we can play more games, and slow eaters will have enough time to eat and play." Next steps: Practice organizing your thoughts before writing. You can use a graphic organizer with boxes for each reason you want to give. This will help you write clearer paragraphs. Also, remember to check your work for punctuation. Try reading your sentences out loud to see where you naturally pause"that's often where a comma goes. Keep working hard, and you'll improve even more! """*

Zero-shot CoT feedback *What you did well: You have a clear opinion and are enthusiastic about your topic. You provided several reasons why lunchtime should be longer, like allowing slow eaters to finish and have playtime, accommodating longer games, and enhancing social interactions. What you need to improve: Some of your sentences are a bit confusing, which makes it hard to understand your points at times. Also, repeating the same*

idea in different ways can make it seem like you don't have enough different reasons. Watch out for run-on sentences and make your examples more specific to strengthen your argument. For punctuation, remember to use capital letters for "I" and after periods. Examples and solutions: Instead of saying, "the so people like to play," you could say, "Some people like to play short games, but others prefer long games that require more time." In terms of structure, organize your paragraphs around different points. For the first paragraph, discuss the issue with slow eaters; in the second, focus on those who like longer games, and so on. Next steps: Revise your essay by creating a clear plan for each paragraph " each one should introduce a new idea. Check your sentences for run-ons and try to make them more precise. Practice using punctuation correctly, like capitals and periods. Lastly, try to give real-life examples that show exactly why more time is needed, which will make your argument stronger. Great job, and keep writing! """

Few-shot dynamic feedback *What you did well: You've chosen a relatable topic and expressed your opinion clearly which is great for persuasive writing! Your use of examples helps to explain why a longer lunchtime at school might be beneficial, and you've addressed your essay to the reader with a question which is a good way to engage them. What you need to improve: There are a few things you can do to make your writing stronger. It's important to organize your ideas into clear paragraphs, each with one main point. For example, one paragraph can focus on why slow eaters need more time, and another can discuss the benefits of longer games during lunchtime. Watch out for your spelling and try to use full words rather than shortcuts, like "l" should be "I." Also, look out for run-on sentences and try to break them into shorter sentences that are easier to read. Lastly, make sure each sentence has a clear subject and verb so your reader knows what you're saying. Next Steps: You've done a good job starting your essay. Now, try to organize your ideas more clearly into separate paragraphs. Read your work out loud or ask someone to read it to you to catch any mistakes or parts that might be unclear. Remember, practice makes perfect, so keep writing and you'll continue to improve! Keep up the good work, and good luck with your revisions! """*

Few-shot dynamic CoT feedback *What you did well: You've chosen an engaging topic that is relevant to your interests and experiences. Your introduction clearly states your position on the issue, which is great for letting readers know your perspective right away. You also offer a variety of reasons why more break time is beneficial, such as making friends, learning from nature, and having fun. What you need to improve: To strengthen your essay, consider providing more detailed examples or evidence to support your arguments. For example, you could include information about studies that show the benefits of breaks for learning and social development. Be careful with contradictions; the conclusion suggests children should have both more and less break time, which is confusing. Next Steps: For your next draft, clarify your conclusion to support your main argument consistently. Try to find specific examples or research to back up your points about the benefits of extra break time. And when you state a new idea, develop it with a sentence or two explaining why it's important. Keep up the thoughtful writing, and remember to stay clear and focused on your main idea throughout your essay. Good job!*