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

Yupei Wang, Renfen Hu, Zhe Zhao*

Introduction

Automated Essay Assessment is the use of AI to evaluate essays, including both scoring and providing feedback.

- Fine-tuned pretrained encoders like BERT can **score essays** consistently with human raters.
- Generative models like GPT-4 can **score essays and give feedback** in natural language.
- The **criteria** used by these models to assess essays are **not explicitly known**.

Motivation

In high-stakes exams, AI must reliably score and give feedback.

- No way to measure AI's adherence to human scoring rubrics.
- No way to verify consistency between AI feedback and scoring.

Datasets

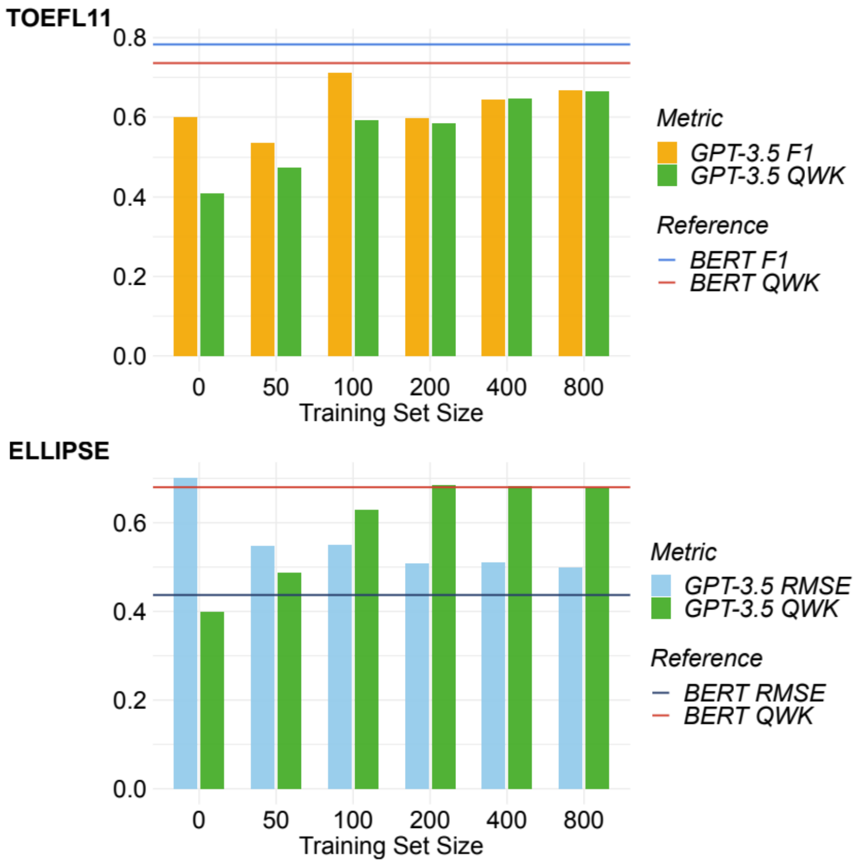
	TOEFL11		**ELLIPSE**	
Total Size	12,100 essays		6,482 essays	
Data Split	Train	9,900	Train Val	3,914
	Val	1,100		
	Test	1,100	Test	2,568
Source	2006-2007 TOEFL exams		8th-12th grade English learners	
Rating Scale	Low/Medium/High		1-5 scale (0.5 increments)	
Eval Metrics	Weighted F1, QWK		RMSE, QWK	

Comparison of TOEFL11 and ELLIPSE datasets

Scoring Performance

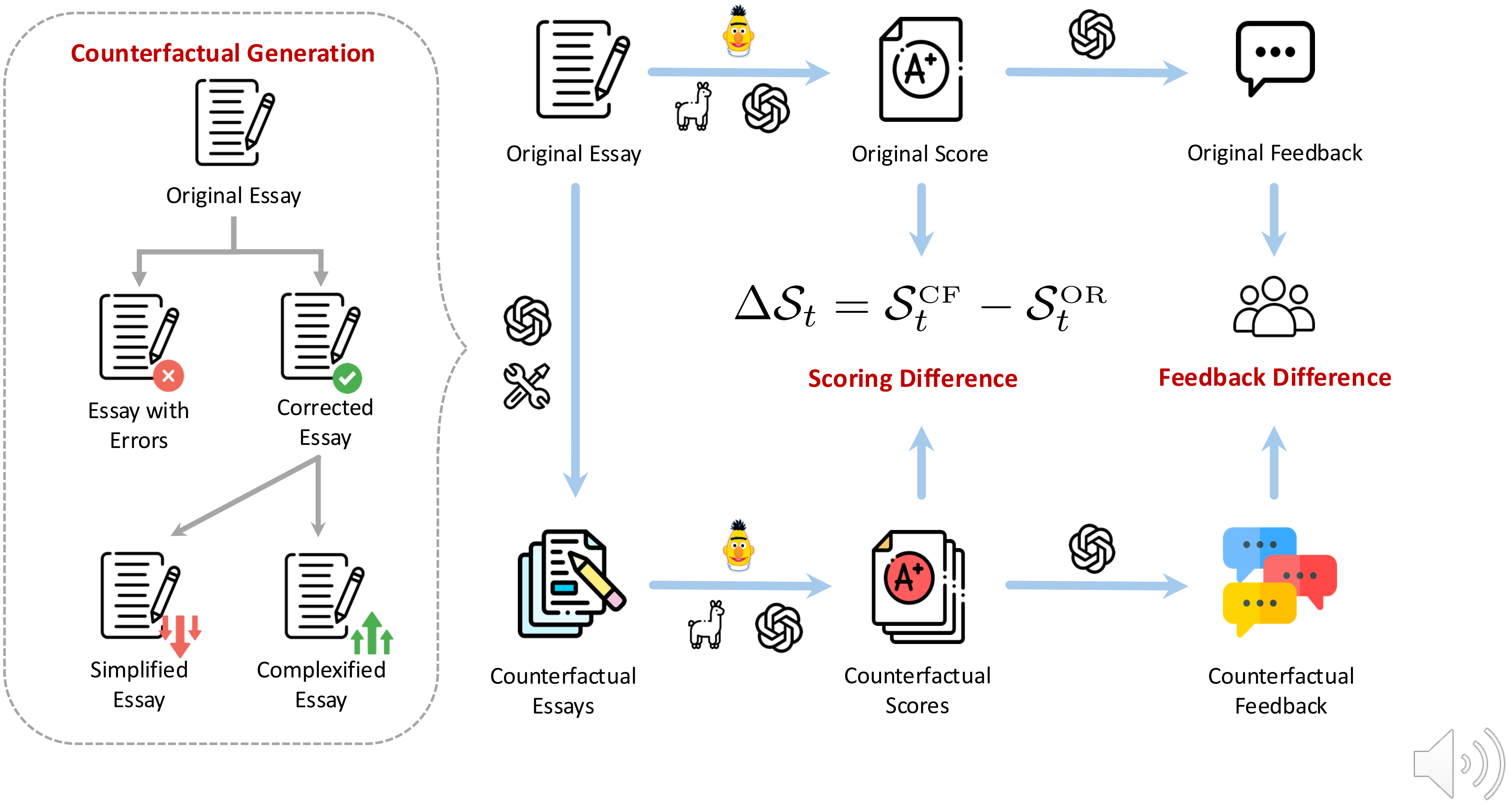
Setting	TOEFL11		ELLIPSE	
	F1 ↑	QWK ↑	RMSE ↓	QWK ↑
BERT	0.783	0.736	0.437	0.680
RoBERTa	**0.795**	0.739	0.430	0.695
DeBERTa	0.790	**0.741**	**0.422**	**0.720**
GPT-3.5-ZSL	0.599	0.408	0.701	0.399
GPT-3.5-FSL	0.546	0.314	0.570	0.378
GPT-3.5-SFT-100	0.710	0.592	0.550	0.629
GPT-4-ZSL	0.368	0.380	0.960	0.261
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	0.609	0.562	0.589	0.503

The scoring agreement performance on both test sets: **best setting** in bold, fine-tuned GPT-3.5 with a green shadow, best off-the-shelf LLMs underlined. Metrics with ↑ indicate that higher values are better, while the one with ↓ indicates that lower values are better.



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.

Note that in our scoring experiments, the model is required to output scores directly without any CoT style reasoning.


Counterfactual Generation
Original Essay
Essay with Errors
Corrected Essay
Simplified Essay
Complexified Essay
Original Essay
Original Score
Original Feedback
$\Delta S_t = S_t^{\mathrm{CF}} - S_t^{\mathrm{OR}}$
Scoring Difference
Feedback Difference
Counterfactual Essays
Counterfactual Scores
Counterfactual Feedback

Example of Counterfactual –Word Order Swapping

Some people think that , **it would be better** to have broad **knowledge of different academic subjects because variations of subjects adds to** human experience and provides possible solutions in many situations .\nOthers think it would be better to concentrate **on one topic and go on** .\nBased on my experience , both opinion are helpful according to my grade .\n\nWhen i was in under graduate school , I was fond of reading and attending different classes which helped me to great extend to pass my exams easily and receive high score .\n\n**On the other hand** , when I was in physical therapy school **I concentrated on only sports medicine classes** which **helped me to apply for master degree** and find convineant job .\n\nFinally both thoughts are important depend on needs and goals .

Some people think that, **would it be better** to have broad of **knowledge different subjects because academic variations of adds subjects** human to experience and provides possible in solutions situations many.\n Think it others be better would to concentrate **on one topic go and on**.\n Based on my experience , both opinion are helpful according to my grade .\n\n When i was in under graduate school , I was fond of reading and attending different classes which helped me to great extend to pass my exams easily and receive high score .\n\n **On other the hand**, when I was in physical therapy school **concentrated I on only sports medicine classes** which **helped me apply to for degree master** and find convineant job.\n\n Finally both thoughts are important depend on needs and goals .

Example of Counterfactual – Error Correction

Some people think **that , it** would be better to have broad knowledge of different academic subjects because variations of subjects **adds** to human experience and provides possible solutions in many situations .\nOthers think it would be better to concentrate on one topic and go on .\nBased on my experience , **both opinion** are helpful according to my grade .\n\nWhen **i was in under graduate school** , I was fond of reading and attending different classes which helped me **to great extend** to pass my exams easily and receive **high score** .\n\nOn the other hand , when I was in physical therapy school I concentrated on only sports medicine classes which helped me to apply for **master degree** and find **convineant** job .\n\nFinally both thoughts are important **depend on** needs and goals .

Some people think **that it** would be better to have a broad knowledge of different academic subjects because variations of subjects **add** to human experience and provide possible solutions in many situations. Others think it would be better to concentrate on one topic and go deep. Based on my experience, **both opinions** are helpful according to my grade.\n\nWhen **I was an undergraduate**, I was fond of reading and attending different classes, which helped me **to a great extent** to pass my exams easily and receive **high scores**.\n\nOn the other hand, when I was in physical therapy school, I concentrated only on sports medicine classes, which helped me to apply for **a master's degree** and find a **convenient** job.\n\nFinally, both approaches are important **depending on** needs and goals.

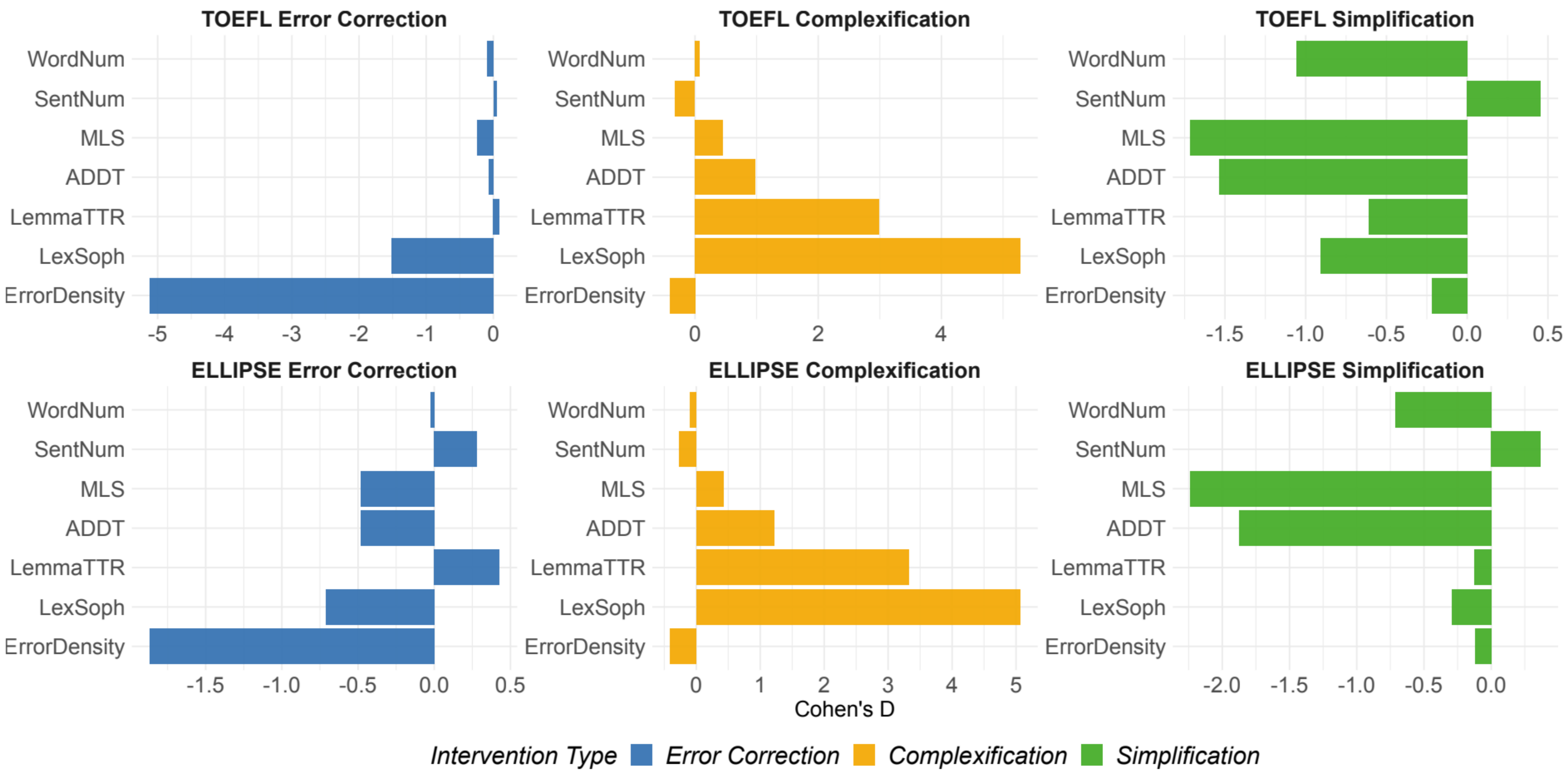
Metric	Description
WordNum	The number of words in an essay.
SentNum	The number of sentences in an essay.
MLS	Mean length of sentences. The length of each sentence is the number of words it has.
ADDT	Average depth of dependency tree for all sentences in an essay.
LemmaTTR	A *lexical diversity* measure 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).

The linguistics metrics used for the evaluation of counterfactual samples.

The Validity of LLM Generated Counterfactuals

We compute Cohen's D to measure the effect size for each linguistic feature:

$$\mathcal{D} = \frac{\bar{x}_{CF} - \bar{x}_{OR}}{s} \qquad s = \sqrt{\frac{\left(n_{OR} - 1\right) s_{OR}^2 + \left(n_{CF} - 1\right) s_{CF}^2}{n_{OR} + n_{CF} - 2}}$$



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

The Validity of LLM Generated Counterfactuals

We assess content preservation by calculating the average cosine similarity between text embeddings of original-counterfactual essay pairs.

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

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

Dataset	Setting	Conventions & Accuracy				Language Complexity		Organization & Development	
		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-SFT-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-SFT-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-SFT-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-SFT-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-SFT-800*	$1.02^{+.085}_{-.085}$	$-0.83^{+.080}_{-.080}$	$-0.23^{+.065}_{-.067}$	$-0.55^{+.070}_{-.070}$	$0.94^{+.055}_{-.055}$	$-0.03^{+.048}_{-.050}$	$-0.14^{+.052}_{-.055}$	$-0.23^{+.060}_{-.062}$

Mean score differences between original and counterfactual groups, with scores ranging from 1 to 5. Results are shown for both full and stratified subsets (stratified results are marked with *). Subscripts and superscripts indicate confidence intervals, obtained through 10,000 bootstrap iterations. Gray values indicate non-significant differences ($p > 0.01$), while coral values represent significant differences contrary to the expected intervention trend. (+) and (−) denote the expected direction of intervention effect.

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.

An Example of Feedback Generation

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

Voting-Based Detection Rates of Original vs. Counterfactual Feedback.
```

Conclusion

- BERT-like models excel at technical language aspects (complexity and conventions) but struggle with higher-level essay elements like **organization and coherence**, while LLMs demonstrate **more comprehensive** response to all aspects of essay evaluation.
- LLMs show **different sensitivity patterns** between scoring and providing feedback for counterfactual interventions, especially in organization and coherence aspects. Some insensitivity in conventions and language complexity exists but may be attributed to ELLIPSE **dataset characteristics**.

Thanks!

View on GitHub