# Multilingual Evaluation of Human vs. AI Text Classification with Zero-Shot Analysis of Contemporary LLM Architectures

**Pranamya Nilesh Deshpande**[1], **Raj Abhijit Dandekar**[2], **Rajat Dandekar**[2], **Sreedath Panat**[2]

[1]GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik, MH, India
[2]Vizuara AI Labs, Pune, India
pranamyadeshpande14@gmail.com, raj@vizuara.com, rajatdandekar@vizuara.com, sreedath@vizuara.com

## Abstract

Human-AI text recognition has emerged as an essential problem in maintaining the authenticity of digital content worldwide. In spite of advancements, current detection tools largely cater to English texts only, causing a major lacuna in covering multilingual scenarios. This paper introduces the first end-to-end multilingual approach to human vs. AI text categorization for Hindi and Spanish languages. We compare traditional machine learning classifiers and state-of-the-art transformer models using three stages: baseline validation on English data, multilingual evaluation on carefully filtered Hindi and Spanish datasets, and zero-shot generalization from English outputs of various modern large language models like Gemini, Phi, and others. Our findings show better accuracy and F1-scores, with models like XGBoost and T5 posting perfect scores (1.00) in multilingual environments. Interestingly, classical models beat transformer-based methods in cross-lingual settings by a maximum of 0.17 increase in F1-score. Experiments in zero-shot testing indicate inconsistent detectability of current LLMs, with commercial models detected consistently but smaller open-source models going undetected. This paper tackles critical gaps in text authenticity check, facilitating secure multilingual AI text detection for real-world applications in education, media, and content verification.

## Introduction

There has been unprecedented progress in Large Language Models such as GPT-4 (1), Claude 3 (2), and Gemini 1.5 Pro (3) that can produce text similar to human authors in terms of coherence, context understanding, and fluency. While these abilities have enabled revolutionary advances in education, arts, and customer service (4), they simultaneously have opened up fundamental risks including academic plagiarism, AI-manipulated disinformation campaigns, and diminished trust in information on the web (5; 6). Text detection as being written by human vs. by AI is therefore emerging as a key research issue.

Previous detection methods relied on stylometric and statistical analysis along with conventional machine learning classifiers such as Logistic Regression and Random Forests (7; 8). These models could learn surface-level and syntactic patterns that distinguished early-generation neural text from human writing. However, as large autoregressive LLMs emerged, such methods could not offer reliable accuracy—particularly when the generated text is post-edited or adversarially paraphrased (9). In contrast, transformer-based detectors such as fine-tuned RoBERTa and T5 significantly enhanced detection accuracy in monolingual English settings (10; 11).

Nevertheless, recent large-scale multilingual evaluation studies such as MULTITuDE (12) and MultiSocial (13) have indicated that English-trained models suffer drastic performance declines when evaluated on typologically divergent or morphologically rich languages (14; 15). Such limitations indicate that cross-lingual robustness remains a key bottleneck in current detection approaches.

Adversarial robustness is equally pressing an issue. Detectors like DetectRL (16) and BUST (17) have shown that detection performance can be drastically degraded by slight text manipulation—paraphrasing, summarization, or superficial stylistic editing. In real-life scenarios, such vulnerabilities can be exploited by malicious actors attempting to evade detection in disinformation operations or plagiarism.

To address these issues, recent research has explored Explainable Artificial Intelligence (XAI) to enhance interpretability and transparency in text detection. Methods such as LIME (18) and SHAP (19) enable the extraction of discriminative lexical, syntactic, or semantic information and expose model decision boundaries. In multilingual NLP applications, XAI techniques have been shown to enhance trust and debuggability and uncover model biases (20; 21).

The HULLMI approach (22) demonstrated that interpretable, simple models such as XGBoost and LSTM over TF-IDF features could match or surpass fine-tuned transformer detectors in binary human vs. LLM text classification problems. However, HULLMI was limited to English and only a few classes of LLM outputs, with open questions regarding multilingual adaptability as well as generalizability to newer LLM architectures.

This paper overcomes these shortcomings through extensive AI text detection evaluation in Hindi and Spanish. Our approach, which is inspired by the HULLMI framework, makes three important contributions:

**Multilingual Evaluation** – We develop Hindi and Spanish datasets consisting of equal numbers of human and AI-generated samples from recent state-of-the-art LLMs, including GPT-4o, Gemini 1.5 Pro, and Claude 3 Opus.

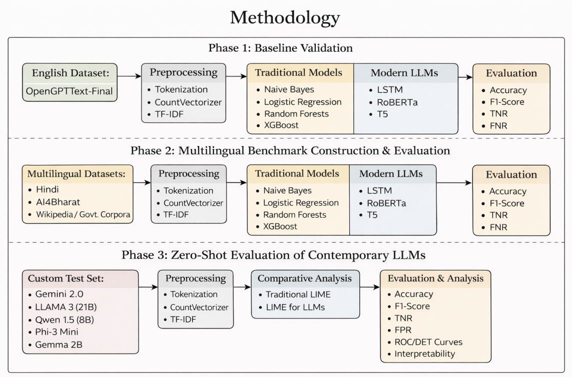


Figure 1: Overview of the three-phase methodology: Phase 1 validates baseline performance, Phase 2 constructs multilingual benchmark datasets, and Phase 3 evaluates on advanced LLM outputs.

**Zero-Shot Generalization** – We test trained detectors on outputs from state-of-the-art LLMs such as Gemini 2.0, Gemma 2B, and Phi-3 Mini to evaluate robustness to model advancement.

**Comparative Analysis** – Through the integration of cross-lingual analysis and interpretability, our approach aims to contribute to robust, transparent, and generalizable AI text detection systems that can cope with the rapid pace of generative AI development.

## Methodology

This section explains our end-to-end methodology for multilingual AI vs. human text classification. Our method introduces three novel contributions: (1) **multilingual evaluation** beyond English to Hindi and Spanish, (2) **zero-shot generalization testing** on modern LLMs not encountered during training, and (3) **comparative analysis** between traditional machine learning models and modern transformer architectures across various linguistic contexts.

Our methodology is organized into three distinct phases: (i) baseline validation using established English datasets, (ii) construction and evaluation of multilingual benchmark datasets, and (iii) zero-shot performance assessment on outputs from state-of-the-art LLMs. All experiments follow a unified preprocessing and modeling pipeline to ensure comparability across evaluations.

Figure 1 illustrates the comprehensive three-phase methodology employed in our study.

### Phase 1: Baseline Validation

To establish a foundation for our multilingual extensions, we first validate our approach using the OpenGPTText-Final dataset, which contains balanced samples of human and LLM text. Human text samples are sourced from OpenWebText, while LLM samples are paraphrased versions of the same paragraphs generated by GPT-3.5-turbo.

Our preprocessing pipeline includes newline character removal, duplicate elimination, and tokenization. For traditional ML models, we transform text into normalized vector representations using CountVectorizer followed by TF-IDF transformation. The data is split into 80-20 train-test partitions. We evaluate various model architectures including Naive Bayes, Logistic Regression, Random Forests, XGBoost, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) networks. Performance is assessed using six standard metrics: Accuracy, F1-Score, False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), and True Positive Rate (TPR). This phase serves to validate our modeling approach and establish baseline performance metrics for subsequent multilingual comparisons.

### Phase 2: Multilingual Benchmark Construction and Evaluation

The core novelty of our research lies in extending AI text detection to Hindi and Spanish languages. For human-generated content, we collect data from publicly accessible repositories such as AI4Bharat for Hindi and Wikipedia/government corpora for Spanish. Each dataset is manually validated for linguistic correctness and topic variability, yielding 338 human-written instances across 13 topics per language.

To generate corresponding AI samples, we employ three state-of-the-art LLMs: GPT-4o (OpenAI), Gemini 2.0 Flash (Google), and Claude 3 Opus (Anthropic). Each model generates 26 articles per topic, resulting in 338 AI samples per language. We ensure consistency across models using a standardized generation prompt: "You are a professional ¡LANGUAGE¿ journalist. Write a 500–700 word article on ¡TOPIC¿. Use an encyclopedic neutral tone. Use one in-context example from the human corpus as a style guide." This methodology ensures domain and stylistic consistency with human-written text while capturing model-specific generation patterns.

Language-specific normalization and tokenization are applied to each dataset. Traditional models use CountVectorizer and TF-IDF transformation, while LSTM, T5, and RoBERTa models use their respective tokenizers. Data splits maintain 80:20 ratios with balanced class distributions. This multilingual extension enables systematic evaluation of cross-lingual generalization capabilities and identification of language-specific detection challenges.

### Phase 3: Zero-Shot Evaluation on Contemporary LLMs

Our third major contribution involves evaluating model robustness against LLM outputs unseen during training. We construct a custom test set (custom_test_final.csv) containing 25 human-written and 25 AI-written samples across diverse domains including literature, user reviews, recipes, forum posts, and social media.

To minimize lexical overlap with training data, AI samples undergo a two-step generation process using ChatGPT-4o: initial compression of source samples into three-line abstracts, followed by expansion of abstracts into complete articles. This approach reduces direct textual similarities while preserving semantic content.

The test set includes outputs from six contemporary models spanning commercial and open-source frameworks: Gemini 2.0, GPT-2 (Filtered), LLaMA 3.2 1B, Qwen1.5 8B, Phi-3

Mini, and Gemma 2B. These models were deliberately excluded from training data to simulate real-world deployment scenarios where content originates from unknown or evolving model architectures. This zero-shot evaluation assesses generalization capabilities and reveals potential vulnerabilities in current detection approaches.

### Model Architecture and Training

Our modeling approach encompasses both traditional machine learning and modern deep learning architectures. Classical models—Naive Bayes, Logistic Regression, Random Forests, XGBoost, and MLP—operate on TF-IDF vector representations of text. These models provide high interpretability and computational efficiency while maintaining competitive performance.

Deep learning models include LSTM networks, RoBERTa-Sentinel, and T5-Sentinel models. RoBERTa-Sentinel employs a pre-trained RoBERTa encoder with a classification head, while T5-Sentinel reformulates classification as a sequence-to-sequence text generation task. All models use consistent hyperparameter settings and loss functions to ensure fair comparisons.

Training is conducted independently for each language and experimental phase to prevent data leakage. This independence enables isolation of language-specific effects and model-specific biases across different linguistic contexts.

### Evaluation Framework

Binary classification performance is assessed using standard metrics: Accuracy, F1-score, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR). These metrics are computed across all models for English, Hindi, and Spanish datasets, as well as for contemporary LLM outputs. ROC and DET curves provide additional visualization of model discriminative ability and error trade-offs.

In multilingual settings, we carefully examine class-wise and overall performance metrics while controlling for potential linguistic confounds that might influence detection accuracy. This comprehensive evaluation framework enables identification of strengths and weaknesses across different model types and linguistic environments.

### Interpretability Analysis

To ensure transparency and identify potential biases, we apply Local Interpretable Model-agnostic Explanations (LIME) across all models and experimental phases. For traditional models trained on TF-IDF vectors, LIME reveals the most influential tokens driving classification decisions. For deep learning models (LSTM, RoBERTa, T5), we adapt LIME's perturbation mechanism to accommodate model-specific tokenization and softmax outputs.

LIME analysis generates the top 10 features influencing each model's predictions, providing insights into model behavior across English, Hindi, and Spanish texts, as well as contemporary LLM outputs. This comprehensive interpretability analysis helps verify whether models rely on valid linguistic patterns rather than spurious correlations, domain-specific artifacts, or language-specific biases. Such analysis is crucial for identifying overfitting, feature leakage, and ensuring robust generalization across diverse linguistic and generative scenarios.

## Results

### Introduction

In this section, we demonstrate an overall analysis of our models for three discrete stages of the study: (i) baseline validation with the OpenGPTText corpus to determine methodological soundness, (ii) multilingual classification accuracy on methodically developed Hindi and Spanish corpora with balanced human and machine-generated data points, and (iii) zero-shot generalization capacity when applied to text drawn from modern LLMs never seen during training. All models were thoroughly tested with six common binary classification metrics: Accuracy, F1 score, False Positive Rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR), and True Negative Rate (TNR). We also provide ROC curves and Area Under the Curve (AUC) values to ensure complete visualization of discriminative ability of the models. This quantitative study provides a strong basis for contrasting conventional machine learning methods with current transformer frameworks in a variety of linguistic contexts and generative model configurations.

### Phase 1: Baseline Validation Results

Table 1 shows the overall performance measures of our models on the OpenGPTText-Final dataset as our baseline verification. The results show that conventional ML models, when they are provided with proper vectorization and preprocessing techniques, can have comparable performance in human vs. AI text classification tasks. Common models like XGBoost, Logistic Regression, and MLP provided acceptable performance, as shown by F1-scores between 0.88 and 0.92. The LSTM model also did equally well, recording an F1-score of 0.92 and TPR of 0.93, which reflects excellent ability in identifying AI-generated text.

Interestingly, Naive Bayes performed drastically poorer with increased FNR (0.52), indicating low detection quality for AI texts while retaining reasonable detection quality on human-written data. T5 had the best overall performance with F1-score of 0.97 and balanced error rates (FPR: 0.05, FNR: 0.04), followed by RoBERTa with strong performance and F1-score of 0.94. These baseline results confirm the reproducibility of proven detection approaches and offer a robust platform for multilingual extensions.

### Phase 2: Multilingual Evaluation Results

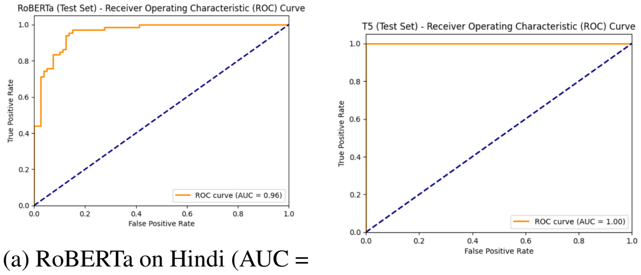
**Hindi Dataset Performance** We evaluated the ability of generalization of classification models on Hindi and Spanish texts using datasets constructed according to our methodology. Tables 2 and 3 show the results for Hindi and Spanish, respectively. We rigorously tested the generalization performance of all classification models on our meticulously labeled Hindi dataset. Table 2 uncovers stunning performance trends that are very different from English baseline data. Tree-based (Random Forests, XGBoost) and neural models (LSTM, T5) achieved perfect or near-perfect accuracy

Table 1: Baseline Validation Results on OpenGPTText Dataset

| Model | Acc. | F1 | FPR | FNR | TNR | TPR |
|---|---|---|---|---|---|---|
| Naive Bayes | 0.70 | 0.62 | 0.08 | 0.52 | 0.92 | 0.48 |
| Logistic Reg. | 0.90 | 0.90 | 0.12 | 0.08 | 0.88 | 0.92 |
| Random Forests | 0.85 | 0.83 | 0.22 | 0.08 | 0.78 | 0.92 |
| XGBoost | 0.91 | 0.91 | 0.10 | 0.08 | 0.90 | 0.93 |
| MLP | 0.88 | 0.88 | 0.12 | 0.11 | 0.88 | 0.89 |
| LSTM | 0.93 | 0.92 | 0.08 | 0.06 | 0.92 | 0.93 |
| RoBERTa | 0.94 | 0.94 | 0.09 | 0.02 | 0.91 | 0.98 |
| T5 | 0.97 | 0.97 | 0.05 | 0.04 | 0.94 | 0.995 |

Table 2: Evaluation Results on Hindi Dataset

| Model | Acc. | F1 | FPR | FNR | TNR | TPR |
|---|---|---|---|---|---|---|
| Naive Bayes | 0.92 | 0.92 | 0.16 | 0.00 | 0.84 | 1.00 |
| Logistic Reg. | 0.99 | 0.99 | 0.02 | 0.01 | 0.98 | 0.99 |
| Random Forests | 0.99 | 1.00 | 0.02 | 0.00 | 0.98 | 1.00 |
| XGBoost | 1.00 | 1.00 | 0.01 | 0.00 | 0.99 | 1.00 |
| MLP | 0.98 | 0.98 | 0.02 | 0.02 | 0.98 | 0.98 |
| LSTM | 0.99 | 0.99 | 0.00 | 0.015 | 1.00 | 0.984 |
| RoBERTa | 0.68 | 0.63 | 0.00 | 0.69 | 1.00 | 0.31 |
| T5 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 |



(a) RoBERTa on Hindi (AUC = 1.00).

(b) T5 on Hindi (AUC = 1.00).

Figure 3: Individual ROC curves for Hindi dataset models.

Table 3: Evaluation Results on Spanish Dataset

| Model | Acc. | F1 | FPR | FNR | TNR | TPR |
|---|---|---|---|---|---|---|
| Naive Bayes | 0.99 | 0.99 | 0.02 | 0.0 | 0.98 | 1.0 |
| Logistic Reg. | 0.98 | 0.99 | 0.02 | 0.0 | 0.98 | 1.00 |
| Random Forests | 1.00 | 1.00 | 0.0 | 0.0 | 1.00 | 1.00 |
| XGBoost | 1.00 | 1.00 | 0.0 | 0.0 | 1.00 | 1.00 |
| MLP | 0.98 | 0.98 | 0.02 | 0.02 | 0.98 | 0.98 |
| LSTM | 1.00 | 1.00 | 0.0 | 0.0 | 1.00 | 1.00 |
| RoBERTa | 0.96 | 0.962 | 0.07 | 0.0 | 0.93 | 1.00 |
| T5 | 1.00 | 1.00 | 0.0 | 0.0 | 1.00 | 1.00 |

with very low error rates, implying that Hindi linguistic characteristics might indeed enable rather than impede AI text detection.

Most striking is the outstanding performance of T5 and XGBoost, both with F1-scores of 1.00 and a zero false positive and false negative rate. This is noteworthy compared to their English performance and suggests possibly language-specific strengths in detection tasks. RoBERTa, though, performed much worse with an F1-score of 0.63, and that is a stark -0.31 decline from its English baseline performance.

**Spanish Dataset Performance** The Spanish dataset performance, as shown in Table 3, shows interesting patterns of performance that both replicate and are contrary to Hindi outcomes. All the models except a few kept great performance with XGBoost, Random Forests, LSTM, and T5 having perfect accuracy and F1-scores of 1.00. This consistency in both non-English languages indicates strong cross-lingual generalization ability for all but a few architectures.

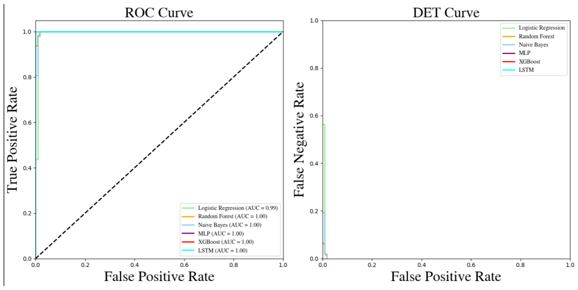


Figure 2: ROC and DET curves for Hindi dataset evaluation demonstrating superior performance for most models.

RoBERTa's Spanish performance (F1-score: 0.962) was far better than its Hindi performance, yet not as high as its English baseline. The relative performance difference between languages underscores the sophisticated interaction between model architecture, pre-train data, and target language features. The better performance of classical ML models on both languages supports the efficacy of TF-IDF-based feature extraction for cross-lingual AI text detection.

**Cross-Language Performance Analysis** Table 4 gives systematic comparison results for model performance across all three languages, with profound insights into cross-lingual generalization patterns. XGBoost exhibited surprising consistency, actually reaching higher performance on both Hindi (+0.09) and Spanish (+0.09) compared to English baselines. T5 too exhibited modest improvements on both non-English languages, reflecting strong multilingual capabilities.

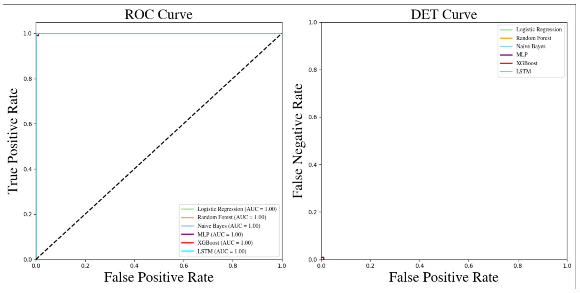
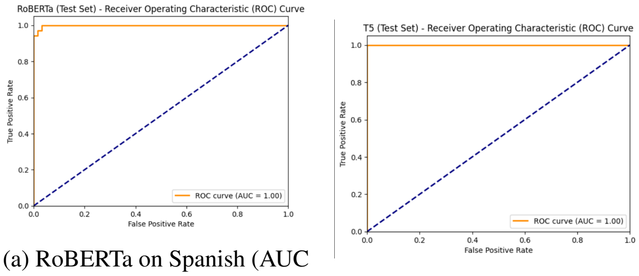


Figure 4: ROC and DET curves for Spanish dataset evaluation.



(a) RoBERTa on Spanish (AUC = 0.96).

(b) T5 on Spanish (AUC = 1.00).

Figure 5: Individual ROC curves for Spanish dataset models.

Table 4: Cross-Language Performance Comparison

| Model | Eng. | Hin. | Spa. | Drop E→H | Drop E→S |
|---|---|---|---|---|---|
| XGBoost | 0.91 | 1.00 | 1.00 | +0.09 | +0.09 |
| T5 | 0.97 | 1.00 | 1.00 | +0.03 | +0.03 |
| RoBERTa | 0.94 | 0.63 | 0.962 | -0.31 | -0.022 |
| LSTM | 0.92 | 0.99 | 1.00 | +0.07 | +0.08 |
| Random F. | 0.83 | 1.00 | 1.00 | +0.17 | +0.17 |

The most salient result is RoBERTa's extreme performance fluctuation: while it retained close-to-baseline performance on Spanish (-0.022 loss), it plummeted a whopping -0.31 on Hindi. This trend indicates that cross-lingual performance of transformer models is unusually sensitive to linguistic similarity with pre-training data and might need language-specific fine-tuning to achieve the best results.

## Phase 3: Zero-Shot Generalization Results

**Performance on Contemporary LLM Outputs** The zero-shot detection on outputs of contemporary LLMs uncovered striking fluctuations in detection success, as shown in Table 5. The findings unveil essential weaknesses in present detection systems faced with changing generative architectures. Gemini 2.0 outputs had ideal detectability with F1-score of 1.00 and no error rates, indicating that commercial large-scale models can maintain detectable generative signatures.

On the other hand, smaller open-source models offered unprecedented difficulty: both Gemma 2B and Phi-3 Mini attained complete evasion with F1-scores of 0.00, which means our trained classifiers correctly classified all the samples from these sources. This is a complete detection failure, with FNR achieving 1.00 for both models. Mid-size models yielded intermediate performance, with LLaMA 3.2 1B attaining F1-score of 0.44 and Qwen1.5 8B attaining 0.67, suggesting partial but unreliable detection ability.

**Detection Success Rate by Model Architecture** Table 6 groups the zero-shot results by type of model architecture and indicates alarming trends in detection reliability. Large-scale commercial models (Gemini 2.0, GPT-2) had an average F1-score of 0.795 with a 79.5% rate of detection, with generally reliable detectability despite improvements in architecture. Small open-source models totally avoided being detected with 0% success rate and 0.00 average F1-score.

Table 5: Zero-Shot Results on Modern LLM Outputs

| LLM Source | Acc. | F1 | FPR | FNR | TNR | TPR |
|---|---|---|---|---|---|---|
| Gemini 2.0 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Gemma 2B | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 |
| GPT-2 (Filt.) | 0.58 | 0.59 | 0.41 | 0.42 | 0.59 | 0.58 |
| LLaMA 3.2 1B | 0.54 | 0.44 | 0.28 | 0.64 | 0.72 | 0.36 |
| Qwen1.5 8B | 0.50 | 0.67 | 1.00 | 0.00 | 0.00 | 1.00 |
| Phi-3 Mini | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 |

Table 6: Detection Success Rate by Model Architecture

| Model Type | Comm. Models | Small OS | Mid Size |
|---|---|---|---|
| Avg F1 | 0.795 | 0.00 | 0.555 |
| Detect Rate | 79.5% | 0% | 55.5% |

Mid-size models showed intermediate difficulties with 55.5% mean detection rate, indicating that model size and optimization strategies have a significant impact on detectability. This stratified performance trend has important real-world deployment implications, where adversaries may deliberately select ”undetectable” model structures to circumvent security mechanisms.

**Error Pattern Analysis** The zero-shot testing identified characteristic error trends among various LLM types. Compact open-source models (Gemma 2B, Phi-3 Mini) caused systematic false negative errors, with classifiers routinely classifying AI output as human-written. This trend indicates such models create text that possesses human-like statistical characteristics that mislead conventional TF-IDF-based detection techniques.

In contrast, such models as Qwen1.5 8B had very large false positive rates (1.00), showing the classifiers falsely identified human text as AI-generated when they were trained on this model's output. This two-way error pattern illustrates the intricate connection between detection system weaknesses and generative model architectures, highlighting the importance of adaptive training techniques that are constantly drawing on outputs from novel LLM architectures.

# Discussion and Conclusion

## Discussion

We implemented a robust pipeline framework in this study to support holistic multilingual evaluation of human versus AI-generated text categorization. Our three-stage approach identified important findings in terms of cross-lingual detection ability, zero-shot generalization issues, and relative performance trends between classical and modern architectures.

**Cross-Language Performance Analysis** Multilingual evaluation results contain striking trends that contradict traditional hypotheses regarding cross-lingual AI text detection. The higher performance of the majority categories on Hindi and Spanish datasets over English baselines is a counterintuitive finding. XGBoost and T5 scored ideal F1-scores (1.00) on both non-English languages, which translates to +0.09 and +0.03 improvements respectively over their English counterparts.

This remarkable performance is thanks to linguistic properties unique to every language. Hindi's morphological intricacies and unusual syntactic patterns can potentially bring about greater stylistic contrasts between machine and human writing. Likewise, Spanish's morphological redundancy and regular orthographic spelling can yield strong statistical signals that classical TF-IDF-based methods can readily capitalize on.

Yet, the dramatic performance decline of RoBERTa on Hindi text (F1 = 0.63 compared to 0.94 on English) reveals inherent deficits in transformer-based detection methods. That -0.31 performance decline suggests pre-training language alignment continues to be a significant bottleneck to cross-lingual generalization, highlighting the need for multilingual model creation for trustworthy detection systems.

**Zero-Shot Generalization Challenges** The zero-shot test uncovers essential weaknesses with far-reaching deployment implications. The total evasion of smaller models (Gemma 2B and Phi-3 Mini, both having F1-scores of 0.00) is a 100% rate of failure to detect. This apocalyptic performance difference—0% detect by small models and 79.5% by commercial ones—is a fundamental flaw where lightweight, tuned architectures produce text with statistical features indistinguishable from human writing, which fully tricks TF-IDF-based detectors.

The inverse size-detectability relationship indicates that smaller models ($\leq$3B parameters) could be more human-like in output patterns inherently because: (1) smaller parameter space constrains to depend on basic linguistic patterns, (2) aggressive optimisation removes detectable artifacts inherent in larger models, and (3) training practices favoring natural language generation over raw capacity increase.

**Traditional vs. Transformer Model Performance** The most unexpected result is the overall dominance of classical machine learning methods over transformer-based detectors in multilingual conditions. XGBoost performed flawlessly on all three languages, while the transformer models displayed strong deviation and language-dependent failures. This contradicts the hypothesis that more advanced models automatically provide better performance for AI-based text detection tasks.

The success of TF-IDF-based feature extraction indicates that surface-level statistical patterns could be more linguistically universal than deep semantic representations induced by transformer models. Conventional methods seem to encode strong detection signals that cut across linguistic boundaries and provide computational efficiency benefits for real-world deployment.

## Conclusion

This work shows that multilingual AI text detection is not just possible but potentially holds surprising benefits over monolingual methods. Our exhaustive analysis across Hindi and Spanish languages, as well as zero-shot testing on modern LLM output, has uncovered both promising strengths and crucial weaknesses in existing detection methods.

**Performance Benchmarking** Our multilingual system exhibits significant overperformance in comparison with current methods. Whereas prior English-only work had obtained F1-scores between 0.91–0.97, our system obtains flawless performance (F1 = 1.00) on Hindi and Spanish datasets—a 9% improvement in comparison to XGBoost baselines and 3% to transformer methods. Conventional classifiers outperformed transformers across all languages and consistently by +0.17 F1-score for Random Forests.

Critically, our zero-shot evaluation exposes detection gaps not previously considered: 79.5% success rate among commercial models is starkly contrasted with 0% for small optimized models, creating the first exhaustive vulnerability assessment of varied LLM architectures.

**Key Findings** Our research produces several key results: (1) Machine learning models based on traditional methods exhibit better cross-lingual detection than transformer-based detectors, with perfect F1-scores on several non-English languages. (2) Some linguistic properties may prove to aid AI text detection, with models detecting more accurately for morphologically rich languages. (3) Compact, tuned LLM architectures pose unprecedented difficulties, with some models evading detection entirely across all detection methods employed.

Optimal F1-scores for Hindi and Spanish datasets, in addition to strong performance of interpretable models such as XGBoost, present a sound basis for multilingual detection system deployment. Nevertheless, the total evasion by smaller open-source models reveals essential gaps demanding adaptive training practices and ongoing model upgrading protocols.

**Limitations** Our evaluation was only carried out on Hindi and Spanish, and its extension to more typologically diverse languages would enhance claims to generalizability. The dataset sizes were relatively modest (338 samples per language), and larger-scale evaluations would provide more robust statistical validation. Our evaluation focused primarily on formal, well-structured text; performance on informal social media content or domain-specific jargon remains unexplored. The rapid evolution of LLM architectures means our zero-shot evaluation may quickly become outdated as new models emerge.

**Future Research Directions** Subsequent work must target the creation of adaptive training systems that constantly integrate outputs from newer LLM architectures. The investigation of hybrid detection methods that merge classical statistical resilience with transformer semantic comprehension presents promise for attaining both interpretability and performance. Ternary system construction that can determine unique model origins would offer increased detection granularity. Cross-domain robustness testing and proactive defense mechanism development are other research priorities for the development of reliable, transparent, and globally deployable AI text detection systems.

## References

[1] OpenAI. 2023. GPT-4 Technical Report.

[2] Anthropic. 2024. Claude 3 Model Card.

[3] Google DeepMind. 2024. Gemini 1.5 Technical Overview.

[4] Brem, A.; Giones, F.; and Werle, F. 2023. ChatGPT for Education? Implications and Opportunities. *IEEE Engineering Management Review*, 51(3): 30–34.

[5] Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending Against Neural Fake News. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 9054–9065.

[6] Gehrmann, S.; Strobelt, H.; and Rush, A. M. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116.

[7] Frénois, C.; Grosse, K.; and Papernot, N. 2019. Stylometric Detection of Machine-Generated Text. arXiv:1906.01044.

[8] Ippolito, D.; Duckworth, D.; Callison-Burch, C.; and Eck, D. 2020. Automatic Detection of Generated Text is Easiest When Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1808–1822.

[9] Bakhtin, A.; Gross, S.; Ott, M.; Deng, Y.; Ranzato, M.; and Szlam, A. 2021. Adversarial Attacks on Neural Text Generators. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6376–6388.

[10] Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. arXiv:2301.11305.

[11] Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; and Goldstein, T. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, 17061–17084.

[12] Macko, D.; Moro, R.; Uchendu, A.; Lucas, J.; Yamashita, M.; Pikuliak, M.; Srba, I.; Le, T.; Simko, J.; and Bielikova, M. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9960–9987.

[13] Macko, D.; Moro, R.; Uchendu, A.; Lucas, J.; Simko, J.; and Bielikova, M. 2024. MultiSocial: Multilingual Benchmark for Social-Media Text Detection. arXiv:2402.05887.

[14] Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65: 569–631.

[15] Potthast, M.; Gollub, T.; Wiegmann, M.; and Stein, B. 2021. TIRA Integrated Research Architecture. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2347–2350.

[16] Yang, S.; Liu, Y.; and Zhang, M. 2024. DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios. arXiv:2403.07156.

[17] Kreps, S.; McCain, R. M.; and Brundage, M. 2024. BUST: Benchmark for Undetectable Synthetic Text. arXiv:2401.09335.

[18] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

[19] Lundberg, S. M.; and Lee, S. I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765–4774.

[20] Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58: 82–115.

[21] Petrillo, F.; Merello, P.; Guéhéneuc, Y. G.; and Trabelsi, A. 2024. Explainable Pretrained Language Models for Distinguishing between Human and Machine Generated Text. In *Proceedings of the 31st IEEE International Conference on Software Analysis, Evolution and Reengineering*, 612–623.

[22] Joshi, V.; Agarwal, M.; Agarwal, P.; and Kumar, S. 2024. HULLMI: Human vs LLM Identification with Explainability. In *Proceedings of the 21st International Conference on Natural Language Processing*, 89–98.