Supplementary Material: Limits of LLM Text Detectors for Measuring Human Contribution in Education

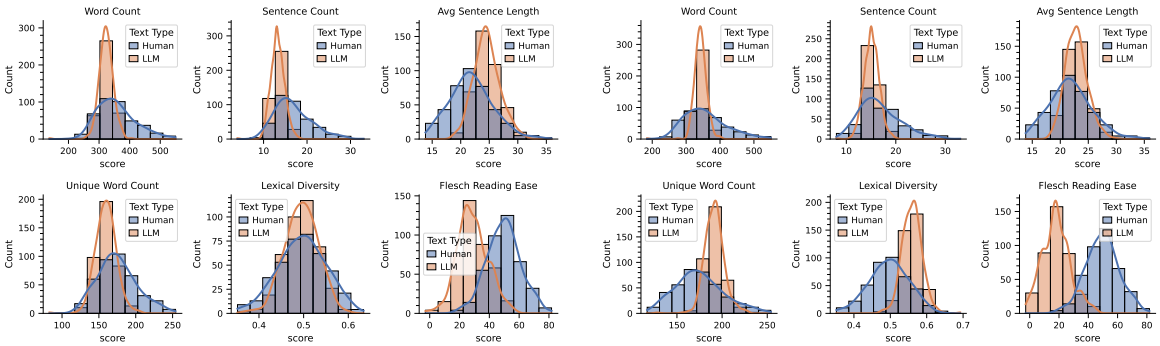
Anonymous authors

1 Dataset Statistics

This section provides supplementary statistics of the corpora used in our experiments. The presented analyses aim to give a more detailed overview of the corpus characteristics underlying the GEDE dataset.

1.1 Text Corpus Statistics

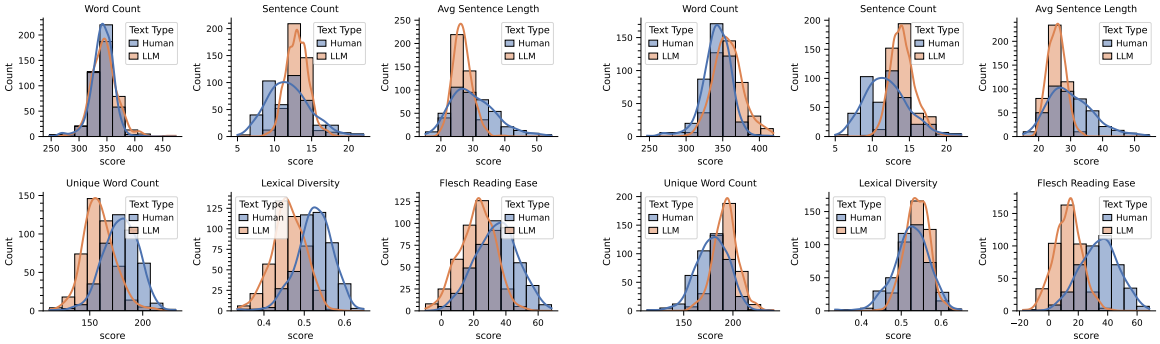
In the following, we report detailed statistics for the AAE [4, 3] (Fig. 1), BAWE [2] (Fig. 2), and PERSUADE [1] (Fig. 3) corpora. For each corpus, we compare essays from the *Task* contribution level with those from the *Human* contribution level.



(a) Llama-3.3-70b-Instruct

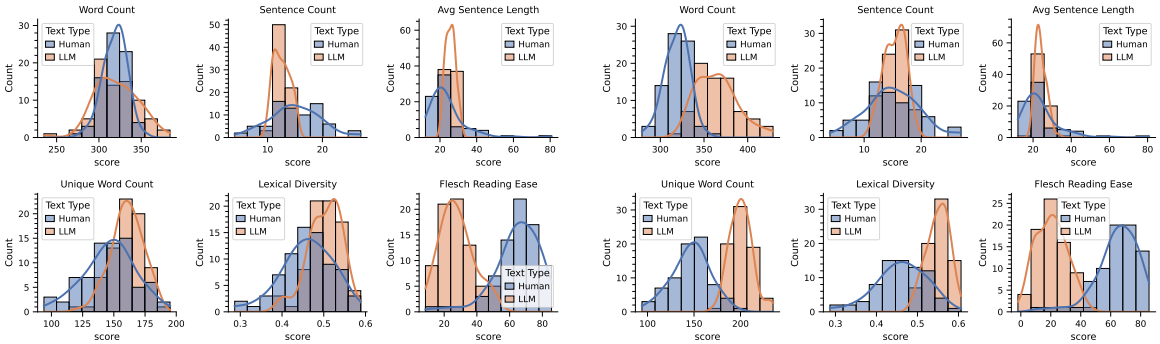
(b) GPT-4o-mini

Figure 1: AAE *Task* vs *Human* statistics



(a) Llama-3.3-70b-Instruct (b) GPT-4o-mini

Figure 2: BAWE *Task* vs *Human* statistics



(a) Llama-3.3-70b-Instruct (b) GPT-4o-mini

Figure 3: PERSUADE *Task* vs *Human* statistics

1.2 BAWE Disciplines

This subsection provides supplementary information on the disciplinary groups included in the BAWE corpus. Table 1 summarizes the disciplinary groups, their relative frequencies, and the mean ROC-AUC scores across all detectors.

Table 1: Discipline ROC-AUC comparison

Discipline Group	Group Mean ± Std	Group Count	Discipline	ROC-AUC (Mean ± Std)	Discipline Count
			Other	0.845 ± 0.08	22
			History	0.840 ± 0.09	25
			Comparative American Studies	0.839 ± 0.10	25
			Archaeology	0.829 ± 0.13	25
AH	0.812 ± 0.11	999	English	0.823 ± 0.10	25
			Classics	0.803 ± 0.11	25
			Linguistics	0.794 ± 0.14	24
			OTHER	0.786 ± 0.17	1
			Philosophy	0.753 ± 0.15	25
			Medicine	0.868 ± 0.10	4
			Biological Sciences	0.862 ± 0.08	4
			Agriculture	0.823 ± 0.18	11
LS	0.823 ± 0.14	999	Psychology	0.805 ± 0.17	25
			Health	0.802 ± 0.15	10
			Food Sciences	0.777 ± 0.15	3
			Chemistry	0.910 ± 0.06	1
			Computer Science	0.880 ± 0.09	3
PS	0.830 ± 0.13	999	Architecture	0.816 ± 0.14	3
			Planning	0.813 ± 0.13	6
			Physics	0.731 ± 0.16	3
			Publishing	0.859 ± 0.08	2
			HLTM	0.857 ± 0.11	17
			Anthropology	0.828 ± 0.15	22
			Business	0.827 ± 0.16	25
SS	0.828 ± 0.12	999	Sociology	0.826 ± 0.14	25
			Law	0.819 ± 0.10	25
			Other	0.815 ± 0.15	3
			Economics	0.814 ± 0.13	25
			Politics	0.805 ± 0.14	25

2 Contribution Level Prompts

This section documents the system and user prompts used to generate the LLM-based contribution levels in the GEDE dataset. Table 2 lists all prompts employed during text generation. The *Human* and *Humanize* levels are included for completeness, although they are not created using system or user prompts.

Table 2: System and User Prompt for each contribution level.

Contribution Level	**System Prompt**	**User Prompt**
Human	-	-
Improve-Human	Your task is to improve a given text. Structure and content of the text should be retained and you should only make small improvements to the grammar and language.	Rewrite this text: {`human-text`}
Rewrite-Human	-	Rewrite this text: {`human-text`}
Summary	You are a student writing an essay on a given topic. Write around 300 words.	Write an essay with the following information: {`summary`}
Summary+ Task	You are a student writing an essay on a given topic. Write around 300 words.	Write an essay for this task: {`task`} Please include the following information: {`summary`}
Task	You are a student writing an essay on a given topic. Write around 300 words.	Write an essay for this task: {`task`}
Rewrite-LLM	-	Rewrite this text: {`llm-text`}
Humanize	-	-

3 Changes to Human-Written Text

This section provides qualitative examples illustrating how large language models modify human-written texts at different contribution levels.

Figure 4a shows an excerpt from a randomly selected human-written text from the BAWE corpus that was improved by GPT-4o-mini for the *Improve-Human* contribution level. Text passages highlighted from orange to blue indicate modifications of existing content, while red denotes removed words and green newly added ones. Although the model introduces several changes, the overall meaning of the text is preserved. In comparison, the improvements produced by LLaMA-3.3 in Figure 4b involve fewer edits to the original text. Nevertheless, both models exhibit similar editing patterns, such as replacing *changes* with *shift* or simplifying expressions like *in order to* to *to*.

In addition, we compute the cosine similarity between the original human-written texts and their improved or rewritten counterparts from the *Improve-Human* and *Rewrite-Human* contribution levels. As shown in Figure 5, texts from the *Improve-Human* category are more similar to the original texts than those from the *Rewrite-Human* category.

The present study aims to discuss the implications for extension practice **of changes → resulting from a shift** in thinking **about extension** from 'transfer of technology' to 'communication to support **innovation'. In order to discuss → innovation.' To explore** this topic, **this → the** study will **present → provide** an overview of what innovation **is, what → entails, the focus of technology transfer, and the reasons for transitioning from a technology** transfer **of technology focus, and the reasons why there is a need to shift from transfer of technology → model** to an innovative approach **in the extension practices context. In → within** the context of **extension practices. For** this study, rural extension is defined as assistance **provided** to individuals living in rural areas **(farmers → (whether farmers** or not) **in helping → to help** them **to** identify and **analyse → analyze** their **problems, and → problems while also** being aware of opportunities for improvement (see Adams, 1982). **Extension was usually → Traditionally, extension has been** associated with increasing food production and **encouraging → promoting** economic development, **having the function of promoting → functioning as a means of disseminating** knowledge and transferring technology between farmers and researchers (or **among** farmers **to farmers) → themselves)** (Leeuwis, 2004:17). ...

(a) Visualization of the changes made by GPT-4o-mini.

The present study aims to **discuss → explore** the implications for extension practice of **changes → the shift** in thinking about extension from 'transfer of technology' to 'communication to support innovation'. **In order to → To** discuss this topic, this study will **present → provide** an overview of what innovation **is, what → entails, the focus of** transfer of **technology focus, → technology,** and the reasons why **there is a need to → a** shift from transfer of technology to an innovative approach **in the extension practices context. → is necessary in the context of extension practices.** In **the context of** this study, rural extension is defined as assistance to individuals living in rural areas **(farmers → (whether farmers** or not) **in helping them to identify and analyse → identifying and analyzing** their problems, and **being → becoming** aware of opportunities for improvement (see Adams, 1982). **Extension → Traditionally, extension** was **usually** associated with increasing food production and **encouraging → promoting** economic development, **having the → with the primary** function of **promoting → disseminating** knowledge and transferring technology between farmers and researchers (or **among** farmers **to farmers) → themselves)** (Leeuwis, 2004:17). ...

(b) Visualization of the changes made by Llama-3.3-70b-Instruct.

Figure 4: Visualization of the changes made by both LLMs to a human-written text from the BAWE subset, belonging to the *Improve-Human* contribution level.

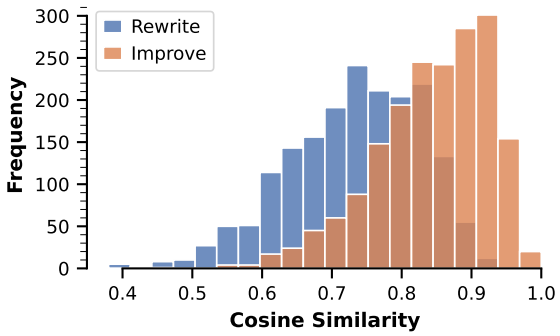


Figure 5: Cosine Similarity between Improve-Human and Rewrite-Human texts.

4 Large Language Model Configurations

Table 3 summarizes the configurations of the language models used to generate the essays in the GEDE dataset. It provides an overview of the model versions and the most relevant inference parameters to ensure transparency and reproducibility.

Table 3: Overview of LLM model configurations

Parameter	**GPT-4o-mini**	**Llama-3.3-70B**
Model version	gpt-4o-mini-2024-07-18	Llama-3.3-70b-Instruct
Provider / Source	OpenAI API	Hugging Face
Quantization	–	4-bit
Temperature	1.0	1.0
Max new tokens	512	512

5 RoBERTa Fine-Tuning Details

We fine-tune the RoBERTa model using the HuggingFace *roberta-base* model[1] on each of the three text corpora. Table 4 shows the hyperparameters we used for fine-tuning RoBERTa on the different subsets. We chose the best epoch based on the evaluation cross-entropy loss. All experiments were conducted with a fixed random seed of 42 to ensure reproducibility.

Table 4: Hyperparameters of RoBERTa fine-tuning on the different text corpora subsets.

Subset	Batch Size	Test Size	Epochs	Best Epoch
AAE	32	0.2	5	2
BAWE	32	0.2	5	4
PERSUADE	32	0.2	8	4

6 Supplementary ROC Curves

This section provides additional ROC curves corresponding to the experiments described in Sections 4.2–4.5 of the main paper.

6.1 Contribution Level

The ROC curves shown in Figure 6 correspond to the contribution-level experiment described in Section 4.2.

[1] https://huggingface.co/FacebookAI/roberta-base

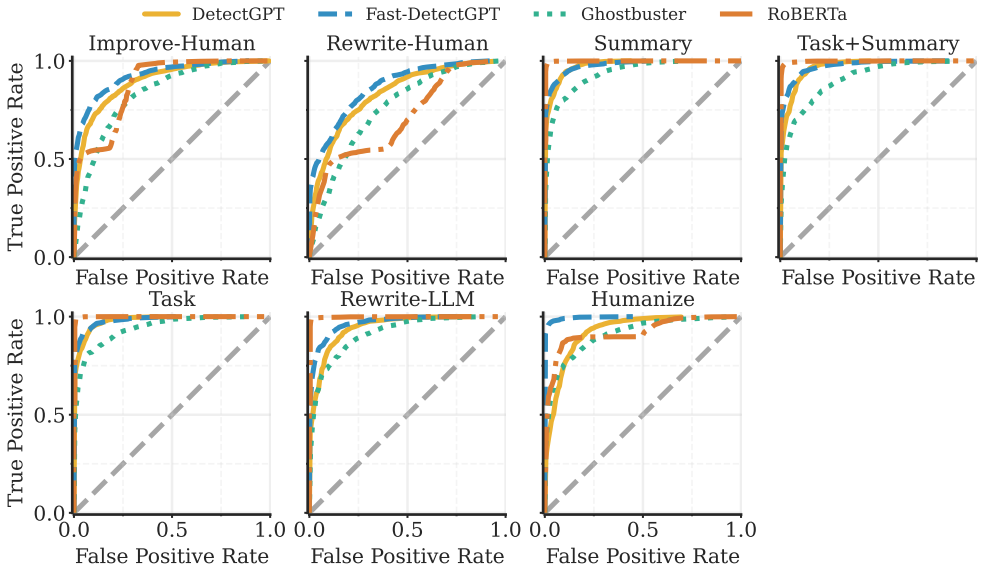


Figure 6: ROC curves for the contribution-level experiment.

6.2 Generative Model

The ROC curves shown in Figure 7 correspond to the generative-model experiment described in Section 4.3.

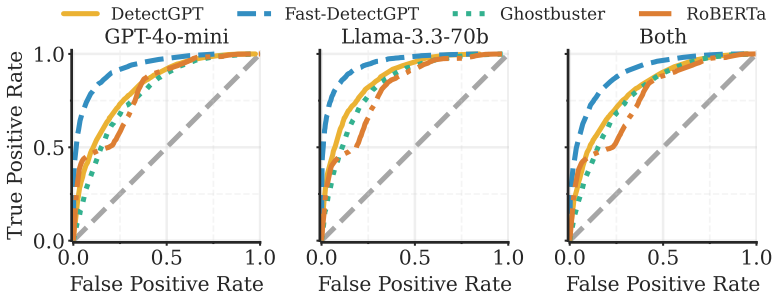


Figure 7: ROC curves for the generative-model experiment.

6.3 Out-of-Distribution Data

The ROC curves shown in Figure 8 correspond to the out-of-distribution data experiment described in Section 4.4.

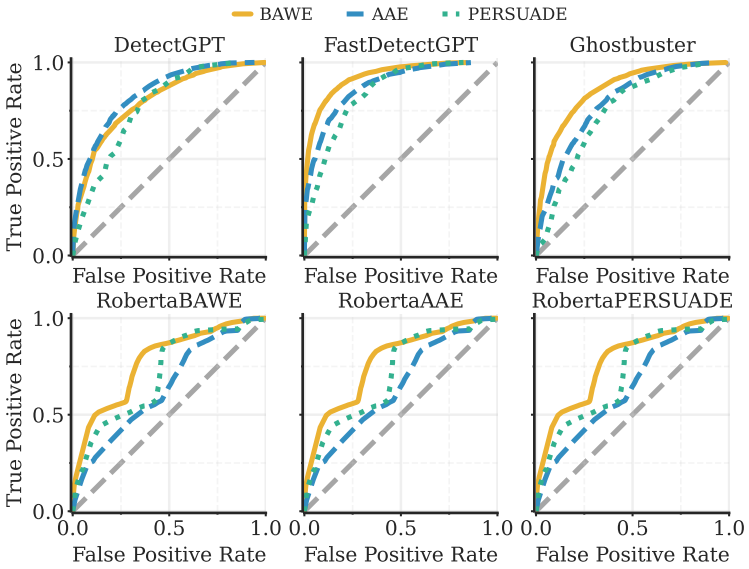


Figure 8: ROC curves for the out-of-distribution data experiment.

6.4 Text Lenght

The ROC curves shown in Figure 9 correspond to the text length experiment described in Section 4.5.

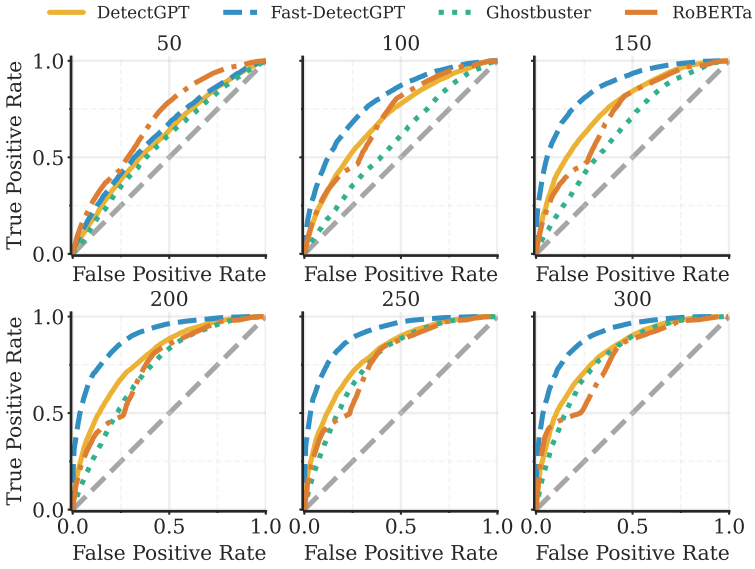


Figure 9: ROC curves for the text length experiment.

References

[1] S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865, 2024. doi:10.1016/j.asw.2024.100865.

[2] Hilary Nesi, Sheena Gardner, Paul Thompson, and Paul Wickens. British academic written english corpus, 2008. URL `http://hdl.handle.net/20.500.14106/2539`. Literary and Linguistic Data Service.

[3] Christian Stab and Iryna Gurevych. Argument annotated essays (version 2), 2017. URL `https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422`.

[4] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017. doi:10.1162/COLI_a_00295.