A Statistical Approach of Distinguishing Patent Abstracts written by Human from those Generated by ChatGPT

C. J. Yue P. Huang Shiuh-Sheng Hsu Chin-Yuan Fan

Abstract

This study investigates statistical differences between human-written patent abstracts and those generated by ChatGPT under controlled conditions. We analyze 500 patent abstracts from Taiwanese applications filed in 2020, a dataset selected to ensure temporal consistency and to minimize potential influence from AI-assisted writing tools. Given that patent abstracts emphasize essential technical features and follow relatively standardized, claim-oriented language, they provide a suitable setting for controlled stylistic analysis.

Using restricted inputs (patent titles or original abstracts), we generate corresponding texts with ChatGPT and examine distributional differences via exploratory and confirmatory statistical analysis. The proposed framework, inspired by ecological diversity, models words and phrases as species and employs diversity-based metrics—such as standardized type-token ratio and entropy—as explanatory variables. Experimental results show that a small set of interpretable statistical features can achieve classification performance comparable to a BERT-based model, while requiring substantially fewer variables and lower computational cost. Importantly, the observed differences reflect distributional characteristics under constrained generation conditions rather than universal distinctions between human and AI-generated text.

This study provides an interpretable, domain-specific baseline for understanding statistical differences between human and AI-generated language, highlighting the role of context, domain, and input constraints in shaping stylistic patterns.

Keywords: Text Analysis, Writing Style, Exploratory Data Analysis, ChatGPT, Large Language Model

1. Introduction

Writing has always been regarded as an important medium for disseminating human culture. Unlike oral traditions, written language transcends the limitations of time and space, allowing for reflection, critique, and improvement by later generations. In recent years, the rapid development of artificial intelligence has led to the widespread application of language models in various fields.

Among them, OpenAI's ChatGPT (Generative Pre-trained Transformer), developed based on a large language model (LLM), has attracted much attention due to its outstanding capabilities in language comprehension and generation. ChatGPT's ability to generate coherent, context-sensitive, and stylistically consistent text has raised concerns about academic integrity, information security, and the authenticity of professional documents. These concerns are particularly prominent in technical fields such as patent documentations, where the accuracy and originality of language are crucial.

Existing methods for distinguishing between AI-generated text and human-written text include linguistic statistics, stylistics, and machine learning-based detection techniques. For example, Desaire et al. (2023) demonstrated that, although their results were obtained in a domain-specific context, combining stylistic features with traditional machine learning techniques can accurately distinguish between AI-generated scientific text and human-written text. Similarly, Alamleh et al. (2023) showed that traditional machine learning models enhanced with language and stylistic features can effectively detect content generated by ChatGPT, demonstrating that high detection performance can be achievable without relying solely on complex deep learning architectures such as BERT.

Based on these insights, this study focuses on distinguishing between authentic Chinese patent abstracts and those generated by ChatGPT. We propose a statistical method inspired by ecological diversity, treating words and phrases as species and using species richness and heterogeneity indicators as explanatory variables. We can apply these features to a logistic regression model to classify patent abstracts and compare the results with a BERT-based model. This approach emphasizes interpretability and

domain specificity, provides a complementary perspective to deep learning methods, and enables quantitative analysis of lexical and syntactic diversity in technical documents.

2. Literature Review

Recent advances in large language models (LLMs) have significantly improved the fluency and coherence of machine-generated text, making it increasingly difficult to distinguish between human-written and AI-generated content. This development has raised concerns regarding academic integrity, misinformation, and content authenticity, thereby motivating extensive research on automated detection methods.

Existing approaches to AI-generated text detection can be broadly categorized into three groups: statistical methods, stylometric approaches, and deep learning-based models. Statistical approaches rely on measurable properties of text, such as entropy, perplexity, and n-gram distributions, to capture differences between human and machine-generated text. These methods are computationally efficient and interpretable, but may be sensitive to distributional shifts across domains and generation settings (Chakraborty et al., 2023; Ippolito et al., 2020; Koppel et al., 2011). Importantly, such approaches may partially rely on superficial structural cues, such as text length or sentence length, which can be influenced by prompt design and generation conditions rather than inherent differences between human and AI writing.

Stylometric approaches focus on capturing writing style through linguistic features at multiple levels, including lexical choice, syntactic structure, and phrase patterns. Prior research has shown that stylometric features can effectively distinguish between human-written and machine-generated text, even when the generated text is highly fluent (Zaitsu et al., 2025). Earlier studies in authorship attribution further demonstrate that stylistic signals—such as function word usage, syntactic patterns, and distributional characteristics—can provide robust indicators of authorship (Lagutina et al., 2019; Juola, 2006). These findings suggest that stylistic differences may remain detectable despite advances in generative models, although their stability may vary across domains and generation conditions.

Deep learning-based approaches, particularly transformer architectures such as BERT and RoBERTa, have also been widely applied to this task. These models are capable of capturing complex contextual and semantic patterns and have demonstrated strong detection performance, with some studies reporting accuracy exceeding 99% on specific datasets (Zellers et al., 2019; Desaire et al., 2023). However, such models often suffer from limited interpretability, high computational cost, and vulnerability to adversarial manipulation, including paraphrasing attacks that can substantially reduce detection accuracy (Sadasivan et al., 2023; Ippolito et al., 2020).

From a theoretical perspective, the feasibility of detecting AI-generated text has been examined using information-theoretic frameworks. Prior work suggests that detection performance depends on the statistical distance between human and machine-generated text distributions, as well as the number of available samples (Chakraborty et al., 2023). As these distributions become increasingly similar, detection becomes inherently more difficult. This perspective highlights that successful detection ultimately relies on identifying stable distributional differences rather than dataset-specific artifacts.

Overall, prior research suggests that both traditional statistical approaches and deep learning models can effectively distinguish human-written from AI-generated text under certain conditions. However, most existing studies focus on general or open-domain text and emphasize classification performance rather than interpretability. In particular, the extent to which existing methods rely on superficial structural features—such as text length or prompt-induced patterns—rather than deeper distributional properties remains insufficiently examined. Furthermore, relatively limited attention has been given to domain-specific corpora and controlled generation settings, where both stylistic variation and generation conditions can be systematically analyzed. These gaps motivate the present study, which aims to provide an interpretable and controlled analysis of distributional differences in a specialized domain.

3. Theoretical Framework

The proposed approach is grounded in three complementary theoretical

perspectives: information theory, stylometry, and ecological diversity. From an information-theoretic perspective, text generation can be viewed as a probabilistic process in which sequences of words are drawn from an underlying distribution. Human-written text is typically produced through complex cognitive processes that involve varying levels of uncertainty and contextual adaptation, which are often associated with greater variability in word choice and structure. This variability may lead to higher entropy and more diverse linguistic patterns in certain contexts. In contrast, large language models generate text by sampling from learned probability distributions optimized for fluency and likelihood, which may result in more regular and predictable patterns of expression.

Prior theoretical work (Chakraborty et al., 2023; Sadasivan et al., 2023) suggests that the detectability of AI-generated text depends on the statistical distance between human and machine-generated distributions. As these distributions become more similar, detection becomes increasingly difficult. This implies that measurable differences in distributional properties can serve as a principled basis for classification.

From a stylometric perspective, writing style reflects systematic patterns in lexical choice, syntactic structure, and phrase usage, and has long been studied in the context of authorship attribution and text classification. Prior research in stylometry suggests that human authors often exhibit greater stylistic variability, including irregular sentence structures, diverse vocabulary, and less predictable phrase patterns. In contrast, AI-generated text may exhibit more uniform stylistic characteristics due to the optimization of language models toward high-probability sequences. These stylistic tendencies provide measurable and theoretically grounded signals for distinguishing between human and machine-generated text.

Building on these perspectives, this study adopts an ecological view of language, treating words and phrases as analogous to species within an ecosystem. In this framework, lexical diversity corresponds to species richness, while the distribution of word frequencies reflects species evenness. Human-written text can often be characterized as a more diverse and heterogeneous ecosystem, with a wider range of lexical items and less concentrated usage patterns. In contrast, AI-generated text tends to exhibit more regular and concentrated distributions, reflecting the underlying

probability-driven generation mechanism.

By integrating these three perspectives, this study proposes that diversity-based metrics—such as standardized type-token ratio (STTR) and entropy—can capture meaningful distributional differences between human-written and AI-generated text. Specifically, this framework suggests that human-written texts are more likely to exhibit higher lexical diversity and greater variability, whereas AI-generated texts may exhibit more concentrated and regular distributions. These metrics therefore provide an interpretable and theoretically grounded basis for classification, particularly in domain-specific contexts where stylistic variation is constrained but distributional differences remain observable.

4. Data and Methodology

We obtained patent application data from the Intellectual Property Office (TIPO) of the Ministry of Economic Affairs of Taiwan. We focused on patent applications filed in 2020, totaling 2,835. This period was chosen to ensure that the original abstracts were unlikely to be influenced by AI-assisted writing tools, as these tools only began to be widely used after 2021. Furthermore, limiting the data set to one year helps maintain consistency in time and style, thereby reducing potential discrepancies due to changes in writing standards or regulatory guidelines.

From the full dataset, 500 patent abstracts were randomly sampled for analysis. Patent abstracts differ from scientific abstracts in purpose and structure: patent abstracts do not summarize research objectives, methods, and results, but rather aim to concisely describe the key technical features of an invention and are usually closely related to the independent claims. Therefore, patent abstracts often employ more standardized, technical, and function-oriented language. This relative homogeneity provides a controlled environment for analyzing stylistic and linguistic differences between human-written and AI-generated text, but may also limit the general applicability of the research results to other text types.

We used sampled abstracts to generate corresponding texts using ChatGPT-4o (released by OpenAI in May 2024) under two input conditions: (1) using only the patent

title ("title-to-abstract"), and (2) using the original abstract as input ("abstract-to-abstract"). These two settings are designed to reflect different levels of contextual information provided to the AI system. In particular, title-based generation represents a constrained input scenario, which may result in lower information richness and potentially amplify stylistic differences from manually written abstracts.

We then compared the generated text with the original, human-written abstracts. We employed two analytical methods: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). EDA identifies potential patterns and generates hypotheses by summarizing key features of the data, while CDA is used to statistically test the differences between the human-generated and AI-generated abstracts. It is important to note that the domain specificity of the patent abstract and the limited input conditions used to generate the text can influence the observed differences. Therefore, the findings of this study should be interpreted within this specific context and may not be directly generalized to other forms of writing or AI-generated text using richer input information.

Tukey (1977), a famous statistician who is considered the first person to propose EDA, emphasized the importance of examining data before formal modeling. Data visualization is crucial in conducting EDA and can be used to study the relationship between two variables. This is especially important when analyzing unstructured data such as text. Modern Chinese is mainly written in vernacular Chinese. Chinese writing uses characters as the basic unit to convey meaning and its writing style resembles an ecosystem, characterized by the diversity and number of species as well as the interactions among them. By analyzing the types of textual elements and their interrelationships, we can explore the distinctive features of the writing.

Thus, we consider the diversity metrics of Chinese characters and evaluate whether they can provide key information to distinguish human-written text from computer-generated text. Note that modern Chinese often uses two-word phrases (rather than single characters) to express meaning, so we will show diversity metrics for both words and phrases. In addition to word count and type of different words, we also consider word richness and heterogeneity. For example, Table 1 shows the basic statistics of words and phrases for human-written and computer-generated abstracts. Since the

number of single words and two-word phrases in human-written abstracts is much larger, it is unfair to directly compare the number of different types. A more appropriate diversity metric is the Standardized Type-token-ratio (STTR), where we randomly sample 500 (or a fixed number) words and count how many unique words there are in them:

$$STTR = \frac{Number\ of\ types}{Number\ of\ tokens} \quad (1)$$

Table 1. Basic Vocabulary Statistics for Human-written & Computer-generated Texts

		Title-generated	Abstract-generated	Original
Words	Counts	100,990	69,084	102,722
	Types	1,310	1,307	1,378
Phrases	Counts	57,599	39,052	60,184
	Types	3,207	3,342	4,054
Word Frequency	Top 100	56.62%	53.41%	54.89%
	Top 500	94.30%	93.64%	93.70%
Phrase Frequency	Top 100	35.49%	33.45%	29.94%
	Top 500	57.16%	53.75%	48.58%
Sentence	Counts	6,650	4,470	5,628
	Average Words	15.19	15.46	20.19
	Ave. Phrases	8.66	8.74	10.69

Another measure of diversity is evenness, that is, whether a small number of words make up a large proportion of the total number. The accumulated probability is one way of measuring evenness and we compute the most common words and phrases in Table 1, for top 100 and 500 words and phrases. The results for two-word phrases are more distinguishable between human- and computer-generated text. There are other types of evenness measure, such as entropy (or Shannon's index):

$$Entropy = -\sum_i p_i \log(p_i) \quad (2)$$

where p_i is the proportion of i^{th} species, or Simpson's index $\sum_i p_i^2$. These two indices usually show similar information and we will only show the results of entropy in this study.

Sentence length is also a diversity indicator considered in this study and can be used to distinguish writing styles (Lagutina ETAL., 2019; Wang et al., 2024). We calculate the average numbers of words and phrases in a sentence and a sentence is defined as the texts between any two of five punctuation marks: period, comma, semicolon, exclamation mark, and question mark (Yue and Lin, 2020). Apparently, there are more words and phrases in a sentence of human-written abstracts, while those of title-generated and abstract-generated abstracts are about the same (Table 1). This result is similar to those of our previous studies for comparing between human-written and computer-generated texts regarding modern Chinese writing.

One of our study goals is to evaluate whether traditional statistical modelling has similar detection performance comparing to that of LLMs, but with fewer variables and less computing time. Thus, we chose three models in the stage of CDA, which are logistic regression and two machine models, Random Forest and XGBoost. The detection results of these models will be compared with those of BERT. These models are popular in the studies of text mining. Logistic regression belongs to the family of Generalized Linear Model and is used to handle the case of binary response. By applying the logit link to the response variable Y,

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \quad (3)$$

the process of selecting feasible logistic regression is similar to that of ordinary rergression, where $X_1 \ldots, X_k$ are the independent variables.

Random Forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) are both powerful ensemble machine learning algorithms for classification and regression tasks. They are based on decision trees but use different techniques to improve model performance, reduce overfitting, and increase accuracy. Random forests are ensembles of decision trees built through a method known as bagging (bootstrap aggregating), where multiple trees are trained independently on random subsets of the data. In contrast, XGBoost is a highly efficient implementation of gradient boosting, an ensemble approach that constructs trees sequentially, with each tree correcting the errors of the previous ones.

In the next section, we will explore the characteristics of human-written and

computer-generated abstracts through the EDA method. Visualization tools can clearly show the differences between the two types of texts, helping us select key variables for distinguish writing styles. The effect of this variable selection can then be seen from the model analysis results.

5. Exploratory Data Analysis

In this study, we randomly selected 500 abstracts from Taiwanese patent applications filed in 2020 and used ChatGPT-4o to generate two sets of computer-generated abstracts, each containing 500 items. We first examined the basic statistics of words and phrases across these three datasets. Figure 1 shows the distribution of word counts per abstract, revealing significant differences among the groups: the title-generated set exhibits the highest mean word count, while the abstract-generated set has the smallest variance. These results suggest that the type of prompt affects ChatGPT's outputs, and that the human-written and abstract-generated abstracts have more similar word counts. Notably, patent abstracts are on average about 100 words shorter than Taiwanese master's theses.

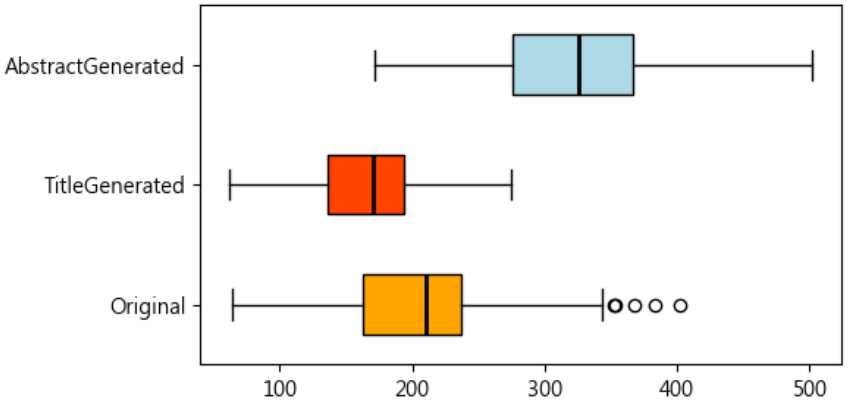

Figure 1. Word Counts of 3 Sets of Abstracts

As mentioned earlier, modern Chinese often uses phrases to express meaning, so we mainly present the results of short words. Also, we use CKIP for Chinese word segmentation, which is a preprocessing step in analyzing Chinese text. CKIP is a word segmentation tool developed by the Chinese Knowledge and Information Processing

(CKIP) Group at Academia Sinica in Taiwan. The system is built upon a large-scale lexical corpus and linguistic rules, and employs statistical methods such as the Hidden Markov Model (HMM) and the Maximum Entropy Model, along with syntactic analysis techniques, to accurately identify word boundaries. It effectively addresses challenges such as word sense disambiguation, unknown word detection, and part-of-speech tagging. Compared to segmentation tools based on simple string matching, the CKIP system offers higher semantic consistency and syntactic sensitivity, making it particularly suitable for Chinese natural language processing tasks that require a high degree of linguistic precision in academic research.

The phrases used in all abstract sentences appear to be slightly skewed to the right, with human-written abstracts being more obvious, resulting in slightly larger averages and variances (Figure 2). This shows that the abstracts written by humans are more diverse, and the difference in word count shows the differences in human writing. This difference is more obvious in the number of phrases per sentence (Figure 3). The human-written sentences are longer and may contain more than 13 phrases in one sentence, while computer-generated sentences have a higher ratio of 6 to 13 words per sentence. The cumulative proportion of phrases also shows the diversity of phrases used in human writing. Figure 4 is the graph of the cumulative proportions of the most common phrases used in 3 types of abstracts and the curve of abstracts written by humans increases more slowly, indicating that the vocabulary usage is less concentrated on certain common phrases. As a double check, the top 500 common phrases account for less than 50% of all phrases, and Table 1 shows similar results.

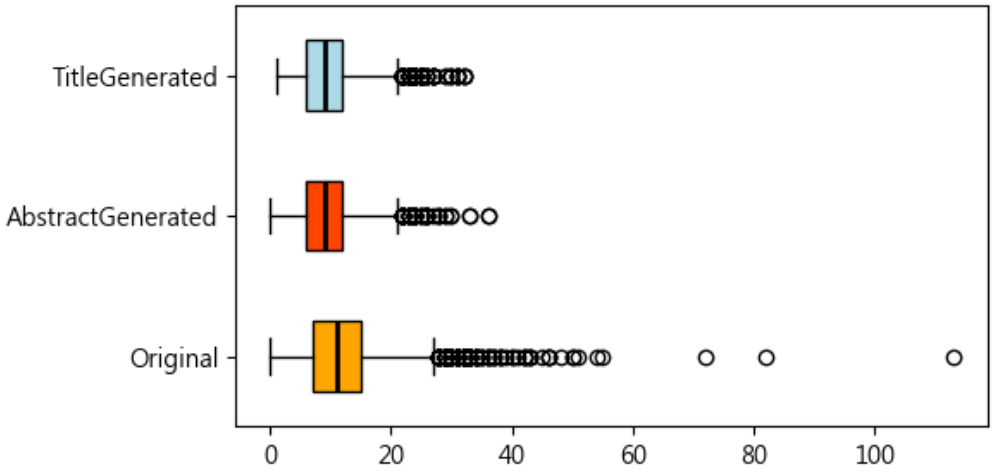

Figure 2. Boxplot of Number of Phrases per sentence in 3 Sets of Abstracts

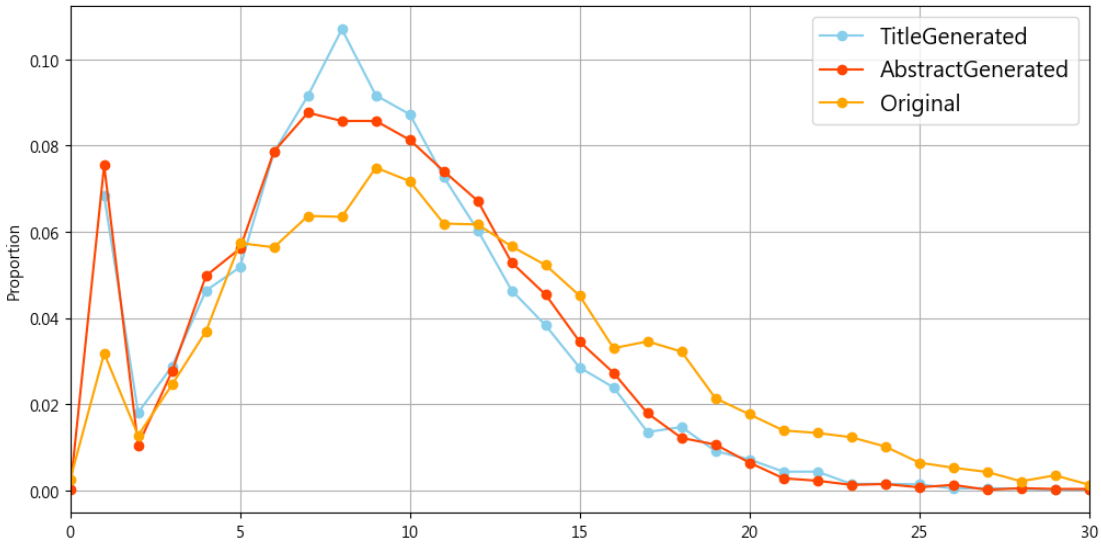

Figure 3. Histogram of Number of Phrases per sentence in 3 Sets of Abstracts

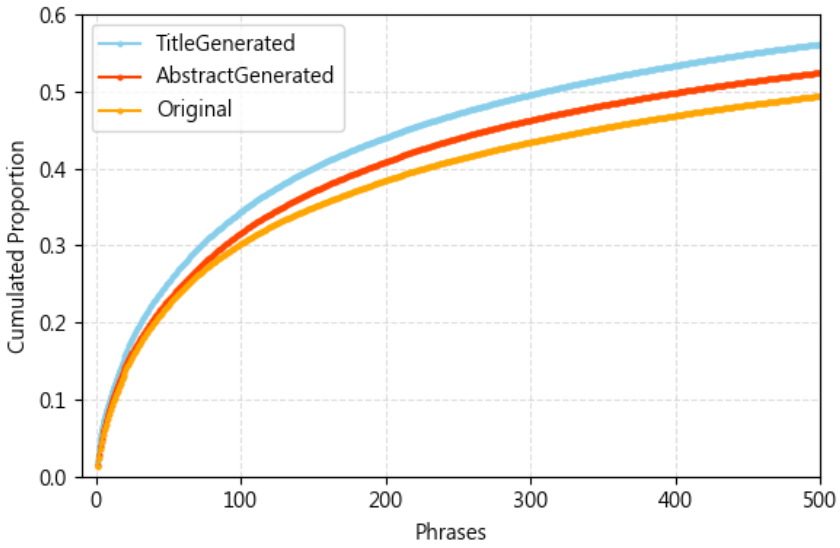

Figure 4. Cumulative Proportions of Phrases in 3 Sets of Abstracts

The lexical diversity in the three abstract data can also be assessed using STTR and entropy. We calculated the mean and standard deviation of STTR and entropy for 1,000 computer simulation runs, with 2,000 and 5,000 phrases randomly selected in each simulation (Table 2). However, only the STTR values of the three groups of data were significantly different (which can be verified by t-test), but those of entropy values were not. Also, unlike the cumulative proportions of words and phrases mentioned above, the group of abstract-generated abstracts tend to have the largest values. The simulation results of randomly sampling words were similar and are omitted here. It should be noted that the simulation results of STTR and entropy of the three groups of data were very different when the number of phrases sampled was different (2,000 phrases vs. 5,000 phrases). Therefore, if you plan to use TTR and entropy to compare the writing style of two groups of data, it is recommended to sample the same number of words/phrases.

Table 2. Phrase Diversity in 3 Sets of Abstracts (Average ± st.d.)

		Title-generated	Abstract-generated	Original
STTR	2,000 Phrases	.2564±.0016	.2588±.0019	.2576±.0015
	5,000 Phrases	.1428±.0006	.1423±.0008	.1411±.0008

Entropy	2,000 Phrases	5.8612±.0179	5.9187±.0165	5.9045±.0167
	5,000 Phrases	6.0120±.0123	6.0719±.0127	6.0396±.0120

In addition to the preceding characteristics of words and phrases, we considered the differences between humans and computers in the use of phrases, and used phrases with larger usage differences as variables in the classification models. To reduce the number of variables, we only considered the 100 most common phrases in human-written and computer-generated abstracts. We kept only those phrases that appeared at least 100 times and were either twice as frequent in humans as in computers or twice as frequent in computers as in humans. For the comparison of human-written and title-generated abstracts, 50 phrases met these requirements (Figure 5, left panel), while 40 phrases met the same criteria for human-written and abstract-generated abstracts (Figure 5, right panel).

Figure 5. Phrases with Different Ratios in Human- & Computer-generated Abstracts

6. Model Fitting

In this section, we will apply the variables chosen in the previous section to classification models and compare the results with those of BERT. The variables considered can be divided into two parts: sentence and lexical structure, and differences in phrase usage. We consider the combinations of these two sets of and variables and phrase differences have a greater impact on classification accuracy. We focus on phrase-related variables in the rest of this study. For comparison, we randomly divided 500 abstracts into a training set and a test set, with a ratio of 80% and 20%, respectively.

We first built a model based on the training set data, and then substituted the test set data into the model to calculate the prediction error. We repeated this process 500 times, and calculated the mean and variance of the error.

The comparison of classification models will be based on four types of indicators, including precision, accuracy, recall, and F1, which are defined as

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = \frac{2\times Precsion\times Accuracy}{Precision+Accuracy} \tag{7}$$

where TN, FN, FP, and TP are defined in the confusion matrix (Table 3). We chose to use F1-score as the overall comparison.

Table 3. Confusion Matrix

		Predicted	
		Negative	Positive
True	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

We first show the comparison results of human-written and title-generated abstracts. Since the title-generated abstracts contain only the information of title, it would be easier to differentiate the two sets of abstracts. The simulation results were consistent with expectations, so we only present the results using the 50 variables that distinguish between human-written and computer-generated abstracts for comparison with those of BERT. Table 4 shows the average accuracy results and their standard deviations. All models achieve very accurate classification results, and the 50 variables appear to be sufficient to distinguish whether an abstract was human-written or title-generated.

Table 4. Classification Results of Human- and Title-generated Abstracts (Average ± st.d.)

Models		Precision	Accuracy	Recall	F1
Logistic Regression	50 Variables	.9970±.0032	.9980±.0040	.9961±.0054	.9970±.0032
Random Forest		.9998±.0012	.9955±.0002	1.0000±.0004	.9998±.0012
XGBoost		.9962±.0038	.9944±.0066	.9998±.0043	,9962±.0038
BERT		.9983±.0041	.9997±.0023	.9970±.0081	.9983±.0043

Table 5. Classification Results of Human- and Abstract-generated Abstracts (Average ± st.d.)

Models		Precision	Accuracy	Recall	F1
Logistic Regression	40 Variables	.9282±.0170	.9411±.0.0228	.9143±.0287	.9271±.0177
Random Forest		.9398±.0154	.9475±.0207	.9316±.0258	.9392±.0158
XGBoost		.9343±.0160	.9387±.0231	.9299±.0251	.9340±.0162
Logistic Regression	95 Variables	.9454±.0147	.9541±.0200	.9365±.0250	.9449±0150
Random Forest		.9520±.0139	.9608±.0189	.9428±.0225	.9515±.0142
XGBoost		.9523±.0133	.9549±.0195	.9500±.0209	.9522±.0133
BERT		.9631±.0285	.9871±.0239	.9396±.0607	.9612±.0327

It is more difficult to distinguish between human-written or abstract-generated abstracts. As mentioned above, the variables related to phrases play a more important role, and we consider adding more variables on top of the 40 variables in Figure 5. The additional variables beyond the 40 variables are divided into two categories: sentence and lexical structure, and a relaxation of the criteria for filtering out phrase-related variables. The first category includes five variables: STTR, entropy, and number of phrases per sentence (average, standard deviation, and outlier). The criteria of phrase-

related variables are 100 most common phrases, occurring at least 70 times, and being 1.6 times more frequent in humans than in computers or 1.6 times more frequent in computers than in humans.

Table 5 shows the 500 simulation results for distinguishing human-written abstracts from abstract-generated abstracts. Compared to distinguishing between human-written abstracts and abstracts generated by titles, these two data sets are indeed more difficult to distinguish, and BERT's accuracy values are about 3% lower than those in Table 4. On the other hand, the accuracy values of the proposed method (using 40 variables) are significantly lower than those of BERT, and thus we add extra 55 variables. The three models considered in this paper achieve very close accuracy to BERT when using 95 variables. Note that the number of variables considered in BERT is around 21,000, which is about 200 times that of the proposed approach. Furthermore, the computation time for 500 simulation runs of all three classification models in this study was less than 5 minutes, compared to 15 minutes for BERT per simulation. In other words, the proposed statistical method offers excellent efficiency without sacrificing accuracy.

To further strengthen our findings, we conducted additional validation using 500 newly collected patent abstracts. In this setting, the original 500 abstracts served as the training dataset, while an additional 500 abstracts randomly extracted from the patent abstracts served as a separate test set. We replicated the entire process, including generating abstracts from titles and abstracts based on ChatGPT, variable selection, and classification. The results were largely consistent with the original analysis (Table 6, showing only accuracy and F1 score). Although distinguishing between human-written and AI-generated abstracts remains challenging, the accuracy and F1 score remained fairly high. This confirms that even using real abstracts as input, the output generated by ChatGPT is still statistically significantly different from manually written text. Importantly, while BERT and our method achieved comparable accuracy, our method requires only 93 features—similar to the 95 variables reported in Table 5—demonstrating its simplicity and robustness on the expanded independent dataset.

Table 6. Classification Results of Human- and Abstract-generated Abstracts (Original Main_500 vs. Newly Validation_500 abstracts)

Model	Main_CV Accuracy	Main_CV F1	Validation_500 Accuracy	Validation_500 F1
Logistic Regression	0.9890	0.9889	0.8050	0.7782
Random Forest	0.9880	0.9880	0.9030	0.9099
XGBoost	0.9880	0.9879	0.9060	0.9118

7. Conclusion and Discussions

This study examined 500 human-written abstracts from Taiwanese patent applications filed in 2020 alongside two matched sets of 500 abstracts generated by ChatGPT-4o under controlled input conditions. The analysis revealed systematic differences in sentence structure and lexical usage, particularly in terms of sentence length distribution and phrase diversity. Computer-generated abstracts tended to exhibit shorter sentences with narrower distributions and introduced lexical choices that differ from those commonly used by Taiwanese authors.

It is important to emphasize that these findings should be interpreted within the controlled experimental setting of this study. The observed differences are influenced by both the domain-specific characteristics of patent abstracts and the constrained input conditions used during text generation. Therefore, the results do not imply that the identified features universally distinguish human-written from AI-generated text across domains, languages, or model configurations.

Within this setting, diversity-based features—particularly those related to phrase usage and sentence structure—proved effective in distinguishing between human-written and AI-generated abstracts with high accuracy. Importantly, this feature-driven approach achieved performance comparable to a BERT-based model while requiring fewer than 100 variables, resulting in substantially lower computational cost. These

results highlight the potential of interpretable statistical features as efficient alternatives to more complex models.

From a methodological perspective, this study prioritizes interpretability and controlled statistical analysis over general-purpose detection performance. The proposed framework should therefore be understood as establishing a baseline for analyzing distributional differences between human and AI-generated text, rather than as a universal detection solution.

While preliminary exploratory tests on other Chinese-language documents suggest that similar patterns may exist, these observations were not systematically evaluated and should be interpreted with caution. Future research should extend this framework to multiple language models, domains, and languages, and incorporate adversarial scenarios—such as paraphrasing and hybrid text generation—to assess robustness more comprehensively.

In summary, this study provides an interpretable, domain-specific statistical framework that captures distributional differences between human-written and AI-generated text under controlled conditions. These findings contribute to a deeper understanding of stylistic and structural variation, while highlighting the importance of context, domain, and generation conditions in evaluating AI-generated content.

References

- Alamleh, H., AlQahtani, A.A.S. and ElSaid, A. (2023). Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning, 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 154-158. doi: 10.1109/SIEDS58326.2023.10137767.
- Breiman, L. (2001). Random Forests, *Machine Learning* 45, 5–32.
- Chakraborty, S., Bedi, A.S., Zhu, S., An, B., Manocha, D., and Huang, F. (2023). On the Possibilities of Ai-generated Text Detection, *arXiv:2304.04736*.
- Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Chen, Z., Huang, L., Yang, W., Meng, P., & Miao, H. (2012). More than Word Frequencies: Authorship Attribution via Natural Frequency Zoned Word Distribution Analysis, arXiv preprint arXiv:1208.3001.
- Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing Academic Science Writing from Humans or ChatGPT with over 99% Accuracy using Off-the-shelf Machine Learning Tools. *Cell Reports Physical Science* 4(6). https://doi.org/10.1016/j.xcrp.2023.101426
- Devlin, J., Chang, M. W., 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*, volume 1 (long and short papers) (pp. 4171-4186).
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N.A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models *arXiv:2009.11462*.
- Hu, X., Wang, Y., & Wu, Q. (2014). Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. *Advances in Adaptive Data Analysis* 6(04), 1450012.
- Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2020). Automatic

Detection of Generated Text is Easiest when Humans are Fooled, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

- Juola, P. (2006). Authorship Attribution, *Foundations and Trends in Information Retrieval* 1(3), 233–334.
- Koppel, M., Schler, J., and Argamon, S. (2011). Authorship Attribution in the Wild, *Language Resources and Evaluation* 45(1), 83–94.
- Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A Survey on Deep Learning for Patent Analysis, *World Patent Information* 65: 102035.
- Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., and Liu, H. (2023). Stylometric Detection of AI-Generated Text in Twitter Timelines, *arXiv:2303.03697*.
- Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., & Demidov, P.G. (2019). A Survey on Stylometric Text Features, 25th Conference of Open Innovations Association, *IEEE*.
- Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z.., Zhang, Y., Huang, X., Zhu, D., Cai, H., Li, Q., Liu, T., & Li, X. (2023). Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study, *JMIR Medical Education* 9: e48904.
- Mihalcea, R. & Tarau, P. (2004). Textrank: Bringing order into Text, In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404-411.
- Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? *Transactions on Machine Learning Research*, arXiv:2303.11156.
- Shalaby, W. & Zadrozny, W. (2019). Patent Retrieval: A Literature Review, *Knowledge and Information Systems*, arXiv:1701.00324v2.
- Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. (2019). Release Strategies and the Social Impacts of Language Models, *arXiv:1908.09203*.
- Tukey, J. W. (1977). Exploratory data analysis, Reading, MA: Addison-Wesley.

- Uchendu, A., Le, T., Shu, K., and Lee, D. (2023). Authorship Attribution for Neural Text Generation, *Transactions of the Association for Computational Linguistics* 11, 849–864.
- Webb, L. & Schönberger, D. (2024). Generative AI and the Problem of Existential Risk, arXiv preprint arXiv:2407.13365.
- Yadagiri, A., Shree, L., Parween, S., Raj, A. Maurya, S. and Pakray, P. (2024). Detecting AI-generated Text with Pre-trained Models using Linguistic Features, *Proceedings of the 21st International Conference on Natural Language*, 188–196.
- Yue, J.C. and Lin, C. (2020). A Statistical Analysis of Chinese Writing Style in Xin Qingnian (New Youth Magazine) in the 1910s and 1920s, *Rivista degli Studi Orientali XCIII(4)*, 167-181.
- Zaitsu, W., Jin, M., Ishihara, S., Tsuge, S., and Inaba, M. (2025). Stylometry can Reveal Artificial Intelligence Authorship, but Humans Struggle: A Comparison of Human and Seven Large Language Models in Japanese, PLoS One 20(10): e03335369.
- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending Against Neural Fake News, *Advances in Neural Information Processing Systems* (*NeurIPS 2019*).

Appendix A. List of Selected Variables

1. Outliers in entropy and TTR (100 words were randomly selected each time in 30 simulation runs)
2. Standard deviation of Entropy and TTR 100 words were randomly selected each time in 30 simulation runs)
3. Average of sentence length
4. Standard deviation of sentence length
5. Short sentence ratio (sentences less than 6 Chinese words)
6. Parentheses usage rate
7. Frequency of 5 function words used (與，之，應，所，了)
8. 20 Common words (words with a frequency ratio that differs by more than double and appear at least 100 times)

Appendix B. Prompt Design and Generation Procedure

In this study, all AI-generated texts were produced using ChatGPT-4o (OpenAI). The generation process was conducted in April 2025 under two input conditions: (1) abstract-based generation and (2) title-based generation.

(1) For abstract-based generation, the following prompt template was used:

“根據以下摘要進行修改：”

[original abstract]

(2) For title-based generation, the following prompt template was used:

“根據以下標題生成 48–376 字的摘要，請直接說明技術內容：”

[patent title]

All samples were generated using a single user message without any system message. Within each dataset, the same prompt template was applied consistently, with only the input content (i.e., patent title or original abstract) varying across samples.

The API parameters (e.g., temperature, top_p, max_tokens, frequency_penalty, presence_penalty) were not explicitly specified in the implementation and therefore default API settings were used. No additional sampling strategies (e.g., multiple generations per input or candidate selection) were applied; each input instance was generated exactly once. If an API request failed or timed out, the error was recorded and the corresponding sample was regenerated manually. Only successfully generated outputs were retained for subsequent analysis.

A minimal post-processing step was applied to remove fixed prefixes (e.g., “修改後摘要：”) that occasionally appeared in the generated outputs. This step was performed solely to eliminate non-content artifacts introduced by the generation interface and did not alter the substantive textual content. The same generation procedure, prompt templates, and parameter settings were applied consistently to both the primary dataset and the additional validation dataset.

It should be noted that prompt design may influence the stylistic properties of generated text. In this study, we intentionally employed simple and standardized prompt templates to reduce variability introduced by prompt engineering and to ensure

comparability across samples. No manual editing or rewriting of generated texts was performed beyond the removal of fixed prefixes.