

Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation

Maja Stahl

Leibniz University Hannover
m.stahl@ai.uni-hannover.de

Leon Biermann

Leibniz University Hannover
leon.biermann@stud.uni-hannover.de

Andreas Nehring

Leibniz University Hannover
nehring@idn.uni-hannover.de

Henning Wachsmuth

Leibniz University Hannover
h.wachsmuth@ai.uni-hannover.de

Abstract

Individual feedback can help students improve their essay writing skills. However, the manual effort required to provide such feedback limits individualization in practice. Automatically-generated essay feedback may serve as an alternative to guide students at their own pace, convenience, and desired frequency. Large language models (LLMs) have demonstrated strong performance in generating coherent and contextually relevant text. Yet, their ability to provide helpful essay feedback is unclear. This work explores several prompting strategies for LLM-based zero-shot and few-shot generation of essay feedback. Inspired by Chain-of-Thought prompting, we study how and to what extent automated essay scoring (AES) can benefit the quality of generated feedback. We evaluate both the AES performance that LLMs can achieve with prompting only and the helpfulness of the generated essay feedback. Our results suggest that tackling AES and feedback generation jointly improves AES performance. However, while our manual evaluation emphasizes the quality of the generated essay feedback, the impact of essay scoring on the generated feedback remains low ultimately.

1 Introduction

Essay writing is a valuable skill that empowers students to communicate effectively, to think critically, and to express their opinions clearly (Bangert-Drowns et al., 2004; Liu et al., 2014; Schendel and Tolmie, 2016). However, learning to write essays of high quality can be challenging (Peloghitis, 2017; Febriani, 2022). Individual feedback to essays is thus critical, so students may recognize and overcome their weaknesses (Riddell, 2015). However, manually writing essay feedback is time- and labor-intensive. Given educators' limited time, this often renders real individual feedback infeasible in practice. As an alternative, automated essay writing support could benefit students by offering guidance at their own pace and convenience.

Essay Prompt

Write a persuasive essay to a newspaper reflecting your views on censorship in libraries. Do you believe that certain materials [...] should be removed from the shelves if they are found offensive?

Student Essay

Everyone has their favorite book. But if it offended someone, should he be allowed to remove it? **Offensive materials should not be removed from shelves.** If we removed books that offended even one person, then no books would remain.

As Americans, we have the right to freedom of speech. Authors use their freedom in their writing, just like musicians use their freedom to make music. But if we denied them their right to put out their creations, we would be denying them their basic rights as an American citizen.

Personally, I hate rap. I don't even consider it to be music, due to the fact that no one actually sings. But I do understand that a large percent of our population does enjoy listening to it. ...

Essay Feedback

Your essay presents a **clear and well-articulated opinion on the issue of library censorship.** However, there are some areas where your essay could be improved.

First, try to stay focused on the topic of library censorship and **avoid bringing in unrelated examples, such as your personal dislike of rap music.** While it's important to acknowledge different perspectives, it's crucial to keep the focus on the main topic.

Second, provide more specific examples and evidence to support your argument. For instance, you could discuss specific instances of censorship in libraries and the negative ...

Figure 1: Exemplary student essay on library censorship from the ASAP dataset (Hamner et al., 2012) along with feedback and essay score generated by one of the methods evaluated in this paper. Explicit connections of the feedback to essay parts are color-coded.

Towards supporting essay writing automatically, much research has focused on quality assessment, mostly in the form of *automated essay scoring* (AES; Ke and Ng, 2019; Ramesh and Sanampudi, 2022). Most existing AES approaches perform a holistic scoring, summarizing the quality of an essay with a single score (Alikaniotis et al., 2016; Vajjala, 2018; Uto et al., 2020; Wang et al., 2023). Alternatively, specific essay quality aspects may be assessed, such as coherence (Li et al., 2018; Farag et al., 2018), grammar (Ajit Tambe and Kulka-rni, 2022), and organization (Persing et al., 2010;

Rahimi et al., 2015). While AES helps assess essay quality and monitor writing skill progress, most approaches cannot explain why a score was predicted, nor guide the student in improving the essay.

Another prominent line of research towards writing support focuses on identifying and correcting grammatical errors (Imamura et al., 2012; Bryant et al., 2017; Rozovskaya and Roth, 2019; Grundkiewicz et al., 2019). However, studies in educational research show that computer-based learning systems lead to higher learning outcomes if elaborated feedback is provided that provides explanations instead of only pointing to errors or providing the solution (Van der Kleij et al., 2015). Therefore, Nagata (2019) introduced the task of *feedback comment generation* in NLP: Given a learner text with a grammatical error, automatically generate a comment with hints and explanations to guide their correction process. Song et al. (2023) extended this task by generating explanations for a broader range of grammatical error types using large language models (LLMs). However, these tasks operate only on the sentence level and are limited to grammatical errors. Generating feedback on the essay level by addressing not only grammatical errors but the essay as a whole remains relatively unexplored.

To foster research in this direction, we tackle the task of *essay feedback generation*: Given a student essay, automatically generate textual feedback that helps students improve their essays. An example is shown in Figure 1. Building on the strong abilities of LLMs in many text-generation tasks, this work examines how well LLMs can generate essay feedback by exploring various prompting strategies in zero- and few-shot settings. Inspired by Chain-of-Thought prompting (Wei et al., 2022), we study whether AES can benefit the performance of essay feedback generation and vice versa.

Our experiments suggest that generating essay feedback by explaining the predicted essay score improves the scoring performance on the widely-used ASAP dataset (Hamner et al., 2012). For essay feedback generation, we deem helpfulness to be the most important quality criterion. Helpful essay feedback should point out and explain mistakes made in an essay in a precise and easy way for students to understand (Shute, 2008; Hattie and Timperley, 2007). We evaluate the helpfulness automatically and manually. Due to the lack of ground-truth essay feedback, we propose using LLMs to automatically judge the essay feedback’s

helpfulness, which turns out to correlate well with human helpfulness judgments. Our manual evaluation also reveals that the generated essay feedback is deemed helpful for students to improve their essay writing skills. However, the impact of scoring the essay remains low ultimately. Altogether, this paper’s main contributions are:

- A comparison of several LLM prompting strategies for automated essay scoring
- An approach and task-specific automatic evaluation strategy for essay feedback generation using LLM prompting
- Empirical insights into the influence of automated essay scoring on generating essay feedback and vice versa¹

2 Related Work

Essay writing is a central task in education to evaluate various skills of students, including logical thinking, critical reasoning, and creativity (Liu et al., 2014; Schendel and Tolmie, 2016). However, manual essay grading is time-consuming and not always consistent within and across raters (Kasim, 2011; Eckes, 2015). Automated essay scoring (AES) aims to alleviate these issues, reducing the effort of graders and, ideally, making grading more consistent and reliable (Ke and Ng, 2019; Uto, 2021; Ramesh and Sanampudi, 2022).

While extensive research exists on AES (Ke and Ng, 2019; Ramesh and Sanampudi, 2022), assessing all important quality aspects (known as *traits*), including the relevance of an essay’s content to the prompt, the development of ideas, cohesion, coherence, and more remains challenging (Ramesh and Sanampudi, 2022). Only few works focus on scoring multiple traits at once (Mathias and Bhat-tacharyya, 2020; Hussein et al., 2020). Instead, the majority of AES research targets holistic essay scoring, that is, summarizing the essay quality in a single score (Alikaniotis et al., 2016; Cozma et al., 2018; Vajjala, 2018; Wang et al., 2023).

State-of-the-art approaches to AES can be divided by their use of the available data into full-data and few- or zero-shot settings (Tao et al., 2022). In the full-data setting, where all labeled data is used for training, most approaches fine-tune pretrained language models, such as BERT (Devlin et al., 2019). Yang et al. (2020) proposed solving the task

¹The code used for our experiments can be found under <https://github.com/webis-de/BEA-24>.

by combining essay scoring and essay ranking, fine-tuning BERT using multiple losses simultaneously. Extending this idea, Xie et al. (2022) combined regression and ranking into a single loss. Rather than fine-tuning a language model, Tao et al. (2022) designed two self-supervised constraints for learning a multi-layer embedding, which prepends the input to a frozen BERT model. They evaluate their approach in the full-data and one-shot setting, outperforming a fine-tuned BERT in the latter. To explore the potential of large language models (LLMs), Mizumoto and Eguchi (2023) prompted GPT-3.5 to score the student essays from the TOEFL11 dataset (Blanchard et al., 2013) in a zero-shot setting, indicating promising scoring performance.

The most straightforward way to provide more detailed feedback for an essay than a holistic score is trait scoring (Jong et al., 2023), that is, to evaluate an essay for different quality aspects. However, the reasoning behind an assigned trait score usually remains unknown to the student. Therefore, Kumar and Boulanger (2020) adopted explainability methods to explain how input features to an AES system influence the trait scores for an essay. While this provides more insights, the pedagogical quality and impact on writing performance remain questionable if no feedback is given together with the scores (Kumar and Boulanger, 2020).

Specific feedback generation tasks have been addressed in educational NLP. Nagata (2019) introduced *feedback comment generation* to explain grammatical errors to a learner on the sentence level. This task has been tackled by combining retrieval and text generation (Hanawa et al., 2021; Ihori et al., 2023), by identifying different feedback types (Stahl and Wachsmuth, 2023), by augmenting the dataset (Babakov et al., 2023; Behzad et al., 2023), and by correcting the error (Jimichi et al., 2023; Koyama et al., 2023), all using fine-tuned language models. For a wider range of grammatical error types, Song et al. (2023) used the LLM GPT-4 to first identify the necessary corrective edit before generating a grammar error explanation using one-shot prompting. In the educational domain, Meyer et al. (2024) showed that LLM-based writing feedback, generated using a single handcrafted prompt, positively impacts students’ text revisions, motivation, and positive emotions.

So far, however, the generation of textual feedback on complete student essays has, to our knowledge, received very little attention. All generation

approaches mentioned above operate on the sentence level and explain grammatical errors only, while our work aims to address all aspects of student essays that may need improvement. The only other work on essay feedback generation tackled the task using Chain-of-Thought prompting using zero-shot learning (Han et al., 2023). The resulting feedback was deemed to be more helpful than the feedback generated using standard prompting, as evaluated by humans.

Motivated by these promising results and the positive effects of LLM-based writing feedback on students, we go beyond previous work by comparing the effectiveness of different prompting strategies for essay feedback generation. We also study how and to what extent AES can benefit essay feedback generation (and vice versa) by addressing the tasks jointly. Following the educational literature on feedback, we aim to generate essay feedback that is specific and elaborate (Shute, 2008) while assessing the current state and instructing on how to improve to achieve the goals (Hattie and Timperley, 2007).

3 Approach

This section describes our approach to essay feedback generation. We propose to tackle essay scoring and feedback generation jointly in order to study how and to what extent AES can benefit essay feedback generation and vice versa. By comparing different prompting strategies for large language models (LLMs), we explore how well the tasks can be solved using in-context learning.

In particular, we test different prompting strategies by systematically varying three main aspects of the prompts, as visualized in Figure 2: (a) the *prompt pattern*, which defines the context and layout; (b) the *task instruction type*, which sets the ordering and phrasing of the tasks to be tackled; and (c) the *in-context learning* approach, which specifies the number of examples provided.

3.1 Prompt Patterns

We compare two different kinds of prompt patterns, which define the context and format of the prompt: (i) a *base pattern* and (ii) different *persona patterns*. All prompt patterns are displayed in Table 1.

Base Pattern The base pattern simply gives the general context and defines the layout and order in which the *essay prompt* (i.e., the task given to the learner writing the essay), the *task instruction*, as

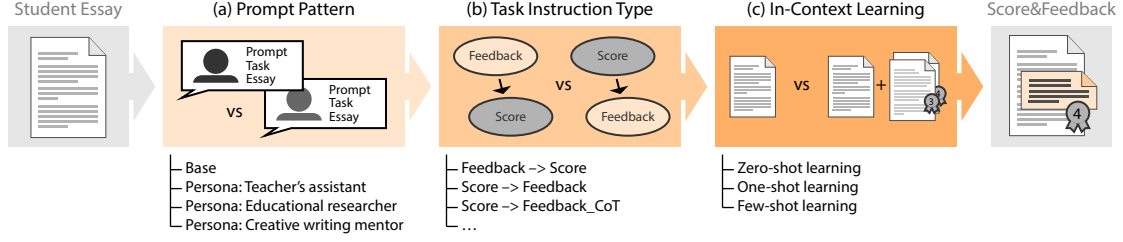


Figure 2: Overview of the main points of variation in our approach to predict a score and to generate feedback for a student essay: (a) Prompt pattern: Use of the base pattern or persona-specific pattern; (b) Task instruction type: Tasks to be tackled and their ordering; (c) In-context learning approach: Number of examples to learn from.

Base: You are given an essay written by a student and the corresponding prompt for the 7th to 10th grade student. ##### Prompt: "{essay_prompt}" #### Task: {task_instruction} ##### Student essay: "{essay}"
Teacher's Assistant: Imagine you are a teacher's assistant in a middle school tasked with reviewing a 7th to 10th grade student's essay. You have the essay and the prompt that was given to the student. ##### Original Prompt Provided to Student: "{essay_prom.}" #### Review Task: {task_instruction} ##### Student's Essay for Review: "{essay}"
Educational Researcher: You are part of an educational research team analyzing the writing skills of students in grades 7 to 10. You have been given a student's essay and the prompt they responded to. ##### Essay Prompt: "{essay_prompt}" #### Analysis Task: {task_instruction} ##### Analyzed Student Essay: "{essay}"
Creative Writing Mentor: You are a creative writing mentor evaluating a piece written by a student in grades 7 to 10. The student's work is based on a specific prompt. ##### Creative Prompt Given: "{essay_prompt}" #### Critique Instructions: {task_instruction} ##### Student's Creative Piece: "{essay}"

Table 1: Prompt patterns: Base pattern and all persona patterns. Brackets indicate placeholders that are filled respectively during the experiments. We removed model-specific pre-/suffixes and line breaks for illustration.

defined by the used task instruction type, and the current learner *essay* will be presented to the model. All inputs are indicated by markdown headings.

Persona Patterns These prompt patterns are inspired by persona prompting (White et al., 2023), giving the LLM a persona or role to play when generating output. This aims to implicitly define the expected type of output. For our task, we compare the three personas, namely, *teacher's assistant*, *educational researcher*, and *creative writing mentor*, by altering the context given in the prompt pattern.

3.2 Task Instruction Types

The task instruction type defines the tasks to be tackled along with their ordering. We differentiate between tackling (i) only essay *scoring*, (ii) essay *scoring and feedback* generation, and (iii) only essay *feedback* generation. This way, we can measure the influence that essay scoring has on feedback generation, and vice versa. We explore the following task instruction types for our tasks:

- **Scoring.** Instruct to score the student essay on a given score range. This serves as a baseline for assessing the essay scoring performance.
- **Feedback.** Instruct to generate essay feedback for the student writer. This serves as a baseline for assessing the feedback performance.
- **Scoring→Feedback.** Instruct to score the essay and then generate feedback for the student writer. This measures the influence of essay scoring on the feedback performance.
- **Feedback→Scoring.** Instructs to first generate feedback before scoring the essay. This evaluates whether feedback generation helps to predict the correct essay score.
- **Scoring→Feedback_CoT.** Instruct to score the essay and to then generate feedback using zero-shot Chain-of-Thought (CoT) prompting, that is, to add the phrase "Let's think step by step.", which has been shown to increase LLM's reasoning performance (Kojima et al., 2022). This might benefit the reasoning needed in feedback generation.
- **Feedback_dCoT→Scoring.** Instruct to first analyze the essay quality using the rubric, to then generate feedback, and to finally score the essay. This is a more detailed variation of CoT that provides task-specific steps to follow before arriving at the final essay score.

Score	Description
3	The response demonstrates an understanding of the complexities of the text. – Addresses the demands of the question – Uses expressed and implied information from the text – Clarifies and extends understanding beyond the literal
2	The response demonstrates a partial or literal understanding of the text. – Addresses the demands of the question, although may not develop all parts equally – Uses some expressed or implied information from the text to demonstrate understanding – May not fully connect the support to a conclusion or assertion made about the text(s)
1	The response shows evidence of a minimal understanding of the text. – May show evidence that some meaning has been derived from the text – May indicate a misreading of the text or the question – May lack information or explanation to support an understanding of the text in relation to the question
0	The response is completely irrelevant or incorrect, or there is no response.

Table 2: Exemplary rubric from essay set 3 of the ASAP dataset (Hamner et al., 2012). The rubrics are provided as additional information within the task instructions.

- *Scoring*→*Explanation*. Instruct to score the essay and to then generate an explanation for the predicted score. This explores whether score explanations as a form of feedback relate to asking for essay feedback specifically.
- *Explanation*→*Scoring*. Instruct to analyze the essay, to then first generate an explanation for an essay score that, in turn, should be generated at the end. This avoids that the LLM predicts an incorrect score and then generates an explanation justifying the incorrect score, as observed by Ye and Durrett (2022).

Task instructions for essay scoring provide the *scoring range* that should be used, while those for feedback generation provide the *rubric*, that is, guidelines including a short description for essays of each quality level and typical elements of such. An exemplary *rubric* can be seen in Table 2.

Since the performance of LLMs is sensitive to the exact wording of a prompt (Leidinger et al., 2023), we create a total of four *task instructions* for each task instruction type by instructing ChatGPT (OpenAI, 2023) to generate three paraphrases of each initial, manually written task instruction. Examples of the latter can be seen in Table 3. We provide all task instructions in Appendix A.

Scoring: Given this essay that was written for the given prompt, grade the essay using those ranges: {*scoring_range*}.

Feedback: Analyze the given essay using the following rubric: {*rubric*}. Provide comprehensive feedback for the student that helps them to achieve better grades in the future.

Scoring→**Feedback:** Grade the given essay using the following rubric: {*rubric*}. Use those score ranges: {*scoring_range*}. Provide comprehensive feedback for the student that helps them to achieve better grades in the future.

Feedback_dCoT→**Scoring:** Analyze the given essay using the following rubric and give helpful feedback to the student: {*rubric*}. Use those score ranges: {*scoring_range*}. Let’s think step by step. First, analyze the quality of the essay in terms of the given rubric. Then, give feedback to the student that explains their mistakes and errors and additionally gives them tips to avoid them in the future. As a final step, output the score at the end.

Scoring→**Explanation:** Grade the given essay using the following rubric: {*rubric*}. Use those score ranges: {*scoring_range*}. Provide an explanation for your score as well.

Table 3: Task instruction types: Examples of the initial, manually written task instructions for five types. Brackets indicate placeholders that are filled with the respective information during the experiments.

One-shot Example: Essay: “{*essay*}”

Reasoning: This is a minimally-developed response with inadequate support and detail. The writer takes the position that computers can be harmful to the eyes and then addresses eye damage to three groups of people (kids, teens, adults). A few specific details are included (sensitive eyes, MySpace), but elaboration is minimal. Some organization is demonstrated but few transitions are used. Overall, the response is sufficiently developed to move into the score point ‘3’ range.

Scores: {Overall: 3}

Table 4: One-shot example consisting of a student essay, a manually written score justification, and the assigned score. The data is taken from the scoring guidelines for essay set 1 of the ASAP dataset (Hamner et al., 2012).

3.3 In-Context Learning

As final point of variation of our approach, we explore how providing one or multiple exemplary essays, together with their score and a reasoning for the score, helps with essay scoring and feedback generation. The data comes from additional material given to human raters. We argue that the reasoning of the score may help with essay scoring, but could also be seen as a form of feedback and may benefit that task as well. We compare (i) *zero-shot*, (ii) *one-shot*, and (iii) *few-shot* learning.

For one-shot, we randomly select an essay with a medium score, as the one in Table 4. For few-shot, we first randomly select examples among the essays with the best and worst scores before covering the

Pattern	Essay Set								Mean
	1	2	3	4	5	6	7	8	
Base	.495	.532	.405	.495	.497	.601	.436	.377	.480
TA	.536	.603	.408	.499	.512	.625	.443	.439	.508
ER	.436	.554	.460	.560	.553	.620	.418	.467	.509
CWM	.484	.588	.382	.434	.507	.596	.471	.352	.477

Table 5: Essay scoring results: Average QWK over all task instructions using zero-shot learning for each prompt pattern: base, teacher’s assistant (TA), educational researcher (ER), and creative writing mentor (CWM). We report the performance for each of the eight essay sets as well as the mean QWK over all sets.

other scores. Due to the limited context length, we restrict the prompt to 5,120 characters and select as many examples that fit this limitation as possible.²

4 Data

Multiple AES datasets are available, with the Automated Student Assessment Prize’s (ASAP) dataset (Hamner et al., 2012) being the most widely used. It comprises 12,980 essays written by school students in grades 7 to 10. All essays were scored manually by two raters. The essays are divided into eight essay sets. The essay sets differ by the essay prompt, i.e., the task description they were written for, the scoring range, and the rubric used by the raters as annotation guidelines. The rubrics provide a short description for essays of each quality level and typical elements of such essays.

Since for the introduced task of essay feedback generation, no parallel dataset is available yet, we use the ASAP dataset as input data and evaluate the generated feedback without supervision.

5 Evaluation

We evaluate the performance of a large language model (LLM) by comparing the proposed prompting strategies on the two tasks: essay scoring and feedback generation. First, we assess the scoring performance and, then, we both automatically and manually evaluate the generated feedback in terms of the helpfulness for the student writer. We aim to study the effects of tackling essay scoring and feedback generation jointly, as well as explore how well LLMs can solve both tasks using prompting.

²For the few-shot variation, the described example selection process led to 3, 2, 4, 5, 8, 6, 4 and 2 examples for the essay sets 1 to 8 respectively. The differences are due to the variation in essay and reasoning length per essay set.

5.1 Essay Scoring

We compare the proposed prompt patterns, task instruction types, and in-context learning approaches, to evaluate the performance of an LLM on the essay scoring task. Also, we measure the influence of feedback generation on the scoring performance.

Approach We use the instruction-following recent LLM Mistral with 7B parameters (*Mistral-7B-Instruct-v0.2*, Jiang et al., 2023) in our experiments, generating each output with greedy decoding.³ We found that instructing the model to generate the essay score in JSON format helps to extract the score from the generated text automatically.⁴ Below, we report the number of essays that still did not receive a score (*Unscored*) and omit them from the performance calculation.

Baselines As a baseline, we report the performance AES-Prompt (Tao et al., 2022), which is, to our knowledge, the best-performing AES approach that is not fully fine-tuned on the ASAP dataset. As an upper bound, we also report the performance of R²BERT (Yang et al., 2020), the state-of-the-art approach fully fine-tuned on the same dataset.

Experimental Setup We automatically assess the essay scoring performance using quadratic weighted kappa (QWK), the most widely adopted metric for automatic essay scoring (Ke and Ng, 2019). Since the test set of the ASAP dataset is not publicly available, we follow Taghipour and Ng (2016) and apply their 5-fold cross-validation split. Since we perform no training, we only use the validation splits to create reasonable initial prompts and report the performance on the test splits.

Results Table 5 presents the scoring performance for each prompt pattern. We report the average QWK of all task instructions using zero-shot learning to measure the influence of the prompt pattern on the scoring performance. Using the personas “educational researcher” (ER) and “teacher’s assistant” (TA) seems beneficial for essay scoring, either of which performs best on all but one essay set, and ER best on average (mean QWK of .509).

To evaluate the influence of the task instruction type, Table 6 shows the performance of the best-

³Initial experiments on essay scoring with Llama-2 (*7b-chat-hf* and *13b-chat-hf*, Touvron et al., 2023) led to lower performance, which halted further testing with Llama-2.

⁴If the score was not generated as instructed, we re-prompted the model to extract the score from its prior response. This was effective when a score was in the initial answer.

Task Instruction Type	Essay Set									Unscored
	1	2	3	4	5	6	7	8	Mean	
Scoring	.448	.585	.479	.596	.557	.649	.438	.481	.529	1
Scoring→Feedback	.510	.615	.439	.530	.489	.621	.449	.481	.517	1
Feedback→Scoring	.388	.561	.484	.600	.622	.630	.385	.545	.527	16
Scoring→Feedback_CoT	.538	.595	.422	.494	.530	.635	.458	.477	.519	19
Feedback_dCoT→Scoring	.546	.564	.424	.558	.581	.628	.477	.489	.533	37
Scoring→Explanation	.466	.580	.472	.565	.541	.639	.420	.417	.513	0
Explanation→Scoring	.470	.553	.488	.636	.571	.675	.384	.484	.533	2

Table 6: Essay scoring results: QWK for the best approach variation per task instruction type in the zero-shot setting. We report the performance per essay set and the average over essay sets. The best results per column are bold.

performing approach variations per task instruction type. We report the combination of prompt pattern and task instruction that performed best on the validation set using zero-shot learning. The results suggest that instructing the LLM to first follow task-specific steps to analyze and give feedback (*Feedback_dCoT→Scoring*) as well as to first generate an explanation for the essay score (*Explanation→Scoring*) particularly help with essay scoring. These two achieve the highest mean QWK (.553). In general, the variations that generate some form of feedback first perform better than their counterparts that perform scoring first.

Finally, we study the influence of in-context learning on the instruction type *Scoring→Feedback* using the prompt pattern and task instruction that performs best on the validation split for a fair comparison to the baselines (Table 7). The results indicate that giving examples of scored essays aid essay scoring. One-shot learning outperforms few-shot learning, but the effect is rather small. Our prompting approaches perform rather competitively to the strong baseline AES-Prompt (Tao et al., 2022).

5.2 Essay Feedback Generation

As with essay scoring, we evaluate the generated feedback by comparing the prompt patterns, task instruction types, and in-context learning approaches. Our goal is to explore how well LLMs perform at generating helpful essay feedback and whether essay scoring can benefit the feedback generation.

Approach We continue using the large language model Mistral (*Mistral-7B-Instruct-v0.2*, Jiang et al., 2023) for the essay feedback generation task since it performed well at the essay scoring task.

Automatic Evaluation Using LLMs to assess the quality of generated texts has been shown to be consistent with human expert annotations for some free-text generation tasks (Chiang and Lee, 2023).

Context	Essay Set								
	1	2	3	4	5	6	7	8	Mean
Zero-shot	.510	.615	.439	.530	.489	.621	.449	.481	.517
One-shot	.565	.619	.523	.600	.606	.665	.509	.233	.540
Few-shot	.558	.586	.515	.586	.618	.671	.472	.297	.538
AES-Pro.	.682	.544	.590	.672	.701	.622	.683	.620	.639
R ² BERT	.817	.719	.698	.845	.841	.847	.839	.744	.794

Table 7: Essay scoring results: QWK per in-context learning approach for *Scoring→Feedback* using the best-performing prompt pattern and task instruction. The baseline AES-Prompt (Tao et al., 2022) also has one shot. R²BERT (Yang et al., 2020) is fully fine-tuned.

Since there are no existing automatic metrics for assessing the quality of generated essay feedback, we follow previous work and use Mistral itself as well as Llama-2 (*Llama-2-13b-chat-hf*, Touvron et al., 2023) for the automatic part of our feedback evaluation. We instruct them to assign an overall helpfulness scores between 1 (not helpful) and 10 (very helpful) for each generated essay feedback. The used prompt can be found in Appendix B.⁵

Our evaluation focuses on helpfulness, which we deem to be the most important quality dimension for essay feedback. We anticipate that other quality aspects, such as faithfulness, are implicitly covered since irrelevant or incorrect feedback would not be helpful for the student author.

Automatic Results Table 8 presents the assigned helpfulness scores for each prompt pattern, averaged over task instructions using zero-shot learning. Both LLMs deemed the feedback generated by a persona pattern to be most helpful, on average: the top helpfulness score is achieved by ER for Mistral (8.26) and CWM for Llama-2 (7.48).

To evaluate the influence of the task instruction

⁵We also experimented with relative comparisons of feedback for automatic helpfulness assessment. However, the correlation to our manual helpfulness annotations was low.

Prompt Pattern	Mistral	Llama-2
Base	7.78 $\pm$ 0.53	6.88 $\pm$ 0.18
Teacher’s assistant (TA)	7.90 $\pm$ 0.39	6.84 $\pm$ 0.19
Educational researcher (ER)	8.26 $\pm$ 0.23	6.87 $\pm$ 0.18
Creative writing mentor (CWM)	7.83 $\pm$ 0.47	7.48 $\pm$ 0.85

Table 8: Automatic feedback generation results: Average helpfulness scores predicted by Mistral and Llama-2 for each prompt pattern over all task instructions using zero-shot learning. The best result per column is bold.

Task Instruction Type	Mistral	Llama-2
Feedback	8.96 $\pm$.25	7.31 $\pm$.19
Scoring→Feedback	8.04 $\pm$.44	7.15 $\pm$.45
Feedback→Scoring	8.27 $\pm$.38	7.27 $\pm$.50
Scoring→Feedback_CoT	7.30 $\pm$.63	6.72 $\pm$.41
Feedback_dCoT→Scoring	8.53 $\pm$.66	7.28 $\pm$.55
Scoring→Explanation	7.22 $\pm$.45	6.68 $\pm$.40
Explanation→Scoring	7.27 $\pm$.63	6.75 $\pm$.36

Table 9: Automatic feedback generation results: Average helpfulness scores predicted by Mistral or Llama-2 for each task instruction type over all task instructions and prompt patterns using zero-shot learning.

type, Table 9 shows the results per type, averaged over prompt patterns and task instructions using zero-shot learning. Both evaluation models gave the highest average scores to performing feedback generation only (*Feedback*). For the other task instruction types, the variations that generate some form of feedback first seem more helpful than their counterparts that perform scoring first.

Finally, we study the influence of each in-context learning approach on the task instruction type *Scoring→Feedback* on average over the prompt patterns and task instructions (Table 10). The results suggest that the reasoning presented in the provided in-context examples positively impacts the feedback helpfulness. Although the effect is small, more examples help more.

Manual Evaluation The proposed automatic evaluation only approximates the quality of the generated essay feedback. Therefore, we conducted a manual annotation study during which 12 annotators manually judged the feedback quality. All annotators have advanced English skills and are not authors of this paper. The annotators were divided into four groups that annotated the same feedback.

In particular, we randomly selected 24 essay feedback texts generated by the three task instruction types that performed best in the automatic evaluation: *Feedback*, *Feedback→Scoring*, and *Feedback_dCoT→Scoring*. Here, we used the best-

In-Context Learning	Mistral	Llama-2
Zero-shot learning	8.04 $\pm$.44	7.15 $\pm$.45
One-shot learning	8.39 $\pm$.54	7.28 $\pm$.47
Few-shot learning	8.42 $\pm$.56	7.30 $\pm$.46

Table 10: Automatic feedback generation results: Average helpfulness scores predicted by Mistral or Llama-2 per in-context learning approach for *Scoring→Feedback* over all prompt patterns and task instructions.

Task Instruction Type	S1	S2	S3	S4	S5
Feedback	5.88	5.71	6.04	5.75	6.08
Feedback→Scoring	5.17	5.04	5.46	5.21	5.08
Feedback_dCoT→Scoring	5.50	4.92	5.29	4.83	5.00

Table 11: Manual feedback generation results: Average scores assigned by the annotators for each approach for statements S1–S5 on a 7-point Likert scale (7 is best).

performing combination of prompt pattern and task instruction. All sampled feedback texts were written for essays from one essay set only to reduce the time the annotators need to read the essay prompt. We chose essay set 4, which covers the most common ASAP task, reading comprehension.

To judge the feedback helpfulness, the annotators received the essay prompt, the student essay, and the generated feedback. Based on this, they were asked to assess to what extent the following statements apply on a 7-point Likert scale (score 1: “I strongly disagree”, score 7: “I fully agree”):

- S1: The feedback clearly points out mistakes that were made in the essay.
- S2: The feedback explains exactly why the errors are errors.
- S3: The feedback is very clear and precise so that the student can understand it.
- S4: The feedback is absolutely suitable for students from 7th to 10th grade.
- S5: Overall, the feedback is very helpful.

Manual Results Table 11 presents the results of the manual annotation study. For all five statements covering different helpfulness aspects, *Feedback* achieved the highest scores on average. Especially the clarity and precision (S3) as well as the overall helpfulness (S5) of *Feedback* were rated with the second-best score of 6 (“I mostly agree”). All compared task instruction types reach an average score above the neutral score of 4, indicating that all feedback is perceived as rather helpful in general. Overall, the generated essay feedback seems

Autom. Evaluation	S1	S2	S3	S4	S5
Mistral	0.29	0.27	0.45	0.25	0.61
Llama-2	-0.11	-0.11	-0.02	0.07	-0.10

Table 12: Pearson correlation of the manual annotations per statement (S1–S5) and the automatic helpfulness scores using Mistral or Llama-2. The top value is bold.

to have the most potential for improvement by better explaining why an error is erroneous (S2) and being more suitable for students (S4). The inter-annotator agreement in terms of Krippendorff’s α on average over the four groups is 0.44.

To evaluate the reliability of our automatic helpfulness evaluation, we show the correlation between manual and automatic helpfulness scores in Table 12. The highest correlation value (0.61) was measured between the manually annotated overall helpfulness (S5) and the automatic helpfulness scores predicted by Mistral. This indicates that using Mistral can be useful for automatically evaluating feedback helpfulness. The helpfulness scores generated by Llama-2 do not correlate with the manual annotation for any statement.

6 Conclusion

Despite the strong text generation abilities of recent LLMs in various tasks, their effectiveness in generating essay feedback that helps student writers improve their essays has remained unclear until now. Also, generating textual feedback that addresses the entire essay has previously only been tackled using one prompting strategy in a zero-shot learning setting. With this work, we go beyond existing work by comparing different LLM prompting strategies for essay feedback generation. We propose tackling essay feedback generation and automated essay scoring (AES) jointly to study whether AES can benefit feedback generation and vice versa. Our experiments suggest that AES can be solved competitively by prompting LLMs, benefiting from tackling feedback generation first. The generated feedback is deemed helpful for students by our automatic and manual evaluation. However, the impact of scoring on the feedback helpfulness remains low ultimately.

7 Limitations

Aside from the still-improvable performance of the presented prompting approaches to automated essay scoring and feedback generation, we see two

notable limitations of our work: the dependence of our feedback approaches on additional data and the pending utilization of the generated essay feedback for real-world essay writing support.

First, we point out that our feedback approaches rely on the availability of a detailed rubric, that is, guidelines including a short description for essays of each quality level, typical elements of such, and textual reasoning as to why example essays received a specific score. Such information might not always be available, which could reduce the transferability of our results to other essay datasets.

Second, while our evaluation suggests that the generated essay feedback is helpful for student writers, it remains unclear whether the student writers also perceive it as such. We encourage future work to utilize our approaches for real-world essay writing support and make it available to students. Feedback from students on such a tool would be useful to guide research on essay feedback generation.

Acknowledgments

We would like to thank the participants of our study and the anonymous reviewers for the valuable feedback and their time. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project ArgSchool, project number 453073654.

References

- Aniket Ajit Tambe and Manasi Kulkarni. 2022. [Automated essay scoring system with grammar score analysis](#). In *2022 Smart Technologies, Communication and Robotics (STCR)*, pages 1–7.
- Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. [Automatic text scoring using neural networks](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.
- Nikolay Babakov, Maria Lysyuk, Alexander Shvets, Lilya Kazakova, and Alexander Panchenko. 2023. [Error syntax aware augmentation of feedback comment generation dataset](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 37–44, Prague, Czechia. Association for Computational Linguistics.
- Robert L Bangert-Drowns, Marlene M Hurley, and Barbara Wilkinson. 2004. [The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis](#). *Review of Educational Research*, 74(1):29–58.

- Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2023. [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.
- Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. [Toefl11: A corpus of non-native english](#). *ETS Research Report Series*, 2013(2):i–15.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. [Automated essay scoring with string kernels and word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Thomas Eckes. 2015. [Introduction to Many-Facet Rasch Measurement](#). Peter Lang Verlag, Berlin, Deutschland.
- Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. [Neural automated essay scoring and coherence modeling for adversarially crafted input](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271, New Orleans, Louisiana. Association for Computational Linguistics.
- Tri Febriani. 2022. [“Writing is challenging”: factors contributing to undergraduate students’ difficulties in writing English essays](#). *Erudita: Journal of English Language Teaching*, 2:83–93.
- Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. [Neural grammatical error correction systems with unsupervised pre-training on synthetic data](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.
- Ben Hamner, Jaison Morgan, Iyannvande, Mark Shermis, and Tom Vander Ark. 2012. [The hewlett foundation: Automated essay scoring](#).
- Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2023. [Fabric: Automated scoring and feedback generation for essays](#).
- 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.
- Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. [A trait-based deep learning automated essay scoring system with adaptive feedback](#). *International Journal of Advanced Computer Science and Applications*, 11(5).
- Mana Ithori, Hiroshi Sato, Tomohiro Tanaka, and Ryo Masumura. 2023. [Retrieval, masking, and generation: Feedback comment generation using masked comment examples](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 60–67, Prague, Czechia. Association for Computational Linguistics.
- Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. [Grammar error correction using pseudo-error sentences and domain adaptation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–392, Jeju Island, Korea. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Kunitaka Jimichi, Kotaro Funakoshi, and Manabu Okumura. 2023. [Feedback comment generation using predicted grammatical terms](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 79–83,

- Prague, Czechia. Association for Computational Linguistics.
- You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. 2023. [Review of feedback in automated essay scoring](#).
- Noor Lide Abu Kassim. 2011. Judging behaviour and rater errors: an application of the many-facet rasch model. *GEMA Online™ Journal of Language Studies*, 179.
- 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.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*.
- Shota Koyama, Hiroya Takamura, and Naoaki Okazaki. 2023. [The Tokyo tech and AIST system at the GenChal 2022 shared task on feedback comment generation](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 74–78, Prague, Czechia. Association for Computational Linguistics.
- Vivekanandan Kumar and David Boulanger. 2020. [Explainable automated essay scoring: Deep learning really has pedagogical value](#). *Frontiers in Education*, 5.
- Alina Leiding, Robert van Rooij, and Ekaterina Shutova. 2023. [The language of prompting: What linguistic properties make a prompt successful?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.
- Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 386–397, Cham. Springer International Publishing.
- Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. 2014. [Assessing critical thinking in higher education: Current state and directions for next-generation assessment](#). *ETS Research Report Series*, 2014(1):1–23.
- Sandeep Mathias and Pushpak Bhattacharyya. 2020. [Can neural networks automatically score essay traits?](#) In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. [Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students’ text revision, motivation, and positive emotions](#). *Computers and Education: Artificial Intelligence*, 6:100199.
- Atsushi Mizumoto and Masaki Eguchi. 2023. [Exploring the potential of using an ai language model for automated essay scoring](#). *Research Methods in Applied Linguistics*, 2(2):100050.
- 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.
- OpenAI. 2023. [ChatGPT \(GPT version: 3.5\)](#). Large language model.
- John Peloghitis. 2017. Difficulties and strategies in argumentative writing: A qualitative analysis. In *Transformation in language education*, Tokyo. JALT.
- Isaac Persing, Alan Davis, and Vincent Ng. 2010. [Modeling organization in student essays](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.
- Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. 2015. [Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing](#). In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30, Denver, Colorado. Association for Computational Linguistics.
- Dadi Ramesh and Suresh Kumar Sanampudi. 2022. [An automated essay scoring systems: a systematic literature review](#). *Artificial Intelligence Review*, 55(3):2495–2527.
- Jessica Riddell. 2015. [Performance, feedback, and revision: Metacognitive approaches to undergraduate essay writing](#). *Collected Essays on Learning and Teaching*, 8:79.
- Alla Rozovskaya and Dan Roth. 2019. [Grammar error correction in morphologically rich languages: The case of Russian](#). *Transactions of the Association for Computational Linguistics*, 7:1–17.
- Rebecca Schendel and Andrew Tolmie. 2016. [Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in rwanda](#). *Assessment & Evaluation in Higher Education*, 42(5):673–689.

- Valerie J. Shute. 2008. [Focus on formative feedback](#). *Review of Educational Research*, 78(1):153–189.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. [Gee! grammar error explanation with large language models](#).
- Maja Stahl and Henning Wachsmuth. 2023. [Identifying feedback types to augment feedback comment generation](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 31–36, Prague, Czechia. Association for Computational Linguistics.
- Kaveh Taghipour and Hwee Tou Ng. 2016. [A neural approach to automated essay scoring](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- Qiuyu Tao, Jiang Zhong, and Rongzhen Li. 2022. [Ae-prompt: Self-supervised constraints for automated essay scoring with prompt tuning](#). In *The 34th International Conference on Software Engineering and Knowledge Engineering, SEKE 2022, KSIR Virtual Conference Center, USA, July 1 - July 10, 2022*, pages 335–340. KSI Research Inc.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Masaki Uto. 2021. [A review of deep-neural automated essay scoring models](#). *Behaviormetrika*, 48(2):459–484.
- Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. [Neural automated essay scoring incorporating hand-crafted features](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Sowmya Vajjala. 2018. [Automated assessment of non-native learner essays: Investigating the role of linguistic features](#). *International Journal of Artificial Intelligence in Education*, 28(1):79–105.
- Fabienne M. Van der Kleij, Remco C. W. Feskens, and Theo J. H. M. Eggen. 2015. [Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis](#). *Review of Educational Research*, 85(4):475–511.
- Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge, and Qing Gu. 2023. [Aggregating multiple heuristic signals as supervision for unsupervised automated essay scoring](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13999–14013, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. [A prompt pattern catalog to enhance prompt engineering with chatgpt](#).
- Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. [Automated essay scoring via pairwise contrastive regression](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. [Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.
- Xi Ye and Greg Durrett. 2022. [The unreliability of explanations in few-shot prompting for textual reasoning](#). In *Advances in Neural Information Processing Systems*.

A Task Instructions

We present all used task instructions in the following list. This includes all paraphrases per task instruction type.

- **Scoring:** (1) Given this essay that was written for the given prompt, grade the essay using those ranges: {*scoring_range*}.
- (2) Review the provided essay in response to the given prompt. Assess its quality and assign a grade according to the following criteria: {*scoring_range*}.

- (3) Examine the essay written in response to the specified prompt. Utilize the following grading ranges to evaluate and score the essay: {*scoring_range*}.
- (4) Analyze the submitted essay that corresponds to the given prompt. Apply these grading standards to determine its score: {*scoring_range*}.
- **Feedback:** (1) Analyze the given essay using the following rubric: {*rubric*}. Provide comprehensive feedback for the student that helps them to achieve better grades in the future.

(2) Please evaluate the essay in accordance with the criteria outlined in: {*rubric*}. Offer detailed and constructive feedback to assist the student in improving their writing skills for future assignments.

(3) Utilize the provided rubric ({*rubric*}) to assess the essay. Your feedback should be thorough, focusing on areas of strength and suggesting improvements to help the student enhance their academic writing.

(4) Conduct an assessment of the submitted essay using this specific rubric: {*rubric*}. Your feedback should be insightful and supportive, guiding the student towards achieving higher grades in their future essays.
 - **Scoring→Feedback:** (1) Grade the given essay using the following rubric: {*rubric*}. Use those score ranges: {*scoring_range*}. Provide comprehensive feedback for the student that helps them to achieve better grades in the future.

(2) Please evaluate the essay in accordance with the criteria outlined in: {*rubric*}. Assign a grade based on these standards: {*scoring_range*}. Offer detailed and constructive feedback to assist the student in improving their writing skills for future assignments.

(3) Utilize the provided rubric ({*rubric*}) to assess the essay. Grade it according to these parameters: {*scoring_range*}. Your feedback should be thorough, focusing on areas of strength and suggesting improvements to help the student enhance their academic writing.

(4) Conduct an assessment of the submitted essay using this specific rubric: {*rubric*}. Apply the grading criteria as per these guidelines: {*scoring_range*}. Your feedback should be insightful and supportive, guiding the student towards achieving higher grades in their future essays.
 - **Feedback→Scoring:** (1) Analyse the given essay using the following rubric: {*rubric*}. Use those score ranges: {*scoring_range*}. To do this, first provide comprehensive feedback for the student that helps them to achieve better grades in the future. Then give the final score.

(2) Begin by carefully reviewing the submitted essay in light of the criteria outlined in {*rubric*}. After your thorough analysis, offer detailed and constructive feedback aimed at guiding the student towards academic improvement. Conclude your review by assigning a score to the essay, adhering to the guidelines specified in {*scoring_range*}.

(3) First, evaluate the essay against the criteria mentioned in {*rubric*}. Your evaluation should include specific, actionable suggestions for the student to enhance their writing skills and essay quality. Following your comprehensive feedback, assign a score to the essay based on the scale provided in {*scoring_range*}.

(4) Commence your assessment by applying the criteria from {*rubric*} to the essay. Focus on delivering in-depth feedback that is both informative and beneficial for the student's future academic endeavors. After providing this feedback, conclude by scoring the essay as per the range defined in {*scoring_range*}.
 - **Scoring→Feedback_CoT:** (1) Analyse the given essay using the following rubric and give helpful feedback to the student: {*rubric*}. Use those score ranges: {*scoring_range*}. Let's think step by step. Make sure to output the score only at the end.

(2) Please evaluate the provided essay according to this specific rubric: {*rubric*}. Scores should be assigned based on these criteria: {*scoring_range*}. Proceed methodically through each step. Conclude your analysis by presenting the final score.

(3) Conduct a thorough assessment of the essay using the rubric below: {*rubric*}. Adhere to the following scoring guidelines: {*scoring_range*}. Break down your analysis into clear steps. Ensure the final score is given at the end of your evaluation.

(4) Examine the student's essay in detail, utilizing the rubric provided: {rubric}. Apply these scoring ranges for evaluation: {scoring_range}. Tackle the analysis in a step-by-step manner. The score should be presented at the conclusion of your feedback.

- **Feedback_dCoT→Scoring:** (1) Analyze the given essay using the following rubric and give helpful feedback to the student: {rubric}. Use those score ranges: {scoring_range}. Let's think step by step. First, analyze the quality of the essay in terms of the given rubric. Then, give feedback to the student that explains their mistakes and errors and additionally gives them tips to avoid them in the future. As a final step, output the score at the end.
(2) Begin by evaluating the essay based on the criteria outlined in the rubric: {rubric}. Consider the scoring guidelines provided: {scoring_range}. First, conduct a thorough analysis of the essay according to the rubric standards. Next, provide constructive feedback to the student, highlighting areas for improvement and suggesting strategies to enhance their writing skills. Conclude with a summary of the essay's strengths and weaknesses. Finally, present the essay's score at the end of your analysis.
(3) Follow these steps to assess the student's essay: First, reference the provided rubric: {rubric}, and apply it to evaluate the essay. Use the scoring ranges given: {scoring_range} for accurate assessment. Provide detailed feedback to the student, pinpointing specific areas of the essay that align or deviate from the rubric, along with advice for future improvement. Your feedback should be clear, constructive, and actionable. After your comprehensive review, conclude by outputting the final score, ensuring this is done only at the very end.
(4) To evaluate the student's essay, proceed as follows: Start with the provided rubric: {rubric}, to assess the essay's attributes. Adhere to the scoring guidelines: {scoring_range} for consistency. Your analysis should first focus on how well the essay meets the criteria in the rubric. Then, craft feedback for the student that is both informative and helpful, addressing any shortcomings

and providing practical advice for future essays. The feedback should be encouraging yet honest. Conclude your evaluation by scoring the essay, presented at the conclusion of your feedback.

- **Scoring→Explanation:** (1) Grade the given essay using the following rubric: {rubric}. Use those score ranges: {scoring_range}. Provide an explanation for your score as well.
(2) Please assess the submitted essay according to the criteria outlined in this rubric: {rubric}. Scores should be allocated based on these guidelines: {scoring_range}. Additionally, include a detailed rationale for the score you assign.
(3) Evaluate the provided essay by referring to the standards specified here: {rubric}. Utilize the following scoring range for your evaluation: {scoring_range}. Also, furnish a comprehensive justification for the grade you determine.
(4) Rate the essay in front of you using these evaluation criteria: {rubric}. Your scoring should align with these parameters: {scoring_range}. Please also give a thorough explanation to support the score you decide upon.
- **Explanation→Scoring:** (1) Analyse the given essay using the following rubric: {rubric}. To do this, first explain using the scoring rubric why you chose the score. After you analysed the essay, give a final grade.
(2) Utilize the provided scoring rubric ({rubric}) to evaluate the essay. Begin by detailing the reasons for your assigned score based on the rubric's criteria. Conclude by summarizing your analysis with a final grade.
(3) Apply the scoring guidelines from {rubric} to assess the essay. Start by discussing how the essay meets or falls short of each criterion in the rubric. Finalize your assessment with a grade.
(4) Use the scoring rubric ({rubric}) as a basis to critically analyze the essay. Explain how the essay aligns with each aspect of the rubric, justifying your evaluation. Conclude with a definitive grade.

B Automatic Helpfulness Assessment

We used the following prompt for our automatic helpfulness evaluation:

“You are given an essay and feedback from a teacher for this essay. Your task is to evaluate the helpfulness of the feedback.

Task: Evaluate the helpfulness of the feedback. Helpful feedback should explain what the errors are, why they are errors, and how to fix them. Give a score between 1 and 10, where 1 means the feedback is not helpful at all, and 10 means the feedback is very helpful.

Provide the output in the following output: {format_instructions}”

C Generated Essay Feedback

Table 13 compares the generated essay feedback for the same essay using three different task instruction types, compared in the manual evaluation.

Task Instruction Type	Generated Essay Feedback
Feedback	Based on the provided essay prompt and the given text, it appears that the student has not fully understood the story "Winter Hibiscus" by Minfong Ho. The student's response, "Saeng would return to her homeland," is not an accurate reflection of the story's conclusion. The author concludes the story with Saeng's determination to retake the driver's test in the spring when the hibiscus is budding and the geese have returned. This decision comes after Saeng's emotional experience in the greenhouse, where she is surrounded by familiar plants and memories, and is struck by a wave of loss for her past life in Vietnam. The author's use of the imagery of the geese returning and the hibiscus budding signifies the cyclical nature of life and the passing of seasons. Saeng's decision to retake the driver's test in the spring shows her resilience and determination to move forward in her new life in the United States while holding on to her past. To improve future essays, the student could focus on closely reading and understanding the text, taking notes, and identifying key themes and symbols. It may also be helpful for the student to practice summarizing and paraphrasing the text in their own words to demonstrate a deeper understanding.
Feedback→Scoring	The essay response is very brief and does not fully address the demands of the question. The statement 'Saeng would return to her homeland' is not explicitly stated in the text, and the essay does not provide any evidence or explanation to support this assertion. To improve, consider reading the text more closely and taking notes to help you understand the main ideas and themes. Use these notes to develop a more detailed and well-supported response that addresses the demands of the question.
Feedback_dCoT→Scoring	Your essay does not fully address the demands of the question. In order to earn a higher score, you should explain why Saeng would return to her homeland based on the details and events in the story. Consider how the hibiscus plant and the memories it evokes for Saeng might be connected to her desire to go back to her homeland. Additionally, you could discuss how the experiences and emotions Saeng encounters in the greenhouse contribute to her decision. Remember to use specific details and examples from the text to support your ideas.

Table 13: Generated essay feedback using the three task instruction types compared in the manual annotation study. The feedback was generated for the same student essay from essay set 4 of the ASAP dataset (Hamner et al., 2012).