

VietBinoculars: A Zero-Shot Approach for Detecting Vietnamese LLM-Generated Text

Trieu Hai Nguyen^{a,†}, Sivaswamy Akilesh^{b,*}

^a *Faculty of Information Technology, Nha Trang University,
02 Nguyen Dinh Chieu Street, North Nha Trang 57100, Vietnam*

^b *Faculty of Business Administration, Swiss School of Business and Management
Geneva, 12 Avenue des Morgines, Geneva 1213, Switzerland*

e-mails: [†]*trieunh@ntu.edu.vn*; ^{*}*akilesh@ssbm.ch*

Abstract

The rapid development research of Large Language Models (LLMs) based on transformer architectures raises key challenges, one of which being the task of distinguishing between human-written text and LLM-generated text. As LLM-generated textual content, becomes increasingly complex over time, and resembles human writing, traditional detection methods are proving less effective, especially when the number and diversity of LLMs continue to grow with new models and versions being released at a rapid pace. This study proposes VietBinoculars, an adaptation of the Binoculars method with optimized global thresholds, to enhance the detection of Vietnamese LLM-generated text. We have constructed new Vietnamese AI-generated datasets to determine the optimal thresholds for VietBinoculars and to enable benchmarking. The results from our experiments show that VietBinoculars achieves over 99% in all two domains of accuracy, F1-score, and AUC on multiple out-of-domain datasets. It outperforms the original Binoculars model, traditional detection methods, and other state-of-the-art approaches, including commercial tools such as ZeroGPT and DetectGPT, especially under specially modified prompting strategies.

Keywords: LLM, Zero-shot Detection, Perplexity, Binoculars, LLM-generated text.

1 Introduction

The text detection tasks performed by LLMs and other AI (Artificial Intelligence) systems play an important role in various real-world applications. AI-generated text can be misused in various contexts, such as students using AI tools to complete assignments, or the automatic creation of misinformation and fake news on social media. These texts are often difficult for humans to distinguish from authentic content, potentially leading to misinformation and public misperception [1]. In recent years, many studies have explored methods to detect if text is written by humans or generated by machines. In this context, Gehrmann et al. [1] introduced GLTR (Giant Language Model Test Room) in 2019 as one of the first tools designed to detect machine-generated texts. Based on statistical methods, GLTR calculates the ranking of the actual next word appearing in the

[†]Corresponding author.
Preprint.

September 30, 2025

input text compared to the model’s prediction and visualizes it using colors. This laid the groundwork for subsequent Zero-shot detection ideas, as the tool does not require to retrain any LLMs while still achieving competitive performance based on statistical features. However, due to its dependence on the model that generates the text, this approach faces challenges in detecting texts generated by commercial LLMs like ChatGPT¹, Claude², and its detection scope is quite limited.

One of the first large language models built on the GPT-2 architecture [2] to generate propaganda news and fake articles was GROVER, proposed by Zellers et al. in 2019 [3]. This model was evaluated to have higher reliability than human-written text. GROVER is particularly dangerous when employed for malicious purposes. In particular, the best way to detect texts generated by GROVER is to use GROVER itself, based on the exposure bias and applying variance reduction algorithms during text generation.

In 2020, Uchendu et al. [4] did not just address the challenge of distinguishing between machine-generated and human-written text, but also formally defined the Authorship Attribution (AA) problem. This task involves identifying the specific Natural Language Generation (NLG) model responsible for producing a given text based on linguistic features such as n-grams, part-of-speech (POS) tags, topic modeling, LIWC, syntactic structure, or stylometric features. They also relied on classical classification methods such as Naive Bayes, SVM, Random Forest, KNN, as well as deep learning architectures like RNN, Stacked CNN, Parallel CNN, CNN-RNN. In 2021 and 2022, open-source autoregressive LLMs such as GPT-Neo, GPT-NeoX-20B [5], and GPT-J-6B [6] were released. These models have facilitated the generation of AI-produced text samples, thereby enhancing the accuracy of detection methods, namely DetectGPT [7] and Fast-DetectGPT [8]. Additionally, GPTZero³ is a commercial application that was specifically developed to detect text generated by ChatGPT in 2023. Another example is Turnitin⁴, which is a well-known commercial platform for detecting AI-generated text, but it particularly focuses on English, Spanish, and Japanese texts.

In the study of DetectGPT [7], Eric Mitchell et al. discovered that text samples generated by LLMs tend to lie in a region of negative curvature of the log-likelihood function of the model. This is also a Zero-shot method, as it does not require retraining or collecting data for classification. Interestingly, compared to the above mentioned methods, DetectGPT analyzes the entire curvature of the log-likelihood probability function instead of relying on the average log-probability threshold per token. Although it is effective in detecting AI-generated text, DetectGPT primarily results in high computational costs due to the need to create numerous perturbed variants of the input text by calling the model or API hundreds of times. An improvement over DetectGPT is Fast-DetectGPT [8], which addresses its limitations by replacing the step of generating perturbed variants of the original text with a more efficient sampling step and uses conditional probability curvature to reduce the number of model calls.

Su et al. [9] proposed another Zero-shot detection method for identifying AI-generated

¹ <https://chat.openai.com>

² <https://claude.ai>

³ <https://gptzero.me>

⁴ <https://www.turnitin.com>

text by analyzing the logarithmic rank of the next token predicted by LLMs. In 2024, Verma et al. introduced an advanced detector named Ghostbuster [10]. Ghostbuster is fine-tuned specifically to detect text generated by ChatGPT, but it is not effective in detecting outputs from other LLMs such as LLaMA [11]. The principle of Ghostbuster allows Human-Machine texts to pass through weaker LLMs than the model being detected, in order to obtain the probability of the next token prediction, which serves as the main component of the extracted features. Finally, these features are passed through a basic classifier like Logistic Regression, to determine whether the text is generated by humans or machines. This method has the advantage of detecting black-box LLMs, such as ChatGPT, as it does not require access to next-token probability distributions from the model.

Building on Zero-Shot Detection, Abhimanyu Hans et al. [12] proposed an additional detection model, referred to as Binoculars. The advantage of Binoculars is that it achieves state-of-the-art accuracy without requiring any training data at the time of its release, outperforming other Zero-Shot Detection methods, including both commercial and open-source approaches. More importantly, Binoculars has practical significance in evaluating detection performance, achieving over 90% accuracy in detecting English text generated by ChatGPT at a False Positive Rate (FPR) of 0.01%, even though it was not trained on ChatGPT. Therefore, in this study, we employ Binoculars to detect Vietnamese text generated by LLMs such as Gemma 3-12B [13], Sailor2-8B [14], and additional models.

In Section 2 of this study, we provide an overview of the Binoculars method proposed by [12]. Then, we present the original Binoculars method adapted for detecting Vietnamese LLM-generated text, referred to as VietBinoculars. Following this, in Section 3, we detail newly constructed Vietnamese AI-generated datasets from the domains of news articles and literary works. Operating on a pair of LLMs, VietBinoculars computes Binoculars scores, which are then used to classify text through our globally optimized thresholds. These thresholds are specifically optimized for the detection of Vietnamese AI-generated and human-written text. Based on our datasets, we evaluate the effectiveness of the VietBinoculars method under various metrics, including accuracy, F1-score, and AUC. We also compare VietBinoculars against the original Binoculars method, other state-of-the-art approaches, and commercial detection tools. Section 4 discusses challenges related to applying the VietBinoculars method to Vietnamese. Finally, Section 5 concludes the paper.

2 Binoculars Method Overview

Currently, there are two main perspectives in academia on detecting whether text is written by humans or generated by AI. First, from the theoretical perspective of Varshney et al. (2020) [15], detecting AI-generated text is considered impossible. However, from an empirical standpoint, any presence of advanced LLMs can be detected if a sufficiently large sample is available for statistical analysis or model training. To validate this perspective, two key issues need to be examined. First, we assess the reliability of the detector

by detecting texts that have been intentionally modified to evade the detection system. Second, for practical applications, it is essential to carefully evaluate the effectiveness of the detector using the FPR metric. False positives are texts created by humans and yet labeled as AI-generated. Only systems based on low FPR criteria are considered good and reliable for commercial applications while minimizing potential harm to users.

Binoculars follows the empirical perspective that it can detect text generated by modified prompts from any LLM at an extremely low FPR (0.01%) without requiring retraining on those specific models. This method outperforms traditional approaches that use the same LLMs as a backbone for fine-tuning and retraining on samples for binary classification tasks [10]. However, Binoculars adopts a Zero-shot approach, utilizing statistical features based on a globally optimal threshold to determine whether text is generated by AI or humans. The advantage of this method is that it does not require retraining on LLMs with hundreds of billions of parameters and numerous released versions while still being capable of detecting them.

In addition, Binoculars offers several advantages over other detectors, such as Ghostbuster, achieving state-of-the-art F1 scores (over 99%) at the time of its release. According to the paper [12], Binoculars demonstrates superior reliability compared to Ghostbuster on out-of-domain datasets. It exhibits high generalization capabilities, being able to detect texts generated by various LLMs such as LLaMa, Falcon, and ChatGPT without requiring detailed modifications for each model. In contrast, Ghostbuster is nearly ineffective at detecting texts generated by LLaMA. Importantly, Binoculars is unaffected by prompting strategies designed to bypass detection systems through prompt modification. Furthermore, Binoculars has commercial value due to its ability to detect black-box LLMs without requiring access to their next-token prediction probabilities unlike DetectGPT, Fast-DetectGPT, and GLTR, which rely on white-box LLMs. Finally, Binoculars performs well not only in English but also in various other languages, whereas well-known commercial tools, for instance Turnitin, largely operate in English. Moreover, it does not exhibit bias against texts generated by non-native English speakers in the task of determining whether English text is generated by humans or AI.

The Binoculars model uses a pair of closely related large language models that share a pre-trained tokenizer to measure the level of surprise (known as perplexity, PPL) on the pair of models and the input text. Binoculars employs a first “observer model” LLM, to calculate $\log(\text{perplexity})$, which serves as a measure of the level of surprise of the input text with respect to the observing model. Next, it uses a second “performer model” LLM, to compute next-token predictions in the input sequence and calculate the surprise score of these predictions from the perspective of the first observing model. This score is referred to as cross-perplexity (X-PPL). According to the research by [12], the Binoculars score is calculated as the ratio of perplexity to cross-perplexity is used as a robust measure for detecting text generated by LLMs or humans.

2.1 VietBinoculars

According to the original Binoculars paper [12], the authors primarily used a pair of Falcon-7B and Falcon-7B-Instruct models [16] to calculate Binoculars scores. However, the Falcon models are fundamentally trained on English and French. The authors noted that for raw text data in non-english languages, the vocabulary may not be well represented in the LLM. Languages such as Vietnamese typically exhibit lower True Positive Rates at low False Positive Rates and are more prone to mispredictions compared to languages that are well represented in the LLM. For our research, we used a pair of custom-designed PhoGPT models for Vietnamese representation, which have nearly 4 billion parameters and a context size of 8,192 [17]. To be specific, we use the PhoGPT-4B version as the observer model and PhoGPT-4B-Chat, a fine-tuned version of the original trained with 70K instructional prompts and 290K dialogue segments, as the performer model. PhoGPT is based on the Transformer decoder architecture for training and employs a BPE tokenizer tailored designed for Vietnamese [18, 19]. In particular, for a

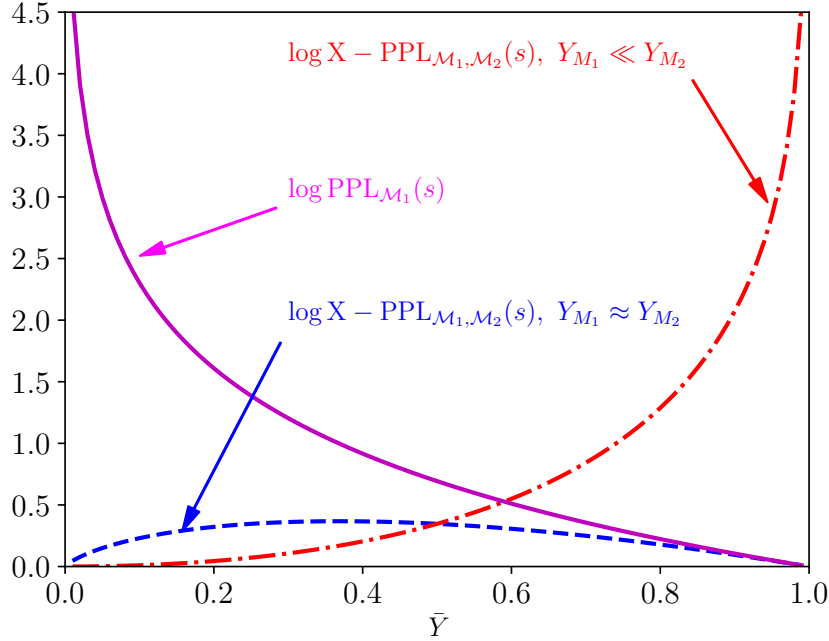


Figure 1: Illustration of log (perplexity) and log (cross-perplexity). The x-axis represents the average next-token prediction probabilities for the same input string s . The solid magenta curve represents $\log \text{PPL}_{M_1}(s)$ and the dashed curves represent $\log X - \text{PPL}_{M_1, M_2}(s)$ for different probability distribution vectors between models M_1 and M_2 . The dashed blue curve illustrates $\log X - \text{PPL}_{M_1, M_2}(s)$ when the observer and performer models are nearly identical. We assume that the observer model M_1 is significantly smaller than the performer model M_2 when the two models differ, as depicted by the dash-dotted red curve.

Vietnamese input string s , the string must be sufficiently long and tokenized using the BPE tokenizer, represented as the vector $\vec{x}$. Using the PhoGPT-4B model, denoted as M_1 , we obtain $\mathcal{M}_1$, a probability distribution vector $\mathbf{Y}$ of the model’s next-token predictions

for each position $i \in [1, L]$ in the string s :

$$\begin{aligned} \text{PhoGPT-4B}(\text{BPE}(s)) &= \mathcal{M}_1(\vec{x}) = \mathbf{Y}, \\ \mathbf{Y}_{ij} &= P_{M_1}(v_j | x_{<i}), \quad \forall j \in \mathcal{V}, \quad i = 1 : L. \end{aligned} \quad (1)$$

Here, $\mathcal{V}$ is the vocabulary of M_1 , L is the number of tokens in the string s , and $\mathbf{Y}_{ij}$ is the probability of the next token v_j based on the preceding tokens $x_{1:i-1}$ in the input string s . We define $\log(\text{perplexity})$ as the average negative log-likelihood of all tokens in s as follows

$$\log \text{PPL}_{\mathcal{M}_1}(s) = -\frac{1}{L} \sum_{i=1}^L \log(\mathbf{Y}_{ix_i}), \quad (2)$$

where $\mathbf{Y}_{ix_i}$ is the predicted probability of the i -th token in the string s , and $\log \text{PPL}_{\mathcal{M}_1}(s)$ measures the surprise of s with respect to the PhoGPT-4B model. According to information theory [20], $\log \text{PPL}_{\mathcal{M}_1}(s)$ can be interpreted as the average cross-entropy of model M_1 over all tokens in s (see Appendix A), given by

$$H(s, \mathcal{M}_1) = \log \text{PPL}_{\mathcal{M}_1}(s). \quad (3)$$

The solid magenta curve in Figure 1 illustrates the relationship between the average next-token prediction probabilities and their $\log(\text{perplexity})$. Clearly, if the predicted probability of the string s according to model M_1 is high, then $\log \text{PPL}_{\mathcal{M}_1}(s)$ will be low. In this case, the observer LLM is not surprised by the input string s because it has been trained to predict similar sequences, allowing model M_1 to predict the next token with high accuracy. Essentially, $\log \text{PPL}_{\mathcal{M}_1}(s)$ is the loss function used to train Causal LLMs for text generation. If $\log \text{PPL}_{\mathcal{M}_1}(s)$ is high, the model struggles to predict the next token, indicating greater surprise because s was not well represented in its training data. Therefore, we can use $\log \text{PPL}_{\mathcal{M}_1}(s)$ to detect text generated by LLMs or humans, based on the intuitive observation that human-written text often has higher perplexity than that produced by LLMs, while LLM-generated text tends to have lower perplexity due to the models’ familiarity with the training data.

Next, the VietBinoculars method uses PhoGPT-4B-Chat as the second performer LLM, denoted as M_2 , on the same string s to measure the output surprise of one model relative to another LLM, defined as cross-perplexity

$$\begin{aligned} \text{PhoGPT-4B-Chat}(\text{BPE}(s)) &= M_2(\text{BPE}(s)) = \mathcal{M}_2(s), \\ \log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s) &= -\frac{1}{L} \sum_{i=1}^L \mathcal{M}_1(s)_i \cdot \log(\mathcal{M}_2(s)_i). \end{aligned} \quad (4)$$

$\mathcal{M}_1(s)_i$ and $\mathcal{M}_2(s)_i$ as shown above, are the probability distribution vectors for next-token predictions by PhoGPT-4B and PhoGPT-4B-Chat, respectively, at position i in the string s . The dot product $\mathcal{M}_1(s)_i \cdot \log(\mathcal{M}_2(s)_i)$ corresponds to cross-entropy, representing the difference between the two probability distributions at the i -th token [12, 20]. The definition of $\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)$ measures the level of “surprise” of the next-token predictions of model M_2 when evaluated under the probability distribution of model M_1 . The

cross-perplexity is illustrated by the dashed curves in Figure 1. If we use two closely related LLMs, the cross-perplexity $\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)$ will be low, as shown by the dashed blue curve, indicating that the performer model M_2 is also unsurprised by the next-token predictions of the observer model M_1 . VietBinoculars also considers cross-perplexity an important quantity for normalizing perplexity. To explain why normalized cross-perplexity is used instead of solely perplexity scores for detection, we need to consider the case where a user intentionally prompts LLMs to generate text by using the question “Can you write a few sentences about a capybara being an astrophysicist?”. Clearly, in this case, $\log \text{PPL}_{\mathcal{M}_1}(s)$ would yield a very high score due to the surprise in the text generated from the next-token predictions for the words “capybara” and “astrophysicist”, leading to the prediction that this text was written by a human. Therefore, VietBinoculars focuses on estimating the baseline of perplexity by comparing it with an expected baseline, namely cross-perplexity. The VietBinoculars score is calculated as follows

$$B_{\mathcal{M}_1, \mathcal{M}_2}(s) = \frac{\log \text{PPL}_{\mathcal{M}_1}(s)}{\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)}. \quad (5)$$

Here $B_{\mathcal{M}_1, \mathcal{M}_2}(s)$ denotes the VietBinoculars score for the input string s when using the observer model M_1 and the performer model M_2 . With the obtained VietBinoculars score, we expect that the next-token predictions for human-written text will yield higher scores than those for machine-generated text. To make it clearer, human-written text tends to be more challenging for next-token prediction (resulting in a lower predicted probability distribution $\mathbf{Y}$), thereby leading to a relatively high value of $\log \text{PPL}_{\mathcal{M}_1}(s)$. In contrast, the next-token predictions of M_2 will not differ significantly from those of M_1 as these two LLMs form a closely related pair that sharing the same tokenizer $T = \text{BPE}$, hence resulting in a relatively low value of $\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)$. Therefore, the VietBinoculars score will be higher in the case of human-written text. Conversely, machine-generated text will exhibit a low value of $\log \text{PPL}_{\mathcal{M}_1}(s)$, while $\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)$ also remains low due to the absence of significant differences in “surprise” between the two models. As a result, the VietBinoculars ratio in this case will be lower than that for human-written text. Moreover, as illustrated by the dashdot red curve in Figure 1, we can clearly observe that M_1 and M_2 must be closely related for $\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)$ to serve as a normalization component, thereby ensuring that the VietBinoculars score remains meaningful.

2.2 Evaluation Metrics for VietBinoculars

In essence, the task of detecting whether text is generated by LLMs or written by Humans, is formulated as a binary classification problem. We can use various binary classification evaluation metrics to assess the effectiveness of the VietBinoculars method, such as Accuracy, Precision, Recall, F1 Score, Confusion Matrix, ROC, or AUC. Most previous studies have employed AUC and F1 as evaluation metrics for detectors [7, 10, 12], whereas VietBinoculars introduces an additional practical evaluation metric that emphasizes the True Positive Rate at a very low False Positive Rate threshold (0.06%) specifically for Vietnamese. In this work, we define the negative class (labeled as 0) as Human-written

text, referring to content authored by Humans, and the positive class (labeled as 1) as AI/LLM/Machine-generated text. Based on the normalized Confusion Matrix, we have:

$$\text{TPR} = \frac{TP}{TP + FN}, \text{FPR} = \frac{FP}{FP + TN}, \quad (6)$$

in which TP is the number of correctly detected LLM-generated texts; FN is the number of LLM-generated texts incorrectly identified as human-written. FP and TN are the number of human-written texts incorrectly predicted as LLM-generated and the number of human-written texts correctly identified, respectively. We compare TPR at a very low FPR reference to minimize the potential harm that the detector may cause to users in practical applications. For example, if a student writes an essay but it is incorrectly labeled as AI/LLM-generated, an error rate of less than or equal to 6 in 10,000 texts is considered acceptable for Vietnamese in practical use.

Based on the criteria mentioned above, we determine the global optimal threshold for classifying texts written by AI or Humans by analyzing the ROC curve using two widely adopted methods: Youden’s J [21] and the Closest Point approach [22]. Note that, instead of using the output as predicted class probabilities, we use the VietBinoculars score to represent points (FPR, TPR) on the graph when varying different thresholds to identify the global optimal threshold. The Youden’s index determines the optimal threshold t^* such that $J(t)$ attains its maximum value:

$$J(t) = \text{TPR}(t) - \text{FPR}(t), \quad t^* = \arg \max_t (J(t)). \quad (7)$$

The optimal threshold, according to Youden’s index, is defined as the point representing the largest vertical distance from the random diagonal in the ROC graph. Next, the Closest Point method finds the point that is nearest to the optimal top-left corner, with coordinates (0,1), based on the Euclidean distance defined as follows

$$d(t) = \sqrt{(\text{FPR}(t) - 0)^2 + (1 - \text{TPR}(t))^2}, \quad t^* = \arg \min_t (d(t)). \quad (8)$$

The optimal threshold t^* in Eq. 8 represents the global optimal threshold determined by the Closest Point approach. This method prioritizes the selection of the optimal point that exhibits a low FPR and a high TPR. In addition, we also employ the method of selecting the optimal point at the 0.06% FPR reference to ensure that this threshold can be applied in practice with the highest level of safety for Vietnamese, by defining the following:

$$t^* = \arg \max_{\text{FPR}(t) \leq 0.06\%} (\text{TPR}(t)). \quad (9)$$

3 Application of VietBinoculars to Vietnamese

To apply the VietBinoculars method to Vietnamese, we first require datasets containing Vietnamese text samples written by both humans and AI. However, most such datasets are

currently unavailable. Thus, we construct datasets of Vietnamese human- and machine-generated text. We then apply the VietBinoculars method to the largest of these datasets to determine the optimal threshold. The detection performance of the method is subsequently evaluated on the remaining datasets, which serve as out-of-domain test sets. Moreover, we compare the performance of the VietBinoculars method with that of other methods on Vietnamese out-of-domain datasets.

3.1 Vietnamese Datasets

First, the Vietnamese texts written by humans are collected from the Vietnamese Online News Dataset on Kaggle [23] and from Vietnamese literary works authored by Vu Trong Phung⁵. The Vietnamese Online News Dataset comprises news articles published up to July 2022 from reputable newspapers in Vietnam. The Vietnamese Literary Works Dataset, which contains both short stories and novels by the Vietnamese author Vu Trong Phung, has been segmented into smaller chunks of fewer than 512 tokens. In the text pre-processing stage for both datasets, we removed samples containing fewer than 50 words, as well as filtered out special characters, HTML tags, and links. Note that, for the VietBinoculars method to perform optimally, the length of news samples should be at least 50 tokens and should not exceed the context window size of the LLM model being used. Longer texts generally exhibit richer statistical features and are therefore easier to detect. After collecting the human-written text samples, we use their first 50 tokens as prompts for

Table 1: Summary of Vietnamese datasets from two domains used in this study. The datasets are categorized into In-Domain and Out-of-Domain types. For the In-Domain category, Sailor2-8B-OptiThreshold-News and Sailor2-8B-Validation-News are used to determine the Vietnamese optimal threshold and to subsequently evaluate the performance of the identified threshold, respectively. The remaining datasets, which belong to the Out-of-Domain category, are employed to assess the generalization capability of VietBinoculars. The number of documents (#Docs) is evenly distributed between the AI-generated and human-written classes. The token counts are determined using the BPE tokenizer of the PhoGPT model.

Domain	Dataset	Target	# Docs	Median tokens/doc	
				Human	AI
News articles ⁶	Sailor2-8B-OptiThreshold-News	In-Domain	82,690	466	576
	Sailor2-8B-Validation-News	In-Domain	38,258	488	627
	Gemma-3-12B-News	Out-of-Domain	14,444	500	592
Literary works ⁷	Sailor2-8B-VuTrongPhung	Out-of-Domain	738	478	532
	Gemma-3-12B-VuTrongPhung	Out-of-Domain	738	478	575

LLMs to generate new text. By removing these initial 50 tokens from the generated output, we obtain the AI-generated texts. The LLMs used in this study are Sailor2-8B-Chat

⁵ <https://doi.org/10.57967/hf/6218>

[14] and Gemma-3-12B-it [13]. Both are large, open-source language models, recently released in 2025, and capable of supporting multiple languages, including Vietnamese. Finally, we obtain the human-written and AI-generated text datasets used in this study for VietBinoculars, as summarized in Table 1. Specifically, we extracted and split the Vietnamese Online News Dataset into three subsets⁶: Sailor2-8B-OptiThreshold-News, Sailor2-8B-Validation-News, and Gemma-3-12B-News, corresponding to the Sailor2-8B-Chat and Gemma-3-12B-it models, respectively. Similarly, from the Vietnamese Literary Works Dataset, we generated two additional datasets⁷, Sailor2-8B-VuTrongPhung and Gemma-3-12B-VuTrongPhung, using the same procedure but with the respective LLMs. Remarkably, among multilingual LLMs with fewer than 10 billion parameters, Sailor2-8B demonstrates exceptional performance for Vietnamese. Therefore, the largest dataset Sailor2-8B-OptiThreshold-News is used to determine the optimal threshold for the VietBinoculars method in Vietnamese. For ease of understanding, Sailor2-8B-OptiThreshold-News can be regarded as the “training” dataset and Sailor2-8B-Validation-News as the “validation” dataset, both classified as In-Domain. The remaining datasets serve as “test” datasets and belong to the Out-of-Domain category. However, this analogy is only for illustrative purposes, as VietBinoculars does not require any training process. Some details of the datasets listed in Table 1 are provided in Appendix B.

3.2 The Global Optimal Threshold for VietBinoculars

In this section, we calculate the VietBinoculars score on the largest dataset, Sailor2-8B-OptiThreshold-News, using the observer model M_1 and the performer model M_2 . The VietBinoculars scores are computed using the formula in Eq. 5. Based on these scores, we determine the global optimal threshold using Youden’s J and Closest Point methods, as well as by selecting the global optimal point at 0.06% FPR on the ROC curve, as shown in Figure 2. The optimal thresholds obtained were 0.86, 0.87, and 0.70 for Youden’s J, Closest Point, and TPR@0.06%FPR, respectively. We can observe that the TPR@0.06%FPR threshold is lower than the other thresholds, since it must ensure that the FPR is approximately zero. Moreover, the Detection AUC of the VietBinoculars method on this dataset is approximately 0.98. With these thresholds, we can achieve a detection accuracy of 96% for the Sailor2-8B-Validation-News dataset. The comparison of detection results, including Precision, Recall, Accuracy, and F1-score on the Sailor2-8B-OptiThreshold-News and Sailor2-8B-Validation-News datasets using the global thresholds identified above is presented in Table 2.

Overall, based on the results in Table 2, the Youden’s J and Closest Point global thresholds yield a detection accuracy and F1-score of 96%. The TPR@0.06%FPR threshold achieves slightly lower performance, with 74% accuracy and 72% F1-score on the Sailor2-8B-Validation-News dataset. Nevertheless, this level of performance provides acceptable confidence for applying VietBinoculars in real-world applications. Additionally, detailed detection results on the Sailor2-8B-Validation-News dataset are presented in the

⁶ <https://doi.org/10.57967/hf/6233> ⁷ <https://doi.org/10.57967/hf/6234>

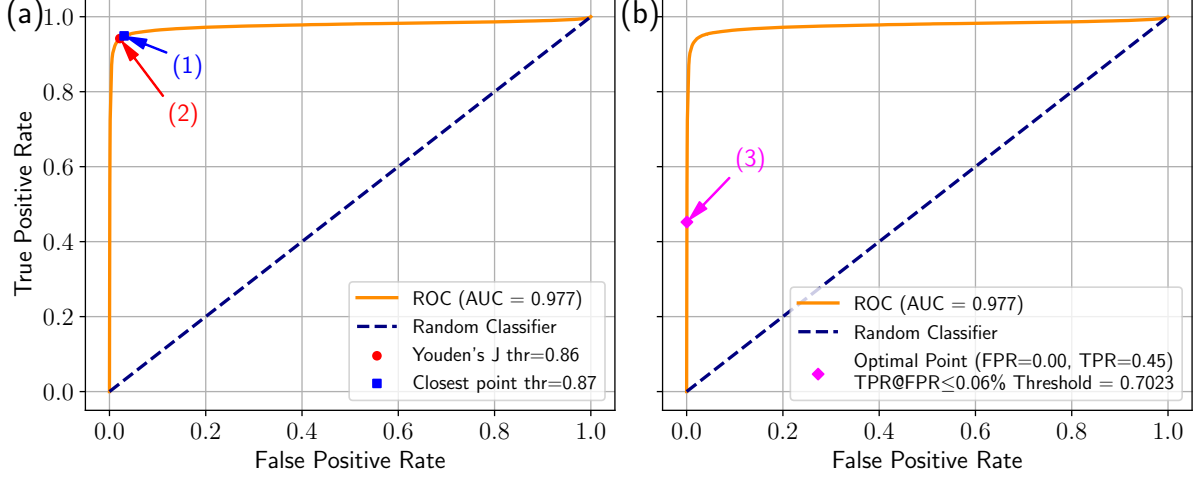


Figure 2: Global optimal threshold points represented on the ROC curve, based on VietBinoculars scores for the Sailor2-8B-OptiThreshold-News dataset: (1)–Youden’s J point, (2)–Closest point and (3)–Optimal point at the 0.06% FPR. In figure (a), Youden’s J and Closest thresholds are marked by a red circle [●] and a blue square [■], respectively. The magenta diamond [◆] denotes the TPR@0.06%FPR threshold in figure (b).

Table 2: Comparison of detection results using global thresholds of VietBinoculars (Figure 2) on the Sailor2-8B-OptiThreshold-News dataset (denoted by [♣]) and the Sailor2-8B-Validation-News dataset (denoted by [♠]). The symbols corresponding to VietBinoculars global thresholds used in the table are: [●] for Youden’s J, [■] for Closest Point, and [◆] for TPR@0.06%FPR. Since the VietBinoculars method is based on Zero-shot learning, there is almost no performance difference between the “training” [♣] and “validation” [♠] datasets, with both achieving the same highest F1-score of 0.96. **Bold** shows the highest F1-score within each column for the in-domain datasets.

Dataset	Class/ Metric	Precision			Recall			F1-score			Support
		[●]	[■]	[◆]	[●]	[■]	[◆]	[●]	[■]	[◆]	
[♣]	Human	0.94	0.95	0.65	0.98	0.97	1.00	0.96	0.96	0.78	41345
	AI	0.98	0.97	1.00	0.94	0.95	0.45	0.96	0.96	0.62	41345
	accuracy							0.96	0.96	0.73	82690
	macro avg	0.96	0.96	0.82	0.96	0.96	0.73	0.96	0.96	0.70	82690
	weighted avg	0.96	0.96	0.82	0.96	0.96	0.73	0.96	0.96	0.70	82690
[♠]	Human	0.95	0.96	0.66	0.98	0.97	1.00	0.96	0.96	0.79	19129
	AI	0.98	0.97	1.00	0.95	0.96	0.48	0.96	0.96	0.65	19129
	accuracy							0.96	0.96	0.74	38258
	macro avg	0.96	0.96	0.83	0.96	0.96	0.74	0.96	0.96	0.72	38258
	weighted avg	0.96	0.96	0.83	0.96	0.96	0.74	0.96	0.96	0.72	38258

Confusion Matrix shown in Figure 3.

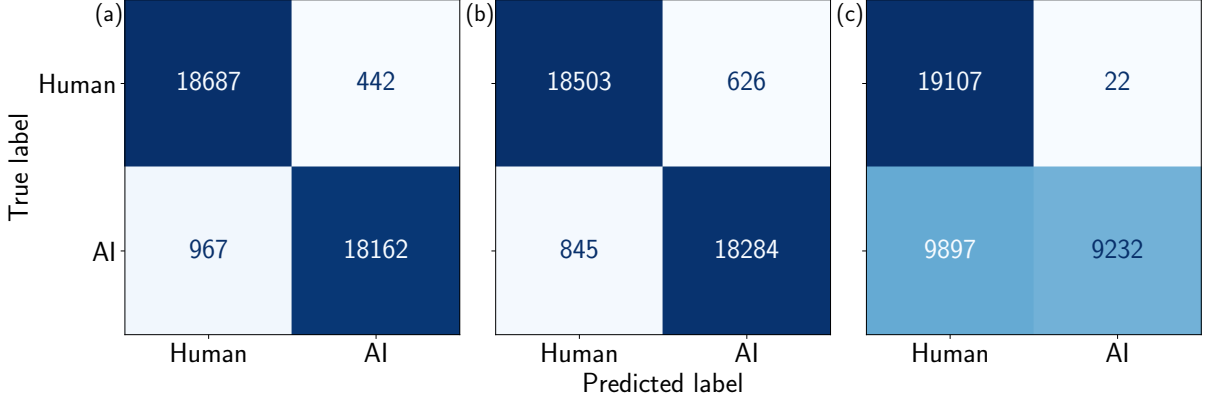


Figure 3: Confusion Matrix of VietBinoculars on the Sailor2-8B-Validation-News: (a)–Youden’s J threshold, (b)–Closest Point threshold, and (c)–TPR@0.06%FPR threshold.

3.3 Applying VietBinoculars to Out-of-Domain Datasets

To generalize VietBinoculars, we evaluate its effectiveness on out-of-domain datasets (different from the dataset used to determine the optimal threshold) using the global optimal thresholds described in Section 3.2. Notably, we use the Gemma-3-12B-News and Gemma-3-12B-VuTrongPhung datasets generated by the latest Gemma-3-12B model, together with the Sailor2-8B-VuTrongPhung dataset generated by Sailor2-8B-Chat, for out-of-domain performance evaluation. Table 3 shows that the VietBinoculars method achieves up to 99% accuracy and F1-score using the Youden’s J threshold on the large out-of-domain dataset (Gemma-3-12B-News). Moreover, VietBinoculars demonstrates

Table 3: Generalization results of VietBinoculars across a variety of out-of-domain datasets (F1-score and Accuracy). Result comparisons are presented according to the VietBinoculars global thresholds, using the following symbols: [●] for Youden’s J, [■] for Closest Point, and [◆] for TPR@0.06%FPR. The datasets used for evaluation include Gemma-3-12B-News, Gemma-3-12B-VuTrongPhung, and Sailor2-8B-VuTrongPhung. **Bold** indicates the best performance achieved by VietBinoculars on each out-of-domain dataset. Values in the table are rounded to two decimal places.

Dataset	F1-score			Accuracy			Support
	[●]	[■]	[◆]	[●]	[■]	[◆]	
Gemma-3-12B-News	0.99	0.98	0.94	0.99	0.98	0.94	14,444
Gemma-3-12B-VuTrongPhung	1.0	1.0	0.95	1.0	1.0	0.95	738
Sailor2-8B-VuTrongPhung	1.0	1.0	0.83	1.0	1.0	0.84	738

strong generalization capabilities on the smaller out-of-domain datasets (Gemma-3-12B-VuTrongPhung and Sailor2-8B-VuTrongPhung), achieving greater than 99% accuracy and F1-score with both the Youden’s J and Closest Point thresholds. In particular, the TPR@0.06%FPR threshold yields reasonable performance in real-world scenarios, with

F1-scores of 94%, 95%, and 83% on the Gemma-3-12B-News, Gemma-3-12B-VuTrongPhung, and Sailor2-8B-VuTrongPhung datasets, respectively.

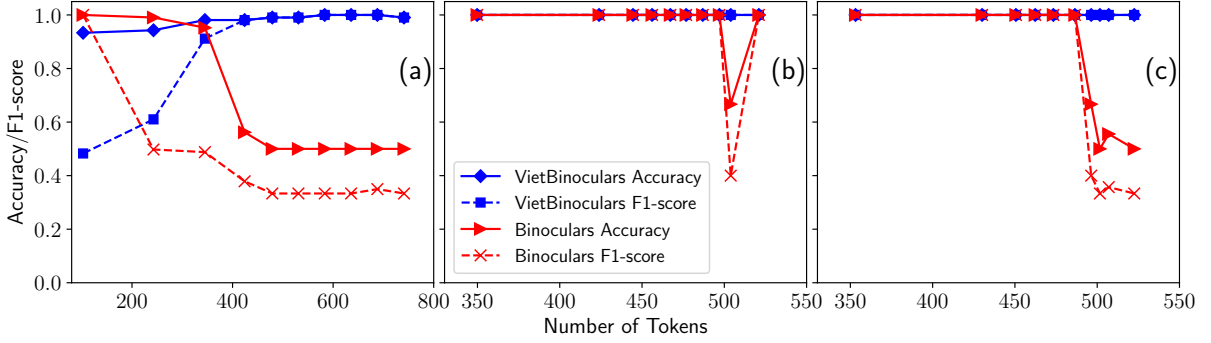


Figure 4: The effect of text length (in tokens) on the detection performance of VietBinoculars and the original Binoculars method on Vietnamese out-of-domain datasets: (a)–Gemma-3-12B-News; (b)–Gemma-3-12B-VuTrongPhung; (c)–Sailor2-8B-VuTrongPhung. The x-axis shows the number of tokens, calculated using the BPE tokenizer. Solid and dashed curves represent Accuracy and F1-score, respectively. Red curves with markers $\blacktriangleright$ and $\times$ correspond to the original Binoculars method, while blue curves with markers $\blacklozenge$ and $\blacksquare$ correspond to VietBinoculars. Youden’s J threshold is used for VietBinoculars to determine the maximum detection performance. Binoculars employs Falcon-7B and Falcon-7B-Instruct as the observer model M_1 and the performer model M_2 , respectively.

In the subsequent step, we study the impact of the number of tokens on the detection performance of VietBinoculars. We conduct experiments on the out-of-domain datasets and compare the results with the original Binoculars, which demonstrates good performance primarily in English. We use Accuracy and F1-score, represented by solid and dashed curves respectively, to compare detection performance. Figure 4 shows that the original Binoculars method performs worse than VietBinoculars on Vietnamese text, particularly when the number of tokens is sufficiently large. Moreover, the results presented in Fig. 4(a) indicate that as the minimum token length increases, the detection performance of VietBinoculars improves significantly on the Gemma-3-12B-News dataset. Similarly, the Binoculars method shows a notable decline in performance once the input length exceeds 480 tokens on the Gemma-3-12B-VuTrongPhung and Sailor2-8B-VuTrongPhung datasets (see Fig. 4(b), (c)). Clearly, longer text samples provide richer statistical features, making them easier to detect using VietBinoculars or other Zero-shot detection methods.

Furthermore, the comparison of Detection AUC between Zero-shot methods and supervised learning methods on Vietnamese out-of-domain datasets is presented in Figure 5. The Zero-shot methods include VietBinoculars, Binoculars [12], Likelihood (average per-token log probability), Rank/LogRank (average observed rank or log-rank of tokens), Entropy (average entropy in the model’s predictive distribution) and DetectGPT [1, 7]. The supervised learning methods include OpenAI’s RoBERTa-based GPT-2 detector models [24, 25]. The results in Figure 5 show that VietBinoculars consistently outperforms the

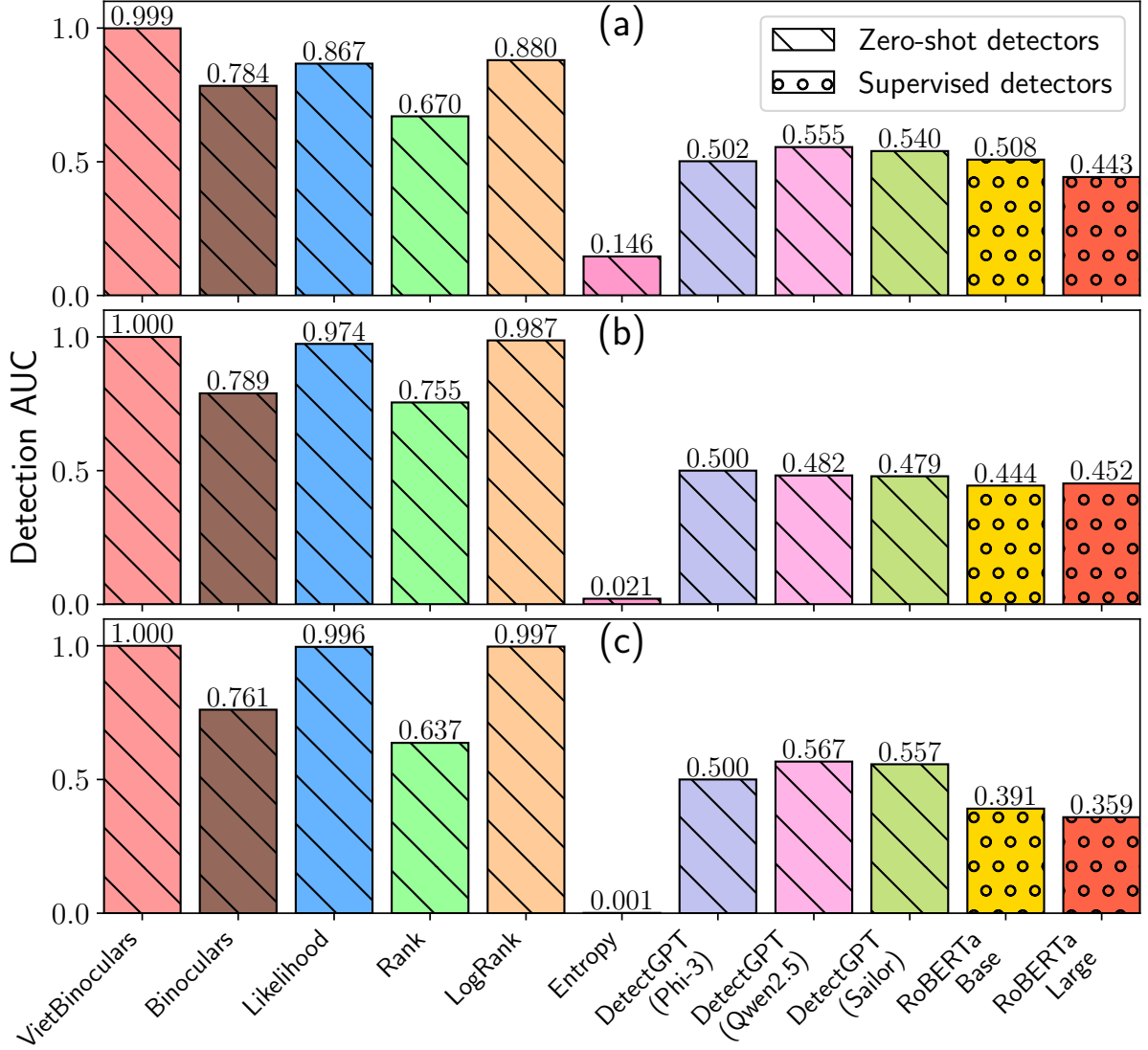


Figure 5: Detection AUROC for various Zero-shot and supervised learning methods on Vietnamese out-of-domain datasets: (a)–Gemma-3-12B-News; (b)–Gemma-3-12B-VuTrongPhung; (c)–Sailor2-8B-VuTrongPhung. Zero-shot detectors are represented with the “\” hatch pattern, while supervised learning detectors are represented with the “o” hatch pattern. DetectGPT uses top- k and top- p sampling with parameters $k = 40$ and $p = 0.96$.

other Zero-shot methods across all datasets. Notably, VietBinoculars achieves a Detection AUC greater than 0.99, which is significantly higher than the approximately 0.76 obtained by the original Binoculars method on Vietnamese out-of-domain datasets. Likewise, VietBinoculars surpasses the performance of other Zero-shot methods, such as Rank, Entropy and several variants of DetectGPT. In this experiment, we use FLAN-T5 Large [26], a pre-trained mask-filling model with support for Vietnamese, to generate 20 perturbed variants of each input for the DetectGPT method. Specifically, we randomly

mask 2-word spans until 15% of the words are replaced. Sailor-7B⁸, Qwen2.5-7B⁹ and Phi-3-mini-4k-instruct¹⁰ are used as source models for detection evaluation. Additionally, VietBinoculars demonstrates competitive performance compared to supervised learning methods, achieving a Detection AUC of 1.000 on the literary works domain, which is comparable to the RoBERTa Base and RoBERTa Large OpenAI detectors. The failure of the OpenAI detector models can be explained by the fact that they were trained only to detect text generated by GPT-2, whereas our out-of-domain dataset was produced by recently released state-of-the-art LLMs. The results of the above experiments show that VietBinoculars, with its Zero-shot approach, has a strong generalization capability for Vietnamese AI-human text detection across different domains and LLMs.

3.4 Capybara Problem in Vietnamese

In this section, we evaluate the effectiveness of the VietBinoculars method in a special case where users intentionally prompt LLMs to generate text, for example: “*Bạn có thể viết một vài câu kể chuyện về một con capybara là một nhà vật lý thiên văn không?*” (English version is “*Can you write a few sentences telling a story about a capybara being an astrophysicist?*”). The LLMs¹¹ used to generate text from the capybara prompt include GPT-4o, GPT-4.1, GPT-4.1-mini, GPT-3.5-Turbo, Llama-3.1-405B, Llama-4-Scout-Nitro, Llama-4-Maverick-T, Grok-3, Gemini-2.5-Flash, Gemini-2.5-Pro, Gemma-3-27B, Claude-3-Sonnet, Claude-3-Haiku, Claude-3.5-Sonnet, Claude-3.7-Sonnet, DeepSeek-R1, DeepSeek-V3, Qwen3-235B-A22B-FW, Qwen-2.5-72B-T, Mistral-Medium, and Mistral-Large-2. In this case, AI-Human detection models based on statistical features tend to misclassify significantly due to the unexpected presence of tokens such as “*capybara*” and “*nhà vật lý thiên văn*” (English version is “*astrophysicist*”). Based on the optimal threshold for Vietnamese determined in Section 3.2, we compare the detection performance of VietBinoculars with other popular detection methods, including RadarTester [27], the commercial products DetectGPT¹² and GPTZero³, as well as Ghostbuster [10]. In order to determine whether a text is more likely AI-generated or human-written using RadarTester, we rely on four backbone models: Dolly V2-3B, Camel-5B, Dolly V1-6B, and Vicuna-7B. The closer the value is to 1, the more likely the text is classified as AI-generated; conversely, the closer it is to 0, the more likely it is classified as human-written. In this experiment, the default prediction threshold for probabilities is typically set at 0.5. If any of the four models produces a value greater than 0.5, the text is classified as AI-generated. However, RadarTester is only effective in detecting capybara texts when they are generated by DeepSeek-R1, while failing to detect those generated by other LLMs. (see Appendix C). Similarly, the commercial detection tools and Ghostbuster also adopt default thresholds of 0.5, classifying a text as AI-generated when the predicted probability is greater than or equal to 0.5. In fact, the default thresholds used by these tools are not really reliable to classify texts generated by LLMs or written by humans. Figure 6

⁸ <https://huggingface.co/sail/Sailor-7B>

⁹ <https://huggingface.co/Qwen/Qwen2.5-7B>

¹⁰ <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct>

¹¹ <https://poe.com>

¹² <https://detectgpt.com>

presents the detection results of VietBinoculars with different optimal thresholds, along with other popular detectors, on the Capybara problem in Vietnamese. We observe that the Binoculars detection model demonstrates superior performance, achieving 80.95% accuracy when using the optimal Closest Point threshold. Even the Binoculars detection model with TPR@0.06%FPR reliability achieved higher detection accuracy (42.86%) than Ghostbuster (33.33%), which recorded the highest accuracy among the comparison tools mentioned above. Meanwhile, the commercial tools GPTZero and DetectGPT only achieve 14.29% accuracy, while Turnitin does not even support Vietnamese.

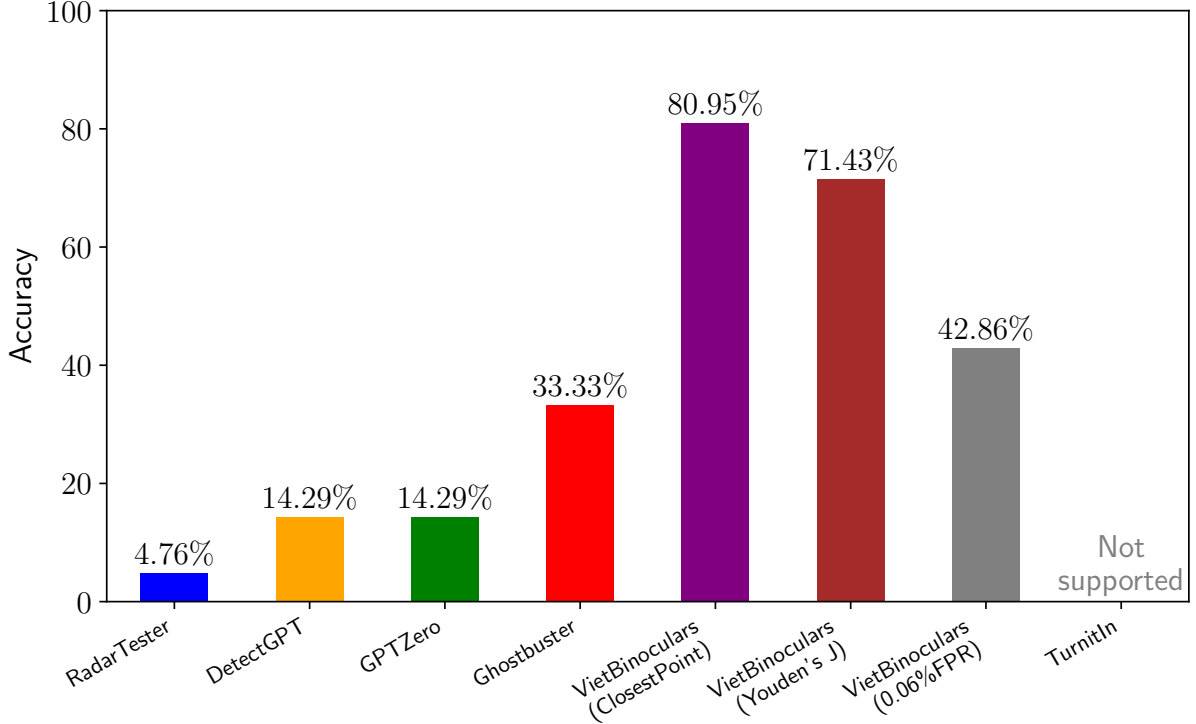


Figure 6: Comparison of VietBinoculars with popular detection methods on the Capybara problem in Vietnamese Language. The x-axis denotes different detection methods, while the y-axis indicates their detection accuracy. VietBinoculars was evaluated using three different global thresholds as described above. Other detection methods, such as RadarTester, which is based on four backbone models (Dolly V2-3B, Camel-5B, Dolly V1-6B, and Vicuna-7B), along with DetectGPT, GPTZero, and Ghostbuster, all apply a default prediction threshold of 0.5 for classification. Commercial tools exhibit poor detection performance on the Vietnamese Capybara problem.

4 Discussion

With detection accuracy and F1-score exceeding 99% on the out-of-domain Vietnamese dataset, the VietBinoculars method demonstrates its effectiveness, simplicity, and resource efficiency, while requiring no retraining of models under Zero-shot Detection. It also

exhibits higher reliability compared to other common methods when applied in practical scenarios. However, in this study, we only used the news domain, with AI-generated texts derived from the Sailor2-8B model, to determine the optimal threshold for VietBinoculars. Due to the limited diversity in real-world data collection across domains and the restricted range of LLMs used to generate AI text, the thresholds we identified may not fully capture the statistical characteristics of Vietnamese. The generated texts in our experiments were required to meet certain minimum length conditions; however, in practice, not all generated texts fall within the range of 300-750 tokens.

In addition, this study has not yet tested or evaluated the effectiveness of VietBinoculars on generated texts produced under different creative parameter settings of LLMs, such as *temperature* and *top_p*. Instead, we conducted experiments only with the default parameter values, as described in Appendix B. This limitation may lead to a decrease in detection performance when users modify the creativity settings of LLMs, resulting in more diverse and less predictable generated texts.

The VietBinoculars method also encounters deployment resource and inference efficiency challenges similar to those of the original Binoculars method. Although VietBinoculars does not require retraining or fine-tuning large language models for binary classification tasks, it still relies on two separate models, an observer and a performer, which increases computational and deployment overhead. Consequently, this effectively doubles the resource requirements for the detection system. For example, several experiments in this study involving LLMs with more than 7 billion parameters per model were executed on an Ubuntu Server 20.04.6 LTS, equipped with a 4-core Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz, 32GB of RAM, and an NVIDIA A100 Tensor Core GPU with 40GB of VRAM. A batch size of 10 was used to ensure stable inference and efficient resource utilization.

This paper has so far only evaluated VietBinoculars on standard Vietnamese texts. In future work, we plan to extend our experiments to other types of content, such as detecting code generation by LLMs. Moreover, VietBinoculars can be applied as a data-cleaning tool to improve the quality of Vietnamese training datasets. By identifying and filtering out AI-generated texts that may have been unintentionally included, we can ensure higher dataset integrity and reliability, ultimately leading to better model performance in downstream tasks. Finally, when deploying VietBinoculars in practice, it is important to acknowledge that errors may occur in detecting Vietnamese AI-generated texts. Therefore, the labeling results should be regarded as reference guidance rather than definitive judgments, and all final decisions should be made under careful human supervision.

5 Conclusion

In this study, we introduced VietBinoculars, a Zero-shot detection method for distinguishing between human and AI-generated text. We provided a theoretical overview of the VietBinoculars method using PhoGPT-4B as the observer model and PhoGPT-4B-

Chat as the performer model in Section 2. Furthermore, in Section 3, we constructed new Vietnamese AI-generated datasets from the domains of news articles and literary works to apply the VietBinoculars method on these datasets. By using the largest dataset, Sailor2-8B-OptiThreshold-News, we determined the global optimal thresholds—Youden’s J, Closest Point, and $\text{TPR}@0.06\%\text{FPR}$ —for detecting whether texts were generated by AI or written by humans. The evaluation results of the VietBinoculars method on out-of-domain datasets such as Gemma-3-12B-News, Gemma-3-12B-VuTrongPhung, and Sailor2-8B-VuTrongPhung show that it achieved performance levels of at least 99% in terms of accuracy, F1-score, and AUC. Moreover, VietBinoculars demonstrated superior detection performance compared to the original Binoculars, as well as other Zero-shot and supervised learning methods, on Vietnamese out-of-domain datasets. We also compared the detection effectiveness of the VietBinoculars method with existing commercial and open-source tools on the challenging Vietnamese Capybara detection case, where VietBinoculars overwhelmingly outperformed the alternatives mentioned in Section 3.4. Finally, we discussed the limitations and practical challenges associated with deploying the VietBinoculars method.

For future work, we plan to explore the application of VietBinoculars in other domains, such as code generation detection, to aid in checking student programming assignments. We also intend to integrate VietBinoculars into the Vietnamese plagiarism checker system. Additionally, we aim to investigate the impact of varying creativity parameters (e.g., *temperature* and *top_p*) on the detection performance of VietBinoculars. Furthermore, we will explore the use of VietBinoculars as a data-cleaning tool to enhance the quality of Vietnamese training datasets by filtering out AI-generated texts. Finally, we will continue to refine the method to minimize detection errors and improve its reliability in practical applications.

Acknowledgment

This research has received no external funding

References

- [1] S. Gehrmann, H. Strobelt, and A. Rush, “GLTR: Statistical detection and visualization of generated text,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, M. R. Costa-jussà and E. Alfonseca, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 111–116. [Online]. Available: <https://aclanthology.org/P19-3019/>
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI*, 2019, accessed: 2025-01-

27. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [3] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, *Defending against neural fake news*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [4] A. Uchendu, T. Le, K. Shu, and D. Lee, “Authorship attribution for neural text generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 8384–8395. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.673/>
- [5] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, “GPT-NeoX-20B: An open-source autoregressive language model,” in *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, A. Fan, S. Ilic, T. Wolf, and M. Gallé, Eds. virtual+Dublin: Association for Computational Linguistics, May 2022, pp. 95–136. [Online]. Available: <https://aclanthology.org/2022.bigscience-1.9/>
- [6] B. Wang and A. Komatsuzaki, “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model,” May 2021. [Online]. Available: <https://github.com/kingoflolz/mesh-transformer-jax>
- [7] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 24 950–24 962. [Online]. Available: <https://proceedings.mlr.press/v202/mitchell23a.html>
- [8] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, “Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [9] J. Su, T. Zhuo, D. Wang, and P. Nakov, “DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 395–12 412. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.827/>
- [10] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, “Ghostbuster: Detecting text ghostwritten by large language models,” in *Proceedings of the 2024 Conference of the*

- North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1702–1717. [Online]. Available: <https://aclanthology.org/2024.naacl-long.95/>
- [11] AI@Meta, “Llama 3 model card,” https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024, accessed: 2024-12-01.
- [12] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein, “Spotting LLMs with binoculars: Zero-shot detection of machine-generated text,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 17 519–17 537. [Online]. Available: <https://proceedings.mlr.press/v235/hans24a.html>
- [13] G. Team and collaborators, “Gemma 3 Technical Report,” *arXiv e-prints*, p. arXiv:2503.19786, Mar. 2025.
- [14] L. Dou, Q. Liu, F. Zhou, C. Chen, Z. Wang, Z. Jin, Z. Liu, T. Zhu, C. Du, P. Yang, H. Wang, J. Liu, Y. Zhao, X. Feng, X. Mao, M. T. Yeung, K. Pipatanakul, F. Koto, M. S. Thu, H. Kydlíček, Z. Liu, Q. Lin, S. Sripaisarnmongkol, K. Sae-Khow, N. Thongchim, T. Konkaew, N. Borijindargoon, A. Dao, M. Maneegard, P. Artkaew, Z.-X. Yong, Q. Nguyen, W. Phatthiyaphaibun, H. H. Tran, M. Zhang, S. Chen, T. Pang, C. Du, X. Wan, W. Lu, and M. Lin, “Sailor2: Sailing in south-east asia with inclusive multilingual llms,” *arXiv e-prints*, 2025. [Online]. Available: <https://arxiv.org/abs/2502.12982>
- [15] L. R. Varshney, N. Shirish Keskar, and R. Socher, “Limits of detecting text generated by large-scale language models,” in *2020 Information Theory and Applications Workshop (ITA)*, 2020, pp. 1–5.
- [16] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocar, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, “The Falcon Series of Open Language Models,” *arXiv e-prints*, p. arXiv:2311.16867, Nov. 2023.
- [17] D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, “PhoGPT: Generative Pre-training for Vietnamese,” *arXiv e-prints*, p. arXiv:2311.02945, Nov. 2023.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

- [19] T. H. Nguyen, T. K. N. Pham, T. H. M. Bui, and T. Q. C. Nguyen, “Clustering vietnamese conversations from facebook page to build training dataset for chatbot,” *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 08, no. 01, pp. 1 – 17, 2022.
- [20] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948. [Online]. Available: <http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>
- [21] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3>
- [22] N. J. Perkins and E. F. Schisterman, “The inconsistency of “optimal” cut-points using two roc based criteria,” *American journal of epidemiology*, vol. 163, no. 7, pp. 670–675, 2006.
- [23] H. T. Quang, “Vietnamese online news dataset,” <https://www.kaggle.com/datasets/haitranquangofficial/vietnamese-online-news-dataset>, 2022, accessed: 2025-09-01.
- [24] I. Solaiman, M. Brundage, J. Clark, A. Askeel, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, “Release Strategies and the Social Impacts of Language Models,” *arXiv e-prints*, p. arXiv:1908.09203, Aug. 2019.
- [25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [26] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tai, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, “Scaling instruction-finetuned language models,” *J. Mach. Learn. Res.*, vol. 25, no. 1, Jan. 2024.
- [27] X. Hu, P.-Y. Chen, and T.-Y. Ho, “Radar: robust ai-text detection via adversarial learning,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2023.

Appendix

A Average Cross-Entropy Representaion for $\log(\textit{perplexity})$

Cross-entropy between two probability distributions P and Q over a set of events X is defined as

$$H(P, Q) = - \sum_{x \in X} P(x) \log Q(x). \quad (10)$$

Where $P(x)$ is the probability of event x according to distribution P , and $Q(x)$ is the probability of event x according to distribution Q . Cross-entropy measures the difference between two probability distributions. In the context of $\log(\textit{perplexity})$, we consider cross-entropy as a measure of the difference between the empirical distribution P of tokens in a sequence s and the predicted distribution P_{M_1} of tokens by a language model M_1 . The empirical distribution P is defined as

$$P(v_j | x_{<i}) = \begin{cases} 1, & \text{if } v_j = x_i \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

The cross-entropy between the empirical distribution P and the predicted distribution P_{M_1} at token x_i in string s is given by

$$H(x_i, P_{M_1}) = - \sum_{v_j \in \mathcal{V}} P(v_j | x_{<i}) \log P_{M_1}(v_j | x_{<i}). \quad (12)$$

Substituting the empirical distribution in Eq.11 into Eq.12 at $v_j = x_i$ yields:

$$H(x_i, P_{M_1}) = -1 \log P_{M_1}(x_i | x_{<i}) = -\log(Y_{ix_i}). \quad (13)$$

Base on Eq.13, the average cross-entropy for the entire sequence s with L tokens can be expressed as

$$H(s, P_{M_1}) = \frac{1}{L} \sum_{i=1}^L H(x_i, P_{M_1}) = -\frac{1}{L} \sum_{i=1}^L \log(Y_{ix_i}) \quad (14)$$

By combining the notations from Eq.1 with the above equation, the average cross-entropy can be rewritten as

$$H(s, \mathcal{M}_1) = \log \text{PPL}_{\mathcal{M}_1}(s). \quad (15)$$

B Illustration of Datasets

B.1 Dataset structure

There are two main keys in our datasets that should be noted. The first key is “text”, which corresponds to the human-written text samples. The second key refers to text samples generated by LLMs, with dataset-specific naming conventions:

- ★ sail-Sailor2-8B-Chat-generated_text for Sailor2-8B-OptiThreshold-News and Sailor2-8B-Validation-News Datasets,
- ★ google-gemma-3-12b-it-generated_text for Gemma-3-12B-News Dataset,
- ★ sail-Sailor2-8B-Chat-hf_generated_text_wo_prompt for Sailor2-8B-VuTrongPhung Dataset,
- ★ google-gemma-3-12b-it-hf_generated_text_wo_prompt for Gemma-3-12B- VuTrongPhung Dataset.

In addition, our datasets also include a “prompt” key, , which corresponds to the prompt used by LLMs to generate the text samples. For example, in the Sailor2-8B-VuTrongPhung dataset, the prompt is a combination of “*system_prompt*” and “*user_prompt*”. The “prompt” is defined as:

Bạn là một trợ lý văn học của tôi, hãy sinh ra một $\{type\}$ dựa vào đoạn văn dưới đây: \n $\{user_prompt\}$

Its English translation is:

As my literary assistant, please generate a $\{type\}$ based on the following text:
 \n $\{user_prompt\}$

Here, *type* refers to the category of Vu Trong Phung’s writing we wish to generate, such as *novel* or *short story*. The “*user_prompt*” consists of the first 50 tokens of each chunk, as described in Section 3.1. We also attach the LLM generation parameters under the “gen_meta” key, which records the configuration used to produce the text samples. Finally, all datasets are stored in the *.jsonl* format. The first line of the Sailor2-8B-OptiThreshold-News dataset is illustrated below:

```

1 {
2   "id": "bf3589f91e473f89",
3   "text": "Theo thống kê của Sở Y tế Đắk Lắk, trong 6 tháng qua, toàn ngành y
            tế tỉnh có 36 viên chức y tế xin thôi việc. Trong đó, có 26 bác sĩ, 4 đ
            iều dưỡng, 1 nhân viên kỹ thuật và 5 viên chức ở các vị trí khác nhau.
            Trong số đó, Bệnh viện Đa khoa vùng Tây Nguyên là đơn vị có nhiều viên
            chức y tế thôi việc (16 bác sĩ). Mỗi trường hợp đều đưa ra lý do nghỉ
            việc khác nhau, tuy nhiên, theo ông Nay Phi La - Giám đốc Sở Y tế Đắk
            Lắk, nguyên nhân viên chức y tế xin thôi việc, bỏ việc xuất phát từ
            thực trạng quá tải công việc, ảnh hưởng của đại dịch COVID-19, môi
            trường làm việc áp lực và một số vị trí làm việc dễ bị lây nhiễm trong
            quá trình tiếp xúc bệnh nhân. Trong khi đó, chế độ tiền lương hưởng theo
            chức danh nghề nghiệp và các chế độ phụ cấp khác tính theo mức lương tối
            thiểu chung còn thấp, chưa tương xứng so với công việc, vị trí công
            tác. Ngoài ra, chính sách đãi ngộ đối với viên chức y tế đã có nhưng
            chưa tạo ra sự khác biệt lớn để khuyến khích cán bộ y tế yên tâm công
```

tác tại các cơ sở y tế công lập dẫn đến tình trạng có nhiều viên chức, nhân viên y tế xin thôi việc, bỏ việc để chuyển công tác đến cơ sở y tế tư nhân, nơi có mức thu nhập cao hơn, ổn định hơn. Ông Nay Phi La thông tin thêm, địa phương đang đẩy mạnh tiến độ tiêm chủng vắc xin phòng COVID-19 và phòng, chống sốt xuất huyết đang diễn biến phức tạp, rất cần nhiều nhân lực. Do đó, loạt bác sĩ, nhân viên y tế nghỉ việc trong thời gian này đã phần nào ảnh hưởng đến công tác chăm sóc sức khỏe tại các cơ sở công lập. Thời gian tới, Sở Y tế tiếp tục tuyển dụng các vị trí còn thiếu trong chỉ tiêu biên chế; cấp tốc bồi dưỡng, đào tạo nguồn nhân lực tại chỗ. Bên cạnh đó, Sở Y tế Đắk Lắk đã có công văn kiến nghị Bộ Y tế đề xuất với Chính phủ xem xét ban hành chính sách tiền lương đặc thù cho đối tượng viên chức, nhân viên y tế; chính sách thu hút, ưu đãi, hỗ trợ đối với cán bộ y tế công tác ở lĩnh vực y tế dự phòng và tuyến y tế cơ sở nhằm tháo gỡ những khó khăn và thu hút viên chức ngành y tế yên tâm làm việc trong hệ thống y tế công lập.",

4 "user_prompt": "Theo thống kê của Sở Y tế Đắk Lắk, trong 6 tháng qua, toàn ngành y tế tỉnh có 36 viên chức y tế xin thôi việc. Trong đó, có 26 bác sĩ,",

5 "topic": "unknown",

6 "sail-Sailor2-8B-Chat-generated_text": "4 điều dưỡng và 6 nhân viên khác.\n\nTóm tắt: Tỉnh Đắk Lắk đã chứng kiến sự gia tăng đáng kể về số lượng nhân viên y tế từ bỏ công việc của họ trong nửa đầu năm nay, với tổng cộng 36 người, bao gồm 26 bác sĩ, 4 điều dưỡng và 6 nhân viên hỗ trợ, theo dữ liệu được cung cấp bởi Sở Y tế địa phương. Tình hình này làm nổi bật những thách thức đang diễn ra mà hệ thống chăm sóc sức khỏe phải đối mặt ở khu vực này.\n\nCác yếu tố góp phần vào sự ra đi hàng loạt:\n\n1. **Lương thấp**: Mức lương cho các chuyên gia y tế thường không cạnh tranh so với chi phí sinh hoạt ngày càng tăng ở Đắk Lắk, dẫn đến cảm giác bất mãn tài chính giữa nhiều nhân viên.\n\n2. **Điều kiện làm việc kém**: Các cơ sở hạ tầng y tế, chẳng hạn như thiết bị lỗi thời, môi trường làm việc không an toàn và thiếu nguồn lực, ảnh hưởng tiêu cực đến tinh thần và hiệu suất của nhân viên.\n\n3. **Thiếu cơ hội phát triển nghề nghiệp**: Nhiều nhân viên y tế báo cáo rằng họ không thể tiếp cận đào tạo liên tục hoặc tiến bộ nghề nghiệp, khiến họ mất động lực và cam kết với vai trò hiện tại của mình.\n\n4. **Sự cân bằng giữa cuộc sống và công việc không phù hợp**: Với khối lượng công việc nặng nề và giờ làm việc kéo dài, nhân viên y tế đấu tranh để duy trì sự hài hòa giữa trách nhiệm cá nhân và chuyên môn của họ.\n\n5. **Hỗ trợ tâm lý xã hội hạn chế**: Thiếu dịch vụ tư vấn và hỗ trợ tâm lý cho nhân viên y tế, đặc biệt là sau khi đối phó với căng thẳng cao độ do COVID-19 gây ra, góp phần vào tỷ lệ nghỉ hưu cao hơn.\n\nNhững hậu quả tiềm năng:\n\nViệc mất mát nhân viên y tế quan trọng này đặt ra mối đe dọa nghiêm trọng đối với chất lượng dịch vụ chăm sóc sức khỏe, khả năng đáp ứng nhu cầu của bệnh nhân và sự ổn định tổng thể của hệ thống y tế. Để

```

giải quyết tình trạng chảy máu chất xám này, cần thực hiện các biện
pháp khẩn cấp, bao gồm:\n\n* **Tăng cường phúc lợi tài chính**  

xét",
7  "gen_meta": {
8    "max_new_tokens": 512,
9    "do_sample": false,
10   "pad_token_id": 151643,
11   "eos_token_id": 151645,
12   "use_cache": true,
13   "num_beams": 1
14 },
15 "model_name": "sail/Sailor2-8B-Chat"
16 }

```

B.2 Distribution of VietBinoculars scores with respect to sequence length on the Gemma-3-12B-News dataset

Figure 7 presents the distribution of VietBinoculars scores with respect to the number of tokens in the input texts from the Gemma-3-12B-News dataset. For visualization purposes, we also plot the three classification thresholds of VietBinoculars: Youden’s J, Closest Point, and TPR@0.06%FPR. Based on the distribution plot and the TPR@0.06%FPR threshold, we can reasonably trust that texts with scores below this threshold are highly likely to be AI-generated.

C Detailed Detection Results for the Capybara Problem in Vietnamese

We employ the LLMs listed in Section 3.4 to generate Vietnamese text from the capybara prompt. RadarTester, DetectGPT, GPTZero, and Ghostbuster are used as popular detection baselines for comparison with VietBinoculars. These detection methods are applied to identify Vietnamese AI-generated texts produced by the aforementioned LLMs. We calculate the detection accuracy of each method using the optimal thresholds determined in Section 3.2 for VietBinoculars, and the default threshold of 0.5 for the remaining methods. The detailed comparison results are presented in Figure 8. Notably, all detection methods fail to identify texts generated by DeepSeek-V3, Claude-3.5-Sonnet, and GPT-4o as AI-generated. VietBinoculars with the Closest Point threshold achieves the highest performance, with prediction failures occurring only for Qwen3-235B-A22B-FW, DeepSeek-V3, Claude-3.5-Sonnet, and GPT-4o.

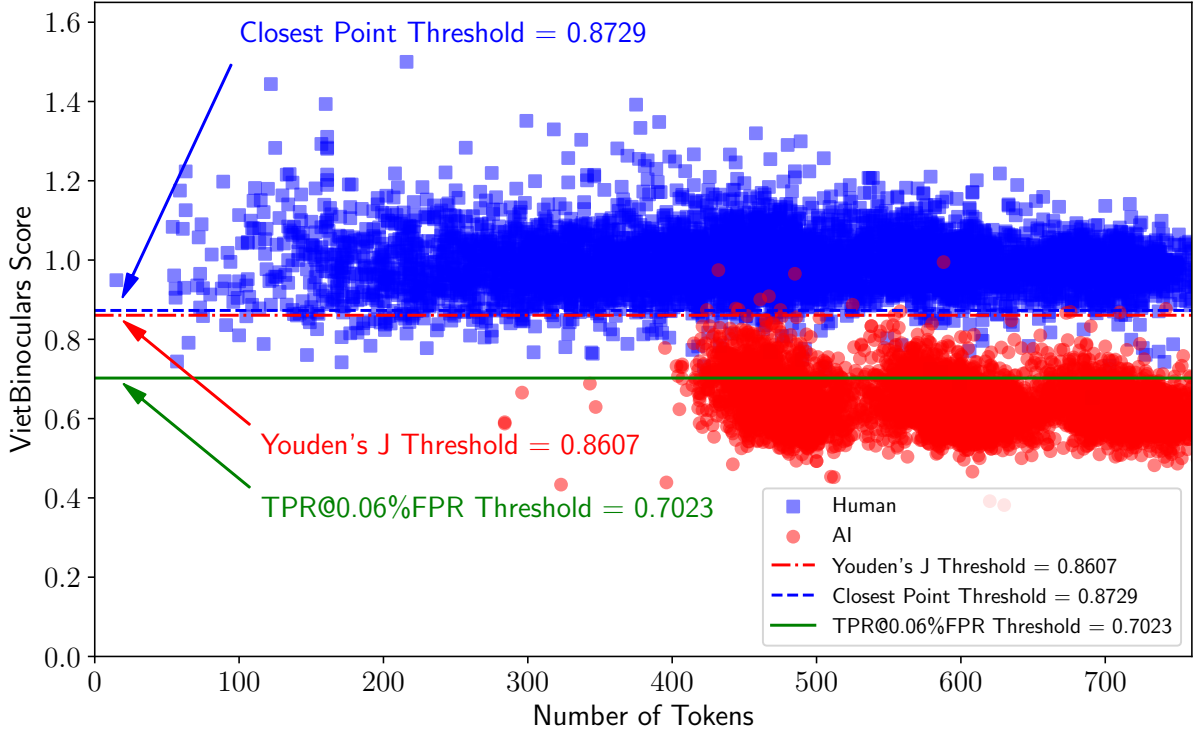


Figure 7: Distribution of VietBinoculars scores using BPE-encoded sequence length on the Gemma-3-12B-News dataset. The plotted lines include the classification thresholds: Youden's J, Closest Point, and TPR@0.06%FPR. The y-axis indicates the VietBinoculars scores. Each point represents a text sample, with blue squares denoting human-written texts and red circles denoting AI-generated texts.

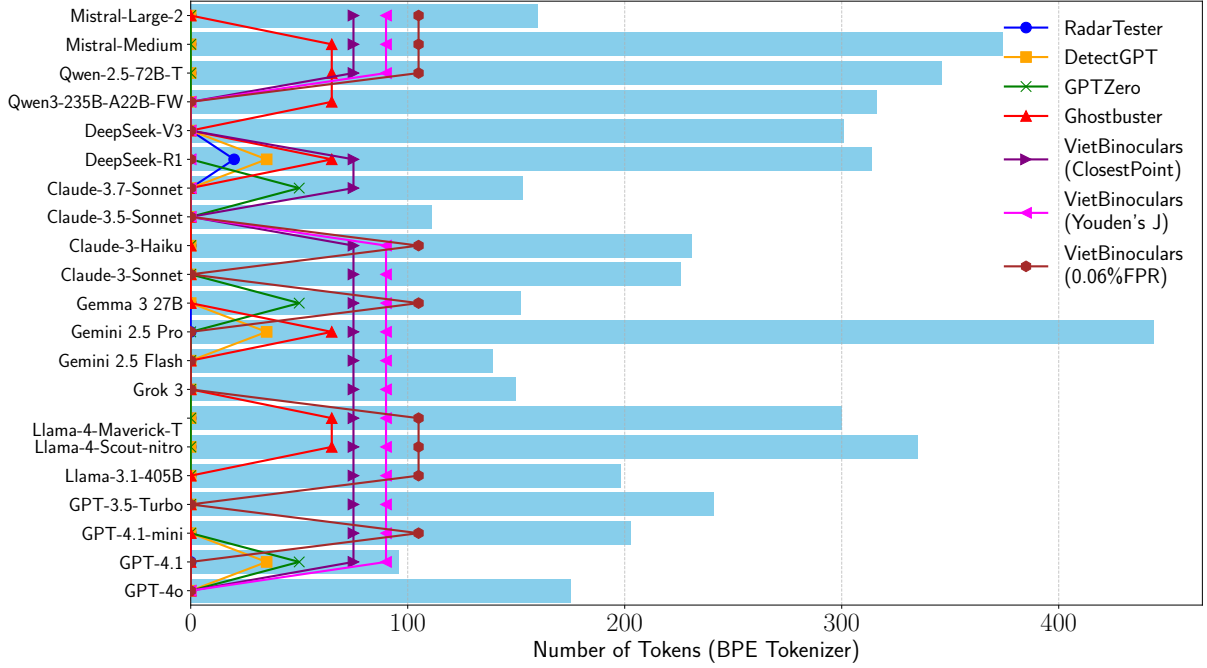


Figure 8: Detection accuracy of RadarTester, DetectGPT, GPTZero, Ghostbuster, and VietBinoculars on Vietnamese texts generated by various LLMs from the capybara prompt. The x-axis represents the number of tokens in text samples generated by LLMs using the capybara prompt, while the y-axis indicates the names of the LLMs used to generate the texts. The lines with markers show the detection accuracy of each method. If a detection method correctly identifies a text as AI-generated, the corresponding marker is placed above zero; otherwise, the marker is positioned at zero.