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

Anonymous EMNLP submission

Abstract

While current automated essay scoring (AES) methods demonstrate high scoring agreement with human raters, their decision-making mechanisms are not fully understood. Our proposed method, using counterfactual intervention assisted by Large Language Models (LLMs), reveals that BERT-like models primarily focus on sentence-level features, whereas LLMs such as GPT-3.5, GPT-4 and Llama-3 are sensitive to conventions & accuracy, language complexity, and organization, indicating a more comprehensive rationale alignment with scoring rubrics. Moreover, LLMs can discern counterfactual interventions when giving feedback on essays. Our approach improves understanding of neural AES methods and can also apply to other domains seeking transparency in model-driven decisions. Access codes and data at anonymous repo during review: `https://anonymous.4open.science/r/beyond-agreement-aes-2024-8321`.

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), enabling a better understanding of the models' rationale. However, current model evaluations mainly use 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. (2023) 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. (2023) 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 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. As shown in Figure 1, we propose a model-agnostic diagnosis method that uses linguistically-informed counterfactuals to scrutinize the scoring behavior of BERT-like models and LLMs. The diagnostic approach closely in-

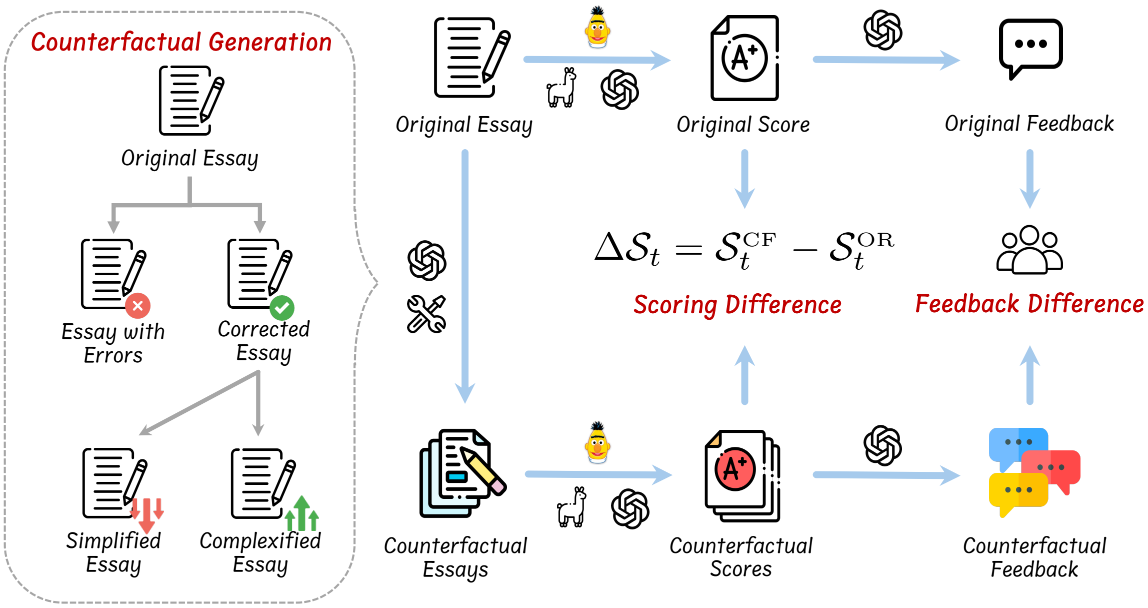

Figure 1: The pipeline of our proposed method.

tegrates linguistic knowledge from scoring rubrics, such as conventions, accuracy, vocabulary, syntax, and coherence, with LLMs employed for fine-grained and controllable counterfactual generation.

Our investigation reveals that: **(1)** BERT-like models can discern differences in conventions and language complexity but struggle to grasp the logical structure and coherence of essays; and **(2)** LLMs, although have lower score agreement than traditional models, display a superior alignment with human experts' reasoning during scoring and can also address counterfactual interventions in their feedback. Through few-shot learning or fine-tuning, LLMs can achieve both high scoring agreement and rationale alignment.

2 Related Work

2.1 AES based on Neural Language Models

Pre-trained neural language models have made significant progress in the field of AES. After fine-tuning on specific datasets, these models can achieve high levels of agreement with human raters (Rodriguez et al., 2019; Yang et al., 2020; Ormerod et al., 2021; Wang et al., 2022). Since the emergence of ChatGPT, the scoring performance of LLMs has garnered considerable attention. Leveraging their powerful language understanding capabilities and in-context learning abilities, LLMs can evaluate essays and assign overall scores or scores for specific dimensions (Naismith et al., 2023). However, research has shown that zero-shot and few-shot LLMs fail to achieve state-of-the-art scoring performance (Mizumoto and Eguchi, 2023), while fine-tuned LLM models exhibit notable superiority (Xiao et al., 2024).

Although the scoring ability of LLMs without fine-tuning is not particularly remarkable, they can provide explainable feedback in natural language. Previously, essay feedback was primarily provided through trait scores (e.g., vocabulary) (Carlile et al., 2018; Hussein et al., 2020; Lee et al., 2023). With the emergence of LLMs, researchers discovered that it is possible to elicit explanations about assessment decisions from the models (Caines et al., 2023). Han et al. (2023) assessed the feedback generated by GPT-3.5 on level of detail, accuracy, relevance, and helpfulness, while Xiao et al. (2024) found that GPT-4 feedback could elevate novice raters to expert levels.

2.2 Interpretability and Robustness of AES Models

In terms of model interpretability in AES research, Fiacco et al. (2023) analyzed the features encoded by transformer models, but this approach provides limited insight into the decision-making rationale of the models. Singla et al. (2023) employed the integrated gradients method (Sundararajan et al., 2017) on neural models to analyze token-level feature importance and discovered that BERT-based models frequently assign substantial importance to stopwords and non-linguistic tokens. This counterintuitive result may stem from the fact that the IG method does not address interactions between tokens, thereby failing to capture abstract linguistic concepts such as cohesion and syntax. Moreover, these methods cannot be directly applied to closed-source models like GPT-3.5 and GPT-4.

Additionally, a line of works have utilized adversarial modifications to diagnose model robustness. Powers et al. (2002) invited human writers to compose essays that would ”trick” the AES system and found that repeating, rewording, and reordering were effective strategies. Bejar et al. (2014) employed the substitution of words with less frequent and longer synonyms. Kabra et al. (2022) used methods such as the addition of irrelevant lines, the introduction of grammatical errors, and the deletion of lines from the responses. Myers and Wilson (2023) evaluated models using a sentence-level randomization approach. It is important to note that these studies aim to expose model vulnerabilities by introducing input perturbations rather than exploring the interpretability of model decisions.

2.3 Counterfactual Analysis

Counterfactuals are hypothetical scenarios created to understand the causal effects of specific interventions in a given situation (Feder et al., 2022). Existing counterfactual generation methods 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. While recent efforts have leveraged LLMs for generating more natural and diverse counterfactuals (Dixit et al., 2022; Chen et al., 2023), most have only exploited LLMs’ powerful language generation capabilities without tapping into their potential to understand and manipulate abstract concepts within texts. Gat et al. (2023) found that LLMs can produce high-quality counterfactuals, which assist in providing strong black-box model explanations. Li et al. (2024) prompted LLMs to identify and modify *causal terms* to generate counterfactuals. Inspired by these works, we decided to combine LLMs with rule-based methods to achieve controlled sample generation in AES.

3 Method

We employed counterfactual interventions to establish causality between target concepts and predicted scores. Typically, counterfactual intervention involves manipulating a specific feature or concept while controlling for others and observing the subsequent effect on the model’s prediction. We firstly extracted target concepts from scoring rubrics for intervention, and then generated counterfactual samples for different concepts using LLMs and heuristic rules.

3.1 Concepts for Intervention

To identify the target concepts for AES scenarios, we reviewed scoring rubrics from major standardized English tests (IELTS, TOEFL iBT, TOEIC, PTE Academic) and the ELLIPSE dataset, which is based on various state and industrial English language proficiency assessments. We conducted a detailed annotation process to identify common linguistic features across the five rubrics. See Appendix A for more information. Through this analysis, we discovered that all the scoring criteria consistently emphasize three key aspects:

Conventions and Accuracy: An essay is considered to adhere to conventions and demonstrate accuracy when it is free from mechanical (spelling, capitalization, and punctuation) mistakes and grammatical inaccuracies.

Language Complexity: An essay demonstrates lexical and syntactic complexity through the use of a broad vocabulary, sophisticated lexical control, and varied sentence structures.

Organization and Development: 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.

3.2 Counterfactual Generation

Let T be an essay, and let C_i denote a specific concept within the essay, which can take on different values $\{v_1, v_2, \ldots, v_i, \ldots\}$. A counterfactual intervention alters the value of concept C_i from v_i to v_i' , while holding the other concepts fixed.

We employ a hibrid approach combining rule-based and LLM-based methods to generate eight types of linguistically informed counterfactuals for diagnostic purposes, as detailed in Table 1. These interventions derive from three aforementioned linguistic concepts and are implemented in both positive and negative directions for conventions and language complexity. As shown in Figure 1, for conventions and accuracy, we introduce errors such

[0]https://github.com/makcedward/nlpaug

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

Table 1: Overview of positive and negative counterfactual intervention methods used.

as spelling, subject-verb agreement, and word order for negative impacts, and use LLMs to correct all errors for positive impacts. Regarding the language complexity, we leverage LLMs to increase and decrease the language complexity along both vocabulary and syntax dimensions, building upon the basis of error correction.. For the organizational aspect, negative interventions include disrupting the sentence order within paragraphs to affect local coherence and across the entire article to impact global coherence. See Appendix B.2 for LLM prompts used to generate counterfactuals.

3.3 The Validity of LLM Generated Counterfactuals

As shown in Table 1, we prompted LLMs to correct errors, complexify, and simplify essays to manipulate their conventions and language complexity. To evaluate counterfactual essays generated by LLMs, we introduced seven linguistic metrics that measure the essay length, lexical diversity, lexical sophistication, syntactic complexity and writing error density, as well as cosine similarity between the text embeddings of original and counterfactual essays to measure the extent of content preservation in counterfactual interventions. The descriptions of these metrics can be seen in Table 2.

For content preservation, we compute the average similarity values of "original-counterfactual" pairs for each of 8 types of interventions, while for linguistic metrics, we compute Cohen's $\mathcal{D}$ (Cohen, 2013) effect size for each metric as follows:

$$\mathcal{D} = \frac{\bar{x}_{CF} - \bar{x}_{OR}}{s} \tag{1}$$

where $\bar{x}_{CF}$ and $\bar{x}_{OR}$ are the mean values of a metric for the counterfactual and original samples, and the pooled standard deviation s is defined as:

$$s = \sqrt{\frac{\left(n_{OR} - 1\right) s_{OR}^2 + \left(n_{CF} - 1\right) s_{CF}^2}{n_{OR} + n_{CF} - 2}} \tag{2}$$

where n_{OR} and n_{CF} are the sample sizes, and s_{OR}^2 and s_{CF}^2 are the variances of the original and counterfactual samples respectively.

4 Experiments

4.1 Settings

Our study utilized TOEFL11 (Blanchard et al., 2013) and ELLIPSE (Crossley et al., 2023) datasets. TOEFL11 includes 12,100 essays from the 2006-2007 TOEFL exams, divided into 9,900 for training, 1,100 for validation, and 1,100 for testing, with essays categorized into low, medium, or high proficiency by human raters. We assessed performance using weighted F1 and quadratic weighted kappa (QWK). The ELLIPSE dataset contains 6,482 essays from 8th to 12th-grade English learners, with 2,568 reserved for testing. Essays were rated on a 1 to 5 scale (with 0.5 increments), adjusted to the nearest 0.5 for QWK calculations, alongside Root Mean Square Error (RMSE) evaluation.

The counterfactuals were generated on the test set using GPT-4 Turbo and Llama-3-70b-Instruct models. Comparative analysis revealed that both models successfully completed the task, but the GPT-4 Turbo model exhibited more stable performance in the aforementioned measures across both datasets (see Appendix B.3 for detailed comparisons). Consequently, we employed the counterfactual essays generated by the GPT-4 Turbo model for subsequent analyses.

For automated scoring, we fine-tuned BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) on the training set. For LLMs, we utilized GPT-3.5 Turbo, GPT-4 Turbo and Llama 3 instruction-fine-tuned models (8B & 70B) in zero-shot learning (ZSL) and few-shot learning (FSL) scenarios, and performed supervised fine-tuning (SFT) on GPT-3.5 Turbo. Detailed fine-tuning and inference settings are provided in Appendix C.

For each essay T in the test set (indexed as t), we generated its multiple types of counterfactuals by altering its values of different concepts $\{C_i\}$, used a certain model $\mathcal{M}$ to predict scores for both original essay and its counterfactuals, and calculated the effect of any specific counterfactual intervention by subtracting the original score from the correspond-

Metric	Description
WordNum	The number of words in an essay.
SentNum	The number of sentences in an essay.
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.
LemmaTTR	A *lexical diversity* measure based on the type-token ratio (TTR) of an essay, where each word is lemmatized.
LexSoph	A *lexical sophistication* measure based on word frequency statistics from the 1980s-2010s COHA corpus (Davies, 2010). For an essay with N words, let $w_1, w_2, \ldots, w_N$ be the individual words (including repetitions), ℓ_i be the lemma of w_i, and $\mathrm{Freq}(\ell_i)$ be the frequency of ℓ_i in the selected COHA subset. LexSoph is defined as: $$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\log(\mathrm{Freq}(\ell_i)+1)}$$
ErrorDensity	Density of writing errors in an essay with N words, defined as $\#\mathrm{error}/N$. Writing error analyses are implemented using `LanguageTool` (Naber et al., 2003).
CosSim	The cosine similarity between two essay embeddings to measure the content preservation of interventions.

Table 2: The linguistics metrics used for the evaluation of counterfactual samples.

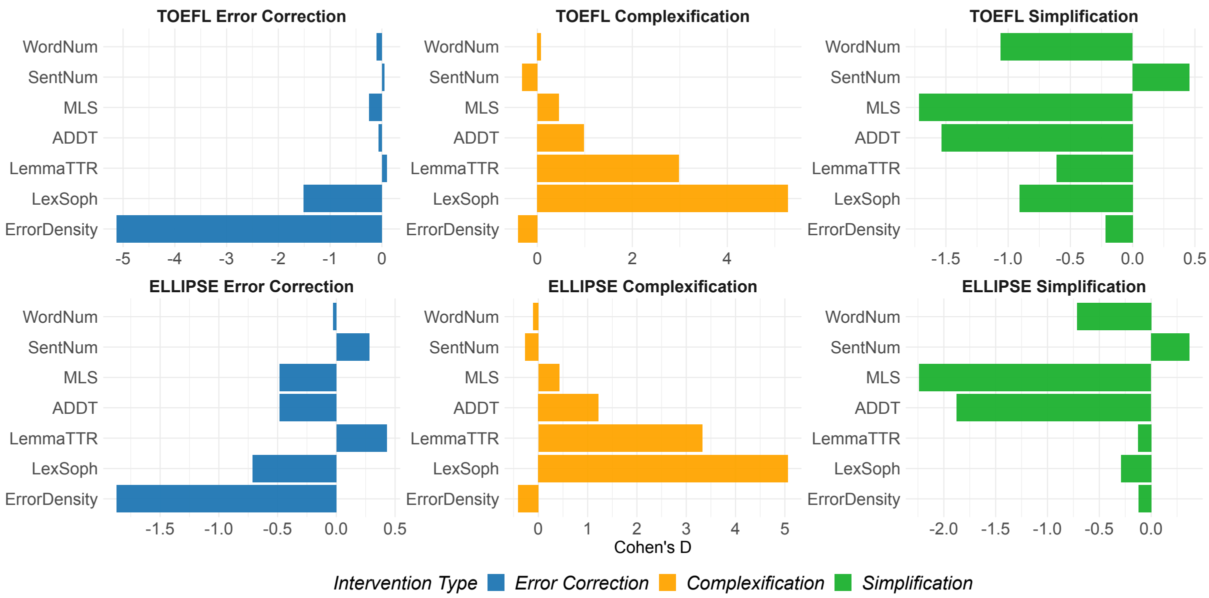

Figure 2: Cohen's $\mathcal{D}$ measured for seven linguistic metrics on three interventions.

ing counterfactual score:

$$\Delta\mathcal{S}_t^{\mathcal{M}}(C_i) = \mathcal{S}_t^{\mathcal{M}}(C_i = v_i') - \mathcal{S}_t^{\mathcal{M}}(C_i = v_i) \quad (3)$$

Since score levels of TOEFL11 essays are labeled as three discrete categories[1]: low, medium and high, we calculated the score delta of BERT-like models after converting the classifier's output probabilities into 1-5 scores for consistency with its rubrics and comparability with ELLIPSE scale:

$$\mathcal{S}_t^{\mathcal{M}} = \sum_{k=1}^{3}(2k-1)\,\mathrm{Pr}^{\mathcal{M}}(\hat{y}_t = k) \quad (4)$$

where $\hat{y}_t$ is the predicted class of essay T.

[1]When few-shot prompting LLMs to rate TOEFL11 essays, score levels of example essays in prompts can only be one of the three categories. Therefore it is not possible to expect numerical predicted scores, which is the reason why in Table 6 we do not show intervention effects of FSL on TOEFL11.

4.2 Counterfactual Validation Results

Figure 2 shows the effect size of three GPT-4-based interventions on seven linguistic metrics across two datasets. Both datasets show similar patterns, with Error Correction notably reducing error density and leaving lexical and syntactic complexity almost untouched. Complexification significantly boosts lexical diversity and sophistication while moderately enhancing syntactic complexity, without substantially changing overall text length and error density. Similarly, Simplification effectively reduces sentence length and complexity while also making corresponding changes to lexical properties..

Table 4 presents the embedding similarities between counterfactuals and original essays. It can be seen that Error Correction almost completely retains the original meaning, and Complexification and Simplification, although inevitably making

Text
Original: In my opinion is better to have a knowledge specialize in one specific subject because this is better to know a thing as well as you can . This is true in all the experiences of the life : refered to the university (the italian university) we can take the example of the of the two years of specialization . \n\n An other example we can find in a great company , infact each people that there are in this have a specific work to do and this bring to an excellent final operation . \n\n A person that are magnifically prepare on one thing will arrive at a sicure result because that `` is your bred " ; we can also observe that the most good professors , scientists , nobels , athlets are all specialize on that they work and do not specialize on many works . \n\n We can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .
Error Correction: In my opinion, it is better to have specialized knowledge in one specific subject because it is better to know one thing as well as you can. This is true in all experiences of life: referring to the university (the Italian university), we can take the example of the two years of specialization. \n\n Another example can be found in a large company, where each person has a specific job to do, and this leads to an excellent final operation. \n\n A person who is magnificently prepared in one area will achieve a sure result because that is 'your bread and butter.' We can also observe that the most distinguished professors, scientists, Nobel laureates, and athletes are all specialized in their respective fields and do not spread themselves too thin across multiple areas. \n\n We can also see that the collaboration of great minds, each specialized in a particular area, is important in many aspects of our lives.
Complexification: In my estimation, possessing profound expertise in a singular domain surpasses a superficial acquaintance with multiple disciplines, as it is advantageous to master one area comprehensively. This principle holds true across various life experiences: consider, for instance, the Italian university system, which underscores the merit of two years dedicated to intensive specialization. Similarly, within the confines of a substantial enterprise, each individual is tasked with a distinct role, culminating in a superlative collective performance. An individual who is exceptionally versed in a discrete field is poised to attain assured success, for such expertise constitutes their primary sustenance. Observation reveals that the most illustrious academics, eminent scientists, Nobel laureates, and elite athletes have all attained distinction by honing their skills within a narrowly defined sphere, eschewing the dilution of their talents over a plethora of subjects. Furthermore, it is evident that the synergy of preeminent intellects, each an authority in a discrete specialization, plays a pivotal role in the advancement of numerous facets of our existence.
Simplification: I think it's best to know a lot about one thing. This is true for all parts of life. For example, at the university in Italy, students focus on one area for two years. In a big company, each worker has their own job. This makes the company work well. A person who knows a lot about one thing will do well. That's because it's what they do all the time. The top teachers, scientists, Nobel winners, and athletes all know a lot about one thing. They don't try to do too many different things. Working together, experts in different things can do a lot of good in our lives.

Table 3: Example of a medium-level TOEFL11 essay and its counterfactual counterparts generated by GPT-4 Turbo.

more changes to the original text, still retain most of the original meaning. For better clarification, Table 3 shows counterfactual examples of a medium level TOEFL essay generated by GPT-4. Examples of rule-based counterfactuals see Appendix B.1.

Intervention	TOEFL11	ELLIPSE
Error Correction	0.935	0.942
Complexification	0.760	0.749
Simplification	0.816	0.849

Table 4: Content preservation for GPT-4-based interventions: text cosine similarities computed by OpenAI `text-embedding-3-large`.

4.3 Scoring Results

Table 5 displays the performance of scoring agreement between models and human on test sets of both datasets. Table 6 shows intervention effects of different types of counterfactual interventions. Based on these results, our findings are as follows:

Firstly, BERT-like models show higher scoring agreement with human raters than LLMs. These models can discern complex concepts (conventions and language complexity). This differs from the phenomenon observed by Singla et al. (2023), where BERT-based models function as a bag-of-words when scoring essays. However, BERT-like models struggle to distinguish interventions on organization and development, showing insensitivity to logical structures and coherence within essays.

Secondly, LLMs respond adequately to all our interventions, suggesting that they align more closely with the criteria outlined in scoring rubrics. It is noteworthy that the scoring agreement between zero-shot LLMs and human raters is relatively low, as these models tend to assign more stringent (lower) scores (see Figure 3 for score distributions). However, introducing FSL and SFT considerably improves their performance while preserving the strength of their rationale alignment, as demonstrated in Table 6. When fine-tuning GPT-3.5, scoring performance improves with an increase in the number of training essays. As shown in Figure 4, With about 400 essays for TOEFL11 and 200 essays for ELLIPSE, its performance nearly stabilizes, achieving performance close to or on par

Setting	TOEFL11		ELLIPSE	
	F1 ↑	QWK ↑	RMSE ↓	QWK ↑
BERT	0.783	0.736	0.437	0.680
RoBERTa	**0.795**	0.739	0.430	0.695
DeBERTa	0.790	**0.741**	**0.422**	**0.720**
GPT-3.5-ZSL	0.599	0.408	0.701	0.399
GPT-3.5-FSL	0.546	0.314	<u>0.570</u>	0.378
GPT-3.5-SFT	*0.710*	*0.592*	*0.550*	*0.629*
GPT-4-ZSL	0.368	0.380	0.960	0.261
GPT-4-FSL	0.490	0.477	0.680	0.466
Llama-3-8B-ZSL	0.558	0.297	0.628	0.345
Llama-3-8B-FSL	0.435	0.441	1.039	0.054
Llama-3-70B-ZSL	0.524	0.390	0.903	0.182
Llama-3-70B-FSL	<u>0.609</u>	<u>0.562</u>	0.589	<u>0.503</u>

Table 5: The scoring agreement performance on both test sets: **best** in bold, *fine-tuned GPT-3.5* in italics, <u>best off-the-shelf LLMs</u> underlined.

Dataset	Setting	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	$0.92^{+.025}_{-.025}$	$-0.80^{+.025}_{-.025}$	$-0.35^{+.021}_{-.021}$	$-0.80^{+.026}_{-.026}$	$0.66^{+.025}_{-.025}$	$-0.24^{+.021}_{-.021}$	$-0.24^{+.018}_{-.017}$	$-0.29^{+.019}_{-.019}$
	Llama-3-8B-ZSL	$0.58^{+.027}_{-.026}$	$-0.37^{+.029}_{-.029}$	$-0.07^{+.018}_{-.018}$	$-0.17^{+.023}_{-.024}$	$0.57^{+.026}_{-.026}$	$-0.11^{+.023}_{-.023}$	$-0.15^{+.024}_{-.024}$	$-0.23^{+.026}_{-.026}$
	Llama-3-70B-ZSL	$0.64^{+.026}_{-.025}$	$-0.56^{+.025}_{-.025}$	$-0.24^{+.021}_{-.022}$	$-0.41^{+.023}_{-.023}$	$1.19^{+.032}_{-.032}$	$-0.17^{+.024}_{-.024}$	$-0.15^{+.019}_{-.019}$	$-0.19^{+.021}_{-.021}$
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-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*	$0.61^{+.052}_{-.048}$	$-0.71^{+.060}_{-.060}$	$-0.27^{+.050}_{-.050}$	$-0.56^{+.048}_{-.050}$	$0.67^{+.055}_{-.052}$	$-0.09^{+.045}_{-.043}$	$-0.14^{+.032}_{-.035}$	$-0.23^{+.042}_{-.045}$
	Llama-3-8B-ZSL	$0.32^{+.017}_{-.016}$	$-0.31^{+.018}_{-.018}$	$-0.06^{+.011}_{-.011}$	$-0.11^{+.013}_{-.014}$	$0.70^{+.013}_{-.013}$	$0.01^{+.009}_{-.010}$	$-0.06^{+.011}_{-.012}$	$-0.10^{+.014}_{-.014}$
	Llama-3-8B-FSL	$0.06^{+.011}_{-.011}$	$-0.11^{+.016}_{-.016}$	$-0.02^{+.008}_{-.008}$	$-0.06^{+.011}_{-.011}$	$0.07^{+.016}_{-.016}$	$-0.00^{+.007}_{-.007}$	$-0.02^{+.010}_{-.010}$	$-0.02^{+.012}_{-.011}$
	Llama-3-70B-ZSL*	$0.51^{+.018}_{-.018}$	$-0.41^{+.011}_{-.011}$	$-0.11^{+.009}_{-.009}$	$-0.19^{+.010}_{-.010}$	$1.63^{+.019}_{-.019}$	$0.03^{+.018}_{-.018}$	$-0.03^{+.007}_{-.007}$	$-0.06^{+.008}_{-.008}$
	Llama-3-70B-FSL*	$0.51^{+.070}_{-.068}$	$-0.54^{+.065}_{-.065}$	$-0.12^{+.033}_{-.035}$	$-0.24^{+.050}_{-.052}$	$1.08^{+.055}_{-.055}$	$-0.04^{+.040}_{-.040}$	$-0.11^{+.040}_{-.042}$	$-0.13^{+.043}_{-.045}$
	GPT-3.5-FSL-50*	$0.83^{+.075}_{-.072}$	$-0.64^{+.077}_{-.080}$	$-0.14^{+.045}_{-.050}$	$-0.34^{+.065}_{-.068}$	$0.96^{+.060}_{-.062}$	$0.08^{+.055}_{-.052}$	$-0.09^{+.045}_{-.045}$	$-0.10^{+.047}_{-.050}$
	GPT-3.5-FSL-100*	$1.12^{+.080}_{-.080}$	$-0.95^{+.080}_{-.080}$	$-0.26^{+.052}_{-.052}$	$-0.58^{+.057}_{-.055}$	$0.88^{+.055}_{-.057}$	$0.05^{+.050}_{-.048}$	$-0.18^{+.047}_{-.050}$	$-0.19^{+.048}_{-.050}$
	GPT-3.5-FSL-200*	$1.03^{+.092}_{-.090}$	$-0.57^{+.087}_{-.090}$	$-0.01^{+.068}_{-.070}$	$-0.32^{+.072}_{-.070}$	$0.79^{+.052}_{-.055}$	$-0.02^{+.037}_{-.037}$	$0.06^{+.060}_{-.060}$	$0.02^{+.062}_{-.062}$
	GPT-3.5-FSL-400*	$1.11^{+.087}_{-.090}$	$-0.95^{+.075}_{-.075}$	$-0.30^{+.060}_{-.060}$	$-0.66^{+.068}_{-.065}$	$0.76^{+.055}_{-.057}$	$-0.03^{+.045}_{-.042}$	$-0.18^{+.052}_{-.052}$	$-0.23^{+.050}_{-.052}$
	GPT-3.5-FSL-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 6: Mean score shifts $\overline{\Delta\mathcal{S}}$ ($\Delta\mathcal{S} = \mathcal{S}_{CF} - \mathcal{S}_{OR}$, where $\mathcal{S}_{OR}, \mathcal{S}_{CF} \in [1, 5]$) 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.

with that of BERT, while consistently maintaining rationale alignment capability.

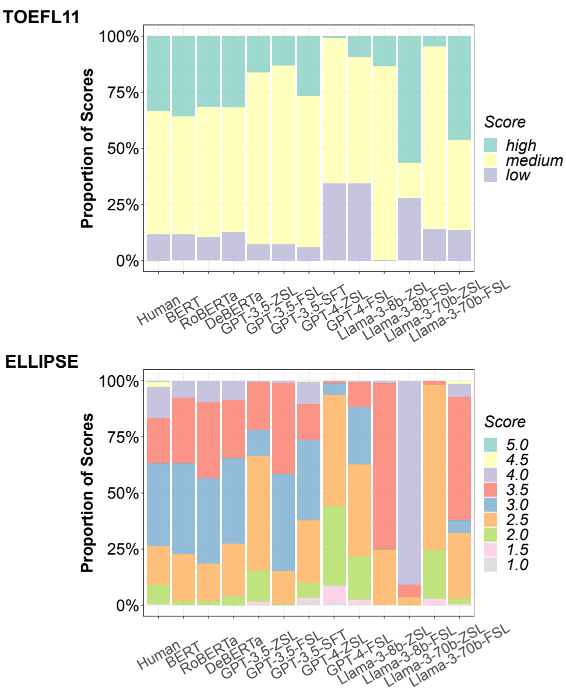

Figure 3: Score Distributions of models' predictions. The fine-tuned BERT-like models exhibit better alignment with human. Most zero-shot LLMs assign scores more stringently with few high scores, while FSL and SFT can mitigate this issue.

4.4 Self-Explanation in Feedback

Han et al. (2023) and Xiao et al. (2024) have proposed that LLMs can provide helpful essay feedback for both writers and evaluators. More importantly, but still largely overlooked, this feedback offers an opportunity to assess the construct validity of models. Therefore we further investigated the feedback differences provided by LLMs regarding the interventions. As Table 6 shows that GPT-4-FSL exhibits comprehensive sensitivity in all types of our interventions, while other models typically show inadequate sensitivity in one way or another, we generated feedback using GPT-4 Turbo for fur-

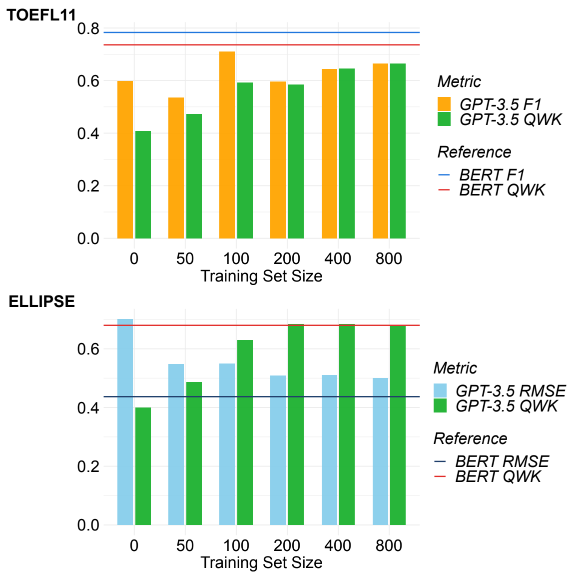

Figure 4: Scoring performance of GPT-3.5 SFT models with varying size of training data. The models' performance improves as the number of training samples increases, reaching comparable or equivalent levels to BERT-like models.

ther analysis of the model's faithfulness.

As shown in Figure 5, after few-shot prompting on scoring task, we continued to ask GPT-4 Turbo to generate feedback based on the scoring rubrics, explaining the scores in terms of aforementioned three concepts. In this way, we obtained feedback for each essay and its counterfactual counterparts[2]. Then, three trained annotators were hired to evaluate the feedback differences within each feedback pair, determining whether counterfactual interventions can be detected without accessing essay content. See the detailed evaluation procedures in Appendix D.

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 of the essay you have just rated (not the example essay):
1. Organization: . . .
2. Language Use: . . .
3. Conventions: . . .
Your response should be a structured JSON object with the following keys:
``` {{
"organization_feedback": "",
"language_use_feedback": "",
"conventions_feedback": ""
}} ```
If possible, include direct citations from the essay to substantiate your feedback.

Figure 5: An Example of Feedback Generation

Table 7 presents the annotator-voted results, demonstrating that a large proportion of counterfactual interventions can be identified simply from the feedback given by the GPT-4 Turbo, especially for complexification, error correction and error introduction except for SVA. On the other hand, simplification and orgnazition interventions are hard to be detected simply from feedback pairs, which is consistent with their relatively smaller absolute effect as shown in Table 6. One possible reason is that the ELLIPSE essays, written by 8th to 12th grade English learners, tend to be simple in vocabulary and syntax, contain some spelling and SVA errors, and exhibit imperfect logic flow and coherence. Consequently, the model frequently identified SVA issues and offered numerous organizational and developmental suggestions both in feedback of original and counterfactual essays, leading to less distinct differences.

[2]We conducted stratified sampling on the ELLIPSE dataset to obtain 200 essay samples and, through two rounds of dialogues, acquired 200 "original-counterfactual" feedback pairs for human evaluation. For the evaluation process, we categorized these pairs based on eight counterfactual interventions and assessed each category of cases accordingly.

| 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 7: Voting-Based Detection Rates of Original vs. Counterfactual Feedback.

## 5 Conclusion

We generated linguistically-informed counterfactuals with an integrated approach combining LLM and rule-based methods, analyzing their impact on essay scoring results of BERT-like models and LLMs. Our findings emphasize that a higher scoring agreement with human raters does not necessarily indicate a better alignment with scoring rubrics, suggesting that a more holistic evaluation approach should consider both aspects. Moreover, our study highlights LLMs' considerable potential in AES domain: while zero-shot prompted LLMs show lower scoring agreement compared to BERT-like models, a major reason for this is that them tend to be conservative or strict when evaluating the essay. FSL and SFT could significantly increase the agreement level with annotated essays serve as anchors to neutralize the conservatism. In the mean time, LLMs demonstrate comprehensive rationale alignment with scoring rubrics. This ability is stably maintained in ZSL, FSL and SFT settings. Lastly, LLMs are not only sensitive to counterfactual interventions when scoring but can also reflect a large part of these differences in their feedback, an advantage beyond the reach of traditional AES methods.

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.
```

6 Limitations

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

Isaac I Bejar, Michael Flor, Yoko Futagi, and Chaintanya Ramineni. 2014. On the vulnerability of automated scoring to construct-irrelevant response strategies (cirs): An illustration. *Assessing Writing*, 22:48–59.

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.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. Disco: distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528.

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

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Mark Davies. 2010. *The corpus of historical American English: COHA*. BYE, Brigham Young University.

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

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Core: A retrieve-then-edit framework for counterfactual data generation. *arXiv preprint arXiv:2210.04873*.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

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.

Mohamed A. Hussein, Hesham A., 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).

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.

Alwyn Vwen Yen Lee, Andrés Carlos Luco, and Seng Chee Tan. 2023. A human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educational Technology & Society*, 26(1):147–159.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13201–13221.

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.

Matthew C Myers and Joshua Wilson. 2023. Evaluating the construct validity of an automated writing evaluation system with a randomization algorithm. *International Journal of Artificial Intelligence in Education*, 33(3):609–634.

Daniel Naber et al. 2003. *A rule-based style and grammar checker*. GRIN Verlag Munich, Germnay.

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

Christopher M. Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models.

Donald E Powers, Jill C Burstein, Martin Chodorow, Mary E Fowles, and Karen Kukich. 2002. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134.

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

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring.

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. 2023. Automatic essay scoring systems are both overstable and oversensitive: explaining why and proposing defenses. *Dialogue & Discourse*, 14(1):1–33.

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*.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation.

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.

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.

Appendix

A Rubrics

To identify the core concepts for intervention, we reviewed five scoring rubrics from IELTS Writing, TOEFL iBT Independent Writing, TOEIC Writing, PTE Academic Writing and the ELLIPSE dataset. We aimed to uncover commonalities across the five rubrics and found that they could be categorized into three dimensions: (1) conventions and accuracy; (2) language complexity; and (3) organization and development. For clarity, this section will present the descriptors for the highest score in each rubric, with **color-coded highlights** to indicate the corresponding dimensions.

A.1 IELTS Writing

- Task achievement: fully satisfies all the requirements of the task; clearly presents a fully developed response.
- Coherence and cohesion: uses cohesion in such a way that it attracts no attention; skillfully manages paragraphing.
- Lexical resource: uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'.
- Grammatical range and accuracy: uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'.

A.2 TOFEL Independent Writing

- Effectively addresses the topic and task.
- Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details.
- Displays unity, progression and coherence.
- Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors.

A.3 TOEIC Writing

- Typically, test takers at level 9 can communicate straightforward information effectively and use reasons, examples, or explanations to support an opinion.
- When using reasons, examples, or explanations to support an opinion, their writing is well-organized and well-developed.
- The use of English is natural, with a variety of sentence structures and appropriate word choice, and is grammatically accurate.
- When giving straightforward information, asking questions, giving instructions, or making requests, their writing is clear, coherent, and effective.

A.4 PTE Academic Writing

- Content: Adequately deals with the prompt.
- Form: Length is between 200 and 300 words.
- Development, Structure & Coherence: Shows good development and logical structure.
- Grammar: Shows consistent grammatical control of complex language. Errors are rare and difficult to spot.
- General Linguistic Range: Exhibits mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No sign that the test taker is restricted in what they want to communicate.
- Vocabulary: Good command of a broad lexical repertoire, idiomatic expressions and colloquialisms.
- Spelling: Correct spelling.

A.5 ELL Dataset

- Overall: 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.
- Cohesion: Text organization consistently well controlled using a variety of effective linguistic features such as reference and transitional words and phrases to connect ideas across sentences and paragraphs; appropriate overlap of ideas.

- Syntax: Flexible and effective use of a full range of syntactic structures including simple, compound, and complex sentences; There may be rare minor and negligible errors in sentence formation.

- Vocabulary: Wide range of vocabulary flexibly and effectively used to convey precise meanings; skillful use of topic-related terms and less common words; rare negligible inaccuracies in word use.

- Phraseology: Flexible and effective use of a variety of phrases, such as idioms, collocations, and lexical bundles, to convey precise and subtle meanings; rare minor inaccuracies that are negligible.

- Grammar: Command of grammar and usage with few or no errors.

- Conventions: Consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors nonexistent or negligible.

B Detail of Counterfactual Generation

In this section, we present the details of our counterfactual generation experiment. This includes examples of rule-based counterfactuals, information on the models used, the prompts provided to the LLMs, and a comparative analysis of various aspects of interest in the counterfactuals generated by GPT-4 Turbo and Llama-3-70b-Instruct.

B.1 Examples of Rule-based Counterfactuals

In this study, all the interventions designed to introduce errors into essays and decrease organization are rule-based. In this subsection, we provide counterfactual examples for each of the rule-based interventions. See Table 8.

B.2 Prompts for Counterfactual Generation

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.

B.2.1 Prompt for Error Correction

System: You are an experienced writing tutor.

User: 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:

B.2.2 Prompt for Complexification

System: You are an experienced writing tutor.

User: 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:

Text
Original: In my opinion is better to have a knowledge specialize in one specific subject because this is better to know a thing as well as you can . This is true in all the experiences of the life : refered to the university (the italian university) we can take the example of the of the two years of specialization . \n\n An other example we can find in a great company , infact each people that there are in this have a specific work to do and this bring to an excellent final operation . \n\n A person that are magnifically prepare on one thing will arrive at a sicure result because that `` is your bred " ; we can also observe that the most good professors , scientists , nobels , athlets are all specialize on that they work and do not specialize on many works . \n\n We can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .
Spelling Error Introduction: In m'i opion is better to have ein knowleges specialize in one spesific subject becueas thes is bitter to know a thing at well as you can.\n This is true in all the experiences of the life : refered to the university (the italian university) we can take the example of the of the two years of specialization .\n\n And other example we can find in and great compa0̆0f1y, infact each pepoles that there are in this have g specific work to di and this brening to en excelant final operassion.\n\n I person that are magnifically prapar on one thing wold arrive at a sicure result bBecause that ' ' is ur bred ' '; we can olso observe taht the mosts good professors, scientists, nobels, athlets are all specialize one that they work and do not specialize on mani works.\n\n We can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .
Subject-verb Agreement Error Introduction: In my opinion is better to have a knowledge specialize in one specific subject because this is better to know a thing as well as you can .\n This is true in all the experiences of the life : refered to the university (the italian university) we can take the example of the of the two years of specialization .\n\n An other example we can find in a great company , infact each people that there is in this have a specific work to do and this bring to an excellent final operation .\n\n A person that are magnifically prepare on one thing will arrive at a sicure result because that " is your bred " ; we can also observe that the most good professors , scientists , nobels , athletes are all specialize on that they work and do not specialize on many works .\n\n We can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .
Word Order Swapping: My in opinion is better to have knowledge a in specialize specific one because subject this better is know to a thing well as as can you.\n This is true in all the experiences of the life : refered to the university (the italian university) we can take the example of the of the two years of specialization .\n\n Other an example we can find in a company, great infact each people there that are in this have a specific work to this do to and bring an excellent final operation.\n\n A person that are magnifically prepare on one thing will arrive at a sicure result because that " is your bred " ; we can also observe that the most good professors , scientists , nobels , athlets are all specialize on that they work and do not specialize on many works .\n\n We can also saw that the of colloboration great brains, each of them specialized a on thing, important is in many ways the of our. life
Intra-paragraph Shuffling: This is true in all the experiences of the life : refered to the university (the italian university) In my opinion is better to have a knowledge specialize in one specific subject because this is better to know a thing as well as you can .\n we can take the example of the of the two years of specialization .\n\nAn other example we can find in a great company , infact each people that there are in this have a specific work to do and this bring to an excellent final operation .\n\nA person that are magnifically prepare on one thing will arrive at a sicure result because that " is your bred " ; we can also observe that the most good professors , scientists , nobels , athlets are all specialize on that they work and do not specialize on many works .\n\nWe can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .
Inter-text Shuffling: In my opinion is better to have a knowledge specialize in one specific subject because this is better to know a thing as well as you can .\nThis is true in all the experiences of the life : refered to the university (the italian university)we can take the example of the of the two years of specialization .\n\nA person that are magnifically prepare on one thing will arrive at a sicure result because that " is your bred " ; we can also observe that the most good professors , scientists , nobels , athlets are all specialize on that they work and do not specialize on many works .\n\nAn other example we can find in a great company , infact each people that there are in this have a specific work to do and this bring to an excellent final operation .\n\nWe can also saw that the colloboration of great brains , each of them specialized on a thing , is important in many ways of the our life .

Table 8: Example of the medium-level TOEFL11 essay used in Table 3 and its rule-based counterfactual counterparts.

```
{"output_essay": "..."}
```
Output:

B.2.3 Prompt for Simplification

System: You are an experienced writing tutor.

User: 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:

B.3 Comparative Performance of Model A and Model B in Counterfactual Generation

Table 9 shows the effect size of three types of interventions performed by both GPT-4 Turbo and Llama-3-70b-Instruct on seven linguistic metrics across two datasets. It can be seen that the impact of the two models on the original essay, across various language metrics of interest during counterfactual interventions, aligns with expectations, albeit with slight variations in degree. In terms of error correction, GPT-4 significantly reduces error density. Meanwhile, for complexification and simplification, GPT-4 intervenes more comprehensively in vocabulary and syntax, with generally smaller changes in length.

Table 4 presents the embedding similarities bewteen counterfactuals and original essays given by both LLMs. Although Llama-3-70b-Instruct retains a higher degree of the original text's meaning than GPT-4 Turbo in most cases, it shows a significant drop when simplifying the ELLIPSE essay, indicating its potential lack of stability.

C The Implementation of AES methods

C.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`.

C.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.

C.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 identify the best learning rate that led to the lowest loss on the validation set (results see Table 11). 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.

C.2 Prompting LLMs to Score Essays

As introduced in Section 4, 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.

C.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}"

Metrics	Model	TOEFL11			ELLIPSE		
		Error Correction	**Complexification**	**Simplification**	**Error Correction**	**Complexification**	**Simplification**
`WordNum`	LLAMA-3-70B-IT	-0.045	-0.170	-1.325	0.065	-0.332	-0.981
	GPT-4 TURBO	-0.098	0.078	-1.060	-0.027	-0.103	-0.714
`SentNum`	LLAMA-3-70B-IT	0.037	-0.508	-0.037	0.215	-0.406	-0.074
	GPT-4 TURBO	0.047	-0.323	0.454	0.280	-0.264	0.367
`MLS`	LLAMA-3-70B-IT	-0.176	0.385	-1.473	-0.354	0.156	-1.816
	GPT-4 TURBO	-0.245	0.449	-1.714	-0.481	0.423	-2.237
`ADDT`	LLAMA-3-70B-IT	-0.030	0.734	-1.359	-0.353	0.817	-1.628
	GPT-4 TURBO	-0.066	0.982	-1.535	-0.481	1.220	-1.875
`LemmaTTR`	LLAMA-3-70B-IT	-0.074	2.130	-0.467	0.020	2.647	-0.009
	GPT-4 TURBO	0.094	2.985	-0.611	0.429	3.323	-0.128
`LexSoph`	LLAMA-3-70B-IT	-1.538	3.596	-0.186	-0.799	3.710	0.301
	GPT-4 TURBO	-1.514	5.277	-0.909	-0.711	5.063	-0.291
`ErrorDensity`	LLAMA-3-70B-IT	-5.015	-0.616	-0.535	-1.887	-0.628	-0.331
	GPT-4 TURBO	-5.122	-0.407	-0.219	-1.869	-0.412	-0.123

Table 9: Cohen's $\mathcal{D}$ for seven linguistic metrics on three interventions of GPT-4 Turbo and Llama-3-70b-Instruct.

Model	Intervention	TOEFL11	ELLIPSE
GPT-4 TURBO	Error Correction	0.935	0.942
	Complexification	0.760	0.749
	Simplification	0.816	0.849
LLAMA-3-70B-IT	Error Correction	0.944	0.957
	Complexification	0.817	0.813
	Simplification	0.853	0.610

Table 10: Mean cosine similarity between original and counterfactual essays for GPT-4 and Llama-3-70b-instruct given by `text-embedding-3-large`.

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.

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

Table 11: 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**.

- 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.

C.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 conventions; 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.

C.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.

C.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.

C.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 Section 3.3 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.3 Fine-tuning GPT-3.5 Turbo

We fine-tuned GPT-3.5 Turbo model using the OpenAI API[3] with the following hyperparameters: 3

[3] https://platform.openai.com/docs/guides/fine-tuning

epochs, a batch size of 1, and a learning rate multiplier of 2. These are the default settings provided by OpenAI, as the size and weight of GPT-3.5 Turbo model are not accessible, a systematic parameter search would be very costly and even impossible.

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

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 12. Table 13 gives an example of a feedback pair where the counterfactual feedback corresponds to a sample obtained by introducing spelling errors to the original sample.

D.3 Ethical Considerations

The three annotators involved in this project were graduate students in linguistics. Prior to assigning them the annotation task, we provided a comprehensive introduction to the content, purpose, and significance of the project. Each annotator was responsible for reviewing 200 feedback pairs and received compensation of $0.42 per annotated pair.

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 12: 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 13: Example feedback pair of original and counterfactual obtained by introducing spelling errors.