

LLM Project: A.I. Generated Code Detection

Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text

Willy VO, Minh Duc NGUYEN, and Matheo Quatreboeufs

Based on Hans et al., ICML 2024

December 14, 2025

Overview

- 1 Introduction
- 2 Literature Review
- 3 Methodology
- 4 Experimental Setup
- 5 Experimental Results
- 6 Baseline Comparison
- 7 Conclusion
- 8 Conclusion
- 9 Conclusion

The Challenge of AI Detection

- **Context:** AI-generated text is increasing (fake reviews, bots, plagiarism, misinformation).
- **Problem:** Distinguishing between human and machine behavior is difficult.
- **Current State:**
 - Existing detectors often rely on supervised training.
 - They tend to fixate on specific models.
 - They fail to transfer to new models or text domains.
- **Sensitivity:** Models are highly sensitive to prompts (tone, style, wording), making outputs vary widely.

The Solution: Binoculars

- **Proposed Framework:** *Binoculars* (Hans et al., ICML 2024).
- **Key Features:**
 - Zero-shot detection.
 - Model-agnostic.
 - Requires only 2 pre-trained LLMs (no training data).
- **Performance:**
 - $> 90\%$ True Positive Rate (TPR).
 - 0.01% False Positive Rate (FPR).
- **Importance:** Low FPR is critical to avoid unjustly labeling human work as machine-generated.

Overview & Hypothesis

- **Concept:** "Look" at text through two lenses (two distinct LLMs).
- **Central Hypothesis:**
 - **AI-Generated:** Two models tend to **agree** (predictions align).
 - **Human-Written:** Two models tend to **diverge** in predictions.
- **Benefit:** Using two models reduces dependence on individual model biases, improving generalization.

Formal Definitions: Perplexity

Let s be a text sequence $\vec{x} = (x_1, \dots, x_L)$.

Log-Perplexity (PPL)

Measures how predictable text is to a single model M .

$$\log\text{PPL}_M(s) = -\frac{1}{L} \sum_{i=1}^L \log P(x_i | x_{<i})$$

- Lower value = Easier to predict.
- Higher value = More "surprising" (typical of human text).

Cross-Perplexity (X-PPL)

Measures how similar predictions of Model M_2 are to Model M_1 .

$$\log\text{X-PPL}_{M_1, M_2}(s) = -\frac{1}{L} \sum_{i=1}^L P_{M_1}(x_i | x_{<i}) \log P_{M_2}(x_i | x_{<i})$$

- Low value = Strong alignment between models.
- High value = Divergence between models.

The Binoculars Score

Binoculars Score Formula

$$B_{M_1, M_2}(s) = \frac{\log \text{PPL}_{M_1}(s)}{\log \text{X-PPL}_{M_1, M_2}(s)}$$

Interpretation Table:

Case	PPL	X-PPL	Score (B)
AI Text	Low	Low	< 1 (Detection)
Human Text	High	Moderate	> 1 (Human)

Datasets Used:

- news_cnn: Aggregated CNN articles.
- train_v2_drcat_02: Argumentative essays (Human & AI).

Process:

- Extracted the first 30 words (seed).
- Used **GPT-4o mini** to complete sentences (approx. 150 words).
- Sample Size: 300 machine samples vs 100 human samples per dataset.

We used three prompts to test detection robustness:

- ① **Standard Continuation:** *"Continue this article..."* (Baseline).
- ② **Whimsical Role (Pirate):** *"Continue using pirate style and vocabulary..."*
- ③ **Adversarial Human Style:** *"Act as a skilled human journalist. Vary sentence structure. Avoid robotic tone. Do not use clichéd transitions."*

Models for Binoculars Calculation

We tested three pairs of scoring models:

1. Mistral (High Performance)

- Ministral-3-3B-Instruct & Mistral-3-3B-Base

2. Gemma (Small Scale)

- gemma-3-270m & gemma-3-270m-it
- Used to investigate impact of model size.

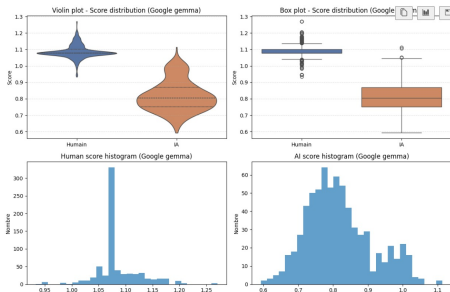
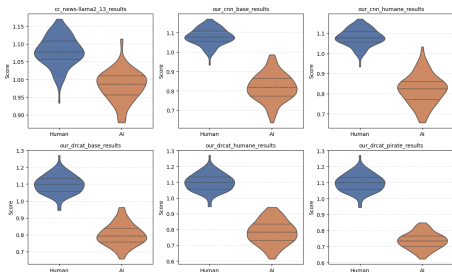
3. Falcon (Replication)

- falcon-7b & falcon-7b-instruct
- Direct comparison to original paper.

- **Standard Metrics:** AUC, F1-score (reported for completeness).
- **Critical Metric:** True Positive Rate (TPR) at **0.01%** False Positive Rate (FPR).
- **Why?** In content moderation, we cannot afford to ban human users mistakenly.
- **Thresholding:** We compute optimal thresholds to minimize FPR while maximizing detection.

Results: Gemma

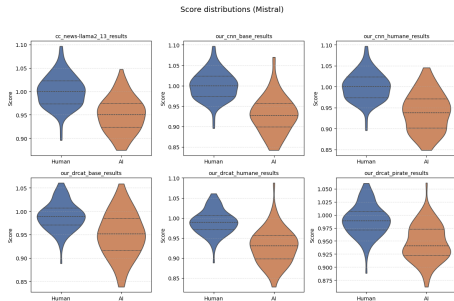
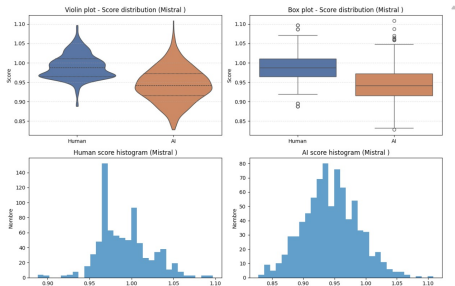
Score distributions (Gemma)



Observations

- Clear separation between Human texts (median ≈ 1.08) and AI-generated texts (median ≈ 0.75 – 0.85).
- Human scores are strongly concentrated at higher values.
- AI-generated scores exhibit a broader distribution tail.

Results: Mistral

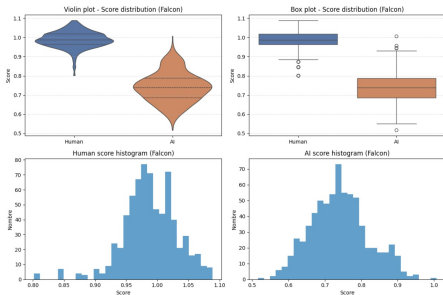
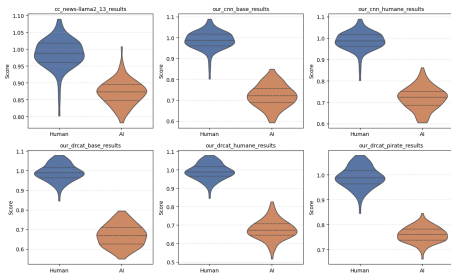


Observations

- Weakest detection performance among the evaluated models.
- Strong overlap between Human and AI score distributions.
- Human peak (≈ 0.96) closely matches AI (≈ 0.92 – 0.97).
- Adversarial prompting significantly widens AI distributions, reducing detectability.

Results: Falcon (Original Paper Model)

Score distributions (Falcon)



Observations

- Good detection performance
- Clear polarization of scores: Human texts around 1.0, AI texts around 0.6.
- Larger model capacity leads to stronger separation under the Binoculars method.

ROC Curves Comparison

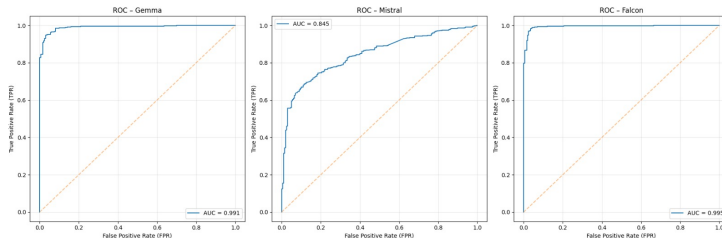


Figure: Left: Gemma (0.991), Center: Mistral (0.845), Right: Falcon (0.995)

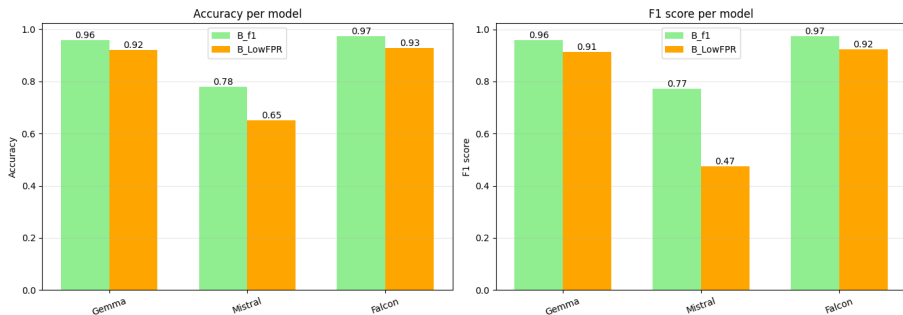
Falcon achieves near-perfect separation (AUC 0.995).

Optimal Thresholds

Model	Low FPR Threshold	F1 Opt. Threshold
Gemma	0.94	1.00
Mistral	0.92	0.97
Falcon	0.85	0.92

Table: Optimal thresholds generally stay near 0.9.

Accuracy Analysis: Threshold Impact



Threshold strategies compared. Low-FPR optimization focuses on minimizing false positives, while Max-F1 optimization aims to maximize the F1-score.

Key finding. Optimizing for the F1-score consistently yields higher overall accuracy than strictly minimizing the false positive rate. This trend is observed across all three evaluated models.

Baseline: Supervised Classifier Implementation

1. Architecture & Training

- **Base Model:**
Solon-embeddings-mini (210M params).
- **Setup:** Frozen backbone + single **Trainable Linear Layer**.
- **Dataset:** HC3 (24k QA pairs).

2. Results (Accuracy)

Dataset	Style	Acc.
CNN	News	83.5%
DRCAT	Essays	54.0%
Pirate	Creative	30.0%

Conclusion: The Generalization Problem

The supervised classifier performs well **only** on data similar to its training set (News). It fails catastrophically on style shifts (Essays, Pirate), confirming that **Binoculars (Zero-Shot)** is far more robust to out-of-distribution text.

Conclusion

- **Robustness:** Being Zero-Shot, the method avoids overfitting and handles "trap" prompts (e.g., Pirate style) where standard classifiers fail.
- **Surprise Finding:** Even tiny models like **Gemma 270M** achieve excellent detection results.
- **Trade-off:** The main drawback is the **high computational cost** (running 2 LLMs) compared to a lightweight supervised model.

Thank you for listening!