Spot the Bot: Detecting AI-Generated Text in the Wild

Corentin Pernot
ENSAE Paris
corentin.pernot@ensae.fr

Jules Chapon
ENSAE Paris
jules.chapon@ensae.fr

1 Introduction

Large Language Models (LLMs) such as GPT or Bard are now capable of generating fluent, coherent, and stylistically rich text across a wide range of domains. While these systems offer new opportunities in content creation and automation, they also introduce important challenges related to authorship attribution, plagiarism detection, and the trustworthiness of online information. As the boundary between human-written and machine-generated text becomes increasingly subtle, the need for reliable detection methods is growing.

In this project, we address the task of **detecting AI-generated text**. We base our study on the *Imitation Game* dataset (1), which contains texts authored by both humans and various LLMs, across multiple genres including poetry, essays, and short stories. This dataset allows us to evaluate detection methods in a realistic and diverse setting.

We explore several detection strategies. We start by analyzing the dataset and implementing simple, interpretable baselines based on **punctuation patterns** and **TF-IDF embeddings**. We then evaluate more advanced approaches, including **perplexity-based scoring** and **RoBERTa-based classification**. These methods are finally combined in a unified model that uses the detection scores from each approach as input features. We also investigate the **robustness** of our final detector by designing adversarial prompts intended to fool it.

The rest of the report is organized as follows: Section 2 introduces the dataset and its structure. Section 3 presents our data analysis and preprocessing steps. Section 4 describes our baseline models. Section 5 reviews relevant state-of-the-art techniques. Section 6 introduces the final model combining all approaches. Section 7 evaluates the model's robustness to targeted attacks. Finally, Section 8 concludes and outlines directions for future work. The code is available on GitHub[1].

2 Dataset

Our study utilizes the *Imitation Game* dataset (1), designed to evaluate the ability to distinguish between human-written and AI-generated texts. This dataset covers three genres: **poetry**, **essays**, and **stories**, with each text authored by either a human or a large language model (LLM) such as GPT or Bard.

Each sample is labeled with both its source (Human, GPT, or Bard) and its genre. In addition to straightforward human- and machine-authored texts, the dataset includes adversarial cases where humans attempt to mimic LLMs and vice versa, increasing the difficulty of the detection task.

Preliminary findings suggest that essays are relatively easier to classify correctly, both for humans and models, whereas stories generated by GPT tend to be more challenging to distinguish from human-written content.

[1] https://github.com/jules-chapon/ml-for-nlp

Genre	Human	GPT	Bard	Total
Poetry	13854	250	250	14354
Essay	2467	200	198	2865
Story	180	25	95	300
Total	16401	475	543	17419

Table 1: Number of texts per genre and author type.

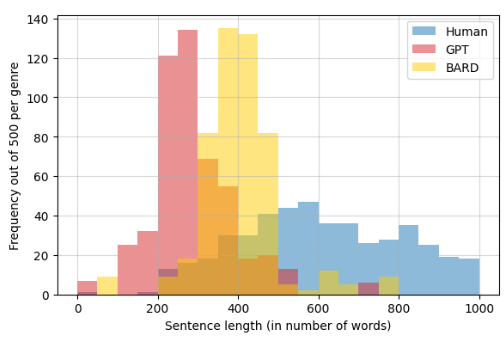
3 Data Analysis and Preprocessing

3.1 Preprocessing and Text Cleaning

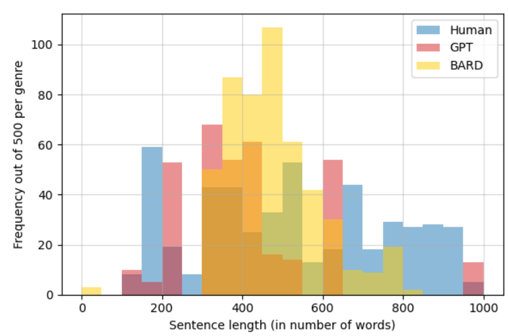
Before feature extraction and modeling, we applied a series of preprocessing steps to the raw texts. In particular, we removed the word *"Sure"* whenever it appeared at the beginning of a response. This artifact was frequently introduced by LLMs (notably Bard), and served as a strong but artificial indicator of machine generation. Its presence could have biased the learning algorithms, so we chose to exclude it to ensure a more robust and fair comparison across sources.

3.2 Text Length Distributions

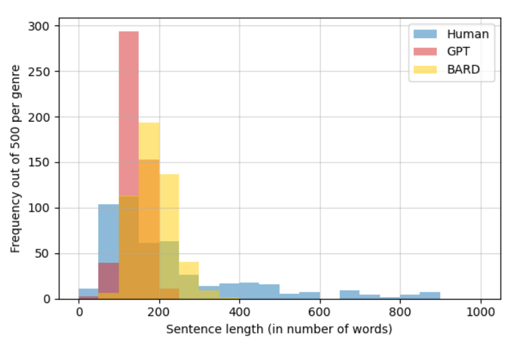
We analyzed the distribution of text lengths (in number of words) for each genre and author type. This provides insight into structural differences between human and machine-generated texts. Figures 1a to 1c present the histograms for essay, story, and poetry samples.



(a) Essay



(b) Story



(c) Poetry

Figure 1: Distribution of text lengths (in number of words) for each genre and generation source.

For stories, human texts tend to be longer and more variable, typically between 400 and 1000 words. GPT displays a bimodal pattern, with peaks around 300 and 600 words, while Bard produces more standardized outputs concentrated around 450–500 words.

For essays, humans write longer texts on average, with most samples ranging from 500 to 800 words. GPT essays are clearly shorter and concentrated around 250–300 words. Bard essays are slightly longer, around 350–400 words, but still more consistent than human productions.

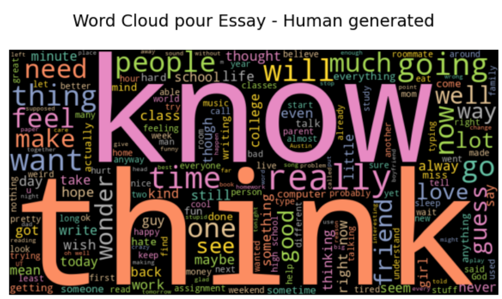
For poetry, the length distributions are generally shorter and more comparable across sources. GPT-generated poems are sharply peaked around 200 words, Bard poems slightly longer around 250, and human poems show slightly more variability, though still mostly below 400 words.

These results suggest that human-authored texts, especially essays and stories, exhibit greater length variability and tend to be longer. In contrast, GPT and Bard tend to produce shorter and more predictable outputs, particularly Bard. Length patterns alone may already provide useful information for classification.

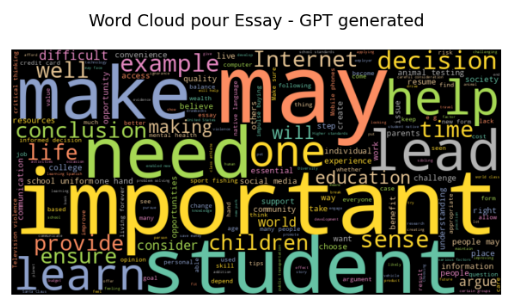
3.3 Word frequency analysis in essays

We constructed word clouds for each generation type in the essay genre to examine whether certain words were more frequently associated with either human or LLM-generated text (Figures 2a to 2c). This exploratory analysis aimed to identify potential stylistic or topical differences.

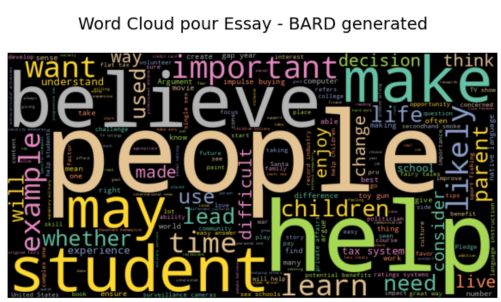
Some lexical differences can be observed: LLM-generated essays tend to use terms such as *student*, *important*, or *help*, while human-written texts more often include verbs like *think*, *feel*, or *want*, which reflect a more personal or reflective tone. However, the overall overlap in vocabulary remains high, suggesting that word frequency alone is unlikely to serve as a reliable feature for distinguishing between sources.



(a) Essay – Human



(b) Essay – GPT



(c) Essay – Bard

Figure 2: Word clouds for the essay genre by generation source (Human, GPT, Bard).

4 Baseline model: classification

Punctuation patterns and word usage are known to be informative for authorship attribution. Despite their simplicity compared to deep contextual models, they offer strong and interpretable baselines for AI text detection. We consider two hand-crafted representations: a punctuation-based embedding capturing structural features, and a TF-IDF vectorization reflecting lexical content.

4.1 Data construction and evaluation setup

To ensure comparability between generation types, we constructed a balanced dataset by selecting, for each genre (essay, story, poetry), an equal number of texts from each class (Human, GPT, Bard), fixed to the smallest available count across sources. A 70/30 train-test split was applied, and the same split was used across all baseline experiments.

All models were trained on the full training set pooled across genres, and evaluated both on the global test set and on genre-specific subsets to assess generalization and sensitivity to writing style.

4.2 Preprocessing and feature extraction

Text preprocessing included basic cleaning operations such as lowercasing and removal of redundant whitespace.

The punctuation-based embedding consists of three handcrafted features: the overall proportion of punctuation marks, the average number of sentence delimiters per word, and the proportion of uncommon punctuation symbols. This representation aims to capture stylistic signals that may differ between human and LLM-generated texts.

In parallel, we built a lexical representation using TF-IDF vectorization with a cap of 1000 features, allowing us to capture frequent and distinctive terms while limiting sparsity and overfitting.

4.3 Model training and hyperparameter search

We trained a Random Forest classifier on each representation. Hyperparameters were tuned using random search over 50 configurations with 3-fold cross-validation. The grid included the number of trees, maximum depth, bootstrap strategy, and minimum samples required for splits and leaves.

4.4 Binary classification

In the binary setting, the goal is to distinguish between human-written and AI-generated texts. We merged GPT and Bard-generated texts under a single "AI" class. The classifiers were evaluated both globally and per genre.

4.4.1 Performance

Table 2 summarizes the precision, recall and F1-score of the punctuation-based and TF-IDF classifiers. TF-IDF consistently outperforms the punctuation-based model across all genres, with near-perfect performance on essays.

Subset	**Punctuation**			**TF-IDF**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Test	0.82	0.79	0.80	0.95	0.92	0.93
Poetry	0.86	0.85	0.86	0.93	0.87	0.89
Essay	0.78	0.73	0.75	1.00	1.00	1.00
Story	0.78	0.76	0.77	0.90	0.89	0.89

Table 2: Performance of Punctuation and TF-IDF classifiers. Best values per row in green.

The confusion matrix (Figure 3) shows that TF-IDF yields very few misclassifications. Only 4 AI-generated texts were predicted as human, and 24 human texts as AI. The model is clearly more confident in detecting AI texts, possibly due to lexical patterns that are less variable and more templated.

4.4.2 Interpretability

To interpret the predictions, we used SHAP values (6) with the `TreeExplainer` on the test set. One visualizations is presented for the TF-IDF model (Figure 4).

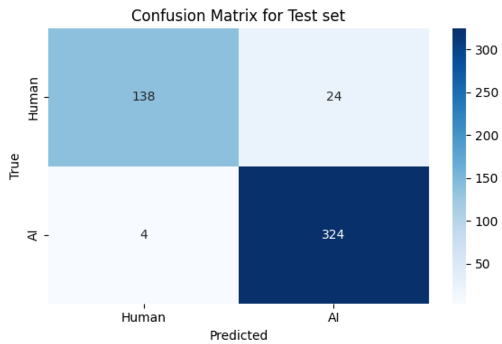


Figure 3: Confusion matrix on the test set (TF-IDF classifier).

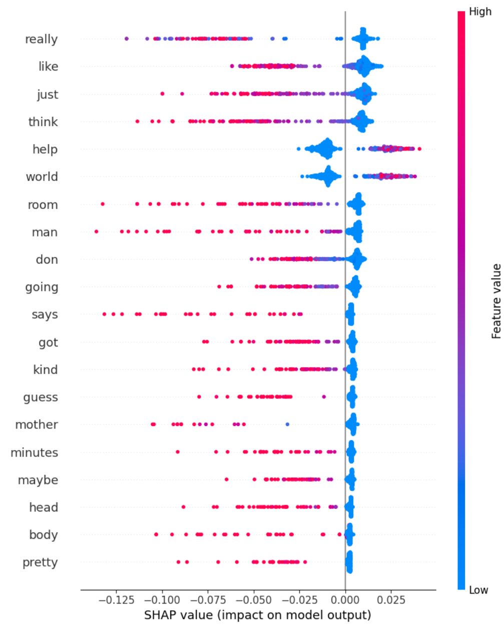


Figure 4: SHAP summary plot for TF-IDF features (top 20).

The SHAP summary plot for the TF-IDF model highlights the most impactful lexical features for predicting class 1 (AI). Positive SHAP values indicate words that increase the probability of an AI prediction, while negative values favor the human class. Terms like *help* and *world* are strongly associated with LLM outputs and consistently push the prediction toward AI. Conversely, words such as *think* and *really* tend to reduce the AI score, reflecting patterns more typical of human writing. These findings are consistent with the word usage differences identified earlier in the dataset analysis.

For the punctuation-based model, the most important feature is the ratio of sentence delimiters per word. Higher values push the prediction toward the human class, suggesting that more varied sentence structuring is indicative of human authorship. Rare punctuation marks also contribute negatively to the AI score, capturing stylistic signals beyond lexical content.

4.5 Multiclass classification

We extend the task to multiclass classification, aiming to assign each text to Human, GPT, or Bard. This setting is more challenging due to stylistic overlap between LLMs.

All metrics are macro-averaged using `sklearn`, computed per class and averaged uniformly, independent of class imbalance.

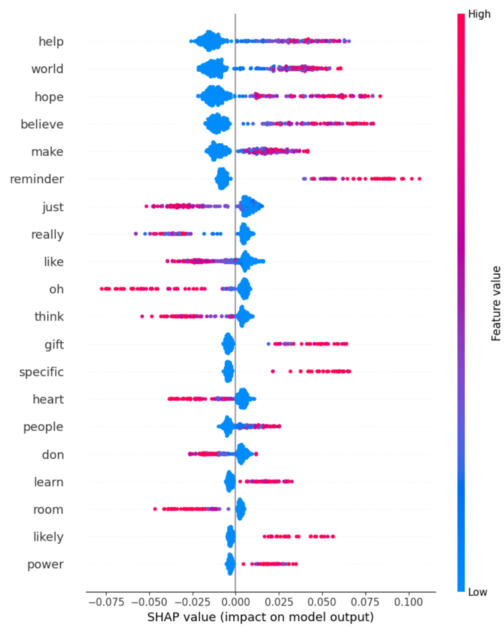
4.5.1 Performances

Subset	**Punctuation**			**TF-IDF**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Test	0.74	0.74	0.74	0.90	0.90	0.90
Poetry	0.80	0.80	0.80	0.90	0.90	0.90
Essay	0.69	0.69	0.68	0.91	0.91	0.91
Story	0.72	0.71	0.71	0.87	0.87	0.87

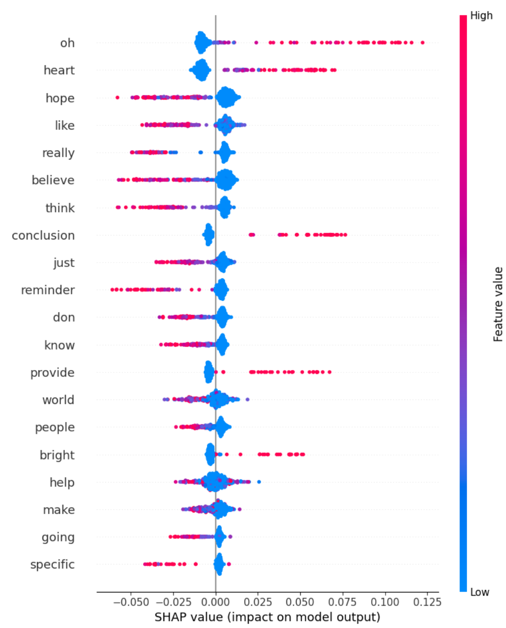
Table 3: Multiclass performance (Human / GPT / Bard) of Punctuation and TF-IDF classifiers. Best values per row in green.

Table 3 reports macro-averaged precision, recall, and F1-score for both models, evaluated on the global test set and separately on each genre. As in the binary setting, TF-IDF significantly outperforms the punctuation-based model across all subsets, with particularly strong performance on essays and poetry.

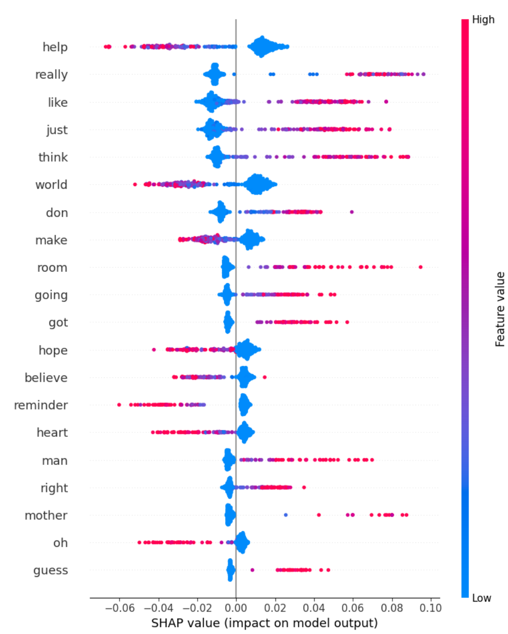
4.5.2 Interpretability



(a) Bard



(b) GPT



(c) Human

Figure 5: SHAP summary plots for each class in the multiclass setting.

To better understand what differentiates GPT and Bard generations, we analyzed SHAP values for each class independently in the multiclass setting. While the aggregated SHAP plot from the binary classification (Figure 4) captures global discriminative patterns between AI and human texts, it does not reflect the specific lexical profiles of each model.

In particular, the SHAP plot for Bard highlights features such as *help*, *world*, *hope*, or *make* as having strong positive contributions to the Bard class. These words, which appeared in the binary setting as indicators of AI-generated text, are here more specifically associated with Bard, suggesting a style that is more general, instructive, and emotionally neutral. This deviates from the aggregated view, where these features were implicitly shared across AI models.

By contrast, GPT predictions rely more on tokens like *oh*, *heart* and *conclusion*, indicating a preference for more narrative or rhetorical structures. The differences in lexical attribution between Bard and GPT suggest that each model introduces its own stylistic biases, which are masked in the binary formulation.

On the Human side, the most influential features remain consistent with earlier findings. Words such as *really*, *think*, or *like* continue to contribute negatively to the AI class score, reinforcing their association with human authorship. The consistency of these markers across both binary and multiclass interpretations strengthens our confidence in their relevance.

5 State-of-the-Art Detection Techniques

In addition to simple handcrafted baselines, several detection methods have recently been proposed to distinguish between human and LLM-generated texts. We briefly review two families of methods that we did not implement but are theoretically relevant, before detailing two others that we evaluated directly.

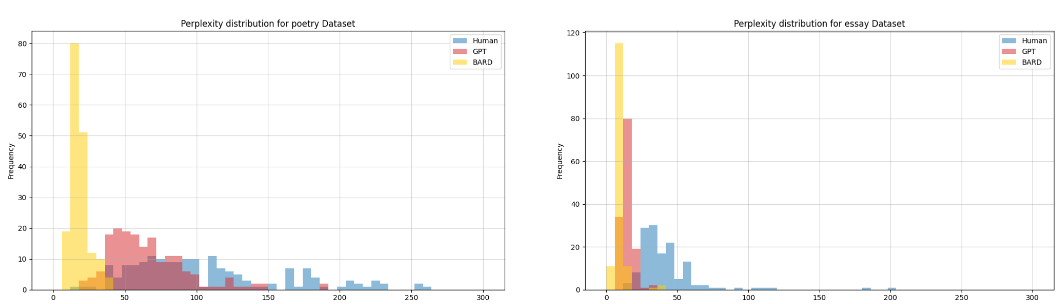
5.1 DetectGPT (not used)

DetectGPT (2) is based on the observation that LLM-generated texts tend to lie near regions of high model likelihood but low local perturbation stability. The method perturbs a candidate text and compares the model log-likelihood of the original versus perturbed versions. A large drop is interpreted as a signal of AI generation. While conceptually interesting, DetectGPT requires repeated forward passes and exact access to the generating model, which makes it impractical in real-world detection pipelines.

5.2 Watermarking Approaches (not used)

Another line of work consists in inserting imperceptible "watermarks" during text generation to make detection easier. The method introduced by Kirchenbauer et al. (3) biases token sampling during generation to embed a statistical signature. Zhao et al. (4) later proposed a provably robust watermarking scheme. While promising, these approaches assume the watermark is present at generation time and are thus not applicable in open-world or retrospective detection settings.

5.3 Perplexity-Based Detection



(a) Perplexity distribution – Poetry

(b) Perplexity distribution – Essay

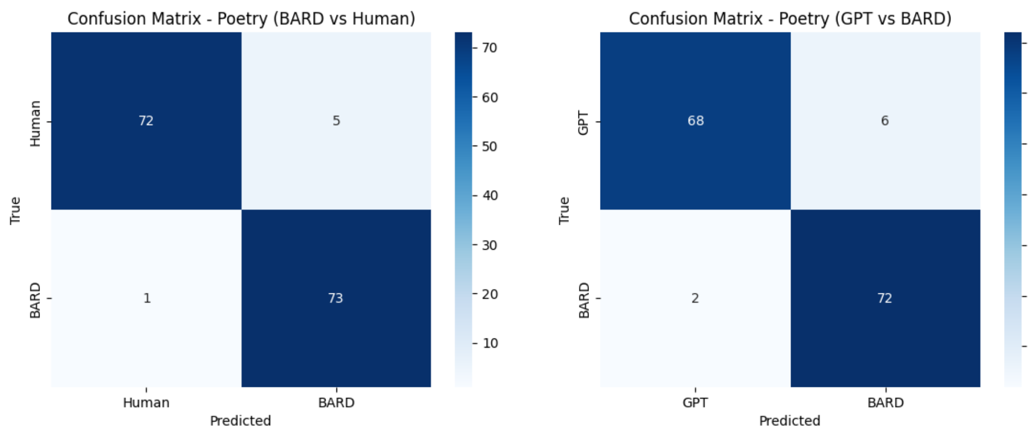
Figure 6: Perplexity scores computed using `distilgpt2` across genres.

Perplexity is a standard metric used to assess how well a language model predicts a sequence. Formally, for a sequence of tokens $x = (x_1, \ldots, x_T)$ and a language model M, the perplexity is defined as:

$$Perplexity(x) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P_M(x_t \mid x_{<t})\right)$$

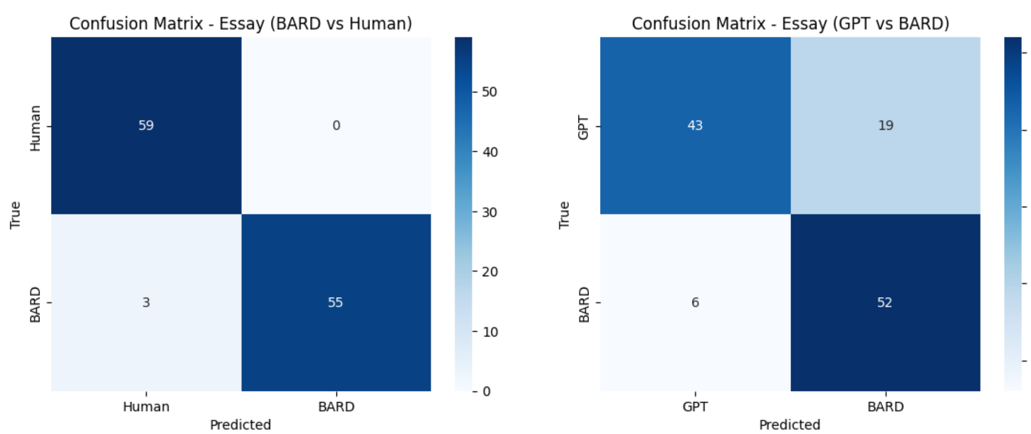
We used the `distilgpt2` model to compute perplexity scores on our dataset. The underlying hypothesis is that LLM-generated text should appear more predictable (i.e., lower perplexity) when evaluated under the same or a similar model. This is clearly seen in Figure 6. A simple threshold, learned on the training set, was used to separate human and AI texts (see Figures 7 and 8).

Applied to the poetry and essay genres, this method yields a clear separation between Bard and Human texts, especially for poetry. Distinction between Bard and GPT remains more challenging due to overlapping distributions, as both are machine-generated and exhibit similar statistical fluency. In particular, GPT and Bard are hard to separate in essays (Figure 8b), while some separation can still be observed in poetry (Figure 7b), suggesting genre-dependent variability in model style.



(a) Bard vs Human – Poetry (threshold = 38.13) (b) Bard vs GPT – Poetry (threshold = 35.12)

Figure 7: Confusion matrices for perplexity-based classification on poetry texts.



(a) Bard vs Human – Essay (threshold = 17.06) (b) Bard vs GPT – Essay (threshold = 12.04)

Figure 8: Confusion matrices for perplexity-based classification on essay texts.

5.4 RoBERTa-Based Classifier

We additionally used the `roberta-base-openai-detector`, a RoBERTa model fine-tuned by OpenAI for human vs. AI text classification. This model was included to provide an external benchmark. It operates by encoding the input text and applying a linear classification head on top of the final [CLS] token representation. We used it in zero-shot mode without further tuning, directly on our test set, to obtain an additional reference score.

6 Final Model: Score Combination

We construct a meta-classifier that combines the outputs of all previously evaluated methods: TF-IDF, punctuation, perplexity, and RoBERTa. Each method provides a scalar score used as input to a second-level classifier.

To prevent data leakage, scores from TF-IDF and punctuation are computed using out-of-fold cross-validation on the training set. Perplexity and RoBERTa scores are obtained directly on the full texts. We additionally include the text genre as a feature, as performance and thresholding have shown to be genre-dependent (see Section 5).

The final model is a Random Forest trained on this combined representation. Results are reported for both binary and multiclass classification. For completeness, we include in the appendix the version without the genre feature, which confirms its relevance.

6.1 Binary Results

Subset	**TF-IDF**			**Perplexity**			**RoBERTa**			**Meta (Final)**		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Test	0.96	0.96	0.96	0.85	0.84	0.84	0.77	0.77	0.77	0.98	0.97	0.97
Poetry	0.95	0.95	0.95	0.78	0.75	0.76	0.60	0.59	0.59	0.98	0.98	0.98
Essay	1.00	1.00	1.00	0.95	0.95	0.95	0.92	0.95	0.93	1.00	1.00	1.00
Story	0.88	0.88	0.88	0.82	0.85	0.83	0.85	0.88	0.86	0.91	0.90	0.91

Table 4: Binary classification performance with genre. Best values per row are highlighted in green.

The final meta-classifier significantly outperforms all individual methods across all subsets. On the test set, it reaches 97% F1-score, compared to 96% for TF-IDF and 84% for perplexity. This confirms that combining weakly correlated signals can lead to strong overall performance.

Genre-specific trends are preserved: performance is highest for essays and lowest for stories, reflecting the intrinsic difficulty of each genre. Notably, the gain from combining methods is most visible for poetry, where single-feature classifiers perform modestly, while the meta-model achieves 98% F1.

As shown in Figure 10, the model relies primarily on the TF-IDF and punctuation scores, while perplexity and RoBERTa-based classifier score provide complementary information. The genre feature, though less important, still contributes to the model's robustness (see Table 6).

6.2 Multiclass

The final multiclass model achieves high precision and recall across all classes, consistently above 90% (see Table 5). As illustrated in Figure 9, it not only separates human from AI-generated texts reliably, but also discriminates well between GPT and Bard outputs—despite their close lexical and structural proximity.

Subset	TF-IDF			Perplexity			RoBERTa			Meta (Final)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Test	0.90	0.90	0.90	0.79	0.79	0.79	0.72	0.72	0.72	0.95	0.94	0.94
Poetry	0.90	0.90	0.90	0.76	0.75	0.75	0.66	0.66	0.66	0.97	0.97	0.97
Essay	0.91	0.91	0.91	0.81	0.82	0.81	0.78	0.78	0.78	0.93	0.93	0.93
Story	0.87	0.87	0.87	0.82	0.81	0.81	0.78	0.79	0.79	0.91	0.91	0.91

Table 5: Multiclass classification performance with genre (Human / GPT / Bard). Best values per row are highlighted in green.

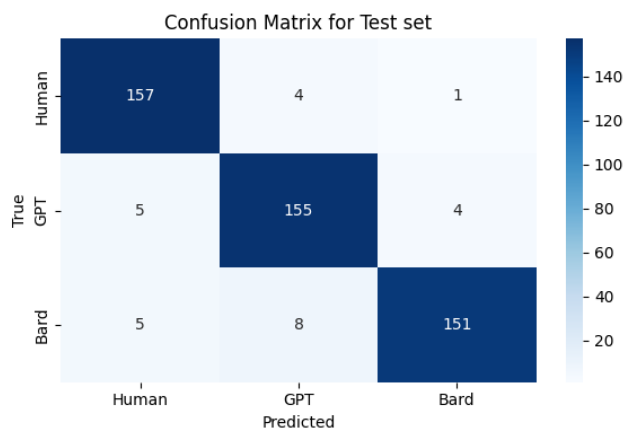


Figure 9: Confusion matrix on the test set for the final multiclass model.

7 Can ChatGPT-4o Fool the Detector?

To evaluate the robustness of our final classifier, we designed a simple attack using ChatGPT-4o. The objective was to assess whether a modern LLM can explicitly control stylistic cues to bypass detection. We provided the following prompt:

```
Give me two poems: one that feels like it was written by a
human, and one that feels like it was generated by an AI.
```

The two generated texts were then fed to our best binary classifier. For the AI-like poem (see A.1), the model correctly predicted the label `AI-GENERATED`, with a high confidence score of 97.5%. However, for the human-like poem (see A.1), the classifier predicted `HUMAN` with 92.3% confidence.

These results suggest that ChatGPT-4o has an implicit understanding of the decision boundary between human and AI writing styles, and can exploit it to produce adversarial examples. While the detection model remains effective on typical generation patterns, this highlights a key limitation: detectors can be fooled when the generation process is explicitly optimized to mimic human traits.

8 Conclusion

In this project, we studied the detection of AI-generated text across three genres using a combination of interpretable baselines (punctuation, TF-IDF) and more advanced methods (perplexity, RoBERTa). A meta-classifier combining these signals achieved the best overall performance. We consistently prioritized interpretability to better understand model behavior and avoid black-box predictions.

However, the project has several limitations. The dataset is restricted to three genres and relies on 2023-era generations from GPT and Bard, which are already outdated. Additionally, we worked under limited computational resources, using off-the-shelf embeddings and simple models such as Random Forests without any fine-tuning.

As a next step, more sophisticated setups could be considered, including prompting large LLMs as zero-shot classifiers (5), or generating adversarial examples through paraphrasing with other models. Such strategies would help further assess the robustness of detectors and address emerging threats from increasingly human-like LLM outputs.

References

[1] Kadhim Hayawi, Sakib Shahriar, and Sujith Samuel Mathew. The Imitation Game: Detecting Human and AI-Generated Texts in the Era of ChatGPT and BARD. *arXiv preprint arXiv:2307.12166*, 2023.

[2] Eric Mitchell, Yoonho Lee, Alexander M. Rush, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv preprint arXiv:2301.11305*, 2023.

[3] Jannick Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. *arXiv preprint arXiv:2301.10226*, 2023.

[4] Xuandong Zhao, et al. Provable Robust Watermarking for AI-Generated Text. *arXiv preprint arXiv:2306.17439*, 2023.

[5] Amrita Bhattacharjee and Huan Liu. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? *arXiv preprint arXiv:2308.01284*, 2023.

[6] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

A Appendix

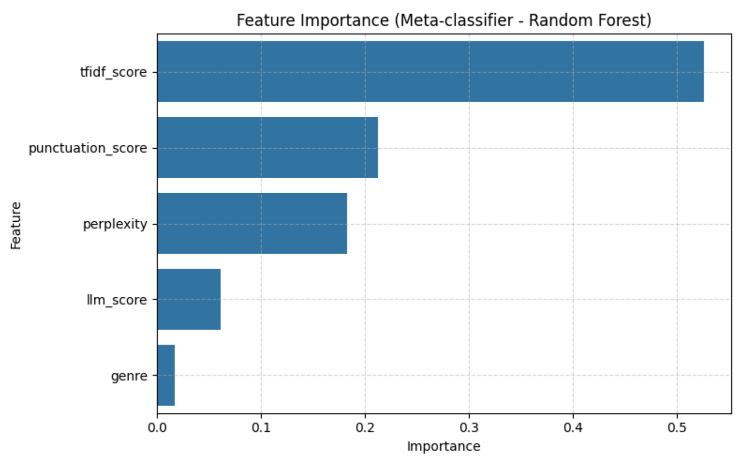


Figure 10: Feature importance from the final Random Forest classifier (binary setting).

Subset	TF-IDF			Perplexity			RoBERTa			Meta (Final)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Test	0.95	0.92	0.93	0.77	0.76	0.76	0.71	0.74	0.72	0.97	0.98	0.97
Poetry	0.93	0.87	0.89	0.72	0.74	0.72	0.55	0.56	0.54	0.97	0.99	0.98
Essay	1.00	1.00	1.00	0.86	0.80	0.82	0.90	0.91	0.91	1.00	1.00	1.00
Story	0.90	0.89	0.89	0.86	0.72	0.75	0.83	0.85	0.83	0.91	0.90	0.91

Table 6: Multiclass performance (Human / GPT / Bard) without genre for TF-IDF, Perplexity, RoBERTa, and the meta-classifier. Best values per row in green.

A.1 Poems Generated by ChatGPT-4o

Prompt. `Give me two poems: one that feels like it was written by a human, and one that feels like it was generated by an AI.`

Human-like poem

```
A Morning Without Reason

I woke up, not sure why-
The sky was gray, but not quite sad.
There might have been wind, or maybe not,
A breath perhaps, or just a thought.

I looked at the wall, the cup, the chair,
Things that don't move but somehow speak.
A silence, worn and a little rough,
Clung softly to the edge of the window.

And I did nothing.  But it felt okay.
As if the moment had forgotten the rest.
No goal, no sound, no proper light-
Yet something quietly...  existed.
```

AI-like poem

```
Reflections at Dawn

The morning breaks with gentle light,
A sky of gold, a world made bright.
The air is still, the silence deep,
A perfect calm, as if in sleep.

The coffee steams, the pages turn,
Soft patterns in the shadows burn.
Each moment flows, precise and clear,
With nothing lost, and nothing near.

A sense of peace, a steady breath,
No fear of time, no thought of death.
Just ordered steps the day shall take,
Until the stars again awake.
```