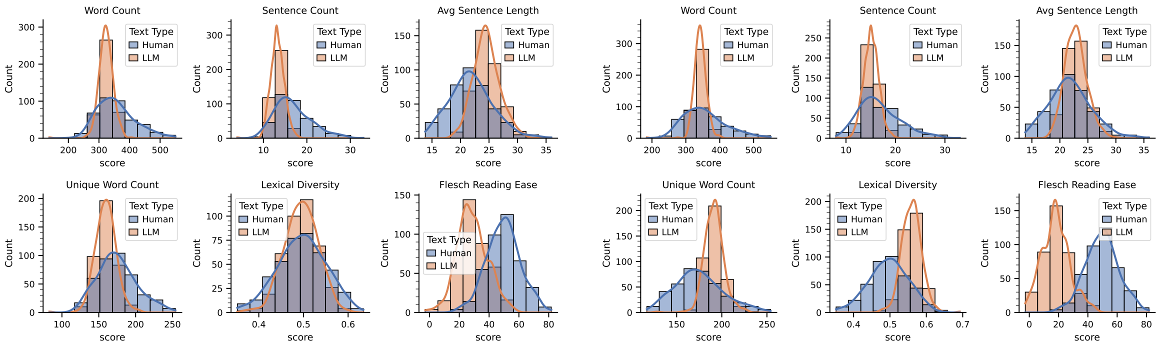
Appendix: Assessing LLM Text Detection in Educational Contexts: Does Human Contribution Affect Detection?

August 8, 2025

1 Dataset Statistics

1.1 Text Corpora Analysis

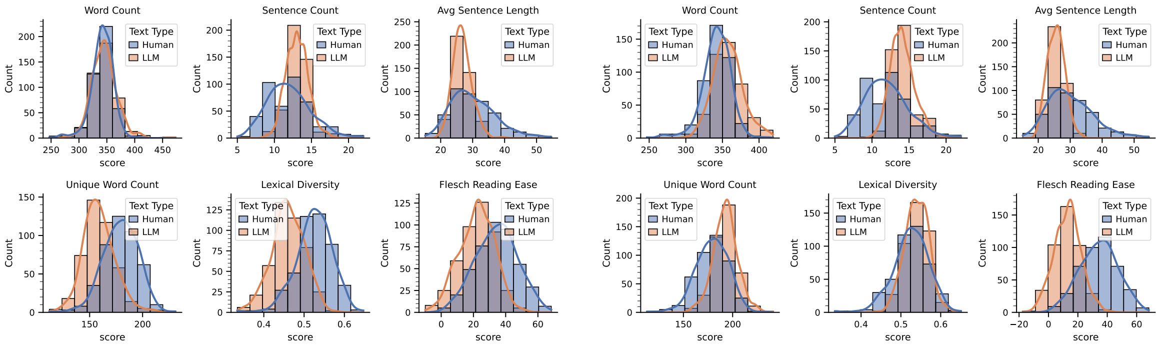
In the following, we provide detailed statistics of the subdatasets AAE [6, 5] (Fig. 1), BAWE [4] (Fig. 2), and PERSUADE [2] (Fig. 3). Note that we compare essays from *Task* contribution level with *Human* contribution level.



(a) Llama-3.3-70B-Instruct

(b) GPT-4o-mini

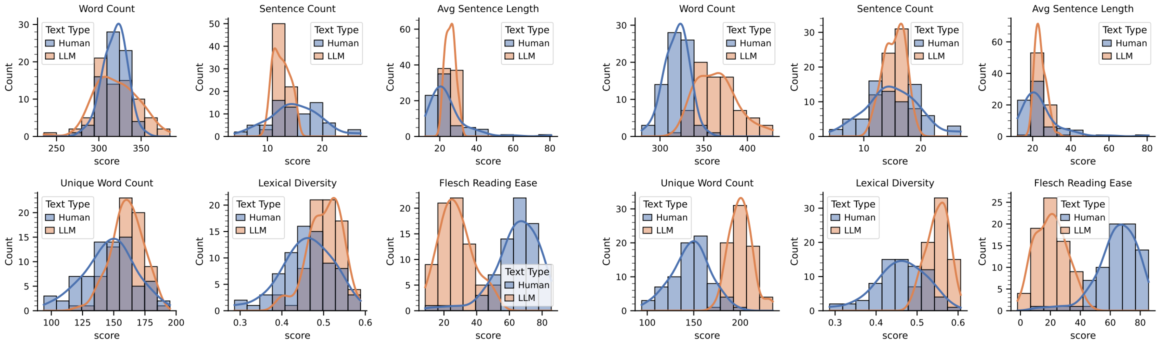
Figure 1: AAE *Task* vs *Human* statistics



(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

Table 1 provides an overview of the disciplinary groups included in the BAWE corpus and their respective 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
AH	0.812 ± 0.11	999	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
			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
LS	0.823 ± 0.14	999	Medicine	0.868 ± 0.10	4
			Biological Sciences	0.862 ± 0.08	4
			Agriculture	0.823 ± 0.18	11
			Psychology	0.805 ± 0.17	25
			Health	0.802 ± 0.15	10
			Food Sciences	0.777 ± 0.15	3
PS	0.830 ± 0.13	999	Chemistry	0.910 ± 0.06	1
			Computer Science	0.880 ± 0.09	3
			Architecture	0.816 ± 0.14	3
			Planning	0.813 ± 0.13	6
			Physics	0.731 ± 0.16	3
SS	0.828 ± 0.12	999	Publishing	0.859 ± 0.08	2
			HLTM	0.857 ± 0.11	17
			Anthropology	0.828 ± 0.15	22
			Business	0.827 ± 0.16	25
			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 Changes to the Human-Written Text in the *Improve-Human* and *Rewrite-Human* Contribution Level

Figure 4a shows an excerpt from a randomly selected human-written text from the BAWE corpus, improved by GPT-4o-mini for the *Improve-Human* contribution level. Text passages highlighted from orange to blue represent changes made by the model to the existing text, while red indicates removed words and green newly added ones. Although the model made some changes to the text, the meaning remains the same. When comparing the improvements made by GPT to Llama-3.3 in Figure 4b, we can observe that Llama makes fewer changes to the original text. However, both LLMs show similar patterns like correcting ’*changes*’ to ’*shift*’ or ’*In order to*’ to ’*to*’.

Furthermore, we compute the cosine similarity between the original human-written text and the corresponding improved/rewritten text from *Improve-Human* and *Rewrite-Human* contribution levels. Figure 5 shows that texts from the *Improve-Human* category are closer to the original texts than the ones from *Rewrite-Human*.

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.

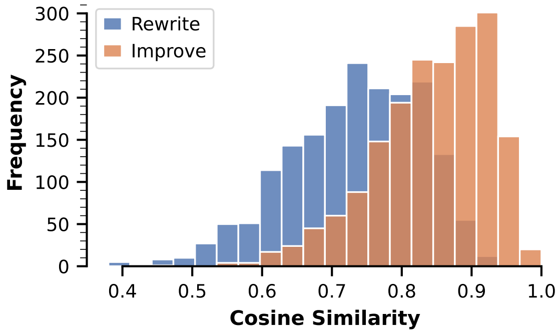


Figure 5: Cosine Similarity between Improve-Human and Rewrite-Human texts.

3 Contribution Level Prompts

Table 2 shows all prompts used to create the LLM-generated essays. The *Human* and *Humanize* levels are included in this table for completeness, although they are not created using system and 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	-	-

4 RoBERTa Fine-Tuning

We fine-tune the RoBERTa model using the HuggingFace *roberta-base* model[1] on each of the three text corpora. Table 3 shows the hyperparameters we used for fine-tuning RoBERTa on

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

the different subsets. We chose the best epoch based on the evaluation cross-entropy loss.

Table 3: Hyperparameter of RoBERTa fine-tuning on the different text corpora subsets.

Subset	Batch Size	Test Size	Epochs	Best Epoch
AAE	32	0.2	5	3
BAWE	32	0.2	5	2
PERSUADE	32	0.2	8	3

5 Contribution Level Performance

Figure 6 shows the performance of all detectors on a single contribution level, corpus, and generative model.

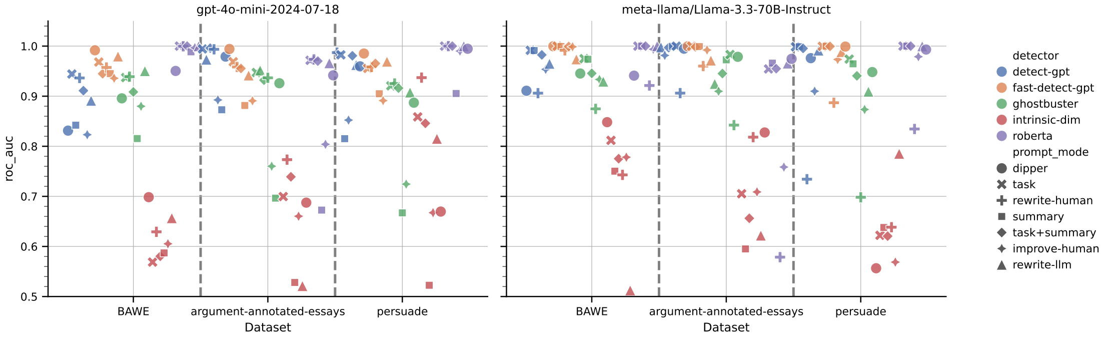


Figure 6: Violin plot of the ROC-AUC over all prompt modes for all detectors on all datasets and generative models.

6 ROC-Curve of Varying Label Boundaries

Figure 7 shows the ROC curves of the different detectors for varying label boundaries. Due to its weak performance on our dataset, IntrinsicDim [7] shows no real difference between the different label boundaries. The RoBERTa [9] model shows multiple inflection points in the ROC curve. This is likely caused by many samples receiving the same prediction score, leading to abrupt changes in the true positive and false positive rates at some thresholds.

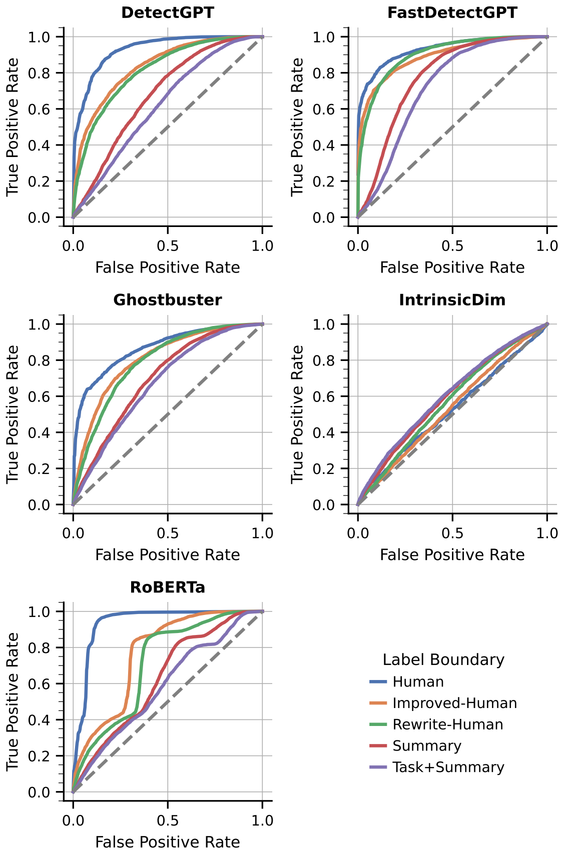


Figure 7: ROC Curve of different detectors at varying human label boundaries.

7 Threshold Optimization

Table 4 provides detailed values for the threshold optimization methods shown in Fig. 3a of the paper.

Table 4: Comparison of different threshold computation methods

Detector	Method	Threshold	Accuracy	Specificity	F1 Score
DetectGPT [3]	F1 Score	0.51	0.76	0.65	0.73
DetectGPT	FPR-based	1.09	0.53	0.95	0.53
DetectGPT	J-Index	0.60	0.74	0.72	0.72
Fast-DetectGPT [1]	F1 Score	2.08	0.84	0.73	0.82
Fast-DetectGPT	FPR-based	3.39	0.69	0.95	0.69
Fast-DetectGPT	J-Index	2.51	0.81	0.83	0.80
Ghostbuster [8]	F1 Score	0.22	0.75	0.62	0.72
Ghostbuster	FPR-based	0.78	0.47	0.95	0.45
Ghostbuster	J-Index	0.27	0.74	0.68	0.72
Ghostbuster	static	0.50	0.63	0.84	0.63
RoBERTa [9]	F1 Score	1.00	0.77	0.59	0.73
RoBERTa	FPR-based	1.00	0.42	0.95	0.40
RoBERTa	J-Index	1.00	0.77	0.60	0.73
RoBERTa	static	0.50	0.72	0.20	0.58

References

[1] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] 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.

[3] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the ICML'23*, 2023. doi:10.5555/3618408.3619446.

[4] 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.

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

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

[7] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. In *Proceedings of the NeurIPS*, pages 39257–39276, 2023. doi:10.5555/3666122.3667828.

[8] Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the NAACL*, pages 1702–1717, 2024. doi:10.18653/v1/2024.naacl-long.95.

[9] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the CCL*, pages 1218–1227, 2021. URL `https://aclanthology.org/2021.ccl-1.108/`.