Beyond Agreement: Diagnosing the Rationale Alignment of Automated Essay Scoring Methods based on Linguistically-informed Counterfactuals

Anonymous ACL submission

Abstract

While current automated essay scoring (AES) methods show high agreement with human raters, their scoring mechanisms are not fully explored. Our proposed method, using counterfactual intervention assisted by Large Language Models (LLMs), reveals that when scoring essays, BERT-like models primarily focus on sentence-level features, while LLMs are attuned to conventions, language complexity, as well as organization, indicating a more comprehensive alignment with scoring rubrics. Moreover, LLMs can discern counterfactual interventions during feedback. Our approach improves understanding of neural AES methods and can also apply to other domains seeking transparency in model-driven decisions. The codes and data will be released at GitHub.

1 Introduction

In recent years, neural approaches to automated essay scoring (AES) have demonstrated remarkable performance (Ke and Ng, 2019; Ramesh and Sanampudi, 2022). The advent of Large Language Models (LLMs) has shifted focus not only towards their scoring capabilities but also towards the potential for providing feedback (Mizumoto and Eguchi, 2023; Caines et al., 2023; Han et al., 2023; Xiao et al., 2024). However, current model evaluations mainly rely on metrics such as Quadratic Weighted Kappa (QWK) to measure agreement with human ratings. This approach leaves the models' underlying reasoning opaque, thereby raising risks and questioning the validity of their use in high-stakes educational tests (Fiacco et al., 2023).

A series of studies have found that neural models can be *right for the wrong reasons*, a concern that persists into the era of LLMs (McCoy et al., 2020; Turpin et al., 2023). To understand the decision-making basis of neural models, researchers have primarily adopted two primary avenues: **what** knowledge a model encodes and **why** a model makes certain predictions (Lyu et al., 2024). Both paradigms have garnered attention in the field of AES. Fiacco et al. (2023) addresses the *what* question by extracting meaningful functional groups from the representations of transformer models and aligning them with human-understandable features. However, a model encodes a myriad of features does not mean that the features are utilized in decision-making (Lyu et al., 2024). To tackle the *why* question, Singla et al. (2021) employed integrated gradients (Sundararajan et al., 2017) to analyze token importance, and discovered that for BERT-based model, most of the attributions are over non-linguistic tokens and stop words. It can be seen that the gradients-based methods only target lower-level token features, thus failing to reveal whether models leverage higher-level linguistic features. Moreover, both Singla et al. (2021) and Kabra et al. (2022) employed adversarial modifications to assess models, but these interventions did not target the linguistic features critical to the AES task, and they did not control for other variables that could affect essay scores during modification. Therefore, even for traditional AES models, reliable explanations of their inner workings remain elusive. Additionally, the explainability of scoring in LLMs is largely unexplored, indicating that considerable work is needed to advance our understanding of model reasoning within this domain.

In this paper, we aim to systematically investigate whether the underlying reasoning of models adheres to scoring rubrics—essentially, whether it aligns with human rationale. Specifically, we propose a model-agnostic diagnosis method that uses linguistically-informed counterfactuals to scrutinize the scoring mechanism of both traditional NLP models and LLMs. The diagnostic approach closely integrates linguistic knowledge from scoring rubrics, such as conventions, vocabulary, syntax, and coherence, with LLMs employed for fine-grained and controllable counterfactual generation.

Concept	Intervention	Description
Conventions	Error Correction $\uparrow$	Prompt GPT-4 to correct spelling, punctuation, and grammar errors.
	Spelling Errors Introduction $\downarrow$	Use `nlpaug` to misspell 30% of words in 50% of sentences.
	Agreement Errors Introduction $\downarrow$	Use `spaCy` to introduce subject-verb agreement (SVA) errors in 50% of sentences.
	Word Order Swapping (WOS) $\downarrow$	Use `nlpaug` to swap 30% of words in 50% of sentences.
Language Complexity	Complexification $\uparrow$	Prompt GPT-4 to enhance vocabulary and sentence structure.
	Simplification $\downarrow$	Prompt GPT-4 to simplify vocabulary and sentence structure.
Organization	Intra-paragraph Shuffling $\downarrow$	Shuffle sentence order *within paragraphs* to disrupt *local* cohesion.
	Inter-text Shuffling $\downarrow$	Shuffle sentence order *across the entire essay* to disrupt *global* cohesion.

Table 1: Overview of the counterfactual generation methods in this study. Note: (1) $\uparrow$ and $\downarrow$ denote positive and negative interventions. (2) Carefully designed GPT-4 prompts preserved the essay content and length while incorporating multidimensional linguistic knowledge in target concepts. We also evaluated the validity of the generated results. See details of the generation and evaluation in **Appendix A**. (3) Language complexity counterfactuals stemmed from the *corrected* samples, as a pilot study indicated that interventions on original essays in language complexity inadvertently corrected writing errors.

Our investigation reveals that: **(1)** BERT-like models can discern knowledge in conventions and language complexity but struggle to grasp the logical structure and coherence of texts; and **(2)** LLMs, although have lower score agreement than traditional models, display a superior inherent alignment with human experts' reasoning. Through few-shot learning or fine-tuning, LLMs can achieve both high score agreement and rationale alignment. Meanwhile, they can discernibly offer writing feedback.

2 Method

As previously mentioned, model explanations have two directions: *what* and *why*. To accurately depict the underlying reasoning of AES models, this study focus on the *why* question. Inspired by Gat et al. (2023), we employ counterfactual intervention method to establish causality between target factors and the prediction. Typically, counterfactual intervention involves perturbing a specific feature or concept while controlling for others and observing the subsequent effect on the model's prediction. We firstly extract target concepts from the essay scoring rubrics for intervention, and then generate counterfactual samples for different concepts using LLMs and heuristic rules.

2.1 Concepts to be Intervened

To select the target concepts applicable to various AES scenarios, we gathered scoring rubrics from the writing tasks of major standardized English tests such as IELTS, TOEFL iBT, TOEIC, and PTE academic. We also incorporated rating rubric from the ELLIPSE dataset, which was developed based on a number of state and industrial English language proficiency assessments. A detailed review of these rubrics allowed us to pinpoint three linguistic concepts that are critical in human assessment:

Conventions: An essay adheres to conventions if it is free from spelling, capitalization, punctuation, and grammar errors, reflecting the writer's ability to follow standard written English rules.

Language Complexity: Language complexity in an essay is demonstrated by a broad vocabulary and sophisticated control of lexical features, coupled with varied sentence structures, ensuring both lexical and syntactic accuracy.

Organization: An essay exhibits effective organization and development by presenting a logical structure with skillful paragraphing and the use of cohesive devices to ensure unity, progression, and seamless connection of thoughts.

2.2 Counterfactual Generation

Given an essay $\mathcal{E}(c_1, c_2, \ldots, c_i, \ldots)$, we generate a counterfactual sample $\mathcal{E}(c_1, c_2, \ldots, c_i', \ldots)$ by changing the value of concept $\mathcal{C}_i$ from c_i to c_i', while holding the other concepts fixed. To isolate the impact of the target concept, the generation process also aims to maintain the content, length, and fluency of the essay.

Existing counterfactual generation utilize keyword replacement (Garg et al., 2019), sentence rewriting (Ross et al., 2021; Wu et al., 2021), and manual editing (Gardner et al., 2020). However, these approaches are often limited to simple local interventions or require costly manual annotation, which hinders the practical estimation of the causal effects of high-level concepts on NLP models. Recognizing the potential of LLMs for generating high-quality counterfactuals to enhance black-box model explanations, as suggested by Gat et al. (2023), we propose a hybrid approach that combines LLMs with rule-based techniques for more controlled and scalable sample generation. As shown in Table 1, we generated eight types of linguistically-informed counterfactual samples for diagnosis. To comprehensively examine the effects of the target concepts on decision-making, we intervened the conventions and language complexity in both positive and negative directions. For organization, interventions were made in local and global cohesion.

3 Experiments

3.1 Settings

We employed two datasets: TOEFL11 (Blanchard et al., 2013) and ELLIPSE[1]. The TOEFL11 dataset contains 12,100 essays from the 2006-2007 TOEFL exams, divided into 9,900 for training, 1,100 for validation, and 1,100 for testing. Human raters have assessed each essay, assigning them to low, medium, or high proficiency categories. We reported weighted F1 and QWK in evaluation. The ELLIPSE dataset consists of 6,482 essays from 8th to 12th-grade English learners (including 2,568 essays for testing), each rated on a scale of 1 to 5 with increments of 0.5. For evaluation purposes, we calculated RMSE and QWK after normalizing the scores to align with the specified increments.

We fine-tuned BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) on the training data. For LLMs, we utilized GPT-3.5 and GPT-4 in zero-shot and few-shot settings and fine-tuned GPT-3.5 as well. See detailed model settings in Appendix B. Counterfactual samples were produced for the test sets. Subsequently, the model's predictions for both the original texts and their counterfactual counterparts were compared.

3.2 Results

Table 2 and Table 3 display the agreement between model and human evaluations on the test set, and the differential scoring on counterfactual samples by the models. Our findings are as follows:

Firstly, BERT-like models exhibit higher agreement with human raters than LLMs, and they can discern complex linguistic concepts such as conventions (spelling, SVA and word order) and language complexity. This challenges the claim by Singla et al. (2021) that BERT-based model functions as bag-of-words. It is important to note, however, that these BERT-like models struggle to differentiate interventions in organization, indicating insensitivity to logical structures and coherence within texts.

Secondly, LLMs, despite their lower agreement scores, are more sensitive to various linguistic aspects. This indicates a closer alignment of LLMs with scoring rubrics. Notably, Figure 1 demonstrates that zero-shot LLMs grade more stringently, resulting in lower alignment with human ratings. Yet, applying few-shot learning and fine-tuning can substantially enhance LLMs' scoring alignment with human ratings without compromising rationale alignment. Notably, the results for the fine-tuned GPT-3.5 in Tables 2 and 3 were obtained with only 100 training samples. Further testing with 50, 200, 400, and 800 samples revealed that fine-tuning GPT-3.5 with only 200 to 400 samples achieves agreement scores comparable to or surpassing BERT's, with consistent rationale alignment capabilities (see Appendix C).

	TOEFL11		ELLIPSE	
	F1 ↑	QWK ↑	RMSE ↓	QWK ↑
BERT	.783	.736	.437	.680
RoBERTa	**.795**	.739	.430	.695
DeBERTa	.790	**.741**	**.422**	**.720**
GPT-3.5-ZSL	.599	.408	.701	.399
GPT-3.5-FSL	.546	.314	.570	.378
GPT-3.5-SFT	*.710*	*.592*	*.550*	*.629*
GPT-4-ZSL	.368	.380	.960	.261
GPT-4-FSL	.490	.477	.680	.466

Table 2: The rating performance on the test sets: **best** in bold, *supervised fine-tuned GPT* in italics, best non-fine-tuned GPT underlined.

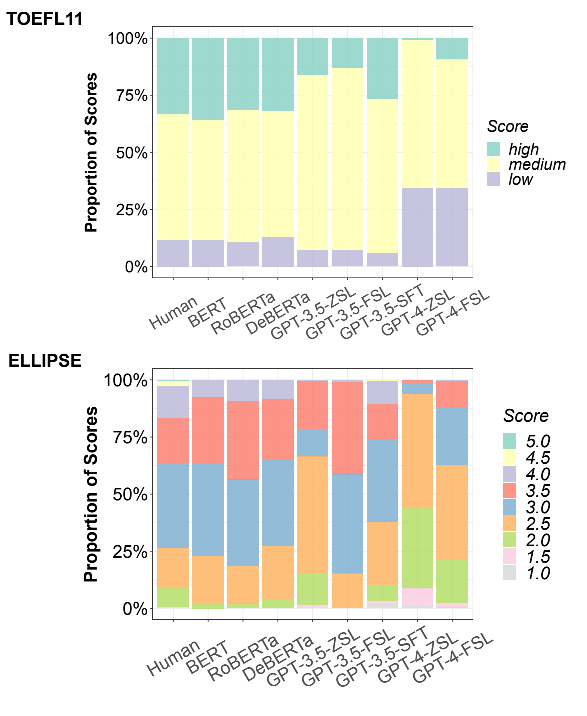

Figure 1: Distribution of score predictions on the test sets by different models.

[1] https://github.com/scrosseye/ELLIPSE-Corpus

Dataset	Model	Conventions				Language Complexity		Organization	
		Error Correction	Error Introduction			Complexification	Simplification	InParaShuffle	InTextShuffle
		–	Spelling	SVA	WOS	–	–	–	–
TOEFL11	BERT	$1.03^{+.043}_{-.041}$	$-0.92^{+.032}_{-.033}$	$-0.22^{+.013}_{-.014}$	$-1.26^{+.033}_{-.032}$	$0.42^{+.035}_{-.035}$	$-0.69^{+.033}_{-.033}$	$-0.01^{+.006}_{-.006}$	$-0.01^{+.006}_{-.006}$
	RoBERTa	$0.99^{+.043}_{-.044}$	$-0.79^{+.033}_{-.032}$	$-0.45^{+.021}_{-.021}$	$-1.13^{+.033}_{-.033}$	$0.24^{+.032}_{-.031}$	$-0.35^{+.025}_{-.025}$	$-0.19^{+.010}_{-.011}$	$-0.02^{+.005}_{-.005}$
	DeBERTa	$1.19^{+.045}_{-.046}$	$-0.92^{+.031}_{-.031}$	$-0.35^{+.016}_{-.016}$	$-1.24^{+.033}_{-.032}$	$0.33^{+.034}_{-.032}$	$-0.27^{+.027}_{-.026}$	$-0.06^{+.005}_{-.005}$	$-0.06^{+.005}_{-.005}$
	GPT-3.5-ZSL	$0.64^{+.032}_{-.031}$	$-0.76^{+.033}_{-.034}$	$-0.20^{+.026}_{-.026}$	$-0.59^{+.032}_{-.030}$	$0.27^{+.025}_{-.024}$	$0.01^{+.019}_{-.020}$	$-0.31^{+.030}_{-.030}$	$-0.42^{+.032}_{-.032}$
	GPT-4-ZSL	$0.92^{+.025}_{-.025}$	$-0.80^{+.025}_{-.025}$	$-0.35^{+.021}_{-.021}$	$-0.80^{+.026}_{-.026}$	$0.66^{+.025}_{-.025}$	$-0.24^{+.021}_{-.021}$	$-0.24^{+.018}_{-.017}$	$-0.29^{+.019}_{-.019}$
ELLIPSE	BERT	$0.84^{+.014}_{-.014}$	$-0.57^{+.011}_{-.011}$	$-0.09^{+.003}_{-.003}$	$-0.57^{+.011}_{-.011}$	$0.31^{+.009}_{-.009}$	$-0.11^{+.008}_{-.008}$	$-0.01^{+.002}_{-.002}$	$-0.02^{+.002}_{-.003}$
	RoBERTa	$0.92^{+.014}_{-.015}$	$-0.50^{+.009}_{-.009}$	$-0.11^{+.003}_{-.003}$	$-0.54^{+.009}_{-.009}$	$0.25^{+.008}_{-.007}$	$-0.05^{+.007}_{-.007}$	$-0.01^{+.002}_{-.002}$	$-0.10^{+.003}_{-.003}$
	DeBERTa	$1.06^{+.016}_{-.016}$	$-0.64^{+.013}_{-.013}$	$-0.20^{+.006}_{-.006}$	$-0.64^{+.013}_{-.013}$	$-0.08^{+.007}_{-.007}$	$0.01^{+.005}_{-.005}$	$-0.02^{+.001}_{-.001}$	$-0.07^{+.002}_{-.002}$
	GPT-3.5-ZSL	$0.77^{+.019}_{-.018}$	$-0.60^{+.019}_{-.018}$	$-0.19^{+.015}_{-.015}$	$-0.35^{+.018}_{-.018}$	$0.48^{+.016}_{-.016}$	$0.08^{+.014}_{-.014}$	$-0.15^{+.015}_{-.014}$	$-0.18^{+.016}_{-.017}$
	GPT-3.5-FSL	$0.35^{+.014}_{-.014}$	$-0.46^{+.015}_{-.015}$	$-0.15^{+.012}_{-.012}$	$-0.31^{+.014}_{-.014}$	$0.36^{+.014}_{-.014}$	$-0.04^{+.012}_{-.012}$	$-0.11^{+.013}_{-.012}$	$-0.16^{+.014}_{-.014}$
	GPT-3.5-SFT	$1.08^{+.021}_{-.021}$	$-1.00^{+.022}_{-.022}$	$-0.30^{+.014}_{-.014}$	$-0.62^{+.018}_{-.017}$	$0.90^{+.017}_{-.017}$	$0.04^{+.014}_{-.013}$	$-0.17^{+.013}_{-.013}$	$-0.23^{+.014}_{-.014}$
	GPT-4-ZSL*	$0.87^{+.060}_{-.058}$	$-0.64^{+.047}_{-.047}$	$-0.30^{+.045}_{-.045}$	$-0.56^{+.045}_{-.045}$	$0.96^{+.065}_{-.065}$	$-0.05^{+.058}_{-.057}$	$-0.10^{+.033}_{-.035}$	$-0.19^{+.037}_{-.040}$
	GPT-4-FSL*	$0.61^{+.052}_{-.048}$	$-0.71^{+.060}_{-.060}$	$-0.27^{+.050}_{-.050}$	$-0.56^{+.048}_{-.050}$	$0.67^{+.055}_{-.052}$	$-0.09^{+.045}_{-.043}$	$-0.14^{+.032}_{-.035}$	$-0.23^{+.042}_{-.045}$

Table 3: Mean score shifts $\overline{\Delta\mathcal{S}}$ ($\Delta\mathcal{S} = \mathcal{S}_{\text{cf}} - \mathcal{S}_{\text{origin}}$, where $\mathcal{S}_{\text{cf}}, \mathcal{S}_{\text{origin}} \in [1, 5]$) after interventions: full and stratified subset results*, with subscripts and superscripts indicating confidence intervals (obtained through 10,000 bootstrap iterations). Gray shading indicates non-significant differences ($p > 0.01$) in scores before and after intervention.

3.3 Self-Explanation in Feedback

Providing feedback is crucial in AES (Ramesh and Sanampudi, 2022), and it also allows the model to explain its scoring decisions. Studies have shown that LLMs can offer useful essay feedback (Han et al., 2023; Xiao et al., 2024), while the faithfulness has not been thoroughly evaluated.

Therefore, after assessing the scoring shifts to interventions, we further investigated whether the feedback provided by LLMs was consistent with specific interventions. We instructed few-shot GPT-4 to generate multi-dimensional feedback based on the scoring rubrics for the essay it had evaluated. Three trained annotators evaluated these feedback, determining if the model provided reasonable feedback for each "original-counterfactual" pair. See detailed evaluation procedures in Appendix D.

Table 4 shows the evaluation results based on a strict majority vote, highlighting that GPT-4 is capable of providing differentiated feedback for both original and counterfactual samples, with particularly strong performance in text complexification, word order swapping, and error correction. However, the feedback distinctions in certain aspects, like SVA, simplification, and sentence shuffling, were not very clear. Further analysis indicated that the ELLIPSE essays, written by 8th to 12th grade English learners, were inherently simple in vocabulary and syntax, contained some SVA errors, and displayed imperfect structural organization. Consequently, the model frequently identified SVA issues and offered numerous organizational and developmental suggestions in its feedback on the original essays, which led to less distinct differences in the feedback for the counterfactual samples.

Category	Counterfactual Type	Detection Rate%
Conventions	Error Correction	72
	Spelling	68
	SVA	48
	WOS	80
Language Complexity	Complexification	100
	Simplification	32
Organization	InParaShuffle	40
	InTextShuffle	20

Table 4: Voting-Based Detection Rates of Original Samples vs. Counterfactual Feedback

4 Conclusion

We generated linguistically-informed counterfactuals by combining LLMs and rule-based techniques, analyzing their impact on essay scoring by BERT-like models and LLMs. Our findings emphasize that a higher agreement with human raters does not necessarily indicate a better alignment with scoring rubrics, suggesting the models' evaluation should consider both aspects. Moreover, our study highlights LLMs' significant potential in AES domain: firstly, while zero-shot LLMs show less agreement compared to BERT-like models, few-shot and fine-tuned LLMs can maintain both high score agreement and rationale alignment. Secondly, LLMs are not only sensitive to interventions in scoring but can also discernibly offer writing feedback, a function beyond the reach of traditional AES systems.

This study sheds light on *why* a neural model assigns specific scores to essays. It unveils how modifying domain-specific concepts in texts to craft counterfactuals enhances transparency in model decisions—a method applicable across multiple fields. With LLMs, counterfactual generation has been greatly empowered, boosting transparency and accountability in machine learning applications.

5 Limitations

We note that GPT-4 shows a strong performance in generating counterfactual samples, yet due to the space limit, we cannot detail the generation methods and evaluation results in the main body of the text.

In addition to conventions, language complexity, and organization, TOEFL independent writing rubrics also emphasize content-related evaluations—namely, assessing relevance to the prompt and fulfillment of task requirements. These aspects, being beyond mere linguistic concepts, were not included in the current scope of our study. This is because counterfactual interventions require modifying a specific aspect while keeping others constant. This is also because we can adjust linguistic features without affecting content, but altering content inevitably impacts the linguistic aspect. However, we acknowledge that task and topic relevance, as important scoring dimensions, warrant future in-depth exploration.

Our experiment demonstrated that LLMs have significant potential in providing feedback. In this paper, we focus on the feedback differences between original and counterfactual samples. A comprehensive evaluation of the LLM-genearated feedback is a crucial step for future research.

References

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.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

James Fiacco, David Adamson, and Carolyn Ros. 2023. Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Anubha Kabra, Mehar Bhatia, Yaman Kumar Singla, Junyi Jessy Li, and Rajiv Ratn Shah. 2022. Evaluation toolkit for robustness testing of automatic essay scoring systems. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD '22, pages 90–99, New York, NY, USA. Association for Computing Machinery.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–70.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

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.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*.

Yaman Kumar Singla, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, and Changyou Chen. 2021. Aes systems are both overstable and oversensitive: Explaining why and proposing defenses.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape.

Appendix

A Details of Counterfactual Sample Generation

A.1 GPT-4 Prompts for Modifying Essays

The counterfactual samples of text correction, complexification and simplification are generated by the `gpt-4-1106-preview` model. When calling OpenAI's APIs, we turn on `JSON` mode to get easier parsing results. For reproducibility, we set the `temperature` parameter to 0 and the `seed` to 42.

A.1.1 Prompt for Error Correction

System: You are an experienced writing tutor.

<u>User:</u> Please fix the spelling, punctuation and grammatical errors in the given essay. Ensure the main idea, the words used, the sentence structure, and the length of the text remain consistent with the original text.

Input Essay:
"{}"

Please return the output essay in JSON format as below:
```
{"output_essay": "..."}
```
Output:

A.1.2 Prompt for Complexification

System: You are an experienced writing tutor.

<u>User:</u> Modify the provided essay to enhance its lexical sophistication and syntactic variety following the instructions below:
1. Expand lexical range: Vary word choice and replace common words with advanced vocabulary when suitable without compromising clarity or meaning. Avoid repeating the same words and capture subtle differences in meaning.
2. Increase syntactic complexity: Incorporate a wider range of sentence structures including compound-complex sentences, varied clause types, subordination and coordination. Use advanced constructions such as non-finite clauses, adverbials, conditionals, inversion and passives where appropriate.
3. Maintain meaning, length and clarity: The revised text should retain the original ideas and conform to the initial length while remaining clear and understandable.

Input Essay:
"{}"

Please return the output essay in JSON format as below:
```
{"output_essay": "..."}
```
Output:

A.1.3 Prompt for Simplification

System: You are an experienced writing tutor.

<u>User:</u> Modify the provided essay to simplify its vocabulary and sentence structure following the instructions below:
1. Simplify vocabulary: Replace advanced words with common everyday equivalents for clear understanding. Limit synonyms to favor those most commonly used.
2. Simplify sentence structure: Break down complex sentences and avoid clauses, conjunctions, and nesting where possible. Favor short, simple subject-verb-object sentences.
3. Maintain meaning, length and clarity: The revised text should retain the original ideas and conform to the initial length while remaining clear and understandable.

Input Essay:
"{}"

Please return the output essay in JSON format as below:
```
{"output_essay": "..."}
```
Output:

Indicator	Description
WordNum	The number of words in an essay. Caculated with spaCy.
SentNum	The number of sentences in an essay. Caculated with spaCy.
MLS	Mean length of sentences. The length of each sentence is the number of words it has.
ADDT	Average depth of dependency tree for all sentences in an essay. Dependency parsing is done using spaCy.
LemmaTTR	A lexical diversity measure based on the type token ratio (TTR) for an essay in which each word is lemmatized.
LexSoph	A lexical sophistication measure based on word frequency statistics from COHA corpus (records from the 1980s to the 2010s). $\lvert\mathcal{W}\rvert^{-1} \cdot \sum_{w \in \mathcal{W}} (\log(\mathrm{Freq}(\mathcal{L}(w)) + 1))^{-1}$, in which $\lvert\mathcal{W}\rvert$ meas WordNum, w means each word in an essay (repeated words are counted for the number of repetitions), $\mathcal{L}()$ means lemmatization and Freq() means the operation of getting the frequency of the lemmatized word from the frequency dictionary from COHA[2].
ErrorDensity	Density of writing errors in an essay, defined as #error/$\lvert\mathcal{W}\rvert$. Writing error analyses are obtained using LanguageTool[3].

Table 5: The linguistics metrics used for the evaluation of counterfactual samples.

A.2 Evaluating the Validity of Counterfactuals

To evaluate the counterfactual samples generated by GPT-4, we introduced seven linguistic metrics that measure the text length, lexical diversity, lexical sophistication, syntactic complexity and writing error density. The descriptions of these metrics can be seen in Table 5.

After generating the counterfactual samples for the essays from the test set, we measured the above seven indices on both original texts and their counterfactual counterparts. Cohen's $\mathcal{D}$ (Cohen, 2013) was employed to quantify the intervention's effect size, which is defined as the difference between two means divided by a standard deviation for the data, i.e.

$$\mathcal{D} = \frac{\bar{x}_{\text{post}} - \bar{x}_{\text{pre}}}{s}.$$

where s, the pooled standard deviation, is defined as:

$$s = \sqrt{\frac{(n_{\text{pre}} - 1)\, s^2_{\text{pre}} + (n_{\text{post}} - 1)\, s^2_{\text{post}}}{n_{\text{pre}} + n_{\text{post}} - 2}}$$

where the variance for the pre-intervention group is defined as

$$s^2_{\text{pre}} = \frac{1}{n_{\text{pre}} - 1} \sum_{i=1}^{n_{\text{pre}}} \left(x^i_{\text{pre}} - \bar{x}_{\text{pre}}\right)^2,$$

and similarly for the post-intervention group.

Figure 2 shows the evaluation results. For error correction, it can be seen that the error density significantly decreased after GPT-4's correction, while the text length (WordNum) remained unchanged. Additionally, there is a minimal shift in lexical and syntactic metrics, largely due to the correction of punctuation, vocabulary and syntax errors. Regarding text complexification, GPT-4 has successfully enhanced lexical diversity and sophistication along with syntactic tree depth, while changes to other metrics are nominal. Lastly, in terms of text simplification, the intervention has significantly streamlined syntactic structures and moderately reduced lexical variety and sophistication. Consequently, the overall text length has been slightly reduced.

B The Implementation of AES methods

B.1 Fine-tuning BERT-like Models

We fine-tuned three commonly used pre-trained transformer-based encoder models, specifically bert-base-uncased, roberta-base, and deberta-v3-base.

B.1.1 Basic Settings

As the essays in the TOEFL11 dataset are categorized into low, medium, and high categories, we developed a three-class classifier using the cross-entropy loss. We use the AutoModelForSequenceClassification class from Hugging Face transformer, setting num_labels=3 to load the pre-training checkpoints. For the ELLIPSE dataset, with scores ranging from 1.0 to 5.0, we model it as a regression problem by setting num_labels=1 and using the mean squared error (MSE) loss function.

B.1.2 Hyperparameters

In our model fine-tuning process, we experimented with four distinct learning rates: 1e-5, 2e-5, 3e-5, and 5e-5, using Hugging Face's Trainer. We

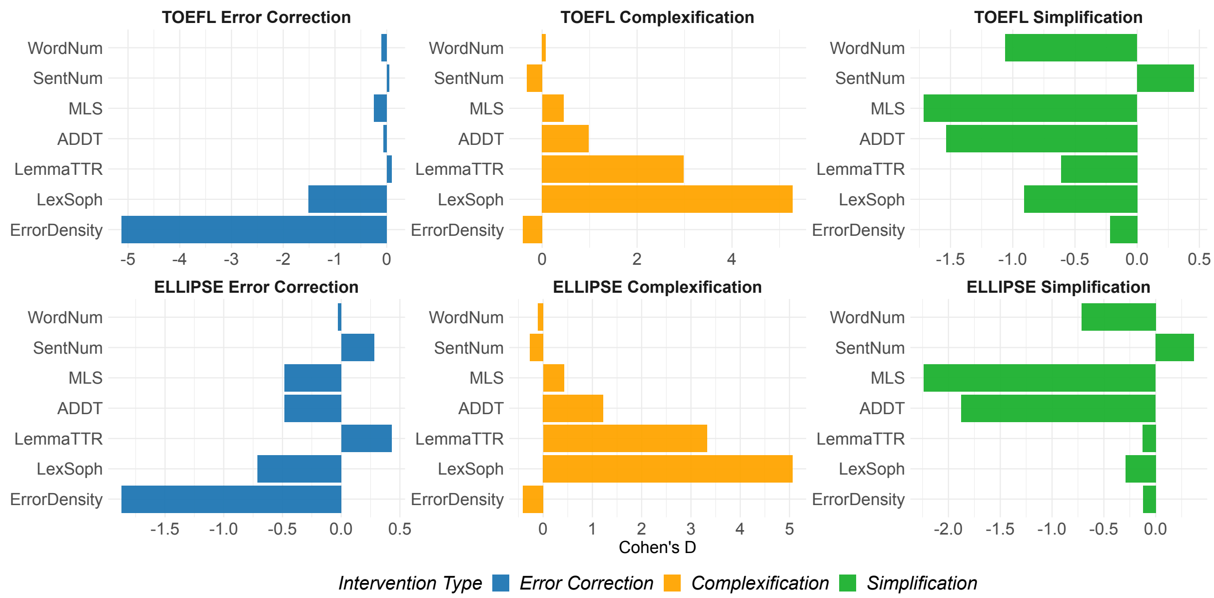

Figure 2: Intervention Effect Sizes. Cohen's $\mathcal{D}$ measured for 7 linguistic indices on three interventions: **Error Correction**, **Complexification**, and **Simplification**.

identify the best learning rate that led to the lowest loss on the validation set (results see Table 6). We used a linear learning rate scheduler that includes a 50-step warm-up phase, where the learning rate initially increases from a lower value to a specified maximum (chosen from the four rates: 1e-5, 2e-5, 3e-5, and 5e-5) and then decreases linearly. This method ensures gradual adaptation of the model's weights, with the peak learning rates being reached at the end of the warm-up.

For other parameters, we used a seed of 42 and a batch size of 16 for both training and evaluation. We aimed for a maximum of 10 epochs, with the actual duration potentially reduced by early stopping, triggered if loss value fails to improve after 5 checks. The approach included a weight decay of 0.01 for overfitting prevention and `FP16` for efficient training. Input lengths were adjusted to 512 tokens through padding and truncation to ensure uniformity across all samples.

B.2 Prompting LLMs to Score Essays

As introduced in Section 3, we also used LLMs for essay scoring, including `gpt-3.5-turbo-1106` and `gpt-4-1106-preview` based on OpenAI's API. We turned on `JSON` mode to get easier parsing results, and set the `temperature` parameter to 0 and the `seed` parameter to 42 for reproducibility.

Dataset	Model	Learning Rate	Early Stop @ Step	Validation Loss ↓
TOEFL11	`bert-base-uncased`	**1e-5**	**450**	**.443**
		2e-5	550	.453
		3e-5	350	.462
		5e-5	150	.482
	`roberta-base`	**1e-5**	**450**	**.403**
		2e-5	450	.424
		3e-5	400	.442
		5e-5	500	.467
	`deberta-v3-base`	**1e-5**	**500**	**.398**
		2e-5	400	.400
		3e-5	250	.416
		5e-5	250	.427
ELLIPSE	`bert-base-uncased`	1e-5	500	.173
		2e-5	**200**	**.172**
		3e-5	300	.179
		5e-5	150	.185
	`roberta-base`	1e-5	250	.196
		2e-5	100	.199
		3e-5	**500**	**.171**
		5e-5	300	.176
	`deberta-v3-base`	**1e-5**	**200**	**.157**
		2e-5	150	.167
		3e-5	200	.160
		5e-5	150	.181

Table 6: Performance of the three models on the validation set after fine-tuning using different learning rates on both TOEFL11 and ELLIPSE datasets. Learning rates for achieving minimum loss in each model for both datasets are **bolded**.

B.2.1 Prompts for Scoring TOEFL11 Essays with Zero-shot Learning

Below is the scoring template for TOEFL11 essays. In the zero-shot setting, we provide the LLMs with the essay prompt, the essay itself, and the scoring rubrics. Notably, while the TOEFL11 dataset only provides low, medium, and high score levels for the essays without specific scores, the TOEFL rating rubric is actually based on a 1 to 5 scale. Consequently, even in zero-shot scenarios without examples or training data, we can still prompt LLMs to assess and score TOEFL11 essays.

System: You are a TOEFL rater specializing in the evaluation of the Independent Writing section.

User: Read and evaluate the essay written in response to the prompt: "{}"

Essay: "{}"

Please assign it a score from 1 to 5 (in increments of 0.5 points) based on rubric below: "{TOEFL11_RUBRICS}"

Your response should be a JSON object containing only one key: 'score', which should be a numeric value representing the score you gave.

TOEFL11 Rubrics

- A 5-point essay effectively addresses all aspects of the topic and task. Well organized and developed with clearly appropriate explanations and details. Displays strong unity, progression and coherence. Shows consistent language facility with syntactic variety, appropriate word choice and idiomaticity. May have minor lexical or grammatical errors.
- A 4-point essay addresses the topic and task well, though some points may not be fully elaborated. Generally well organized and developed with appropriate and sufficient explanations, exemplifications and details. Displays unity, progression and coherence, though may contain occasional redundancy, digression or unclear connections. Demonstrates syntactic variety and vocabulary range. May have occasional minor errors that do not interfere with meaning.
- A 3-point essay addresses the topic and task with somewhat developed explanations, exemplifications and details. Displays unity, progression and coherence, though connection of ideas may be occasionally obscured. May demonstrate inconsistent language facility resulting in lack of clarity and obscured meaning. May display accurate but limited structures and vocabulary.
- A 2-point essay shows limited development in response to the topic and task. Inadequate organization or connection of ideas. Insufficient or inappropriate exemplifications, explanations or details to support generalizations. Noticeable inappropriate word choices or word forms. An accumulation of errors in sentence structure and/or usage.
- A 1-point essay is seriously flawed due to disorganization, underdevelopment, little or no supporting detail, and unresponsiveness to the task. Contains serious and frequent errors in sentence structure or usage.

B.2.2 Prompts for Scoring ELLIPSE Essays with Zero-shot Learning

Below is the scoring template for ELLIPSE essays. Since the ELLIPSE's rubrics do not require adherence to a specific prompt or fulfillment of task requirements. We only provide the LLMs with the essay to be rated and the scoring rubrics.

System: You are an essay rater specializing in the evaluation of essays written by students from 8th to 12th grade who are learning English as a second language.

User: Read and evaluate the essay: "{}"

Assign it a score from 1 to 5, in increments of 0.5, based on this rubric:
"{ELLIPSE_RUBRICS}"

Your response should be a JSON object containing only one key: 'score', which should be a numeric value representing the score you gave.

ELLIPSE Rubrics

- A 5-point essay demonstrates native-like facility in the use of language with syntactic variety, appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.
- A 4-point essay demonstrates facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and con-

ventions; occasional language inaccuracies that rarely impede communication.
- A 3-point essay demonstrates facility limited to the use of common structures and generic vocabulary; organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases.
- A 2-point essay demonstrates inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies.
- A 1-point essay demonstrates a limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies.

B.2.3 Prompts for Scoring TOEFL11 Essays with Few-shot Learning

For few-shot learning on TOEFL11 dataset, we gave three examples from the low, medium and high categories, and asked the models to return the score level as well. See the prompt below.

System: You are a TOEFL rater specializing in the evaluation of the Independent Writing section.

User: Read and evaluate the essay written in response to the prompt: "{}"

Example essay 1 of score level "High":
"{A_REPRESENTATIVE_HIGH_SCORE_ESSAY}"

Example Essay 2 of score level "Medium":
"{A_REPRESENTATIVE_MEDIUM_SCORE_ESSAY}"

Example Essay 3 of score level "Low":
"{A_REPRESENTATIVE_LOW_SCORE_ESSAY}"

Essay to score:
"{}"

Please note:
- Low corresponds to scores of 1.0 - 2.0
- Medium corresponds to scores of 2.5 - 3.5
- High corresponds to scores of 4.0 - 5.0

Assign the essay a score level of Low, Medium or High based on the criteria in the rubric below:
"{TOEFL11_RUBRICS}"

Your response should be a JSON object with the key "`score_level`" set to either "`Low`", "`Medium`", or "`High`" representing the level you determined for this essay.

B.2.4 Prompts for Scoring ELLIPSE Essays with Few-shot Learning

To align with the process of rating TOEFL11 essays, we also provide three example essays from the ELLIPSE dataset, representing low, medium, and high score levels. However, we give the specific scores of these examples and require the model to return numerical scores as well. Refer to the following prompt. For information on how to select samples, see the next section.

System: You are an essay rater specializing in the evaluation of essays written by students from 8th to 12th grade who are learning English as a second language.

User: Read and evaluate the essay:

Example essay 1 of score "4.0":
"{A_REPRESENTATIVE_HIGH_SCORE_ESSAY}"

Example Essay 2 of score "3.0":
"{A_REPRESENTATIVE_MEDIUM_SCORE_ESSAY}"

Example Essay 3 of score "2.0":

"{A_REPRESENTATIVE_LOW_SCORE_ESSAY}"

Essay to score:
"{}"

Assign it a score from 1 to 5, in increments of 0.5, based on this rubric:
"{ELLIPSE_RUBRICS}"

Your response should be a JSON object containing only one key: 'score', which should be a numeric value representing the score you gave.

B.2.5 Few-shot Example Selection

We use a linguistic-based approach to select the representative examples for few-shot learning by following the steps:

1. **Calculate Metrics:** Calculate and normalize the seven linguistic metrics mentioned in Appendix A.2 for training sets of both TOEFL11 and ELLIPSE datasets.

2. **Process Data:** Apply Principal Component Analysis (PCA) to identify the top five components that explain 95% of the variance, representing essential linguistic features.

3. **Represent Samples:** Utilize these principal components to represent the linguistic features of all training samples.

4. **Determine Medoids:** Categorize samples into proficiency levels (low, medium, high) and find the medoid of each group using Euclidean distance.

Note that a medoid is an object within a dataset that minimally differs from all other objects in the dataset, according to a given distance metric. It is similar to the concept of a centroid, but while a centroid may not be an actual data point, a medoid is always a member of the dataset.

C Data Size Effect in Fine-tuning LLM

To investigate the impact of training set size on scoring performance and counterfactual intervention responses, we fine-tuned GPT-3.5 using 50, 200, 400, and 800 examples[4] from the training sets on two datasets. We then conducted a counterfactual analysis on the stratified subset of the ELLIPSE dataset used in Section 3.2. Figure 3 illustrates the test set performance for both datasets. On ELLIPSE, the performance plateaus when the training set size reaches approximately 200 while TOEFL11 shows a deceleration in improvement after 400 examples but suggests potential for further increment. Table 7 presents the impact of counterfactual interventions on the ELLIPSE subset, demonstrating that fine-tuning data size has a limited impact on the model's response to these interventions.

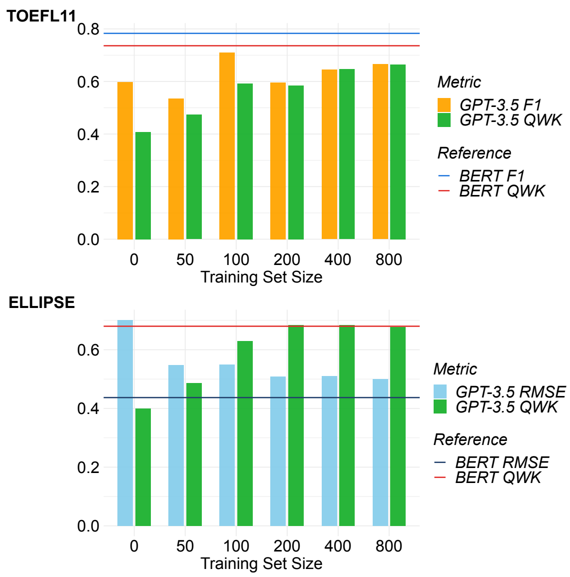

Figure 3: Scoring performance of GPT-3.5 SFT models on both datasets for different training data sizes. The zero tick on the horizontal axis represents the zero-shot learning condition.

D Details for Feedback Generation and Evaluation

D.1 Feedback Generation

Given the stable performance of few-shot GPT-4 in handling a variety of counterfactual interventions, we conducted the manual evaluations on this model. As shown in Figure 4, we prompted the few-shot GPT-4 to provide writing feedback to the essay it just scored. The experiments were conducted on a stratified subset of ELLIPSE. For 200 samples in the subset, we requested GPT-4 to provide feedback

[4]These different sized subsets of the training set were obtained by random stratified sampling on the overall training set.

Training Set Size	Conventions				Language Complexity		Organization	
	Error Correction	Error Introduction			Complexification	Simplification	InParaShuffle	InTextShuffle
	–	Spelling	SVA	WOS	–	–	–	–
50	$0.83^{+.075}_{-.072}$	$-0.64^{+.077}_{-.080}$	$-0.14^{+.045}_{-.050}$	$-0.34^{+.065}_{-.068}$	$0.96^{+.060}_{-.062}$	$0.08^{+.055}_{-.052}$	$-0.09^{+.045}_{-.045}$	$-0.10^{+.047}_{-.050}$
100	$1.12^{+.080}_{-.080}$	$-0.95^{+.080}_{-.080}$	$-0.26^{+.052}_{-.052}$	$-0.58^{+.057}_{-.055}$	$0.88^{+.055}_{-.057}$	$0.05^{+.050}_{-.048}$	$-0.18^{+.047}_{-.050}$	$-0.19^{+.048}_{-.050}$
200	$1.03^{+.092}_{-.090}$	$-0.57^{+.087}_{-.090}$	$-0.01^{+.068}_{-.070}$	$-0.32^{+.072}_{-.070}$	$0.79^{+.052}_{-.055}$	$-0.02^{+.037}_{-.037}$	$0.06^{+.060}_{-.060}$	$0.02^{+.062}_{-.062}$
400	$1.11^{+.087}_{-.090}$	$-0.95^{+.075}_{-.075}$	$-0.30^{+.060}_{-.060}$	$-0.66^{+.068}_{-.065}$	$0.76^{+.055}_{-.057}$	$-0.03^{+.045}_{-.042}$	$-0.18^{+.052}_{-.052}$	$-0.23^{+.050}_{-.052}$
800	$1.02^{+.085}_{-.085}$	$-0.83^{+.080}_{-.080}$	$-0.23^{+.065}_{-.067}$	$-0.55^{+.070}_{-.070}$	$0.94^{+.055}_{-.055}$	$-0.03^{+.048}_{-.050}$	$-0.14^{+.052}_{-.055}$	$-0.23^{+.060}_{-.062}$

Table 7: Mean score shifts $\overline{\Delta S}$ of GPT-3.5 SFT models of different training set sizes on the stratified ELLIPSE subset, with subscripts and superscripts indicating confidence intervals (obtained through 10,000 bootstrap iterations). Gray shading indicates non-significant differences ($p > 0.01$) in scores before and after intervention.

respectively on each of the original samples and their specific counterfactual counterparts.

Session 1: Essay Scoring

User: Read and evaluate the essay: ...

Assistant: {'score': 3.0}

Session 2: Providing Feedback

User: Please provide balanced and constructive feedback on the following aspects ...

Figure 4: An Example of Feedback Generation

Full Prompt Instructing GPT-4 to Provide Feedback

User: Please provide balanced and constructive feedback on the following aspects of the essay you have just rated (not the example essay):
1. Organization:
- Evaluate how effectively ideas are communicated and organized. Identify any issues with the logical flow, transitions between ideas, and clarity in conveying concepts. Comment on the introduction's setup, idea development throughout the body, and the conclusiveness of the ending.
2. Language Use:
- Morphology: Identify errors in word formation and structure, focusing on verb tenses, irregular verbs, plurals, possessives, affixes, agreement, and gerund/participle usage.
- Syntax: Comment on the arrangement of words and phrases to create well-formed sentences, coherence in sentence construction, and the complexity and variety of sentence types.
- Vocabulary: Assess the appropriateness of word choice, the diversity and sophistication of vocabulary employed, and note any imprecise use of words where more accurate or specific terms could be used.
3. Conventions:
- Highlight any errors in spelling, capitalization, and punctuation.
Your response should be a structured JSON object with the following keys:
```json
{{
"organization_feedback": "",
"language_use_feedback": "",
"conventions_feedback": ""
}}
```
If possible, include direct citations from the essay to substantiate your feedback.

D.2 Feedback Evaluation

After collecting 200 "original-counterfactual" feedback pairs, we divided them into 8 equal portions, with each portion corresponding to one counterfactual intervention for manual evaluation. We firstly prepared a dataset with concealed labels to hide which feedback was "original" or "counterfactual" from annotators. Three graduate students major in Linguistics were tasked with identifying the feedback difference within each pair. Outcomes were categorized as correct, incorrect (opposite of actual labels), or indeterminate. A majority vote among the annotators decided each pair's final classification. Pairs without a majority, due to each annotator choosing a different outcome, were labeled "indeterminate". Results are in Table 8. Table 9 gives an example of a feedback pair where the counterfactual feedback corresponds to a sample obtained by introducing spelling errors to the original sample.

Category	Counterfactual Type	Correct%	Incorrect%	Indeterminate%
Conventions	Error Correction	72	4	24
	Spelling	68	0	32
	SVA	48	0	52
	WOS	80	0	20
Language Complexity	Complexification	100	0	0
	Simplification	32	0	68
Orgnization	InParaShuffle	40	4	56
	InTextShuffle	20	8	72

Table 8: Feedback Type Voting Results by Annotators

Feedback for the Original	Feedback for the Counterfactual
organization_feedback The essay begins with a clear topic sentence that introduces the importance of parks for improving mood, but the development of ideas throughout the body is somewhat disjointed. The writer provides personal anecdotes and general observations to support the argument, but transitions between ideas are abrupt, and there is a lack of clear logical flow. The conclusion does not effectively summarize the main points or provide a strong final thought on the topic. The essay would benefit from a more structured approach, with each paragraph focusing on a single aspect of the importance of parks and using transitions to guide the reader through the argument.	**organization_feedback** The essay's organization needs improvement. Ideas about the importance of parks and their impact on mood and childhood experiences are presented, but the logical flow is disjointed. Transitions between ideas are abrupt or missing, making it difficult to follow the argument. The introduction lacks a clear thesis statement, and the body paragraphs do not consistently develop the initial idea. The conclusion is not clearly defined, and the essay ends abruptly without summarizing the main points or providing a final thought on the importance of parks for children.
language_use_feedback Morphology: There are several errors in verb tense consistency ('i use to just go outside' should be 'I used to just go outside') and subject-verb agreement ('kids finds parks activity enjoyable' should be 'kids find park activities enjoyable'). Syntax: Sentence structures are often simplistic and repetitive, lacking variety and complexity. There are also instances of awkward phrasing ('my mom always straggle with tell my brothers to go outside' should be 'my mom always struggles to tell my brothers to go outside'). Vocabulary: The vocabulary is generally appropriate but lacks diversity and sophistication. There are instances of imprecise language ('be open minded' could be 'become open-minded', 'enjoy life' could be 'enjoy the outdoors').	**language_use_feedback** Morphology: There are several errors in verb tenses and subject-verb agreement (e.g., 'my mm takes him ty the park and ha play' should be 'my mom takes him to the park and he plays'). Syntax: Sentence structures are often incomplete or incoherent, which disrupts the flow of information (e.g., 'Know i barely even go outside i go ti school and then i come back is just' is difficult to understand). Vocabulary: Word choice is sometimes inaccurate or too simplistic, and there is a lack of diversity in vocabulary. More precise language could enhance the clarity of the essay (e.g., 'bouring' should be 'boring', 'actvity enjoyble' should be 'activities enjoyable').
conventions_feedback The essay contains several spelling errors ('straggle' should be 'struggle', 'rob' should be 'rope'), incorrect capitalization ('iPod', 'iPhone', 'I', and the first word of sentences should be capitalized), and punctuation mistakes (missing commas in compound sentences, incorrect spacing around punctuation marks). Attention to these details would improve the readability and professionalism of the essay.	**conventions_feedback** The essay contains numerous spelling errors (e.g., 'iMy', 'btother', 'ty', 'ha play', 'lf', 'alls', 'jn', 'tie', 'suing', 'actvity', 'mint', 'whi offen', 'leke', 'ext', 'litter', 'bi on m'i phone', 'utside', 'tou', 'weill', 'offen mindend', 'thingh', 'dayl', 'field', 'wak ta giong tte park', 'nown', 'mer', 'estay', 'gou'). Capitalization is inconsistent, and punctuation is often missing or used incorrectly, making it challenging to understand the intended meaning of sentences.

Table 9: Example feedback pair of original and counterfactual obtained by introducing spelling errors.