Big Data and Information Analytics

BDIA, 9: 350–379.
DOI: 10.3934/bdia.2025016
Received: 11 September 2025
Revised: 13 November 2025
Accepted: 18 November 2025
Published: 05 December 2025

https://www.aimspress.com/journal/BDIA

Research article

Research on AI-generated Chinese text detection method based on deep learning

Chang Su[1], Yaqi Jiang[1], Jianlin Wang[2,*] and Junfang Zhao[1]

[1] School of Science, China University of Geosciences, Beijing 100089, China
[2] Research Institute of Highway Ministry of Transport, Beijing 100088, China

* Correspondence: Email: Wangjl0072000@163.com.

Abstract: This paper proposes a dual stream feature fusion model by integrating RoBERTa semantic encoding with manually designed text statistical features using deep learning to fuse statistical features. A cross domain hybrid multi-source corpus is constructed to train and validate the model's detection performance. We developed a multi-domain corpus encompassing the HC3 dataset, ChatGPT detection dataset, and self-constructed academic abstract and literary work datasets. By integrating academic abstracts from CNKI with texts generated by three LLMs (DeepSeek R1, Phi4, and Qwen 2.5), we constructed a dataset containing human-written and machine-generated texts. Experiments show that the RoBERTa-text model achieves optimal detection performance for Phi4-generated texts (recall: 100%), while Qwen 2.5-generated texts present greater challenges due to their human-like writing patterns (accuracy: 88.91%). Through categorizing texts into true positive, false positive, true negative, and false negative groups, we conducted statistical and linguistic analyses. Texts characterized by limited sentence length variation and dense punctuation were more likely to be identified as AI-generated. Comparative analysis of word frequency distribution and semantic perplexity revealed that LLM-generated texts exhibit repetitive lexical selection patterns, whereas human-written texts demonstrate more diverse vocabulary usage. We elucidate decision-making rationale and provide novel perspectives for AI-generated text detection research.

Keywords: artificial intelligence; RoBERTa; text features; semantic perplexity

1. Introduction

This chapter focuses on the Chinese context and sorts out the research status and challenges in the field of AI-generated text detection. Methods based on text features achieve detection through statistical patterns such as vocabulary distribution and syntactic complexity, while neural language models (e.g., BERT, RoBERTa) rely on deep semantic representation to improve discrimination accuracy. However, existing research still faces issues including scarce training data, insufficient detection effectiveness for short texts, and limited model interpretability. To address these existing problems, this study attempts to integrate statistical features with deep learning models to construct a detection framework adapted to the characteristics of the Chinese language, and optimizes the model's generalization ability through a cross-domain hybrid dataset. This study aims to provide new technical insights for the detection of AI-generated Chinese texts and contribute to practical needs such as the maintenance of academic norms and the guarantee of authenticity. The universality of the relevant methods still needs to be continuously verified and improved in more complex scenarios.

1.1. Research background

In 2022, OpenAI released ChatGPT, a model capable of interactive dialogue, which attracted widespread attention and discussion worldwide [1]. However, in the current era of rapid development of self-media and deep integration of AI technology into information dissemination, false information generated by AI [2] may bring a series of serious harms.

From the perspective of public opinion, false news can throw the public opinion field into chaos. In the academic field, the application of AI-generated texts may trigger academic integrity issues [3]. The issue of copyright ownership of AI-generated texts has long been a focus of debate in the academic and legal circles [4]. Since the process of AI-generated texts involves a large amount of data and algorithms, it is difficult to clearly define the ownership of their copyright.

Under this background, analyzing the content of AI-generated texts can, on the one hand, help researchers understand the characteristics of AI-generated texts, thereby recognizing the advantages and disadvantages of content produced by AI tools and further using intelligent tools properly. On the other hand, to address the issue of false news, supervision and review of AI-generated content can be strengthened to ensure the authenticity and accuracy of information. In the academic field, more stringent academic norms and evaluation systems can be established to prevent AI-generated texts from undermining academic integrity. Regarding copyright issues, it is necessary to clarify the copyright ownership of AI-generated texts, protect the legitimate rights and interests of creators, promote the enthusiasm for creation and innovation, and prevent the misuse and abuse of AI-generated texts [5].

1.2. Research status

To date, numerous scholars and institutions have attempted various methods to distinguish AI-generated texts [6,7], achieving significant detection results especially for AI-generated English texts [8], while research on detecting AI-generated Chinese texts remains relatively limited. Based on different research methodologies, AI-generated text detection methods can be categorized into text feature-based detection methods and neural language model-based detection methods [9]. A summary of relevant literature is provided below.

1.2.1. Text feature-based detection methods

In natural language processing, text feature statistical techniques can be used to extract statistical features from AI-generated texts, which are then classified using downstream classification algorithms support vector machine (SVM), random forest (RF) to obtain classification results [9].

In the field of natural language processing, research on lexical and phrase features in texts is of great significance. Some scholars have attempted to find effective ways to distinguish AI-generated texts from human-generated texts through in-depth analysis of vocabulary and phrases in texts. In their 2017 study [10]. conducted research based on Zipf's law [11]. Zipf's law states that in texts written by humans, the most frequently occurring word appears approximately twice as often as the second most frequent word and three times as often as the third most frequent word. Experimental results showed that the SVM (Stochastic Gradient Descent) algorithm achieved the best classification performance with complex phrase features, reaching an accuracy rate of 89.0%.

Syntactic and semantic features are also important components of text features. In 2020 study [12], Uchendu et al. explored syntactic and semantic features of texts by comparing Flesch reading ease [13] and Flesch-Kincaid grade [14] scores between AI-generated texts and human-created texts, and using simple neural network models (RNN-CNN) for detection. The study found that not all neural network generation methods can produce high-quality human-like texts; among them, texts generated by three language models (GPT2, GROVER, FAIR) were of higher quality than those generated by other models and could confuse machine classifiers.

1.2.2. Neural network model-based detection methods

Detecting AI-generated texts using neural language models is often highly effective, especially detection methods incorporating Transformer models [15–18]. Various studies have shown that Transformer models achieve state-of-the-art performance in diverse natural language processing tasks.

Pre-trained language models such as BERT and RoBERTa have demonstrated great potential in the field of natural language processing. BERT [15] is a pre-trained language model based on the Transformer architecture. Through unsupervised learning on large-scale text data, it can acquire extensive linguistic knowledge and semantic representations. RoBERTa [17] is an enhanced version of BERT. It improves performance through strategies such as longer training steps, larger batch sizes, more training data, removal of the next sentence prediction (NSP) training objective, and adoption of dynamic masking. In AI-generated text detection tasks, RoBERTa exhibits higher accuracy and better generalization ability.

The detection principle of pre-trained language models primarily relies on their ability to understand and represent language. These models can capture deep-level information and contextual details in texts, and establish statistical language models through learning from massive text data. When detecting AI-generated texts, the model judges whether the input text conforms to the habits and rules of human language the learned linguistic patterns and features. If the text contains patterns that deviate from human language norms, the model may classify it as AI-generated.

Transformer models have achieved remarkable results in natural language processing tasks. Their core lies in the attention mechanism, which allows the model to focus on information at different positions when processing texts, thereby better capturing the semantic and syntactic structures of the text.

The advantage of this method lies in its ability to analyze texts from an entirely new perspective. Traditional detection methods primarily focus on surface-level features such as vocabulary and syntax, whereas the topological data analysis-based method delves into the internal mechanisms of the model. It extracts deep-level information by analyzing the model's text processing process, enabling it to capture subtle, deep-seated textual features that are difficult to identify using traditional methods—thereby enhancing detection accuracy and reliability. Furthermore, the interpretability of topological features allows researchers to better understand the model's decision-making process, providing strong support for further optimizing detection methods. When facing evolving AI generation technologies, this model-internal-mechanism-based detection method demonstrates better adaptability and can more effectively address the challenges posed by newly emerging types of AI-generated texts.

Black-box detection is limited to application programming interface (API)-level. It mainly relies on collecting text samples from human and AI sources to train a classification model, which is then used to distinguish between AI-generated and human-generated texts. In black-box detection, researchers cannot directly access the internal structure and parameters of the model; instead, they can only analyze by inputting texts and observing the output results. The advantage of black-box detection is that it is relatively simple to implement, does not require in-depth knowledge of the model's interior, and can utilize existing datasets and models for training. However, its black-box nature severely limits interpretability, and researchers usually need to use explanation tools to understand the results of model detection.

White-box detection, on the other hand, has full access to and can control the model's generation behavior to achieve traceability of generated texts. In white-box detection, researchers can gain in-depth knowledge of the model's internal structure and parameters. By intervening in and analyzing the model's generation process, they can detect and trace the generated texts. However, white-box detection is more difficult to implement, requires in-depth understanding and permission control of the model, and may involve issues related to privacy and security.

1.2.3. Existing research achievements and shortcomings

In terms of detection accuracy, although pre-trained language models in existing research have achieved good results on specific datasets, in practical applications, facing complex and diverse texts and the continuously evolving AI generation technology, the detection accuracy still needs to be improved. Different detection methods also show significant differences when applied to different types of texts. For some complex and highly specialized texts, the detection results are often unsatisfactory.

Although some deep learning-based detection models can achieve good detection results, due to their black-box nature, researchers find it difficult to understand the decision-making process and basis of the models, which limits the application and improvement of the models to a certain extent. In scenarios where the detection results need to be explained, such as academic integrity reviews and legal disputes, models lacking interpretability cannot provide effective support.

1.3. Research content

In response to the deficiencies of existing AI-generated text detection methods in terms of accuracy, generalization ability, and interpretability, this research proposes a dual-stream feature detection model (RoBERTa-text) based on the integration of deep learning and statistical features. This

model integrates the deep semantic features extracted by the RoBERTa language model with manually designed text statistical features (such as punctuation density, standard deviation of sentence length, etc.) to achieve the dynamic fusion of deep semantic features and statistical features. Two sets of ablation experiments are set up to explore the impact of the structure and different text features on the model's detection performance.

At the data level, to enhance the model's generalization ability, this study constructs a multi-domain corpus covering the HC3 dataset, ChatGPT detection dataset, self-constructed datasets of academic abstracts and literary works, ensuring that the training data can cover texts from different fields, scenarios, and types.

To further verify the practicality of the model, this study has developed a public detection platform that provides free AI-generated Chinese text detection services to the public. The platform can detect the texts uploaded by users and generate detailed detection reports, including the text's generation probability, confidence interval, and overall statistical information.

2. Model construction

This chapter describes the systematic process of the Chinese generated text detection model, covering the entire workflow from the integration of multi-source datasets, the design of hierarchical feature fusion architecture, to the verification of model performance. Based on the HC3 dataset, ChatGPT detection dataset, and self-constructed corpus, a hybrid training set spanning multiple domains (such as encyclopedias, news, and social media) is built. Data cleaning and balanced sampling ensure the diversity and balance of training samples. A dual-path feature extraction network is designed, combining semantic encoding with statistical feature engineering, and a gated attention mechanism is adopted to realize dynamic weighted fusion of multi-dimensional features. Ablation experiments verify the effectiveness of the text feature module and the impact of classifier depth on performance, while optimizing the probability threshold division strategy to improve detection reliability. At the application level, an automatic detection report generation module is developed to realize visual output and traceability analysis of detection results, providing technical support for academic review and content supervision.

2.1. Data collection

In this study, to ensure the robustness and generalization ability of model training, a comparative dataset of AI-generated Chinese texts and human-generated texts was specifically constructed. The dataset follows the principle of data balance, ensuring a 1 : 1 ratio between human-written and AI-generated texts to effectively prevent overfitting during model training. To further enhance the model's generalization ability, data collected from multiple domains to ensure the model learns the diversity and complexity of language. This cross-domain data collection method enables the model to better understand and process texts in various linguistic contexts, thereby demonstrating higher adaptability and accuracy in practical applications.

2.1.1. HC3 Dataset

The human-ChatGPT comparison corpus (HC3) [18] dataset is a multi-domain comparative

corpus released by the Hello-SimpleAI team in early 2023. It consists of human-ChatGPT answer pairs for given questions, aiming to study differences between human expert responses and ChatGPT responses, and to support large model training. The team constructed the dataset through two data sources: for public question-answering datasets (which typically provide questions and human expert answers), questions were directly input into ChatGPT to collect its responses; additionally, high-quality concepts and explanations from Wikipedia and Baidu Encyclopedia were crawled, and ChatGPT was queried using prompts (as shown in the table below) to collect its answers. The dataset contains bilingual (Chinese and English) data covering multiple domains such as encyclopedias, law, and medicine, with three fields: "question", "human_answers" (list of human responses), and "ChatGpt_answers" (list of ChatGPT responses). In this study, only the Chinese text data was used, with 25,706 samples, including 12,853 human-generated texts and 12,853 ChatGPT-generated texts.

Table 1. HC3 dataset example.

question	I have a computer-related question, please answer in Chinese, what is the control bus
human_answers	Control bus (CB) is abbreviated CB. The control bus is mainly used to transmit control signals and timing signals.
chatgpt_answers	The control bus is a bus used to transmit control information in a computer. It connects the computer's processor, memory, and...

2.1.2. ChatGPT detection dataset

This dataset is a large-scale, diverse ChatGPT detection dataset constructed by Kang [19] and others in late 2024. It contains 180,000 pieces of ChatGPT-generated texts and 180,000 pieces of human-generated texts, both in Chinese and English, covering multiple corpus domains such as news, social media, and user reviews. In this study, only the Chinese data was used: the human-sourced texts for news data were from the THUCE news dataset, the human-sourced texts for social media data were from Weibo data, and the human-sourced texts for user review data were collected from online shopping review data.

Table 2. Distribution of ChatGPT detection datasets.

Data category	Data name	Source	Sample size
News	THUCE News	Sina Weibo	39691
Social Media	Weibo	Sina Weibo	29790
User Reviews	Online Shopping Reviews	ChineseNIpCorpus	39704

Table 3. Prompt words used in each dataset of ChatGPT detection dataset.

Dataset Name	Tips:
THUENews	As a new news editor, please help me re-edit the following news...
Weibo	Please help me rewrite and polish the following Weibo post...
Online Shopping Reviews	Please help me rewrite and polish the following user comment...

2.1.3. Self-constructed dataset

To expand the scope of text data used in this study and enhance the generalization ability of the trained model, this section constructs a hybrid text detection dataset covering two domains: academic papers and literary works. For academic abstracts, a human-authored corpus is built 4151 documents from China National Knowledge Infrastructure (CNKI), and an equal amount of AI-generated text is produced using ChatGPT-4.0. In the field of literary works, 8900 samples are extracted from classic works (such as *Records of the Grand Historian* and *Dawn Blossoms Plucked at Dusk*) and online literature, covering four genres: classical Chinese prose, modern prose, poetry, and novels. Meanwhile, the iFlytek Spark Large Model (Spark Max) is used to generate imitative texts to further improve the model's generalization ability. At the data processing stage, the key task is to address heterogeneity: academic dissertation identifiers are removed via regular expressions, and paragraph indentation and blank line formats are standardized to eliminate source-specific features.

Academic abstract texts

In this study, web crawler technology is used to retrieve data from the CNKI database. With 21 keywords as screening criteria, 4151 academic abstracts in relevant fields are successfully crawled, covering a wide range of disciplines. Simultaneously, by calling the ChatGPT-4.0 model, corresponding text content is generated for each crawled abstract. The generation process is shown in the table below:

Table4. Summary dataset example.

keywords	Title	Original Abstract	Prompt	AI-generated summary
public health	Optimization of Port State Public Health Risk Assessment Rules	[Abstract]: During the COVID-19 outbreak, port states should respond to public...	Please write a 300–400 word abstract on the topic of "Review and Optimization of Port State Public Health Risk Assessment Rules."	In the context of globalization, reviewing and optimizing port state public health risk assessment rules is particularly important...

Literary works texts

In this study, in addition to analyzing the academic abstract text data, considering that AI-generated literary works also have high significance for detection, text data of multiple literary genres—including prose, novels, poetry, and classical Chinese prose—were also collected. Specifically, for texts of classical Chinese prose, prose, and novels, paragraphs with no more than 400 characters were randomly extracted as sample data; for poetry texts, complete poems were randomly selected as samples. Through the methods, a total of 8900 sample data entries were successfully extracted and organized for subsequent analysis.

Given the wide variety of extracted text data and the lack of a unified theme in this study, adjustments were made to the prompts. Each time a request was sent to the model for text generation, the existing human-generated text was first input, followed by a requirement for the AI to imitate the aforementioned content and generate corresponding text. This approach was adopted to improve the

AI's performance in text generation tasks, bringing its output closer to the level of human creation and thereby providing richer and higher-quality text content. The specific generation process is shown in the Table 5:

Table 5. Distribution of sample sizes of different literary works.

Subject Type	Sources	Sample size
Classical Chinese	*Book of Han*	
	Book of the Later Han	
	Records of the Three Kingdoms	2000
	Records of the Grand Historian	
	Zi zhi Tong jian	
Prose	*A Village of One*	
	Reminiscences	
	Morning Blossoms Plucked at Dusk	2400
	Cultural Journey	
	Random Notes on a Journey to Hunan	
	Elegant House Sketches	
Poetry	*Tang Poetry*	
	Song Ci	2100
	Yuan Opera	
Novel	*Dou luo Da lu*	
	A Mortal's Journey to Immortality	
	In Search of the Gods	2400
	Star Changes	
	Zi Chuan	
	Zhu Xian	

Table 6. Examples of literary works datasets.

Subject Type	Text source	Original Text	Prompt	AI-generated text
Prose	*A Village of One*	As I grow older, I find it increasingly difficult to distinguish which past life situations are real...	As I age, I find it increasingly difficult to distinguish which past life situations are real...	As time goes by, it becomes increasingly difficult for me to distinguish which past life situations are real...

Given the diversified development trend of the current LLM technology ecosystem, research focusing solely on detecting texts generated by a single model can hardly meet the needs of industrial practice. On the basis of the already constructed ChatGPT-4.0 generated dataset, this study additionally uses the iFlytek Spark Large Model to generate texts, aiming to enhance the model's ability to detect texts generated by different. In this study, tagged features and features that are not suitable for model

learning in both human-generated texts and AI-generated texts were processed and cleaned, with specific measures as follows:

(1) Since all abstract texts crawled from CNKI start with the identifier Abstract, while abstracts generated do not have this identifier, this phrase at the beginning of all CNKI-crawled abstracts was removed as an abnormal term to avoid overfitting during subsequent model training.

(2) Observations of the collected texts revealed that many CNKI-crawled literature abstracts are derived from academic dissertations, so terms such as Degree-Awarding Institution and Degree Level often appear at the end of these abstracts.

(3) Beyond the removal of special identifiers, data processing was also conducted from the perspective of text formatting. Texts obtained typically have blank lines between paragraphs and no first-line indentation, whereas texts crawled from CNKI usually have first-line indentation and no blank lines between paragraphs. This ensured the formatting of the two text types was consistent without altering the text content.

Table 7. Summary of all model datasets.

Data Source	Text Subject	Number of human texts	Number of AI-generated texts
HC3 Dataset	Open	3293	3293
	Law	372	372
	Encyclopedia	4617	4617
	Medical	1074	1074
	School Knowledge	1709	1709
	Psychological Issues	1099	1099
	Financial Issues	689	689
ChatGPT Detection Dataset	News	39691	39691
	Weibo	29790	29790
	Shopping Reviews	39704	39704
Self-built Dataset	Academic Abstracts	4151	4151
	Classic Chinese	1982	1982
	Fiction	2376	2376
	Poetry	2088	2088
	Prose	2217	2217

2.2. Model architecture

This study adopts a dual-stream feature fusion architecture. It extracts deep semantic features using a pre-trained language model, combines them with manually designed text statistical features, and finally implements classification decisions through a multi-layer perceptron (MLP). The model is mathematically expressed as follows:

$$\hat{y} = MLP(RoBERTa(x) \oplus S(x)) \tag{1}$$

where $\oplus$ denotes the feature concatenation operation, and $S(x)$ represents the statistical feature vector.

2.2.1. Feature extraction module

This study employs the RoBERTa-wwm-ext-large pre-trained model. As the Large version of the RoBERTa model, it consists of 24 Transformer layers and 16 attention heads, with a hidden layer dimension of 1024. Suitable for processing long texts, it supports a maximum input of 512 tokens, making it applicable to this research.

Statistical feature design

The statistical feature design in this study is based on the behavioral characteristic hypothesis of language generation models: AI-generated texts exhibit quantifiable statistical biases in aspects such as punctuation distribution, sentence structure complexity, and language rhythm control. Therefore, six text statistical features are extracted in the study, namely the occurrence counts of commas, periods, semicolons, and enumeration commas (dunhao), as well as the number of sentences and the standard deviation of sentence lengths.

$$s(x) = [\tau_c, \tau_p, \tau_s, \tau_t, n_s, \sigma_l]^T \tag{2}$$

Table 8. Text feature calculation method.

Characteristic symbol	Characteristic name	Mathematical expression
τ_c	Comma density	$\tau_c = \sum_{i=1}^{n} \prod (c_i = ',')$
τ_p	Period density	$\tau_p = \sum_{i=1}^{n} \prod (c_i = \{'。','.'\})$
τ_s	Semicolon density	$\tau_s = \sum_{i=1}^{n} \prod (c_i = ';')$
τ_t	Comma density	$\tau_t = \sum_{i=1}^{n} \prod (c_i = '、')$
n_s	Number of sentences	$n_s = \|S\|, S = \{s_j \mid s_j \in split(text, [。 ！ ？ .])\}$
σ_l	Standard deviation of sentence length	$\sigma_l = \sqrt{\frac{1}{n_s - 1} \sum_{j-1}^{n_s} (l_j - \bar{l})^2}$

2.2.2. Feature fusion and classification

Multi-source feature fusion is achieved through tensor concatenation and non-linear transformation:

The semantic vector $\mathrm{h}_{\mathrm{cls}}$ and statistical vector S are concatenated along the feature dimension. This preserves the independence of the two feature spaces while establishing cross-modal correlations, thereby constructing a joint representation:

$$h_{fusion} = Concat(h_{cls}, S) \in R^{d+6} \tag{3}$$

The refined 512-dimensional features are mapped to a binary class space through fully connected layers, generating unnormalized class scores. This process learns the contribution weights of different feature dimensions to the classification boundary and establishes a mapping relationship between feature combinations and class labels. The Softmax function is then used to convert the class scores into a probability distribution, ensuring the output conforms to probability properties. Additionally, classification sensitivity can be flexibly adjusted by setting thresholds, achieving a balance between precision and recall.

2.2.3. Model performance evaluation

In this study, human-written texts are labeled as 0, and AI-generated texts are labeled as 1. The performance of the classification model is quantified using four core elements of the confusion matrix:

- True positive (TP): The number of AI-generated texts correctly identified.
- False positive (FP): The number of human-written texts incorrectly classified as AI-generated texts.
- True negative (TN): The number of human-written texts correctly identified.
- False negative (FN): The number of AI-generated texts incorrectly classified as human-written texts.

The structure of the confusion matrix in this study is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision: quantifies the reliability of the model's predictions for positive-class samples. It is defined as the proportion of correctly predicted positive-class samples among all samples predicted as positive-class. In this experiment, it represents the proportion of samples correctly detected as AI-generated among all samples detected as AI-generated.

Recall evaluates the completeness of the model in identifying truly positive samples. It is calculated as the proportion of correctly predicted positive samples among all actual positive samples. In this experiment, it represents the proportion of samples correctly detected as AI-generated among all actual AI-generated samples. Its calculation formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The F1-score synthesizes precision and recall through a harmonic mean, providing a single comprehensive evaluation metric. The harmonic mean assigns higher weights to smaller values, and when either metric approaches 0, the F1-score decays rapidly. Its calculation formula is as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

In this following, using all data in the dataset, the model combining RoBERTa with text features (hereafter abbreviated as RoBERTa-text) is trained and tested on the dataset (with 80% of the data as

the training set and 20% as the test set). The training results of this model are compared with the test results of different detection models such as CNN and RNN. The comparison results are shown in the table below:

Table 9. Comparison of different detection models.

Model	Accuracy	Precision	Recall	F1
RoBERTa-text	96.94%	96.85%	97.61%	97.23%
RoBERTa	93.09%	92%	95.77%	93.85%
BERT	91.36%	88.46%	96.93%	92.50%
CNN	88.50%	90.94%	87.80%	89.34%
RNN	87.63%	88.17%	89.47%	88.81%
LSTM	87.40%	87.85%	89.39%	88.62%
GRU (gated recurrent unit)	87.90%	86.95%	91.71%	89.27%

As shown in the table above, the RoBERTa-text model achieves the optimal values across all four core metrics. Among these, its accuracy (96.94%) and F1-score (97.23%) are 3.85 percentage points and 3.38 percentage points higher than those of the RoBERTa model, respectively. In contrast, the accuracy of traditional sequence models (RNN/LSTM/GRU) is generally below 88%, showing a significant gap compared with pre-trained models based on the Transformer architecture (such as BERT and RoBERTa). This indicates that traditional sequence models have limitations in capturing the deep semantic features of generated texts. Notably, although the BERT model performs well in terms of recall (96.93%), its precision (88.46%) is significantly lower than that of RoBERTa-text (96.85%), revealing an obvious defect in the BERT model regarding false positive control.

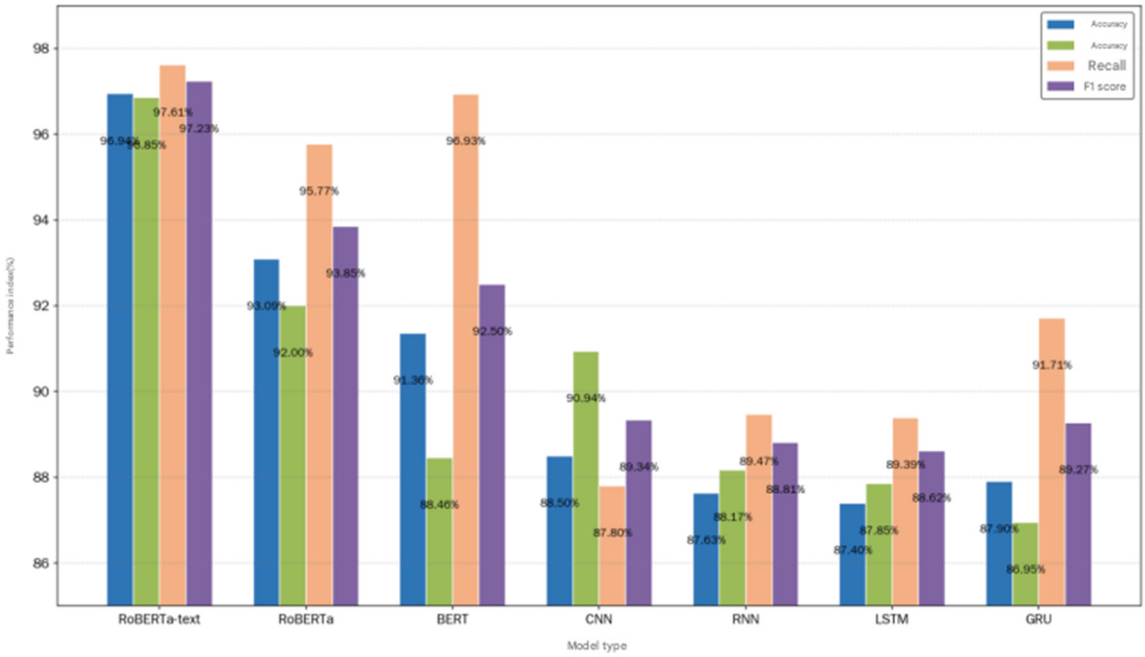

Figure 1. Histogram of performance of each model.

2.3. Ablation study

An ablation study aims to verify the contribution of each model component to the final performance by systematically removing or modifying key components of the model. In this section, two groups of ablation experiments are conducted on the proposed RoBERTa-text fusion model to verify: 1) the impact of the text feature module and classifier depth on performance, and 2) the impact of the six text features on performance.

2.3.1. Impact of the text feature module and classifier depth

In this experiment, the previously used dataset is adopted for model training and testing. The division of the training set and test set strictly follows the model training process described in Section 3.2, ensuring the consistency and reproducibility of the experiment. The experimental configuration is as follows:

Table 10. Experiment (I)

Experimental Groups	Model Configuration	Theoretical Assumptions
Baseline Model	RoBERTa + 6-dimensional statistical features + 2-layer classifier	Complete model performance benchmark
Experimental Group 1	RoBERTa + 2-layer classifier	Verify the effectiveness of text features
Experimental Group 2	RoBERTa + 6-dimensional statistical features	Analyze the feature fusion capabilities of deep structures

The baseline model integrates the RoBERTa-text pre-trained architecture, six-dimensional text statistical features, and a two-layer classifier, aiming to establish a performance benchmark for the synergistic effect of cross-modal features. Experimental Group 1 removes the manually designed text statistical features to focus on verifying the independent contribution of punctuation distribution rules and sentence structure features to generated text detection, and analyzes the value of traditional text statistical features in deep learning frameworks. Experimental Group 2 uses a linear classifier to replace the deep network structure, analyzing the key role of non-linear transformation in the hidden layer in feature fusion. The results obtained after the experiment are shown in the table:

Table 11. Ablation experiment (I) model performance.

Model	Accuracy	Precision	Recall	F1
Baseline Model	96.94%	96.85%	97.61%	97.23%
Experimental Group 1	94.26%	94.59%	95%	94.80%
Experimental Group 2	92.80%	91.26%	96.10%	93.62%

The results of the ablation study show that statistical features and the deep classifier structure have differential impacts on performance, and there is a synergistic enhancement effect between the two.

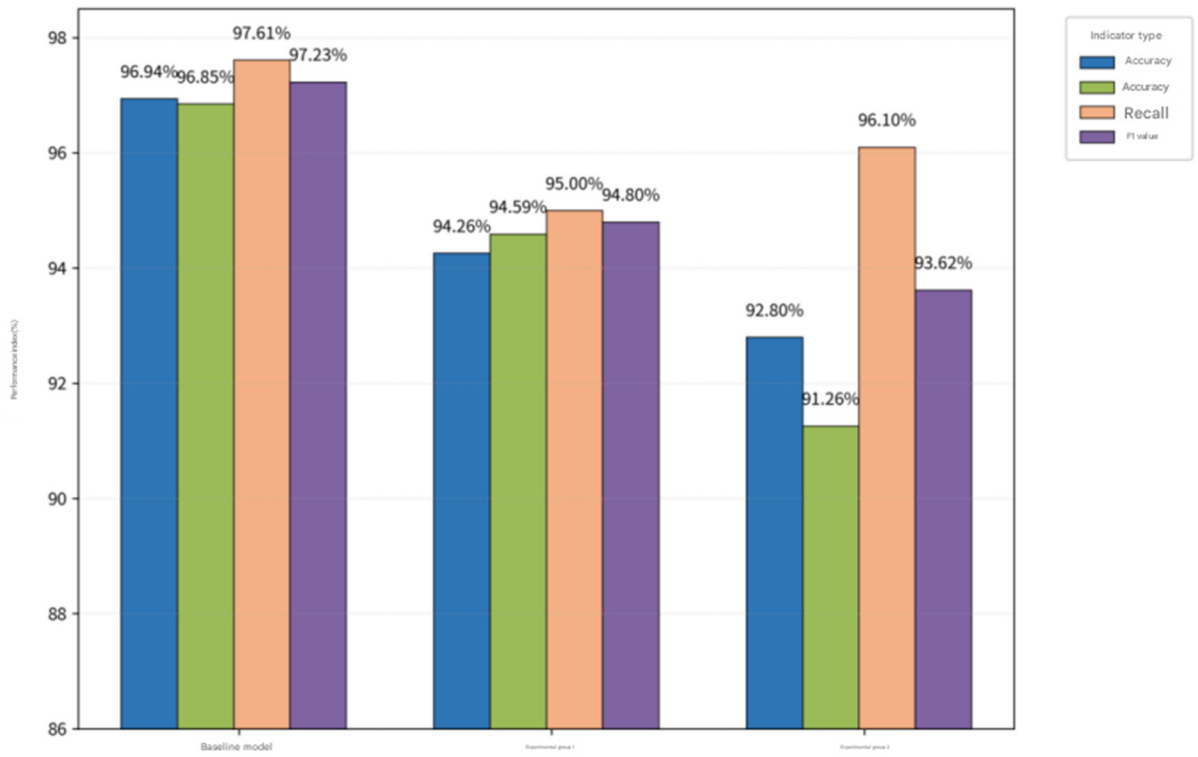

Figure 2. Experiment (I).

Compared with the baseline model, Experimental Group 1 shows a 2.26% decrease in precision and a 2.61% decrease in recall. This proves that statistical features can simultaneously improve the model's discrimination accuracy for positive and negative samples through the detection of punctuation distribution and sentence structure analysis. In terms of the deep classifier dimension, the precision of Experimental Group 2 decreases significantly by 5.59%, indicating that the deep network can effectively learn the combination patterns of semantic features and statistical features through ReLU (rectified linear unit, an activation function) activation functions and multi-layer non-linear transformations, thereby suppressing the misjudgment of false positive samples.

2.3.2. Importance analysis of multi-dimensional statistical features

In this experiment, the control variable method is used to construct multi-dimensional ablation experiments, aiming to systematically evaluate the contribution of six types of Chinese text statistical features to the generated text detection model. The experimental system includes 1 baseline model and 6 ablation experimental groups. In each group of experiments, specific statistical features are removed through a dynamic feature exclusion mechanism, and the experimental parameters are strictly kept consistent to ensure the comparability of results.

By constructing six groups of control experiments, the independent contribution of text statistical features to the generated text detection model is systematically evaluated. The baseline model integrates the RoBERTa pre-trained language model and six-dimensional manual features (comma density, period density, semicolon density, enumeration comma (dunhao) density, number of sentences, and standard deviation of sentence lengths), serving as a reference benchmark for performance evaluation.

Table 12. Experiment (II)

Ablation features	Accuracy	Precision	Recall	F1
Baseline model	96.94%	96.85%	97.61%	97.23%
Number of commas	96.46%	97.55%	95.98%	96.76%
Number of periods	96.42%	96.32%	97.20%	96.76%
Number of semicolons	96.33%	96.08%	97.29%	96.69%
Number of commas	96.17%	96.15%	96.92%	96.53%
Number of sentences	96.01%	96.19%	96.58%	96.38%
Standard deviation of sentence length	94.12%	92.15%	97.61%	94.81%

The experiments indicate that punctuation density features have differential impacts on model performance: removing the comma density feature leads to a 1.63-percentage-point decrease in recall (dropping to 95.98%), while precision increases by 0.70 percentage points (rising to 97.55%), resulting in a trade-off between the two metrics. The ablation of the period density feature causes a slight 0.53-percentage-point decrease in precision (falling to 96.32%), yet recall remains at a high level of 97.20%. The removal of semicolon density and enumeration comma (dun hao) density features results in a 0.54-percentage-point and 0.70-percentage-point decrease in F1-score.

Figure 3. Experiment (II).

In terms of sentence structure features, the ablation of the sentence count feature caused the model's accuracy to decrease by 0.93 percentage points to 96.01%. Meanwhile, the removal of the sentence length standard deviation feature triggered a systemic performance decline, with accuracy

plummeting by 2.82 percentage points to 94.12% and precision dropping by 4.7 percentage points, although recall remained stable at the baseline level. This parameter variation phenomenon suggests that as an indicator of text complexity, the standard deviation plays a crucial discriminative role in the feature space. Its ablation significantly weakens the model's ability to perceive variability.

These experimental results indicate that the sentence length standard deviation feature plays a key role in enhancing the model's ability to discriminate generated texts, and its absence significantly affects the model's accuracy and precision. In contrast, although comma, period, semicolon, and enumeration comma (dunhao) density features contribute to model performance (in the order of comma > period > semicolon > dunhao), their impact is relatively moderate. The sentence count feature also has a limited impact on performance.

2.4. Model application

The trained AI-generated Chinese text detection model, this section builds a public website platform (freeaidetect.top) to provide free AI-generated Chinese text detection services. The platform can segment the text uploaded by users, implement text detection, and generate complete detection reports. It has provided detection services for more than 1400 times.

2.4.1. Probability threshold division

To enhance the detail and comprehensiveness of detection reports, this study performed sigmoid normalization on the output results. This process maps the output of the linear layer to the interval of 0 to 1 and converts it into probability values, thereby transforming the model's original output classification results, i.e., AI-generated (label 1) and human-generated (label 0), into corresponding generation probability expressions.

After normalization, this study further introduced the concept of confidence thresholds, using three intervals of 70%–80%, 80%–90%, and 90%–100% to distinguish the sources of text generation under different confidence levels. By setting thresholds, high-confidence AI-generated texts can be screened out, thereby improving the accuracy of detection.

2.4.2. Establishment of detection system

This study established a deep learning-based AIGC detection report generation system, whose architecture includes four major modules: text preprocessing, model inference, visualization, and statistics generation. By integrating semantic feature analysis and text statistical features, the system constructs a multi-dimensional detection system and finally generates beautiful and professional detection reports.

The report presents the segmentation results of the detected text, with each sentence followed by the probability (determined by the model) that the text may be AI-generated. The confidence threshold, the system classifies the text into categories corresponding to different confidence intervals, such as "highly suspected AI-generated", "moderately suspected AI-generated", or "mildly suspected AI-generated", allowing users to intuitively understand the source of the text's generation and its credibility.

In addition, the report summarizes statistical information of the overall detection results, including the total number of detected characters, the number of manually written characters, the proportion of manual content, the number of characters suspected to be AI-generated (AIGC), the proportion of suspected AIGC characters in the entire text, and the distribution of text across different confidence intervals. Furthermore, texts belonging to different confidence interval categories are marked with different colors in the original text: mild suspicion is marked in light yellow, moderate suspicion in dark yellow, and high suspicion in red.

Users can conduct further analysis and processing of the text based on the detailed information in the detection report. For instance, regarding text paragraphs marked as "Highly Suspected AI-Generated", users can conduct more in-depth reviews or use other auxiliary methods for verification to ensure the authenticity and accuracy of the text.

3. Analysis of multi-model generated text detection

This focuses on researching the generalization ability of AI-generated text detection models and the characteristic differences between texts from different sources. Deepseek R1, Phi4, and Qwen 2.5 (all in their 14B parameter versions)—were selected to generate texts based on 4,000 academic abstracts from CNKI, constructing a dataset containing human-written texts and texts generated by the three models. After preprocessing steps such as format unification and noise removal, the dataset is used to compare the word frequency distributions between CNKI human-written texts and the three types of model-generated texts. With the help of Jieba word segmentation and text perplexity analysis, differences in vocabulary usage habits and semantic expression features among texts from different sources are revealed. The RoBERTa-text detection model is employed to evaluate its detection performance on texts from different sources using metrics such as accuracy and precision, and differences in detection performance are analyzed through confusion matrix heatmaps. For true positive, false positive, true negative, and false negative texts in the detection results, statistical analysis of text features (e.g., punctuation usage, sentence complexity) and linguistic analyses (e.g., part-of-speech tagging, dependency syntax) are conducted to identify key factors contributing to misjudgments.

3.1. Data collection and processing

3.1.1. Data collection

In the model training process described in Chapter 2, a large number of AI-generated texts from ChatGPT and iFlytek Spark Max (Xinghuo) were used. To test the model's generalization ability and avoid over-sensitivity to texts generated, three other well-known large language models were selected in this section for text generation and detection research.

Deepseek R1: Developed by DeepSeek, DeepSeek-R1 is designed for search-enhanced scenarios, built on a trillion-parameter architecture. It adopts a multi-stage progressive training strategy and deeply integrates retrieval-augmented generation (RAG) technology, demonstrating precise information localization and reasoning capabilities in tasks such as open-domain question answering and fact-checking.

Phi4: Developed by Microsoft Research, Phi-4 is positioned as an efficient lightweight language model with an innovative modular sparse architecture.

Qwen 2.5: Qwen2.5 is the second-generation large model launched by Alibaba Cloud's Tongyi Laboratory, with its parameter scale expanded to 72 billion. Its technical highlights include an explainability enhancement module (XAI), a dynamic risk control mechanism, and support for three deployment modes: API, software development kit (SDK), and privatization.

Given the extensive application value and practical significance of AI-generated text detection technology in the field of academic literature detection, academic abstracts were chosen as the experimental objects. Therefore, this study crawled 4000 academic paper abstracts from the CNKI database and used the three different to generate 4000 corresponding academic paper abstracts.

Table 13. Comparison of data examples in this experiment.

Source	Data example
CN Academic Abstract	Abstract: To predict the evolution of public health events, a warning model for secondary derivative events of public health events based on a causal graph is proposed. A two-layer ontology is used to represent structured scenarios. Event extraction is achieved by combining a pre-trained model with a deep neural network. Pattern matching is used to extract logical relationships between events.
Prompt	Please write a 300–400 word abstract on the topic of "Research on the Construction of Early Warning Model for Secondary Derivative Events of Public Health Events Based on the Schema of Reasoning"
Deepseek R1 (14 B)	Research on the Construction of a Warning Model for Secondary Derivative Events of Public Health Events Based on the Schema of Reason With the frequent occurrence of public health events around the world, the impact of secondary derivative events (such as social unrest, economic impact, and the spread of health crises) on society is becoming increasingly significant...
Phi4 (14 B)	With globalization and the development of information technology, public health incidents are becoming increasingly complex and frequent. Facing this challenge...
Qwen 2.5 (14 B)	With rapid socioeconomic development and increased population mobility, public health incidents are occurring more frequently and their impacts are becoming increasingly complex. Traditional public health emergency management systems are struggling to cope with the challenges posed by these emergencies...

3.1.2. Data processing

As can be seen from the data samples presented, both the original abstract texts obtained from CNKI and the abstract texts generated by different large language models exhibit their own unique characteristics. To ensure these characteristics do not interfere with model testing, all text data were preprocessed, with specific measures as follows: the original abstracts crawled from CNKI: uniformly prefixes such as Abstract from the beginning of the texts. For the abstract texts generated by Deepseek R1: uniformly remove the prefix Paper Title and blank lines within the texts. For the abstract texts generated by Phi4: uniformly the prefix Abstract and blank lines within the texts. For the abstract texts generated by Qwen2.5: uniformly remove blank lines within the texts.

3.2. Comparative analysis of texts from different sources

To investigate the differences between texts generated from different sources, this section takes four groups of texts as experimental groups—abstracts crawled from CNKI, abstracts generated by Deepseek R1, abstracts generated by Qwen2.5, and abstracts generated by Phi4. Frequency analysis and text perplexity analysis are conducted on these groups, aiming to explore differences in vocabulary usage habits and semantic expression among texts from different sources.

3.2.1. Experimental design and process

This experiment consists of four experimental groups, with the grouping method shown in Table 14–16 below. Through this grouping, the differences between human-written texts and texts generated are compared and analyzed, and the differences between texts generated are also studied.

Table 14. Experimental minutes and indicator calculation method.

Experimental Groups	Data source	Number of samples (items)	frequency distribution calculation method	Text perplexity calculation model
Experimental Group 1	CNKI	4000	wordcloud+jieba	Wenzhong2.0-GPT2-110M
Experimental Group 2	Deepseek R1 (14 B)	4000		
Experimental Group 3	Qwen2.5 (14 B)	4000		
Experimental Group 4	Phi4 (14 B)	4000		

By comparing and analyzing the word frequency distribution characteristics of real abstracts from CNKI and abstracts generated by AI models such as Deepseek, Qwen2.5, and Phi4, differences between human creation patterns and machine generation patterns of academic texts can be identified in terms of linguistic structure, semantic expression, and knowledge organization.

Perplexity is a metric used to evaluate how well a language model predicts a sample. A lower perplexity value indicates that the model predicts the text more accurately, meaning the model has stronger ability to understand and generate the text; conversely, a higher perplexity value means greater uncertainty in the model's predictions and weaker ability to process the text. In the experiment, by calculating the perplexity of texts from different sources (CNKI abstracts and texts generated), the quality of texts from different sources can be compared.

3.2.2. Word frequency comparative analysis

In this experiment, the text data of the four experimental groups were segmented using jieba word segmentation technology. Meanwhile, the Harbin Institute of Technology (HIT) stopword list (hit_stopwords) was used to identify and remove noise words from the texts. The word cloud

visualization tool (wordcloud) was employed to generate the word cloud diagrams.

By analyzing and comparing the high-frequency keyword distribution charts of each experimental group, it can be clearly observed that the occurrence frequency of high-frequency words in Experimental Groups 2, 3, and 4 is significantly higher than that in Experimental Group 1. Specifically, in these three groups (Experimental Groups 2, 3, and 4), even the high-frequency word with the lowest occurrence count (3202 times) exceeds the highest occurrence count of high-frequency words in Experimental Group 1 (3072 times). More notably, the highest occurrence counts of high-frequency words in Experimental Groups 2, 3, and 4 even reaches 11,706 times.

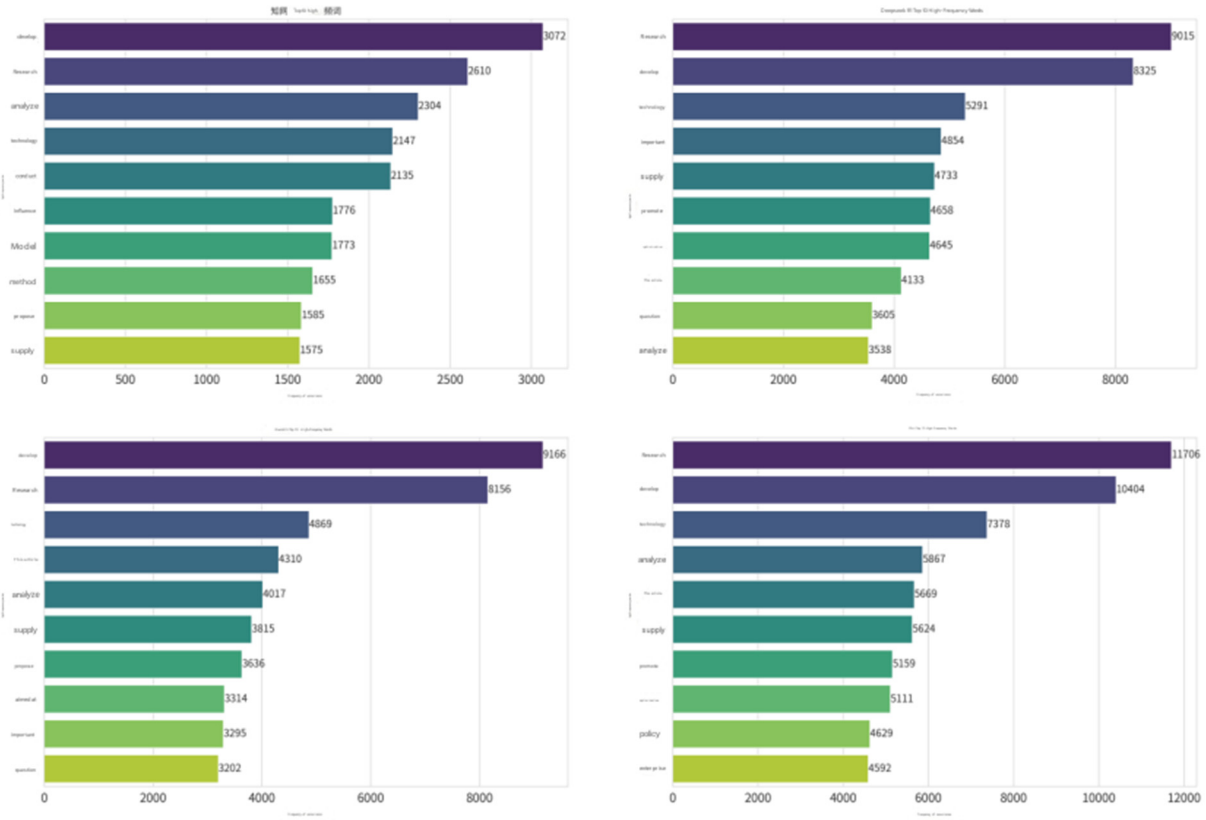

Figure 4. Comparison of the distribution of the top 10 high-frequency keywords in each experimental group.

3.2.3. Comparative analysis of text perplexity

In this study, accuracy is defined as the degree to which the generated text conforms to the semantic characteristics and linguistic patterns of human-written academic abstracts. To quantitatively evaluate accuracy, we adopted a dual evaluation approach:

Perplexity-based accuracy: perplexity (PPL) serves as an indirect measure of text accuracy. Lower perplexity values indicate higher text fluency and predictability, which are key indicators of text accuracy in language generation tasks.

Semantic similarity analysis: we employed cosine similarity to measure the semantic similarity between generated texts and human-written abstracts. The BERT-base model was used to extract

semantic embeddings, and the average cosine similarity score across all text pairs was calculated.

To further explore the differences in semantic expression among abstract texts from different sources, this experiment adopted a text perplexity calculation method to analyze the perplexity of the abstract text data from the four experimental groups. The "Wenzhong-GPT2-110M" model was used in the experiment to calculate text perplexity.

As shown in Figure 5, the four probability density function (PDF) plots clearly illustrate the distribution differences of different generation models in terms of perplexity. Specifically, the perplexity of abstract texts from CNKI shows a relatively flat distribution in the 5–50 PPL interval, with the peak of its probability density concentrated around 15. In contrast, the PPL curve distributions of texts generated by Deepseek R1 and Qwen2.5 are relatively concentrated; compared with CNKI abstracts, their perplexity peaks shift leftward to around 10, and the values in the tail do not exceed 20. For the perplexity of abstract texts generated by Phi4, as shown in Plot (b), it exhibits a high degree of overlap with that of CNKI abstract texts. However, its overall distribution trend is slightly left of CNKI texts, with a peak around 13, showing a relatively flat distribution in the 0–45 interval.

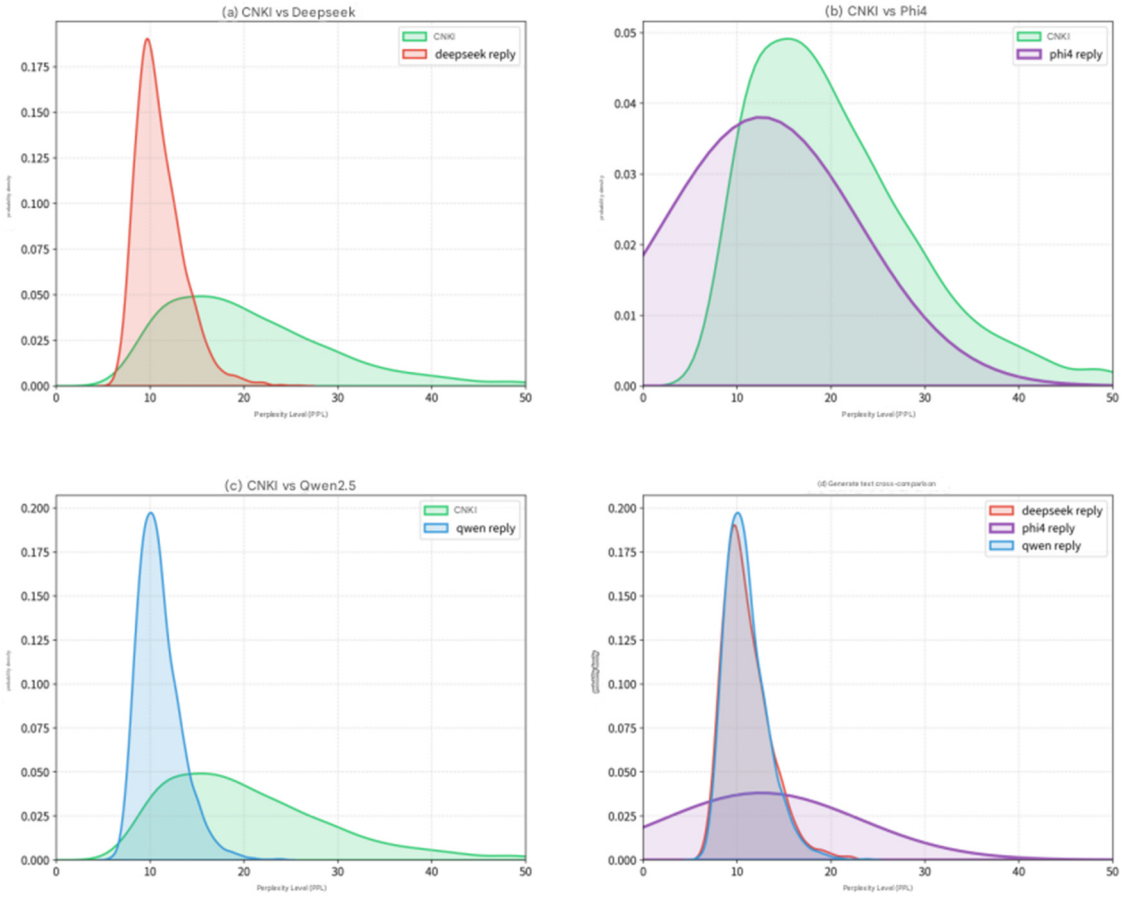

Figure 5. Comparative distribution curve based on perplexity.

When comparing the perplexity distribution between AI model-generated abstracts and CNKI abstracts, significant differences can be observed in the overlap rate of perplexity probability density between the abstracts generated by each model and the CNKI abstracts. Among them, the abstracts generated by Deepseek R1 and Qwen2.5 have a low overlap rate (only 32%) with CNKI abstracts in

terms of perplexity distribution, indicating a large degree of difference. In contrast, the abstracts generated by Phi4 have a high overlap rate with CNKI abstracts in perplexity distribution, which suggests a high similarity to CNKI abstracts in semantic expression.

When comparing the perplexity distribution of abstracts generated, the abstracts generated by different models, the abstracts generated Deepseek R1 and Qwen2.5 show high similarity in text perplexity, with an extremely high overlap rate between them. This implies that these two models may adopt similar strategies or algorithms when generating texts. However, the perplexity distribution of abstracts generated by Phi4 appears smoother.

These experimental results reveal significant differences in semantic expression between different and human-written texts during the text generation process. Specifically, the texts generated by Deepseek R1 and Qwen2.5 show a relatively concentrated trend in the perplexity metric, with overall low perplexity values. This indicates that these two models perform quite well in terms of semantic coherence and predictability when generating texts. On the other hand, the perplexity distribution of texts generated by Phi4 is more similar to that of CNKI abstracts.

3.3. Detection analysis of texts from different sources

In this study, the RoBERTa-text detection model was used to detect texts generated by the three. The detection results, the quality of texts generated by different models was analyzed, and the generalization performance of the detection model was evaluated.

3.3.1. Experimental design and process

Three experimental groups were constructed in this study, corresponding to the three different. The performance of the RoBERTa-text model in detecting texts from different generation sources was evaluated using four evaluation metrics: accuracy, precision, recall, and F1-score. Each experimental group consists of 4000 abstracts generated by the corresponding and 4000 abstracts crawled from CNKI, aiming to balance the number of samples in different label categories.

Table 15. Experimental groups and configurations.

Experimental Groups	AI generated text source	human text source	Detection model	Hardware configuration
Experimental Group 1	Deepseek R1 (14 B) 4000	CNKI crawled abstracts 4000	RoBERTa-text model	RTX 4090 (24 GB) PyTorch 2.5.1
Experimental Group 2	Phi4 (14 B) 4000			
Experimental Group 3	Qwen 2.5 (14 B) 4000			

3.3.2. Comparison of detection performance for texts from different sources

Invoking the RoBERTa-text model to detect texts generated by different, the detection results are

shown in the table below. The results indicate that the three experimental groups exhibit differences in performance when detecting texts generated by different:

Experimental Group 2 achieves the highest values across all metrics (accuracy, precision, recall, and F1-score). In particular, its recall rate reaches 100%, which means the model does not miss any texts generated by Phi4 and can identify such texts comprehensively and accurately. Experimental Group 1 has a recall rate second only to Experimental Group 2, demonstrating strong AI text recognition capability. However, its precision is slightly lower, resulting in an overall balanced performance.

Experimental Group 3 shows the lowest values across all metrics: an accuracy of 88.91%, a precision of 86.06%, a recall of 92.88%, and an F1-score of 89.34%. This reflects that the model faces significant challenges in detecting this type of text, with both missed judgments and misjudgments occurring.

The RoBERTa-text model exhibits differences in its generalized detection ability for texts generated by different. Specifically, the model achieves the best detection performance for abstracts generated by Phi4, followed by those generated by DeepSeek R1, while its detection performance for texts generated by Qwen 2.5 is relatively weaker.

Table 16. Multi-model generated text detection results.

Experimental Groups	Accuracy	Precision	Recall	F1
Experimental Group 1	91.15%	86.61%	97.35%	91.67%
Experimental Group 2	92.48%	86.92%	100.00%	93.00%
Experimental Group 3	88.91%	86.06%	92.88%	89.34%

This can also be interpreted as follows: among the abstracts generated by these three, the Chinese abstracts generated by Qwen 2.5 are the closest to human-generated abstracts, followed by those generated by Deepseek R1, and finally those generated by Phi4.

3.4. Inter-group difference analysis based on classification results

To further improve the classification performance of the trained model in subsequent work, this section four experimental groups using the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) texts from the model detection results in Section 4.2. The goal is to study the RoBERTa-text model's sensitivity to different types of texts and the specific causes of misjudgments. By conducting an in-depth analysis of the features of these texts, we aim to identify the key factors leading to model misjudgments, thereby optimizing the structure or training strategy in a targeted manner.

3.4.1. Experimental design and process

A total of four experimental groups set up in this study: first group focuses on TP texts—texts accurately identified as AI-generated. By conducting an in-depth analysis of the linguistic features, perplexity, and other metrics of these texts, this group aims to explore how the model effectively recognizes the specific characteristics of AI-generated texts.

The second group focuses on FP texts—human-written texts incorrectly classified as AI-

generated. The purpose of this group is to identify the potential causes of model misjudgments, which may involve similarities between certain features in these texts and those of AI-generated texts, or the model's insufficient understanding of the diversity of human-written texts.

The third group focuses on TN texts—texts accurately identified as human-created. Analyzing the features of TN texts helps deepen the understanding of the model's ability to distinguish between human-written and AI-generated texts, and provides a basis for further model optimization.

The fourth group focuses on FN texts—AI-generated texts incorrectly classified as human-created. This group is particularly important for improving the model's detection accuracy. In this study, an in-depth analysis will be conducted on the features of FN texts that lead to misjudgments.

Given that six different text features (such as the number of commas) already used in the process of training the detection model, this experiment first conducts statistical and comparative on the text features of the four experimental groups. Meanwhile, to more deeply distinguish the differences between texts in different experimental groups, linguistic analysis—specifically part-of-speech tagging analysis and dependency parsing analysis—is performed on them.

3.4.2. Statistical analysis of text features

From the text feature statistics presented in the table, it can be observed that there are significant differences between the experimental groups in terms of punctuation usage and sentence complexity:

The average number of commas, periods, enumeration commas (dunhao), and sentences in the texts of the TP group is significantly higher than that in the other three groups, while the average standard deviation of sentence length is significantly lower. This indicates that the texts in the TP group have more complex sentence structures but relatively stable sentence lengths.

Table 17. Differences in text features between experimental groups.

Experimental group	Number of commas	Number of periods	Number of semicolons	Number of colons	Number of sentences	Standard deviation of sentence length
TP group	18.01	10.67	0.85	4.34	10.93	18.11
FP group	10.09	4.99	0.52	3.24	6.04	26.00
TN group	7.69	3.63	0.47	3.44	4.64	32.97
FN group	11.81	6.38	1.57	3.57	6.50	26.57

The average standard deviation of sentence length in the TN group is the largest among all experimental groups, which means the texts in the TN group have highly variable sentence lengths and more flexible sentence patterns.

The average number of semicolons in the FN group is significantly higher than that in the other three groups, suggesting that the texts in the FN group express more complex logical relationships between clauses.

To further intuitively understand and analyze the differences in text features among the four experimental groups, the following heatmap generated. As can be seen from the figure, the experimental groups show a significant stratification phenomenon in terms of syntactic complexity

indicators. Among these indicators, the standard deviation of sentence length has a large difference (ranging from 18.1 to 33.0); in contrast, the discriminative contribution of the number of semicolons (ranging from 0.5 to 1.6) and enumeration commas (dunhao) (ranging from 3.2 to 4.3) is relatively weak.

The TP and FP groups exhibit significant punctuation density and sentence length regularity in terms of syntactic features. The number of commas (TP: 18.0, FP: 10.1) and periods (TP: 10.7, FP: 5.0) in the texts of these two groups are both higher than the average value of all samples. Especially in the true positive (TP) samples, the comma density is much higher than that of human-written texts. There is a significant gap in the standard deviation of sentence length between the TP and FP groups, indicating that the model may take moderate sentence length fluctuations (ranging from 18.1 to 26.0) as a basis for judging AI-generated texts.

The TN and FN groups are characterized by syntactic diversity and punctuation dispersion. The standard deviation of sentence length of true negative (TN) samples the maximum value in the entire range (33.0), which confirms the inherent characteristic of random fluctuations in sentence length in natural language creation. In terms of the number of periods used and the number of sentences, the TN and FN groups are generally lower than the TP and FP groups.

3.4.3. Linguistic analysis

In this experiment, the Chinese multi-task language analysis tool of the Han language processing (HanLP) framework was used, and the efficiently learning an encoder that classifies token replacements accurately (ELECTRA) pre-trained model was adopted to implement part-of-speech tagging and dependency parsing.

HanLP is an open-source natural language processing toolkit focused on Chinese information processing tasks. Its core advantages lie in its multi-task collaborative processing architecture and modular design. ELECTRA, on the other hand, is a pre-trained language model based on the Transformer architecture this toolkit. Through an efficient replacement detection pre-training mechanism and multi-task collaborative learning, it significantly enhances the contextual sensitivity and complex structure parsing capabilities of Chinese part-of-speech tagging and syntactic analysis.

As can be seen from the part-of-speech tagging distribution histogram in Figure 6, nouns are the most frequently occurring words in all experimental groups, accounting for approximately 45% of the total. This phenomenon is related to the theme: text data in this experiment are literature abstracts, which involve a large number of nouns. This distribution pattern is consistent with the typical characteristics of academic texts. In addition, there are no significant differences in the distribution of other major word classes (such as verbs) among the four experimental groups, indicating that the part-of-speech distribution does not have a systematic impact on the model's classification results.

However, as can be seen from the dependency syntax distribution histogram below, the distribution proportions of various dependency syntax relationships among the four experimental groups are basically consistent. Nevertheless, the proportion of compound noun modification relationships in the text data of Experimental Group 3 and Experimental Group 4 is significantly higher than that in Experimental Group 1 and Experimental Group 2. In academic texts, compound noun modification relationships are usually used to describe concepts or things accurately.

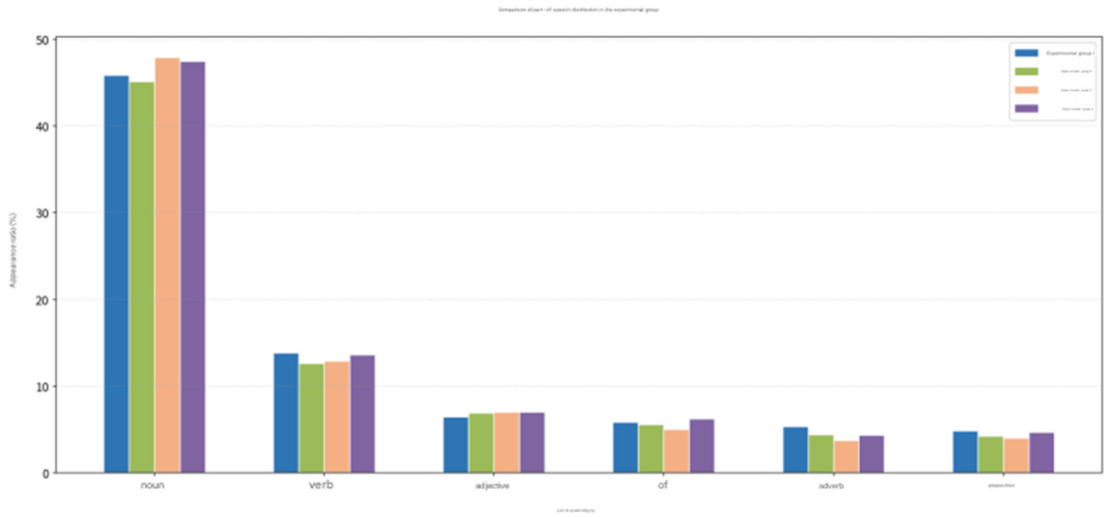

Figure 6. Part-of-speech tagging distribution histogram.

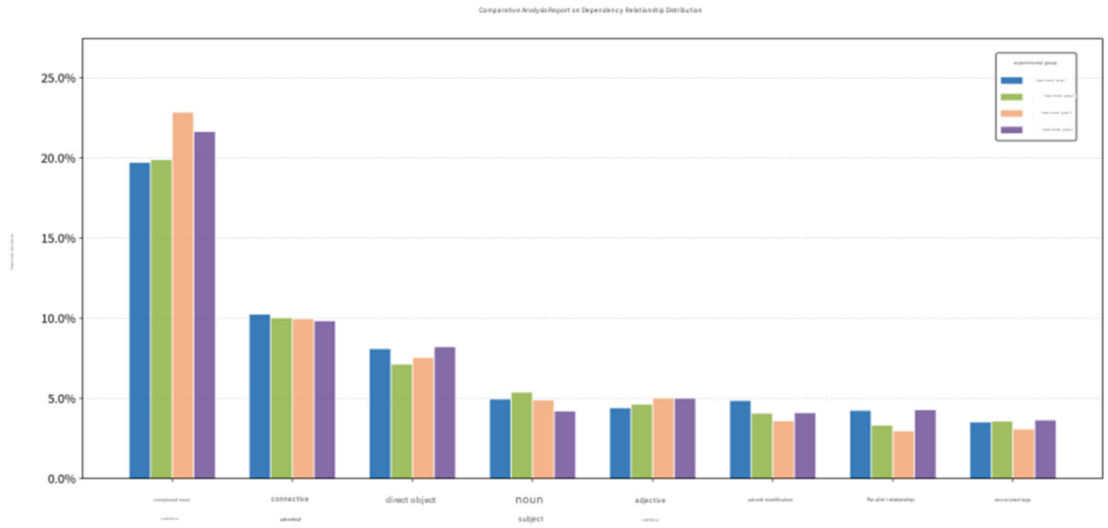

Figure 7. Dependency syntax distribution histogram.

This phenomenon reflects that the detection model has a specific preference for dependency syntactic relationships when judging the source of text generation—especially for noun modification relationships. When a text contains less compound noun modification relationships, the model is more likely to classify it as human-generated; conversely, it tends to classify the text as AI-generated.

By analyzing and comparing the experimental data of the four groups derived from the model's classification results using different methods, this section finds that the part-of-speech distribution has little impact on the model's classification results. However, dependency parsing can provide certain useful clues for understanding the classification mechanism of the RoBERTa-text detection model. The results of dependency parsing show that the model has a specific preference for noun modification relationships when determining the source of text generation. This finding deepens the understanding of the working principle of the RoBERTa-text detection model and also provides a direction for further optimizing the model's performance. Future research can further explore the impact of other linguistic

features on the model's classification results, as well as how to use these features to improve the accuracy and efficiency of AI text detection.

4. Conclusions and outlook

This study explores a deep learning-based detection method for AI-generated Chinese text. Focusing on the differences in multi-dimensional statistical features between AI-generated and human-generated texts, it proposes a detection framework that integrates the RoBERTa pre-trained language model with manually designed text statistical features. Covering key aspects of AI-generated text detection—from theoretical construction to practical verification—the research includes data collection, model architecture design, feature fusion strategy, performance evaluation, ablation experiments, and analysis of the model's generalized detection capability.

At the data construction stage, the study integrates a cross-domain hybrid multi-source corpus:

The HC3 dataset, a comparative corpus, includes an expert-level response text subset in Chinese covering professional fields such as medicine and law.

The ChatGPT detection dataset comprises texts from mass communication scenarios, including news bulletins, social media, and product reviews.

A specially constructed academic abstract dataset: 4151 publicly published academic paper abstracts were crawled from CNKI using web scraping technology. The ChatGPT-4.0 model was then used to generate academic abstracts corresponding to different paper titles, ensuring alignment between academic fields and professional depth.

The literary dataset: a genre-balanced sampling strategy was adopted to cover literary such as classical poetry, modern prose, and online novels. The iFlytek Spark (Xinghuo) large model was used for style imitation, resulting in a balanced dataset with 8900 samples.

In the data processing phase, a regex-based cleaning method was used to effectively remove format identifiers and typesetting noise from the text, ensuring data quality.

The study proposes a detection framework that combines the RoBERTa pre-trained language model with manually designed statistical features. The RoBERTa pre-trained model, equipped with 24 Transformer layers, 16 attention heads, and a hidden layer dimension of 1024, is well-suited for processing long texts. For statistical feature design, six text statistical features were extracted: the number of commas, periods, semicolons, and enumeration commas (dunhao), as well as the number of sentences and the standard deviation of sentence lengths.

Experimental results show that the model outperforms both single pre-trained models and traditional sequence models in accuracy (96.94%), precision (96.85%), and recall (97.61%). Two sets of ablation experiments further that both the depth of the classifier and text statistical features have a significant impact on the model's detection performance.

The study selected academic abstracts generated by three representative—DeepSeek R1, Phi4, and Qwen 2.5—and conducted comparative analyses of their word frequency distribution and perplexity features. Using jieba word segmentation technology for text segmentation, the Harbin Institute of Technology (HIT) stopword list (hit_stopwords) to identify and remove noise words, and the wordcloud tool to generate word clouds, differences in vocabulary usage habits among texts from different sources were analyzed.

Results indicate that texts generated models exhibit a centralized trend in vocabulary selection (with some words used frequently), while human-written texts show more scattered vocabulary usage

and richer lexical diversity. Additionally, by calculating text perplexity, differences in semantic coherence and predictability were observed among texts generated by different.

To evaluate the generalization performance of the RoBERTa-text model, texts generated by three (DeepSeek R1, Phi4, and Qwen 2.5) were used for detection. Results show that the model's detection performance varies across texts from different sources: the best performance in detecting Phi4-generated texts (all Phi4-generated texts were effectively identified) and relatively weaker performance in detecting Qwen 2.5-generated texts (though the recall rate still reaches 92.88%).

By grouping the model's classification results and conducting statistical feature and linguistic analyses, it was found that the RoBERTa-text detection model is sensitive to the number of text features: texts with small sentence length variations and dense punctuation are more likely to be classified as AI-generated. Additionally, the model exhibits a specific preference for dependency syntactic relationships—especially noun modification relationships. Texts with more compound noun modification relationships are more likely to be judged as human-generated, while the opposite is true for AI-generated texts. This finding provides important references for optimizing model detection strategies in future work.

Theoretically, this study enriches the feature analysis system for AI-generated text detection and proposes a hybrid detection framework that integrates statistical features with deep semantics. Practically, a public detection platform (freeaidetect.top) was built to realize the social application of the model, providing technical support for fields such as education, news, and academic publishing.

This study also has several limitations: The research focuses primarily on Chinese text detection and has not conducted in-depth exploration of text detection in other languages. With the continuous development of AI technology, text generation and detection will become an important research direction. Further research is needed to develop cross-lingual text detection models and achieve effective between texts.

Although a hybrid detection framework integrating statistical features and deep semantics is proposed, the feature fusion strategy still needs further optimization. Future work should explore how to more effectively fuse text features from different dimensions to improve detection performance while reducing computational costs.

While the dataset covers texts from multiple fields, types, and sources, the quantity and quality of samples still need improvement. The monotonous prompt structure limits the diversity of generated texts. Future research can use more diverse prompts and generation conditions to enhance the diversity and complexity of the dataset, thereby further improving performance.

The research on AI-generated text detection technology will face many new challenges and opportunities in the future:

First, texts generated by different have unique characteristics, requiring future detection models to more intelligently identify and adapt to these differences. For example, some models may prioritize grammatical correctness, while others may tend to use specific vocabulary or expressions. Developing a unified detection framework with higher generalization ability—capable of automatically identifying and adapting to the generation characteristics of different models—will be a key research direction.

Second, as AI technology advances, users can generate personalized texts using more rigorous and detailed prompts, increasing the difficulty of detecting AI-generated texts. Future research needs to explore how to extract more discriminative features from personalized texts to improve detection accuracy.

Additionally, detecting AI-generated texts polished by humans is an urgent issue to address. Such

texts integrate AI generation patterns with human creative styles, making them difficult to identify using traditional detection methods. Future research can combine technologies such as generative adversarial networks to simulate the human polishing process and enhance the model's ability to detect such hybrid texts.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. (2023) ChatGPT: Jack of all trades, master of none. I*nf Fusion* 99: 101861. https://doi.org/10.1016/j.inffus.2023.101861
2. Nasiri S, Hashemzadeh A, (2025) The evolution of disinformation from fake news propaganda to AI-driven narratives as deepfake. *J Cyberspace Stud* 2025, 9: 229–250. https://doi.org/10.22059/jcss.2025.387249.1119
3. Yeo MA, (2023) Academic integrity in the age of artificial intelligence (AI) authoring apps. *Tesol J* 14: e716. https://doi.org/10.1002/tesj.716
4. Verma A, (2023) The copyright problem with emerging generative AI. *J Intell Prot Stud* 7: 69. https://doi.org/10.2139/ssrn.4537389
5. Ufuk F, (2023) The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. *Radiology* 307: e230276. https://doi.org/10.1148/radiol.230276
6. Gehrmann S, Strobelt H, Rush AM, Gltr: Statistical detection and visualization of generated text, preprint, arXiv:1906.04043. https://doi.org/10.48550/arXiv.1906.04043
7. Fröhling L, Zubiaga A, (2021) Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Comput Sci* 7: e443. https://doi.org/10.7717/peerj-cs.443
8. Corizzo R, Leal-Arenas S, (2023) One-class learning for AI-generated essay detection. *Appl Sci* 13: 7901. https://doi.org/10.3390/app13137901
9. Crothers EN, Japkowicz N, Viktor HL, (2023) Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*11: 70977–71002. https://doi.org/10.1109/ACCESS.2023.3294090
10. Nguyen-Son HQ, Tieu ND, Nguyen HH, Yamagishi J, Zen IE, (2017) Identifying computer-generated text using statistical analysis, In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017: 1504–1511. https://doi.org/10.1109/APSIPA.2017.8282270
11. Li W, (2002) Zipf's law everywhere. *Glottometrics* 5: 14–21.
12. Uchendu A, Le T, Shu K, Lee D, (2020) Authorship attribution for neural text generation, In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020: 8384–8395. https://doi.org/10.18653/v1/2020.emnlp-main.673

13. Flesch R, (1952) “Simplification of Flesch reading ease formula”: Reply. *J Appl Psychol* 36: 54–55. https://psycnet.apa.org/doi/10.1037/h0051965
14. Solnyshkina M, Zamaletdinov R, Gorodetskaya L, Gabitov A, (2017) Evaluating text complexity and Flesch-Kincaid grade level. *J Soc Stud Educ Res* 8: 238–248.
15. Devlin J, Chang MW, Lee K, Toutanova K, (2019) Bert: Pre-training of deep bidirectional transformers for language understanding, In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1: 4171–4186. https://doi.org/10.18653/v1/N19-1423
16. Cui Y, Che W, Liu T, Qin B, Yang Z, Pre-training with whole word masking for Chinese Bert. *IEEE/ACM Trans Audio Speech Lang Process* 29: 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365
17. Liu Z, Lin W, Shi Y, Jun Z, (2021) A robustly optimized BERT pre-training approach with post-training. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 2021: 1218–1227. https://doi.org/10.1007/978-3-030-84186-7_31
18. Crothers E, Japkowicz N, Viktor H, Branco P, Adversarial robustness of neural-statistical features in detection of generative transformers. In: *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022: 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892269
19. Xu K, Hui ZL, Dong ZJ, Cai PH, Lu LQ, (2024) Construction of an automatic detection dataset for open-domain texts generated by ChatGPT. *J Chin Inf Process* 38: 39–53. https://doi.org/10.3969/j.issn.1003-0077.2024.12.005

AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)