A Method for Distinguishing Model Generated Text and Human Written Text

Hinari Shimada * and Masaomi Kimura

College of Engineering, Computer Science and Engineering, Shibaura Institute of Technology, Koto, Japan
Email: al21822@shibaura-it.ac.jp (H.S.); masaomi@shibaura-it.ac.jp (M.K.)
*Corresponding author

***Abstract*—With the rapid development of Large Language Models (LLMs), such as ChatGPT, it is extremely difficult for humans to accurately detect whether sentences are written by LLMs. Especially in academic fields, there is a need to assist human evaluators by discriminating sentences to recognize differences. Assignments such as essays and theses typically require human authors to write content. However, there is a risk of effortlessly generating text using advanced LLMs such as ChatGPT, potentially allowing the completion of class assignments without human effort. As it has a significant impact on the fair evaluation of students, we need to distinguish between text generated by model (model generated text) and written by human (human written text). Detection using existing statistical measures, such as log likelihoods, does not perform well for black-boxed models, such as ChatGPT, because it requires access to the internals of the models. Therefore, we propose a new approach that captures text from two different perspectives using log likelihoods and sentence embeddings with multiple LLMs. In experiments using data, including those generated by the black-box model ChatGPT, our proposed method demonstrated superior accuracy compared to existing approaches.**

***Keywords*—Large Language Models (LLMs), model generated text, human written text, log likelihoods, sentence embeddings**

I. Introduction

Large Language Models (LLMs) are being used around the world with the advent of ChatGPT. Although it may sometimes respond with false content that differs from the facts [1], its ability to understand the content of sentences and to respond in a natural and human-like manner has attracted significant attention. It is expected to be used effectively in a wide range of applications, including sentence generation and application to programming. However, the development of generation technology has made it difficult to discriminate between model generated text and human written text. Without the use of detectors, humans can only discern model generated text to the same extent as they were randomly selected [2]. In addition, evaluations by models and humans are performed from different perspectives and the detection abilities of human may be inferior to those of detectors in some situations [3]. Therefore, particularly in the academic domain, if there is suspicion that a text written by a student may be generated by a LLM, it becomes challenging for human evaluators to make accurate judgments. The absence of a detector can potentially impact the fair assessment of students' work in class assignments. Hence, there is a need for the development of model generated text detectors to assist human evaluators.

As a representative method to detect model generated text, the discrimination of sentences using statistical indices, such as likelihoods, has yielded excellent results. Log likelihood is the probability of selecting the next word from the preceding words. In the previous study, the detection rate of model generated text by humans was significantly improved by visualizing the features of sentences based on various statistical indices [2]. Nevertheless, these indicators assume to access the internals of the specific models and knowledge of the model internals, which requires ingenuity to detect sentences written by black-boxed LLMs. ChatGPT is a proprietary and black-boxed LLMs, and obtaining log likelihood directly from the API is not feasible. As a result, previous research [4] was carried out by substituting other models, such as GPT-2, where log likelihood is obtainable. However, ChatGPT, compared to GPT-2, is a more advanced model that incorporates state-of-the-art technology, and it differs in terms of training data and parameter counts. In the future, with the emergence of even more advanced and black-boxed LLMs, relying solely on GPT-2 for detection may lead to inaccurate results due to performance differences. Hence, we hypothesize that preparing multiple LLMs with varying training data and parameter counts can address this issue by enabling a relative evaluation of their outcomes.

Therefore, we propose a model for distinguishing whether a given text is generated by LLMs. In this detection process, we consider two different perspectives: log likelihoods, which consider the occurrence of preceding words, and sentence embeddings, which consider the entire text. In addition, we capture these perspectives by evaluating the output values of multiple LLMs with diverse training data and characteristics. This approach enables the robust identification of texts generated by black-box models.

Manuscript received January 11, 2024; revised February 1, 2024; accepted February 22, 2024; published June 13, 2024.

The contributions are as follows:

- We propose a novel model for detecting model generated text by calculating log likelihoods and sentence embeddings using multiple LLMs. This model can detect model generated text even for black-box models like ChatGPT, and it achieves the highest accuracy compared to several previous research.
- We analyze the differences between model generated text and human written text from the perspectives of log likelihoods and sentence embeddings. Human written text possesses a creativity that LLMs cannot replicate, evident in the choice of words, content, and writing style.
- We examine the impact of using multiple LLMs. The use of multiple LLMs allows for more robust text detection, avoiding the influence of individual model-specific results.

The rest of the paper is structured as follows: Section II presents a review of related works. Section III describes our proposed methodology. Section IV provides information about the experiments and Section V describes the result. Section VI discusses the experimental result effectiveness of the proposed methodology. Finally, Section VII presents the conclusion of this paper.

II. Related Work

With the advancement of LLMs, such as ChatGPT, there is a growing challenge in developing methods to distinguish between model generated text and human written text. Fundamentally, the discrimination of text generated by LLMs is primarily treated as a classification problem using statistical metrics as a representative approach [3]. Germann *et al.* [2] analyzed the differences between model generated text and human written text based on three statistical metrics: the probability of generating the next word, the absolute rank representing the difference with the predicted token as the next word, and the entropy of the prediction distribution. Mitchell *et al.* [4] proposed DetectGPT, a zero-shot method based on the hypothesis that while the log likelihoods of model generated text tend to be higher than the average log likelihoods of randomly rewritten text, this tendency is not observed in human written text. Su *et al.* [5] also proposed a zero-shot method that effectively uses log rank metrics, focusing on the ratio of log likelihoods to log rank and the difference in log rank after perturbing the sentence. Liang *et al.* [6] used perplexity, a measure of uncertainty in predicting the next word, and tested it on a variety of sentences.

Explainable AI is necessary to support humans when making decisions. Lundberg *et al.* [7] proposed a method for interpreting model predictions and aligning them with human intuition using the Shapley value, a fair measure of a player's contribution to game theory. They proposed a method to interpret model predictions and reconcile them with human intuition. Yang *et al.* [8] created a dataset containing scientific abstracts and quantified the degree of ChatGPT involvement in the text and the originality of the text using Jaccard Distance and Levenshtein Distance.

Supervised learning methods are also frequently used. Guo *et al.* [9] used HC3, a dataset of tens of thousands of responses from human experts and ChatGPT in a variety of fields. Pretraining can also be domain-specific. Yu *et al.* [10] created a large dataset considering situations in which ChatGPT is used in academia and trained with existing models to achieve high accuracy. Liu *et al.* [11] created The GPABenchmark Dataset specifically for academic fields. They used it to train their proposed model, called CheckGPT, on their dataset, which recorded high accuracy on scientific abstracts.

Although all these previous studies have produced excellent results, misclassification is undesirable for actual use in educational situations, and a further improvement in accuracy is needed.

III. Proposed Method

In this section, we describe the detection of sentences generated from LLMs using the proposed method. We define X as a certain dataset and x_i is one of the sentences in X. First, each sentence x_i is input into n GPT-based LLMs to obtain n log likelihoods. Next, the entire dataset X is input into m BERT-based LLMs to obtain 768-dimensional sentence embeddings. Because sentence embeddings are high-dimensional, we use t-SNE [12], which represents the similarity between data points through joint probabilities and achieves dimensionality reduction by calculating similarity in the low-dimensional space based on the t-distribution. This method is characterized by placing similar points close together and different points far apart to reduce the dimensionality to a 2-dimensional vector. By reducing the dimensions to a 2-dimensional vector, it becomes possible to distribute and visually capture the differences between model generated text and human written text.

Therefore, 2m-dimensional vectors are obtained for each sentence, which are combined and input into A Feed Forward Neural Network (FFNN) for supervised learning to determine whether the sentence is model generated text or human written text. The FFNN has 12 input layers, 144, 12 hidden layers, and 1 output layer. BCELoss is used for the loss function, and the stochastic gradient descent was used as the optimizer. Moreover, log likelihoods and sentence embeddings are computed using Generative Pre-trained Transformers (GPT)-based and Bidirectional Encoder Representations from Transformers (BERT)-based models, respectively. The log likelihoods are calculated based on the previous words, whereas sentence embeddings are calculated from the entire text. Using both these elements, each sentence can be viewed from multiple perspectives. Furthermore, multiple models use different training data and parameters. If a single LLM is used, the detection is based solely on the perspective of that model. However, using multiple LLMs, detection results can be obtained from a wide range of perspectives. This allows for robust detection that applies to a diverse range of texts generated by various LLMs, including black box models. An overview of this diagram is presented in Fig. 1.

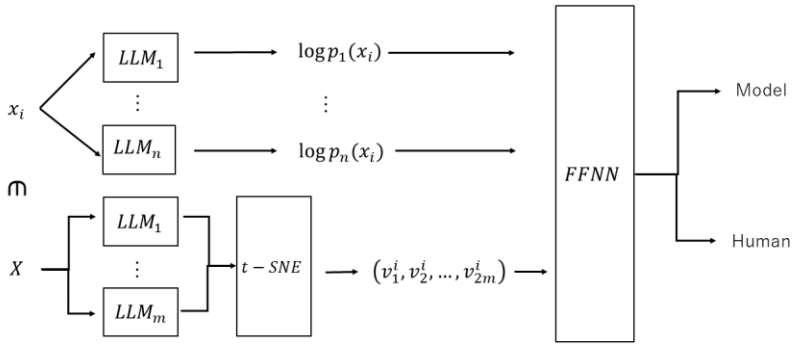


Fig. 1. Overview diagram of the proposed method.

A. The Reason to Use Log Likelihoods

The log likelihood is calculated from the input data. LLMs tend to generate words with a high probability of being followed by the next word based on all previous words. LLMs learn a large amount of data in advance and then compute the appropriate words from the training results. However, when humans write sentences, they tend to pick words unconsciously based on their own experiences and emotions. Therefore, we believe that there is a difference in word selection between model generated text and human written text.

In this study, the log likelihoods are calculated in the following way. First, texts are tokenized and input into LLMs. The LLMs then calculate the conditional probability that the next token will be generated based on all previous tokens. As the probability that a sentence is generated is regarded as the product of the conditional probabilities obtained from all tokens, the log likelihoods of a sentence are the sum of the logarithms of the conditional probabilities.

B. The Reason to Use Sentence Embeddings

Next, we focus on the expression and structure of the sentence output by LLMs. LLMs generate text based on pretrained data, lacking specific individuality, whereas humans create text from various background knowledge and experiences, imparting uniqueness and creativity to their writings. Consequently, even on the same topics, there should be differences in the content and writing styles between model generated text and human written text. Additionally, Ma *et al*. [13] demonstrated that in scientific papers, differences in writing style and syntax appear in both texts. Therefore, emphasizing disparities in content and writing style, we compute sentence embeddings representing the content of the text using BERT-based models.

The method for computing the sentence embeddings is as follows. First, the sentences are divided into tokens, and the tokens are input to the BERT-based models. The BERT-based models have a limit of 512 tokens, but in this case, since we are specifically focusing on the style and delivery of the text, it may not necessarily encompass the entire sentence. Therefore, regardless of the length of the text, we input up to the maximum limit of 512 tokens for the BERT-based models and obtained a vector. The vector of the last hidden layer is then obtained and averaged into a single vector. In addition, because this vector usually has 768 dimensions, the values obtained are used after dimensionality reduction to two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) [12].

IV. Experiment

A. Datasets

The ChatGPT-Detector-Bias dataset [6] available on GitHub, was used for the experiment. This is a mixed dataset that includes both model generated text and human written text. The dataset contains 925 sentences, including not only simply generated and written texts but also prompt-engineered and polished sentences. The CDB dataset was originally used to verify that the difficulty of text detection significantly varies with English proficiency. The CDB dataset was selected as the experimental data because, in comparison to other existing datasets, it mainly contains essays with a higher degree of freedom in writing rather than academically formal texts with limited freedom. While previous research on creating datasets focusing on formal academic writing is abundant in the academic fields, there is still limited research on datasets targeting informal writing styles, such as essays, mainly because of the challenges in collecting data. Therefore, in this experiment, we used the CDB dataset, which includes scientific abstracts using the Stanford CS224N final project report (CS224N), as well as essays such as the 2022-2023 US Common App college admission essay (College Essay), TOEFL essay (TOEFL), and US 8th-grade essays sourced from the Hewlett Foundation's Automated Student Assessment Prize dataset (Hewlett Student Essay). The CS224N and College Essay datasets contain not only sentences simply generated by ChatGPT using title or essay prompt and written by humans but also sentences generated by using prompt engineering and self-editing. In this experiment, we labeled these sequentially as "*LLM*" and "*LLM_prompt*". Additionally, the Hewlett Student Essay and TOEFL datasets contain texts that have been polished by ChatGPT to resemble the writing styles of native and non-native speakers, respectively. These are labeled as "*LLM_to_native*" and "*LLM_to_nonnative*". Furthermore, all data with human written text were categorized as "*Human*".

The aim of this experiment was to determine whether the text was generated by the ChatGPT or by humans. Therefore, the texts finally generated by ChatGPT, namely "*LLM*", "*LLM_prompt*", "*LLM_to_native*", and "*LLM_to_nonnative*", are considered model generated text, and the classification is conducted between these, and human written text labeled as "*Human*". Overall, there were 531 model generated text and 394 human written text in the dataset. Table I summarizes the individual datasets and their respective data counts.

TABLE I. DETAILS OF THE CDB DATASET

Data	LLM	LLM_prompt	LLM_to_nonnative	LLM_to_native	Human	All
College Essay	31	31	-	-	70	132
CS224N	145	145	-	-	145	435
TOFEL	-	-	91	-	91	182
Hewlett Student Essay	-	-	-	88	88	176
All Data (CDB Dataset)	176	176	91	88	394	925

B. Settings

In our experiments, we used four models to calculate log likelihoods: "EleutherAI/gpt-neo-2.7B" [14], "databricks/dolly-v2-3b" [15], "facebook/opt-2.7b" [16], and "EleutherAI/pythia-1.4b" [17].

In the experimental results of Liu *et al.* [18], it was observed that more advanced GPT based LLMs tended to diversify syntax and vocabulary. Models with higher parameter counts tended to have higher performance. Considering that the model generated text in this experiment was produced by the advanced model ChatGPT, we selected models with larger parameter sizes as the focus.

In addition, the computation of sentence embeddings can be done using "bert-base-uncased" [19], "roberta-base" [20], "xlnet-base-cased" [21] and "microsoft/deberta-v3-base" [22, 23], which serve as standard aspects of the model. During training, our proposed model set learning rate of 5e-4, the batch size 16, and the number of epochs is 2000. Furthermore, we used the DetectGPT [4], Roberta-HTTP [8] and CheckGPT [11] models for comparison with previous studies. For DetectGPT, we used "gpt2-medium" [24] as the model to calculate the likelihood and "t5-base" [25] as the model to provide perturbations. CheckGPT was tested using a model trained with a batch size of 98, a learning rate of 7.5e-5, and 30 epochs. The accuracy was defined as the average of the five-part cross-validation results.

V. RESULT AND DISCUSSION

A. Result

Table II summarizes the accuracy of the proposed method and that of previous studies when experiments were conducted on the CDB dataset. Table II shows that our model scored best on the CDB dataset. Compared to DetectGPT with the zero-shot method, the three models with supervised learning recorded similarly high accuracy, confirming the improvement in accuracy due to pretraining. The proposed method and CheckGPT trained on the CDB dataset also achieved higher accuracies, with a difference of only 0.0054.

TABLE II. ACCURACY OF EACH MODEL

Model	Accuracy
DetectGPT	0.7048
Roberta-HPPT	0.8660
CheckGPT	0.9795
Proposed model (ours)	**0.9849**

B. Log Likelihoods from GPT-Based Models

Fig. 2 shows a box-and-whisker diagram summarizing the log likelihoods of each sentence for each LLM. The vertical axis represents the log likelihoods; the higher the value, the more likely the sentence is to be generated by the model. Sentences with "*LLM*" at the beginning are model generated text, and those with "*Human*" are human generated text. The two features shown in Fig. 2 can be considered.

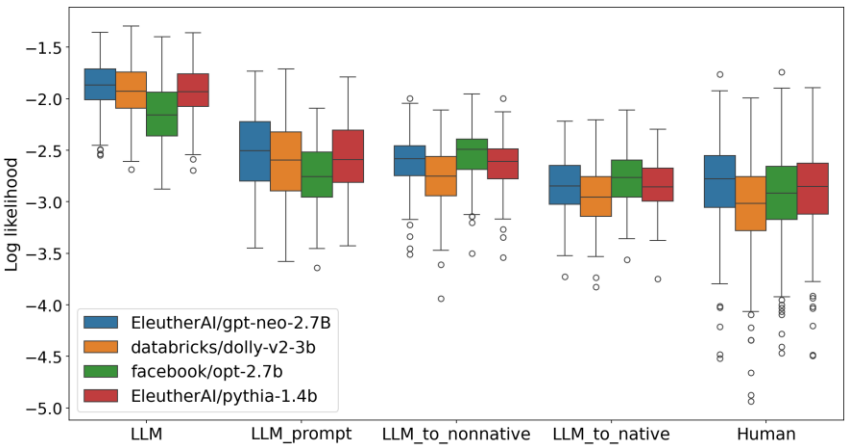


Fig. 2. Log likelihoods of the CDB dataset.

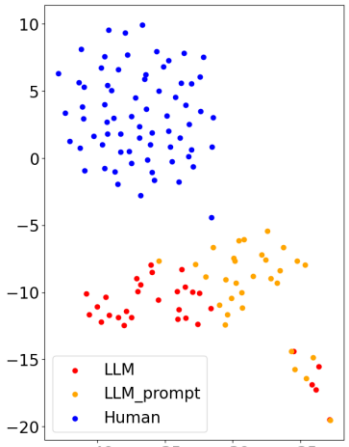
First, even when ChatGPT is used to perform prompt engineering or polishing, there is a change in the log likelihood. The "*LLM_prompt*", which is a prompt-engineered sentence, has a lower value than the "*LLM*" which simply generated by ChatGPT, indicating that prompt engineering changes the choice of words and expressions and makes sentences closer to human written text. The "*LLM_to_nonnative*" sentences, which were written with non-native speakers in mind, have slightly higher values than the "*Human*" sentences. However, the "*LLM_to_native*" sentence, which was written for a native speaker, has a similar value compared to the "*Human*" sentence. This is in general agreement with the results of a previous study [6] that the more word richness is restricted by ChatGPT, the higher the log likelihoods, suggesting that sentences written by non-native speakers are more likely to be misclassified by the detector. This indicates that, even when sentences are prompted by ChatGPT, the overall selection of words is similar to that of human writing. However, looking at the range of box-and-whisker diagram, not only the range of "*LLM_prompt*" but also the range of "*LLM_to_native*" and "*LLM_to_nonnative*" are also smaller than those of "*Human*", and the minimum log likelihoods values are also different. This indicates that human written text contains words and expressions that are difficult to generated by ChatGPT, even though texts seem to be similar. Therefore, it can be concluded that the use of ChatGPT disappears the originality of sentences. Given this feature, it is possible to find differences between model generated text and human written text using log likelihoods.

Second, it is possible to capture sentence characteristics using multiple LLMs. The sentence "*LLM_prompt*" generated by ChatGPT with the prompting engineering has lower log likelihoods than the sentence "LLM" simply generated by ChatGPT, and is closer to a human written text. This indicates that, if only one LLM is used, the detector may misclassify the sentence as written by a human. However, looking at the positional relationship of several LLMs used to calculate log likelihoods, the positional relationship between "databricks/dolly-v2-3b" [15] and "facebook/opt-2.7b" [16] is different for "*LLM*", "*LLM_prompt*" and other labels. "*LLM*", "*LLM_prompt*" are sentences generated entirely by ChatGPT, while the other labels "*LLM_to_native*", "*LLM_to_nonnative*", and "*Human*" are partially or completely written by humans. This suggests that a relative comparison of multiple LLMs can be used to indicate the degree of human involvement in a text.

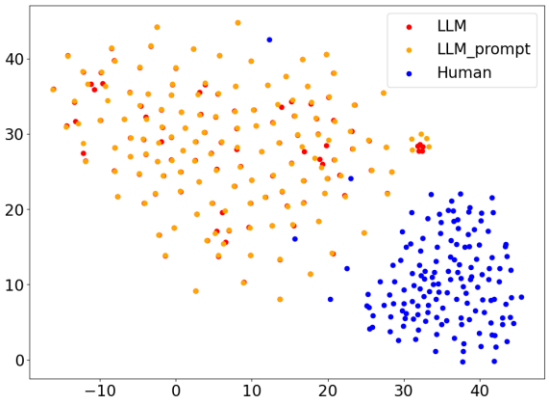
Thus, it was seen that there is a difference between model generated text and human written text in terms of log likelihoods. On the other hand, the median value of "*Human*" is similar to that of "*LLM_prompt*", "*LLM_to_native*", and "*LLM_to_nonnative*". In addition, the maximum value for "*Human*" is close to the median value for "*LLM*". Therefore, there is a potential risk of misclassifying text, especially human written text with high log likelihoods or model generated text with low log likelihoods, even when multiple LLMs are employed. Although there is a clear tendency for log likelihoods, it is not sufficient to determine whether a sentence was written by LLMs or humans.

C. Sentence Embeddings from BERT-Based Models

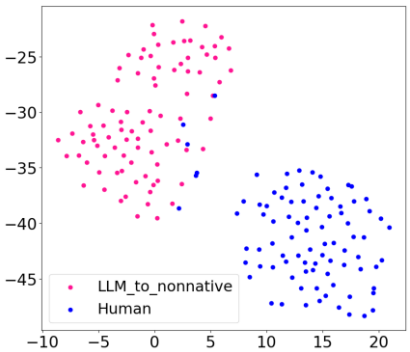
Fig. 3 visualizes the dimensionality reduction of the generated 768-dimensional vectors to two dimensions using t-SNE [12], particularly in BERT [19], for each individual dataset in the CDB dataset. Because the dimensionality reduction in t-SNE [12] was input for all datasets together, some points may overlap for each dataset when viewed as a whole. In addition, the model received only the coordinate values as the input, and the colors were not included in the input.



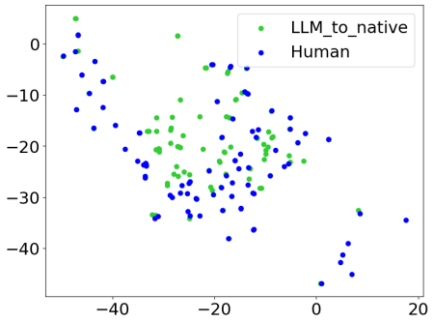
(a)



(b)



(c)



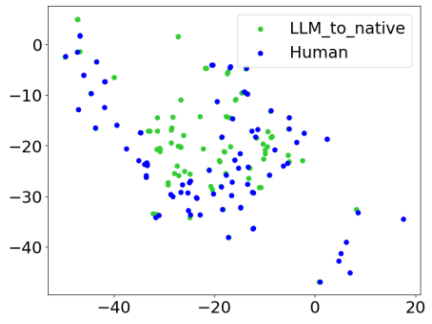
(d)

Fig. 3. Sentence embeddings visualization in BERT [19].
(a) College Essay. (b) CS224N. (c) TOFEL. (d) Hewlett Student Essay.

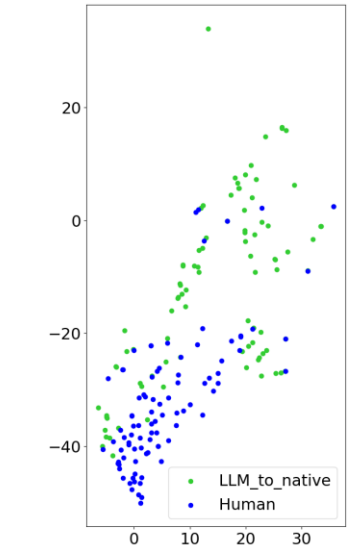
Fig. 3(a) and (b) show the results of the visualizations in College Essay and CS224N. They show that "*LLM*" and "*LLM_prompt*" have close values, whereas "*Human*" has distant values. Therefore, it can be said that ChatGPT's self-edit of sentences did not significantly change the sentences themselves, as the authorship of the sentences remained with ChatGPT.

Furthermore, Fig. 3(c) and (d) show visualizations of the sentences produced by ChatGPT in the TOFEL and Hewlett Student Essay, controlling for the richness of linguistic expressions. Fig. 3(c) shows that "*LLM_to_nonnative*" is plotted far from "*Human*". It can be inferred that the content and structure of the sentences have changed since before polishing by ChatGPT to limit the linguistic expressions. Therefore, in Fig. 3(c), each label is classified as shown in Fig. 3(a) and 3(b). By contrast, Fig. 3(d) shows that "*LLM_to_native*" belongs to a point close to "*Human*". This indicates that each label is not classified because the sentences did not change significantly due to the variety of expressions, as the sentences were polished using ChatGPT to enrich the linguistic expressions.

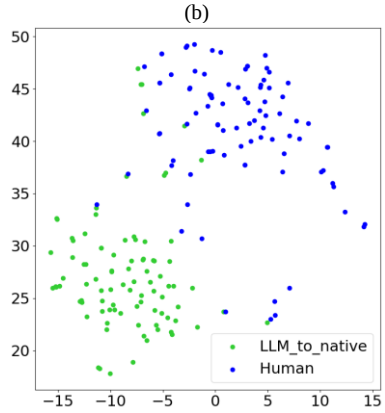
Fig. 4 displays the visualization results for all the BERT-based models focused on the data of Hewlett Student Essay. It can be seen that in Fig. 4(a) (same as in Fig. 3(d)), points of "*LLM_to_native*" and "*Human*" belong to similar positions respectively, but Fig. 4(b), (c), and (d), there is generally a difference between both two labeled sentences. If only one LLMs is used, detections can be made based on discrimination results specific to some LLMs, as shown in Fig. 4(a). By using multiple LLMs in the proposed method, we can prevent misjudgment.



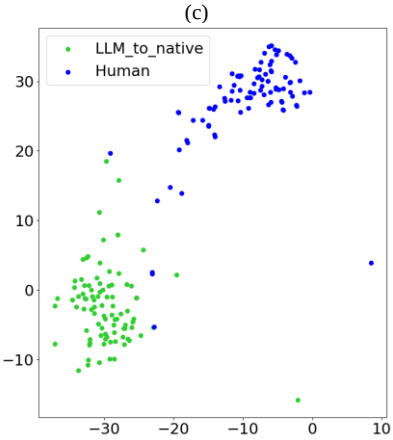
(a)



(b)



(c)



(d)

Fig. 4. Sentence embeddings visualization in Hewlett Student Essay dataset. (a) BERT [19]. (b) XLNet [21]. (c) RoBERTa [20]. (d) DeBERTa [22, 23].

While our approach allows for effective discrimination in this experiment, it is conceivable that datasets with a tendency for model generated text and human written text to exhibit similar patterns, as shown in Fig. 3(d), may pose challenges. The proposed method operates under the assumption of a single discriminator capable of handling diverse datasets. Therefore, relying solely on a discrimination method based on sentence embeddings may prove to be insufficient, and it is imperative to

consider additional perspectives for a comprehensive approach.

D. Choice of LLMs and Impact on Results

In this experiment, four GPT-based models and four BERT-based models were utilized. To investigate the impact of the selected LLMs on the results, the detection performance was examined by varying the parameters of the same selected LLMs. This approach was based on Liu *et al.*'s research [18], which indicated that more advanced GPT-based LLMs tended to diversify syntax and vocabulary. Additionally, we take into consideration the observed trend that the performance of LLMs improves as the number of parameters increases.

First, experiments were conducted by increasing the parameter count:

(1) Using the same LLMs as the experiment in Section IV.
(2) GPT-based model's parameter count was adjusted to be around 3 billion. The Pythia model from EleutherAI was replaced with "EleutherAI/pythia-2.8b" [17].
(3) All BERT-based models were upgraded from the base version to the large version.
(4) Combination of (2) and (3).

Next, experiments were conducted by decreasing the parameter count:

(5) GPT-based model's parameter count was adjusted to be around 1 billion. The GPT-Neo model from EleutherAI and the Open Pre-trained Transformer Language Models (OPT) from Facebook were replaced with "EleutherAI/gpt-neo-1.3B" [14] and "facebook/opt-1.3b" [16], respectively. Additionally, as there were no smaller parameter models available for "databricks/dolly-v2-3b" [15], the same model was used.
(6) GPT-based model's parameter count was further reduced. Similar to (5), "databricks/dolly-v2-3b" [15] was used, and the other models are replaced with "EleutherAI/gpt-neo-125m" [14], "facebook/opt-125m" [16], and "EleutherAI/pythia-410m" [17].
(7) One BERT-based model was changed to small. Among the selected models, only the DeBERTa model had a small version available. Therefore, "microsoft/deberta-v3-small" [22, 23] was used.

The experimental results were presented in Tables III and IV, and a comparison was made with (1), which serves as the baseline. As a result, models with an increased number of parameters exhibit accuracy beyond the baseline, while those with reduced parameters show accuracy below the baseline. Moreover, whether using GPT-based or BERT-based models, higher parameter counts corresponded to improved accuracy. Hence, both elements contributed positively to the discriminative performance. Furthermore, when all BERT-based models were switched to the large version, the accuracy reached its highest point. This suggests that, in terms of log likelihood and sentence embeddings, sentence embeddings have a stronger impact on the discrimination in this experiment. From this validation, it can be inferred that as the selected LLM's performance improves, the accuracy also enhanced.

TABLE III. THE CHANGE UNDER INCREASING THE PARAMETERS

Model	Accuracy
(1)	0.9849
(2)	0.9849
(3)	**0.9903**
(4)	**0.9903**

TABLE IV. THE CHANGE UNDER DECREASING THE PARAMETERS

Model	Accuracy
(1)	**0.9849**
(5)	**0.9849**
(6)	0.9795
(7)	0.9838

E. Potential Bias

The potential biases inherent in the individually selected model are evident in Fig. 3. While it is apparent that Fig. 3(a), (b), and (c) successfully distinguish between model generated text and human written text, Fig. 3(d) shows that "*LLM_to_native*" belongs to a point close to "*Human*". This indicates that each label is not classified because the sentences did not change significantly due to the variety of expressions, as the sentences were polished using ChatGPT to enrich the linguistic expressions. By contrast, examining the visualization results by all BERT-based LLMs in Fig. 4, it is observed that, except for Fig. 4(a) (same as in Fig. 3(d)), it has a difference between model generated text and human written text for the other BERT-based LLMs. When employing a single LLM, detection relies on specific biases inherent to that LLM, as shown in Fig. 4(a). However, by employing multiple LLMs in the proposed method, misjudgments can be prevented. Based on the above observations, it can be concluded that the individual bias inherent in each LLM does not significantly impact the experimental results.

However, proficiency in English is a bias that commonly exists across many LLMs. The CDB dataset [6] used in this experiment was designed with a focus on the varying difficulty of text detection based on English proficiency. Fig. 2 illustrates that sentences supposed to be written by non-native speakers tend to have higher overall likelihoods when compared to sentences assumed to be written by native speakers or by humans in general. In addition, this trend was observed across all LLMs. Experimental results of Liang *et al.* [6] similarly indicated a tendency for texts written by non-native speakers to be misclassified as model generated text. The proposed method aligns with these findings. The consistent occurrence of such results across multiple LLMs suggests the presence of a common potential bias in many LLMs, particularly indicating the existence of biases towards non-native speakers.

The influence of this bias is important, especially in the educational domain, where fair evaluation of writings by international students or those speaking dialects may be

compromised. Considering that the richness of word expressions influences English proficiency, using the log likelihood based on the choice of the next words may not offer a solution to this issue. However, it can be inferred that models, learned solely from textual data, may differ from humans with life experiences in expressing the overall content and manner of writing, such as through sentence embeddings.

Additionally, while models are generally less prone to basic grammatical errors, non-native speakers may make human errors or use expressions in their writing that, while semantically understandable, are not commonly employed by native speakers. There is a considerable possibility that non-native speakers and models may differ in the way they compose sentences in this regard. Thus, it is hypothesized that incorporating new metrics related to the overall content of the text and the level of human involvement in the writing process, independent of word choice, could moderate this bias.

F. Application of the Method

Based on the discussions in Figs. 3 and 4, it is evident that, unlike log likelihoods, sentence embeddings can generally distinguish between human written text and model generated text. This highlights that even LLMs with a vastly larger dataset than humans unconsciously incorporate electronic watermarks during sentence creation, manifesting model-specific uniqueness in writing style and content. Consequently, mimicking human written text entirely proves challenging for advanced LLMs, suggesting that, as LLMs continue to evolve, differences in content and writing style compared to human written text may persist, making sentence embeddings a potentially effective means of discrimination.

In addition, The CDB dataset used in this study primarily targeted essay formats with a high degree of freedom. This choice was motivated by the fact that, in the case of compositions, there is no predetermined structure for writing, making the author's uniqueness more evident in the text. It was characteristic that could lead to noticeable differences from the perspectives of log likelihoods and sentence embeddings. Therefore, additional verification is deemed necessary, especially for texts such as scientific paper abstracts, which follow a more formal structure, and content that involves more advanced and specialized subjects. In cases of formal writing, where word choices may be more restricted compared to free-form descriptions, differences in log likelihoods might be less likely to occur. Additionally, the overall argument of the text may be more similar between model generated text and human written text, potentially leading to less variation in sentence embeddings differences. As a possible solution, as discussed above, it is considered essential to introduce a new metric that quantifies the level of human involvement in the text.

Moreover, using various LLMs employed in this experiment, consistently achieved accuracy records of over 97%. However, considering the rapid advancement of LLMs and the findings of Liu *et al*. [6], there is a possibility that the selected LLMs in our research may become insufficient when faced with LLMs capable of generating even more sophisticated text. Therefore, it is imperative to carefully consider the choice of LLMs when more advanced models are developed in the future.

Furthermore, practical considerations in real-world applications raise concerns about the current high cost of computing log likelihoods and sentence embeddings. Hence, there is a need to develop a more cost-effective method for these calculations to facilitate practical implementation. While this experiment focused on English text, there is room for investigating the generalizability of the methodology to texts in other languages. Currently, there is a limited number of studies targeting languages other than English. Given the emergence of LLMs specialized in specific languages, it is worthwhile to explore their utility for assessing cross-linguistic generality.

VI. Conclusion

The goal of this study is to propose a model for distinguishing whether a given text is generated by LLMs or written by a human. The proposed method focuses on calculating the log likelihoods that consider all previous words and sentence embeddings that consider the entire text. These calculations were performed using multiple LLMs, and the values were inputted into an FFNN for classification. Experimental results using the CDB dataset showed that the proposed model achieved the highest accuracy compared to previous models. Additionally, both log likelihoods and sentence embeddings were effective in distinguishing between model generated text and human written text. Furthermore, relying solely on the results of certain LLMs may lead to misjudgments owing to differences in training data and characteristics. These, comprehensive evaluations of the results from the multiple LLMs were effective. In conclusion, by combining two elements that are effective in detecting model generated text, our model achieves higher accuracy even in identifying text generated by black-box models such as ChatGPT.

Future work will aim to further demonstrate the effectiveness of the proposed method for classifying scholarly texts by conducting experiments on a wide range of data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

H. Shimada conducted all experiments and analyzed data. All research activities were conducted under the supervision and coordination of M. Kimura. All authors had approved the final version.

References

[1] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022, pp. 3214–3252.
[2] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proc. the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Italy, 2019, pp. 111–116.
[3] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proc. the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 1808–1822.
[4] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. the 40th International Conference on Machine Learning*, USA, 2023, pp. 24950–24962.
[5] J. Su, T. Zhuo, D. Wang, and P. Nakov, "DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text," in *Proc. the Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 12395–12412.
[6] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native English writers," *Patterns*, vol. 4, no. 7, 100779, July 2023.
[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. the 31st International Conference on Neural Information Processing Systems*, USA, 2017, pp. 4768–4777.
[8] L. Yang, F. Jiang, and H. Li, "Is ChatGPT involved in texts? Measure the polish ratio to detect ChatGPT-generated text," arXiv preprint, arXiv:2307.11380, 2023.
[9] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," arXiv preprint, arXiv:2301.07597, 2023.
[10] P. Yu, J. Chen, X. Feng, and Z. Xia, "CHEAT: A large-scale dataset for detecting ChatGPT-writtEn AbsTracts," arXiv preprint, arXiv:2304.12008, 2023.
[11] Z. Liu, Z. Yao, F. Li, and B. Luo, "Check me if you can: Detecting ChatGPT-generated academic writing using CheckGPT," arXiv preprint, arXiv:2306.05524, 2023.
[12] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[13] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, "AI vs. Human—Differentiation analysis of scientific content generation," arXiv preprint, arXiv:2301.10416, 2023.
[14] S. Black, G. Leo, P. Wang, C. Leahy, and S. Biderman. (2021). GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. [Online]. Available: https://zenodo.org/records/5297715
[15] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. (2023). Free dolly: Introducing the world's first truly open instruction-tuned LLM. *Company Blog*. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm
[16] S. Zhang *et al*., "OPT: Open pre-trained transformer language models," arXiv preprint, arXiv:2205.01068, 2022.
[17] B. Stella *et al*., "Pythia: A suite for analyzing large language models across training and scaling," in *Proc. the 40th International Conference on Machine Learning*, USA, 2023, pp. 2397–2430.
[18] Y. Liu *et al*., "ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models," arXiv preprint, arXiv:2304.07666, 2023.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minnesota, 2019, vol. 1 (Long and Short Papers), pp. 4171–4186.
[20] Y. Liu *et al*., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint, arXiv:1907.11692, 2019.
[21] Z. Yang *et al*., "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. the 33rd International Conference on Neural Information Processing Systems*, Canada, 2019, no. 517, pp. 5753–5763.
[22] P. He *et al*., "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *Proc. the International Conference on Learning Representations*, Austria, May 2021.
[23] P. He, J. Gao and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style pre-training with gradient-disentangled embedding sharing," arXiv preprint, arXiv:2111.09543, 2021.
[24] R. Alec *et al*., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[25] C. Raffel *et al*., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.