Extending "Spotting LLMs With Binoculars" for Low-Resource Languages: A Reproducibility Study in Haitian Creole

Angeline Dorvil
University of South Florida
angelinedorvil@usf.edu
Code: https://github.com/angelinedorvil/Binoculars.git

Abstract

Large language models (LLMs) increasingly produce text that is difficult to distinguish from human writing, raising growing concerns regarding authenticity, academic integrity, and responsible AI use. Spotting LLMs With Binoculars (Hans et al., 2024) introduced a zero-shot, training-free detection method based on the divergence between two related language models and demonstrated strong performance on English datasets using Falcon-7B and Falcon-7B-Instruct. This project aims to reproduce those results and investigate their robustness in a low resource language setting.

Reproducibility was evaluated on the original three domains: CC-News, CNN, and PubMed using the Binoculars framework with limited hardware constraints (minimal batch configuration and 1024 token evaluation). While performance remained high (AUC 0.98-0.99), detection sensitivity at low false positive rates was mildly reduced compared to reported values, likely due to truncation effects and resource limitations.

To explore generalizability, the method was extended to Haitian Creole, a low resource language with limited LLM training representation. Machine-generated texts were produced using Aya 101 and GPT4, with and without explicit language prompting. Results showed a notable degradation in detection performance relative to English, and prompted generation led to significantly worse scores. Non Prompted/continuation generation moderately improved performance, suggesting that generation quality directly affects detection robustness.

To further evaluate applicability for low resource languages, a multilingual model pair (Falcon-H1 base and deep base) was tested as an alternative detector configuration. Despite using a constrained setup (single batch, 1024 tokens), results showed improved AUC and F1 scores relative to the English only models, suggesting potential advantages from multilingual alignment. However, detection sensitivity at very low false positive rates (TPR@FPR=0.01) declined substantially, indicating that improved calibration alone may not guaranty robustness under strict decision thresholds. These preliminary findings highlight multilingual pairing as a promising but non definitive direction, warranting further study with larger scale evaluation, enhanced token coverage, and multilingual generator alignment.

1 Introduction

Large language models (LLMs) such as GPT-4, Claude, LLaMA-2, and Falcon increasingly produce text that is fluent, contextually coherent, and often indistinguishable from human writing. As their accessibility improves, concerns surrounding academic integrity, misinformation, and authorship verification have prompted research into reliable methods for detecting AI-generated text. Most existing detectors rely on supervised classification techniques trained using labeled machine-generated samples, which makes them data dependent, model specific, and often brittle when applied to unseen domains or generator models.

To address these limitations, Hans et al. (2024) proposed Binoculars, a zero-shot, training free detection method that analyzes discrepancies between the predictive behaviors of two closely related pretrained LLMs: an observer and a performer. The method computes standard perplexity using the observer and a cross-perplexity score that captures how surprising the performer's token distribution is to the observer. A ratio based score between the two provides a classification signal that is robust to calibration issues and prompt complexity, without requiring any supervised training.

The intuition behind this approach is that LLMs tend to produce stylistically similar token distributions when generating text, whereas human writing introduces more variability, causing a stronger divergence between models. The original experiments conducted across English datasets such as news articles, student essays, and blog-style writing demonstrated strong performance, notably high true positive rates even at strict false positive thresholds, outperforming several supervised detectors. The paper further explored multilingual settings and noted that performance degraded in low-resource languages or morphologically complex settings, highlighting an open question regarding generalization beyond English.

That limitation directly motivated the second phase of this study: extending Binoculars to assess its validity in Haitian Creole, a low-resource language with evolving orthographic conventions and limited representation in existing LLM training data. Haitian Creole (Kreyòl Ayisyen) is primarily an oral language and is often underrepresented in multilingual NLP resources, making it an ideal stress test for Binoculars' zero-shot assumption.

Therefore, this project addressed the following research questions (RQ):

Reproducibility: Can Binoculars' original results in English be reproduced under similar conditions using openly available models and resources?

Low-resource generalization: Does Binoculars' zero-shot detection method generalize to Haitian Creole, despite sparse pretraining coverage and potential tokenizer fragmentation?

2 Related Works

Various bodies of research referenced in Hans et al. investigated the same core RQ from different perspectives: supervised pipelines like Ghostbuster learn a classifier over features from weaker LMs and set a strong baseline across news/essays/creative writing, especially on the very datasets later reused by Binoculars; in contrast, DetectGPT and Fast-DetectGPT are training free, leveraging probability curvature signals to flag machine text but at differing compute costs; DNA-GPT is also training free, differentiating machine vs. human text by regenerating continuations and measuring divergence.

2.1 Ghostbuster (Verma et al., 2023)

Ghostbuster is a supervised detection framework that trains a classifier on features extracted from several smaller language models to determine whether a text was written by a human or by an LLM. The authors released three benchmark datasets: news articles, creative writing, and student essays, which later became standard references for detectors including Binoculars. Their approach achieved strong performance across domains and models by leveraging ensemble representations, but it required labeled data and retraining whenever new generators emerged.

In this project, these datasets were used to reproduce the English experiments of Binoculars before extending the evaluation to Haitian Creole.

2.2 DetectGPT (Mitchell et al., 2023)

DetectGPT introduced the idea of probability curvature as a signal for machine-generated text. It computes log likelihoods under a reference LM and observes that machine-generated text tends to reside in regions of negative curvature, meaning its likelihood decreases more sharply than human text when perturbed. This zero-shot method avoids explicit training and can work across different generators, though it is computationally expensive due to requiring multiple forward passes per sample.

DetectGPT demonstrated that geometric differences in probability space can separate certain human and machine distributions without supervision.

2.3 Fast-DetectGPT (Bao et al., 2024)

Fast-DetectGPT builds on DetectGPT by introducing a conditional probability curvature approximation that removes the need for explicit perturbations. This design preserves most of the original method's accuracy while drastically reducing runtime, making zero-shot detection more efficient. However, it still depends on access to log probabilities from the underlying LM, limiting its use with black-box models.

2.4 DNA-GPT (Yang et al., 2023)

DNA-GPT (Divergent N-Gram Analysis) approaches detection through regeneration and divergence. It truncates a document mid-way, prompts an LM to complete it, and then measures how much the regenerated continuation diverges from the original. Large divergence suggests human authorship, while smaller divergence indicates machine-generated text. The method can operate in black-box (n-gram) or white-box (probability) modes and has proven robust against mild paraphrasing.

While this idea is conceptually similar to Binoculars' "two-model comparison," it performs regeneration rather than scoring existing text, which increases computational demands. Its robustness to editing provides a useful contrast when assessing how Binoculars handles orthographic variation or code-switching in Haitian Creole.

2.5 Work Summary and Connection to This Project

Among these methods, Binoculars presents a unique opportunity for scalable, training free detection without requiring labeled data or explicit access to the generator model. The following section outlines how the Binoculars framework was replicated using Falcon-7B and extended to evaluate Haitian Creole.

3 Methodology

In this project, I follow the original Binoculars framework to detect machine generated text using a trainin free, zero-shot scoring mechanism. The core assumption is that language models exhibit highly similar next-token probability distributions when generating AI text, whereas human writing causes greater divergence between such models.

The method operates using two pre-trained autoregressive models: the observer and performer models. The Observer model estimates standard perplexity, and the performer model provides token-level probability distributions for cross-perplexity. Let x be the input sequence. The observer's perplexity formula is showed in figure 1. The cross perplexity formula is showed in figure 2. The Binoculars score is the log ratio of the two where lower scores indicate texts more likely to be AI-generated. Inputs are classified by comparing this score to a predefined threshold, originally calibrated using Falcon-7B and Falcon-7B-Instruct.

$$\mathrm{ppl}(x) = \exp\left(-\frac{1}{|x|}\sum_i \log p_{M_o}(x_i|x_{<i})\right)$$

Figure 1: perplexity

$$x\text{-}\mathrm{ppl}(x) = \exp\left(-\frac{1}{|x|}\sum_i \log p_{M_o}(x_i|M_p(x_{<i}))\right)$$

Figure 2: cross perplexity

To evaluate detection performance, the project follows the same metrics used in Hans et al. (2024). The primary measure is the Area Under the Receiver Operating Characteristic Curve (AUC), which provides an aggregate representation of classifier separability across threshold values. A higher AUC indicates stronger ability to distinguish between human-written and machine-generated text.

In addition to AUC, the F1 score is reported to measure the balance between precision and recall at the classification threshold. Since practical applications often require minimizing false positives, the True Positive Rate at a fixed False Positive Rate of 1 percent (TPR@FPR = 0.01) is also computed. This metric reflects detection sensitivity under high-confidence conditions. The threshold used for classification is selected empirically based on the best performing value for each model run.

All experiments were conducted using Google Colab Pro, configured with a single NVIDIA A100 GPU with 80 GB of VRAM and the high-memory runtime option enabled. The implementation of Binoculars relies on PyTorch operating in inference mode, without any gradient computation or model fine-tuning. All models were loaded via the Hugging Face Transformers library, using their default configurations unless otherwise specified.

A maximum sequence length of 1024 tokens was used for all experiments included in the final report. Batch size was adjusted dynamically to avoid exceeding GPU memory limits during inference. No local modifications, caching overrides, or precision adjustments were applied beyond those inherently managed by the library.

3.1 Reproduction

For the reproduction phase of this project, the experimental setup closely followed that described by Hans et al. (2024). The original authors evaluated Binoculars using three English-language datasets originating from Ghostbuster (Verma et al., 2023): CC-News, CNN, and PubMed. These preprocessed datasets were directly retrieved from the official Binoculars GitHub repository and were used without modification; Two separate machine generated variants were tested, with continuations produced by Falcon-7B-Instruct and LLaMA2-13B respectively. Each sample consisted of paired human written and machine generated text, allowing for a direct comparison under zero-shot detection conditions.

As in the original implementation, the scoring models used were Falcon-7B (as the observer) and Falcon-7B-Instruct (as the performer), which were selected due to their shared pretraining corpus. All experiments were executed using the official Binoculars scoring logic, preserving the same computation of perplexity and cross-perplexity. A maximum token length of 1024 tokens per input and batch size of 2 were used for all final reported runs. This configuration was selected following an internal evaluation of memory constraints on the available GPU, where larger token or batch sizes led to instability.

3.2 Extension

Haitian Creole (Kreyòl Ayisyen) is a predominantly oral language with evolving written conventions, derived primarily from 18th-century French with lexical influences from Spanish and several West African languages. For text generation, two models were also used: Aya 101, a multilingual model trained on Haitian Creole data, and GPT-4. Generation was performed under two conditions: non-prompted, where the system continued a human written Creole text without explicit language instruction, matching the methodology of Hans et al.; Prompted, where the model was directly instructed to "write in Haitian Creole" to improve linguistic coherence and reduce token fragmentation. This additional condition was included specifically due to the low-resource nature of Haitian Creole, with the hypothesis that clearer prompting might enhance generation quality and potentially improve classification behavior.

Each text type: news, poetry, and blogs/Reddit style posts, was segmented into approximately 4,000 character chunks, allowing token stable scoring across models. Overall, over 5k chunks were evaluated. Slight variation was permitted to preserve word integrity at chunk boundaries. For detection, two configurations were evaluated: the original observer performer pairing Falcon 7B base vs. Falcon 7B Instruct, which is English only and therefore not aligned with Haitian Creole; A multilingual alternative using Falcon H1 Deep-Base vs. Falcon H1 Deep-Instruct, selected because they share tokenizers and multilingual pretraining including French and Spanish. Only a limited number of samples were evaluated using the multilingual model due to computational cost.

4 Results

Scores were computed for AUC, F1, and the True Positive Rate at a fixed False Positive Rate of 1 percent (TPR@FPR = 0.01), using the optimal classification threshold determined independently for each dataset.

4.1 Reproduction

The reproduced results closely aligned with those reported in the original paper. For CNN and PubMed, both AUC and F1 nearly match the original findings, with deviations limited to the range of ±0.002 across metrics. The TPR@FPR=0.01 for both datasets also demonstrates strong alignment, differing from the paper by less than one percentage point. The only substantial discrepancy appears in CC-News, where the reproduced TPR value decreased by approximately 0.25, despite high AUC and F1 scores. This deviation is likely attributable to the lower token length used during reproduction (1024 vs. 2048 tokens in the original paper). However, this remained within an acceptable range given the overall reproducibility of the model's detection behavior.

These results confirmed that the core contributions of Binoculars zero-shot, training free detection with strong cross domain performance, can be reliably reproduced using publicly available resources and under constrained computational conditions. View results in Table 1: reproduction (R) and Hans at al. (H).

4.2 Extension

Falcon 7B most reliably detected GPT4 non-prompted samples, particularly for News articles (TPR = 0.549 at FPR = 0.01) and Blogs datasets (TPR = 0.458). Prompted GPT4 samples consistently reduced detectability across domains, likely because prompting encouraged more structured or stylistically human like generation. While absolute AUC scores remain high for GPT4 outputs, detectability under realistic tolerance constraints (TPR@0.01) is substantially lower than in English. This indicates that the Binoculars ratio is less effective in low-resource settings, even when surface characteristics appear favorable. Aya101 generated text was consistently the most challenging to detect, despite Haitian Creole being among its training languages, suggesting structural instability or insufficient domain coverage in low-resource generation pipelines.

Despite modest improvements in F1 (particularly for GPT4 non-prompted data), Falcon H1 failed to improve detection under strict tolerance constraints, with TPR@0.01 remaining very low. This suggests that broader language coverage alone is insufficient for stable zero-shot detection. Future research may explore evaluating multilingual model pairs using larger token windows and additional batch runs, as this work was limited to a single batch at 1024 tokens.

View entension results in table 2 and 3.

References

The Internet Archive. Internet archive - blog and forum content (haitian creole). https://archive.org/search?query=subject%3A%22Krey%C3%B2l+Ayisyen%22. Accessed: 2025-11-09.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Preprint*, arXiv:2310.05130.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Le journal Rezo Nòdwès. Rezo nòdwès. https://rezonodwes.com/. Accessed: 2025-11-09.

Lavwadlamerik. Voa nouvel. https://www.voanouvel.com/. Accessed: 2025-11-09.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Potomitan. Potomitan - platform for haitian creole literature and poetry. https://www.potomitan.info/. Accessed: 2025-11-09.

REKA. Kreyol.org - haitian creole cultural and linguistic forum. https://kreyol.org/. Accessed: 2025-11-09.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. *Preprint*, arXiv:2305.15047.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *Preprint*, arXiv:2305.17359.

Dataset	**AUC (R)**	**F1 (R)**	**TPR (R)**	**AUC (H)**	**F1 (H)**	**TPR (H)**
CC-News	0.9817	0.9517	0.7263	0.9992	0.9821	0.978
CNN	0.9972	0.9789	0.9556	0.9985	0.9813	0.965
PubMed	0.9962	0.9770	0.9549	0.9959	0.9794	0.942

Table 1: Performance comparison between reproduced results and those reported in Hans et al. (2024).

Dataset	**Model Type**	**AUC**	**F1**	**TPR@0.01**
News	Aya101 (Non-Prompted)	0.263	0.665	0.153
	GPT-4 (Non-Prompted)	0.864	0.803	0.549
	GPT-4 (Prompted)	0.779	0.724	0.208
Essays	Aya101 (Non-Prompted)	0.479	0.663	0.210
	GPT-4 (Non-Prompted)	0.824	0.757	0.289
	GPT-4 (Prompted)	0.753	0.715	0.081
Blogs	GPT-4 (Non-Prompted)	0.886	0.812	0.458
	GPT-4 (Prompted)	0.918	0.843	0.426

Table 2: Extension performance using Falcon 7B for Haitian Creole datasets.

Generation Source	**Detector Model**	**AUC**	**F1**	**TPR@0.01 FPR**
GPT4 (Prompted)	Falcon 7B	0.7794	0.03	1.75%
	Falcon H1	0.7692	0.47	1.08%
GPT4 (Non-Prompted)	Falcon 7B	0.8644	0.47	35.09%
	Falcon H1	0.7678	0.62	11.61%
Aya101 (Non-Prompted)	Falcon 7B	0.2628	0.23	13.16%
	Falcon H1	0.3060	0.24	10.31%

Table 3: Comparison of monolingual (Falcon 7B) and multilingual (Falcon H1) detection across model conditions for Haitian Creole.