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

While the recent advancements in Large Language Models (LLMs) present new possibilities for content creation and automation, they also raise significant concerns regarding authorship attribution, plagiarism detection, and the reliability of online information. Indeed, as the distinction between human-written and machine-generated text becomes increasingly blurred, being able to **distinguish AI-generated from human-written texts** is harder and more crucial at the same time.

This is exactly the aim of our project. It is based on the Imitation Game dataset (1), which includes texts written by both humans and various LLMs across different genres such as poetry, essays, and short stories. This dataset enables us to assess detection methods in a realistic and diverse context.

We explore several detection strategies. Initially, we analyze the dataset and implement straightforward, interpretable baselines using punctuation patterns and TF-IDF embeddings. We then evaluate more sophisticated approaches, including perplexity-based scoring and RoBERTa-based classification. These methods are ultimately integrated into a unified model that utilizes the detection scores from each approach as input features. Additionally, we examine the robustness of our final detector by creating adversarial prompts designed to deceive it. I was more involved in the preprocessing and the first solutions, while Corentin focused on the more recent approaches.

The remainder of the report is structured as follows: Section 2 presents the dataset, its preprocessing and its analysis. Section 3 describes and presents the results of our baseline models. Section 4 reviews relevant state-of-the-art techniques. Section 5 presents the final model that combines all approaches. Section 6 assesses the model's robustness against targeted attacks. Finally, Section 7 concludes and suggests directions for future work. The code is available on GitHub[1].

2 Data

2.1 Finding the Good Dataset

Finding a good dataset is a key step in Machine Learning. Having access to reliable, diverse and qualitative examples to train our models on is very important - but yet hard to do, as most of the datasets that are avaiable online are not good enough.

After searching for a long time on Kaggle and HuggingFace, we found the *Imitation Game* dataset (1) as the best dataset. With a decent amount of examples and an article that we could cite, it was also one of the few to provide examples from different categories and LLMs.

Indeed, this dataset is designed to evaluate the ability to distinguish between human-written and AI-generated texts. This dataset covers four genres: **poetry**, **essays**, **code** and **stories**, with each text authored by either a human or one among two LLMs: BARD or GPT. Nonetheless, as explained later in the preprocessing, we did not work on the **code** dataset.

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

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.

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.

2.2 Preprocessing

Before feature extraction and modeling, we needed to clean the dataset to make it suitable for our project.

First, we had to remove the **Code** dataset, as most of the examples only contained the name of the function. Furthermore, we wanted to focus on text, and code was likely to make it very hard for the model to understand, so we preferred to remove it.

Then, we remarked that often, LLMs used the sentence "Sure, here is <INSERT WHAT YOU ASKED FOR>:". Hence, it was a strong but unwanted indicator of AI-generated text, so we removed it to make a fair comparison between LLMs and humans.

Finally, we also decided to clear all "\n" and "\r" that were mostly present in AI-generated texts, probably because of the output format of many LLMs. We wanted to compare LLMs and human beings on the tcontent, not on the form.

2.3 Analysis

First, we analyzed the different text lengths to see if there was any difference between genre and sources. Those text lengths (in number of words) are presented in figure 1.

As we can see, humans seem to produce longer and more diverse texts, as distributions for human-written texts tend to have higher mean and variance than AI-generated ones.

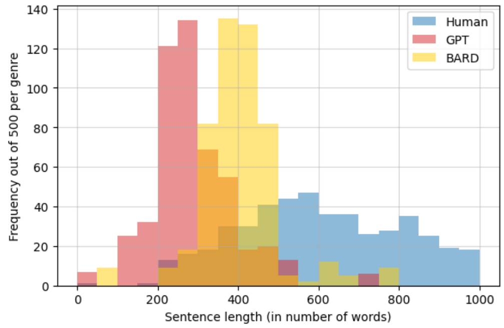
Nevertheless, this is mostly due to the constaints imposed on LLMs, that may not output more than a given number of tokens. By taking this into account, we decided not to include text length as a feature into our models, as it would be an unwanted bias.

Then, 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.

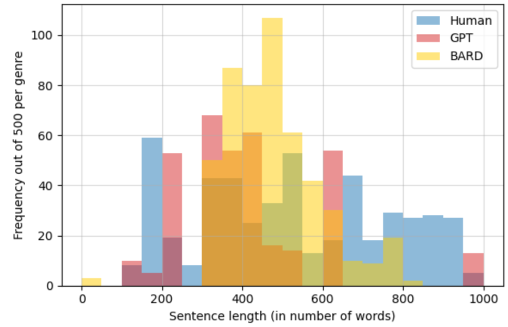
Some lexical differences can be observed: LLM-generated essays tend to use terms such as *student*, *important*, or *help*, while human-written texts more often include verbs like *think*, *feel*, or *want*, which reflect a more personal or reflective tone. However, the overall overlap in vocabulary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

3 Baseline model

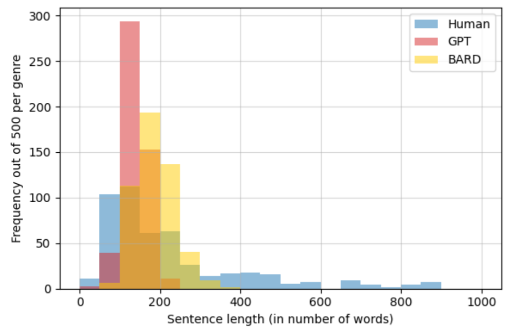
First, we decided to use basic approaches to tackle this classification problem. LLMs are known for being "enthusiastic", and are thus more prone to use special punctuation characters like "!". Moreover,



(a) Essay



(b) Story



(c) Poetry

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

when using LLMs like ChatGPT or LeChat, we realized that those models were often using the same words or patterns in their answers.

Hence, it seemed quite logical to focus on punctuation and vocabulary as our baseline approaches. The first model embeds a text using punctuation features. All of these features represent a ratio between the number of specific punctuation signs and the total number of characters to avoid being biased by the total length of the text. The second model embeds a text using a TF-IDF that has been fitted on the training set,

3.1 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 in all baseline experiments.

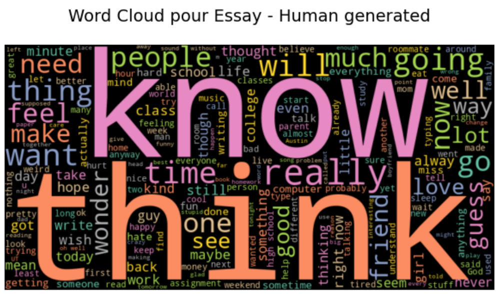
All models were trained on the full training set pooled across genres and evaluated both on the global test set and on genre-specific subsets to assess generalization and sensitivity to writing style.

3.2 Preprocessing and feature extraction

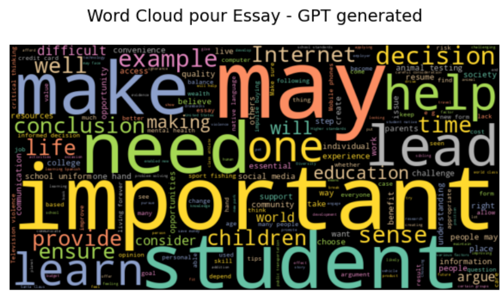
Text preprocessing included basic cleaning operations such as lowercasing and removal of redundant whitespace.

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

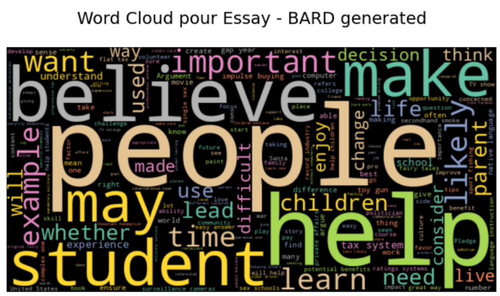
In parallel, we built a lexical representation using TF-IDF vectorization with a cap of 1000 features, allowing us to capture frequent and distinctive terms while limiting sparsity and overfitting.



(a) Essay – Human



(b) Essay – GPT



(c) Essay – BARD

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

3.3 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.

3.3.1 Performance

We trained different models on different sub-tasks.

For the first task, we decided to gather GPT and BARD classes among on "AI" class, and to perform binary classification between AI-generated and human-written texts. 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 binary classifiers.

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.

For the second task, we split the GPT and BARD classes. The goal of the model is now a multiclass classification, where not only it has to determine whether or not the text was AI-generated, but by which model it was. This task is more challenging as different LLMs might share similar patterns. 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. As we can see, results are worse in this setting than in the previous one, which confirms that GPT and BARD share common features that makes it harder for the model to distinguish.

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.

3.3.2 Interpretability

Having a performant model is a good thing, but understanding why a model makes a decision rather than another can be even more important in some cases. However, as modern architectures are becoming more and more complex, it has become harder to understand why those models take such decisions.

But with those two basic models, it is possible to understand their decisions and leverage this to understand what differentiates an AI-generated text from a human-written one. Indeed, regarding the TF-IDF embedding, each feature of the embedding is associated to a word. By determining the contribution of each feature to the prediction, it is thus possible to compute the importance of different words in determining whether or not the text was Ai-generated or not.

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

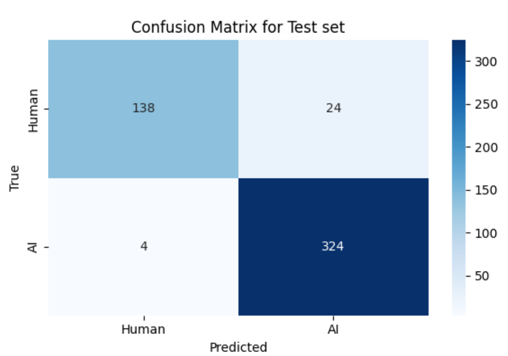


Figure 3: Confusion matrix on the test set (TF-IDF classifier).

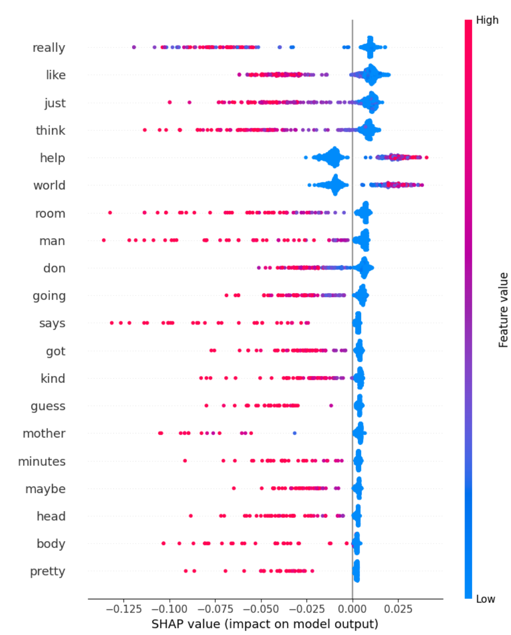


Figure 4: SHAP summary plot for TF-IDF features (top 20).

The SHAP summary plot generated for the TF-IDF model reveals which lexical features most strongly influence the prediction of class 1 (AI). Features with positive SHAP values increase the likelihood of assigning a text to the AI category, while those with negative values lean the prediction toward the human class. Words such as *help* and *world* are frequently linked to outputs produced by LLMs and systematically nudge the classifier toward an AI prediction. Conversely, terms like *think* and *really* tend to lower the AI score, highlighting patterns more characteristic of human writing. These observations align with the word frequency differences previously noted during the dataset exploration.

In the punctuation-based model, the dominant feature is the proportion of sentences. A higher proportion pushes predictions toward the human category, indicating that more complex or varied sentence structuring is a marker of human authorship. Additionally, the presence of infrequent punctuation marks contributes negatively to the AI prediction score, capturing stylistic cues that go beyond pure lexical content.

An equivalent SHAP analysis was carried out within the multiclass framework. The patterns observed closely mirror those from the binary classification setting, with results detailed in the Appendix (Figure 9).

4 State-of-the-Art 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.

4.1 DetectGPT (not used)

DetectGPT (2) leverages a key insight: texts produced by large language models (LLMs) often occupy regions characterized by high model-assigned likelihood yet display limited robustness to small, local perturbations. To exploit this, the approach introduces controlled perturbations to a given text and evaluates how the model's log-likelihood responds. A substantial decline in likelihood following perturbation is taken as evidence that the text may have been AI-generated. Although the underlying idea is compelling, the method's reliance on multiple forward passes and the necessity for direct access to the original model significantly limit its feasibility for deployment in practical detection systems.

4.2 Watermarking Approaches (not used)

Watermarking is used at inference by incorporating some special "markers" that could easily be detected by a model. This kind of method was introduced by Kirchenbauer et al. (3) and improved by Zhao et al. (4). However, for the model to detect it, it must know those specific patterns. Hence, only the companies that built the specific LLM are able to use this watermarking technique.

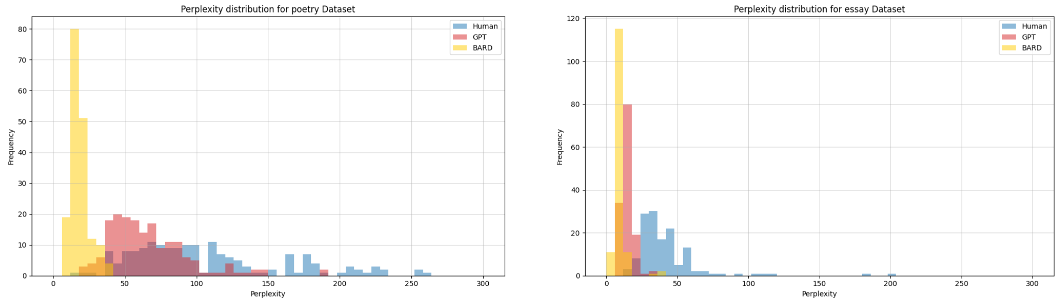
4.3 Perplexity-Based Detection

Perplexity is a standard metric used to assess how well a language model predicts a sequence. Formally, for a sequence of tokens $x = (x_1, \dots, x_T)$ 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 5. A simple threshold, learned on the training set, was used to separate human and AI texts (see Figures 6 and 7).

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

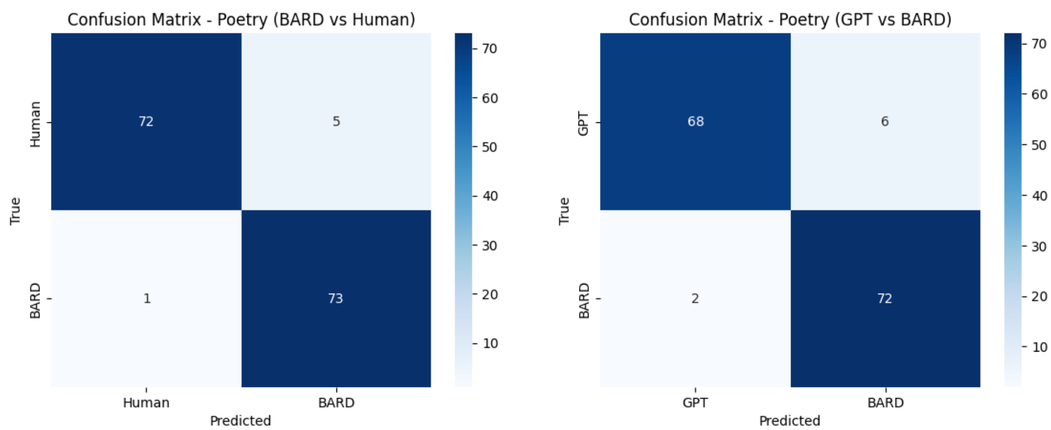


(a) Perplexity distribution – Poetry

(b) Perplexity distribution – Essay

Figure 5: Perplexity scores computed using `distilgpt2` across genres.

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 7b), while some separation can still be observed in poetry (Figure 6b), suggesting genre-dependent variability in model style.



(a) BARD vs Human – Poetry (threshold = 38.13)

(b) BARD vs GPT – Poetry (threshold = 35.12)

Figure 6: Confusion matrices for perplexity-based classification on poetry texts.

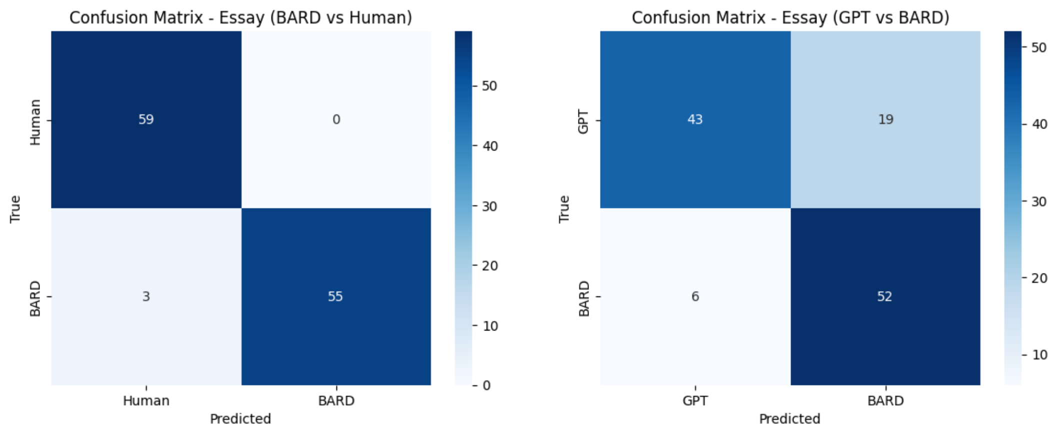
4.4 RoBERTa-Based Classifier

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.

5 Final Model

We developed a meta-classifier that integrates the outputs of various methods: TF-IDF, punctuation, perplexity, and RoBERTa. Each method contributes a scalar score, which serves as input to a second-level classifier.

To avoid data leakage, we calculated the TF-IDF and punctuation scores using out-of-fold cross-validation on the training set. The perplexity and RoBERTa scores were derived directly from the full texts. Additionally, we incorporated the text genre as a feature, given that performance and thresholding have been observed to vary by genre (see Section 4).



(a) BARD vs Human – Essay (threshold = 17.06)

(b) BARD vs GPT – Essay (threshold = 12.04)

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

The final model is a Random Forest trained on this combined set of features. We present results for both binary and multiclass classification tasks. For thoroughness, the appendix includes a version of the model that excludes the genre feature, further underscoring its importance.

5.1 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.

The final meta-classifier demonstrates a substantial improvement over all individual methods across every subset. On the test set, it achieves a 97% F1-score, surpassing the 96% of TF-IDF and the 84% of perplexity. This outcome underscores the effectiveness of integrating weakly correlated signals to enhance overall performance.

The genre-specific performance trends remain consistent: the highest performance is observed for essays, while stories present the greatest challenge, reflecting the inherent complexity of each genre. Remarkably, the benefit of combining methods is most pronounced for poetry, where individual feature classifiers show modest performance, but the meta-model attains a 98% F1-score.

As depicted in Figure 10, the model primarily leverages TF-IDF and punctuation scores, with perplexity and RoBERTa-based classifier scores offering supplementary insights. Although the genre feature is less influential, it still enhances the model's robustness (see Table 6).

The final multiclass model exhibits high precision and recall across all categories, consistently exceeding 90% (see Table 5). As shown in Figure 8, it not only reliably distinguishes between human and AI-generated texts but also effectively differentiates between outputs from GPT and BARD, despite their close lexical and structural similarities.

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).

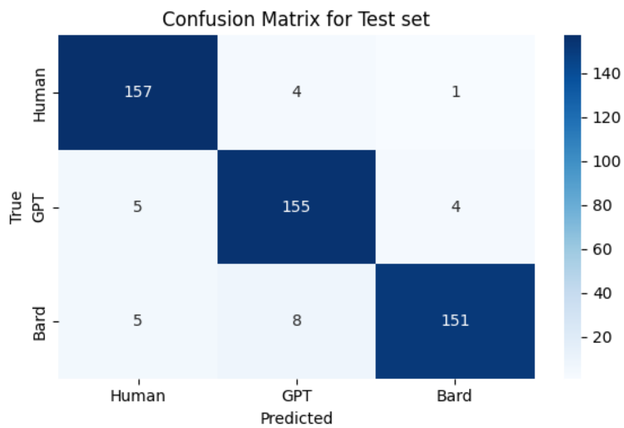


Figure 8: Confusion matrix on the test set for the final multiclass model.

6 To go further

To evaluate the robustness of our final classifier, we designed a simple attack using ChatGPT-4o. The first objective was to assess whether a modern LLM can explicitly control stylistic cues to bypass detection. The second one is to determine whether our model is able to generalize to new LLMs model it was not trained on. 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.3), the model correctly predicted the label `AI-GENERATED`, with a high confidence score of 97.5%. However, for the human-like poem (see A.3), 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.

7 Conclusion

In this project, we explored the identification of AI-generated text across three different genres by employing a mix of interpretable baseline techniques (such as punctuation and TF-IDF) and more sophisticated methods (like perplexity and RoBERTa). A meta-classifier that integrated these various approaches demonstrated the highest overall performance. Throughout our work, we placed a strong emphasis on interpretability to gain a deeper understanding of model behavior and to steer clear of opaque predictions.

Nonetheless, the project has certain limitations. Our dataset is confined to just three genres and utilizes text generated by GPT and BARD from 2023, which is now outdated. Furthermore, we operated under constrained computational resources, relying on pre-existing embeddings and straightforward models like Random Forests without any fine-tuning.

For future work, more advanced configurations could be explored. This includes using large LLMs as zero-shot classifiers through prompting (5), or creating adversarial examples by paraphrasing with alternative models. These approaches would contribute to a more thorough evaluation of detector robustness.

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

A.1 SHAP values for multiclass classification

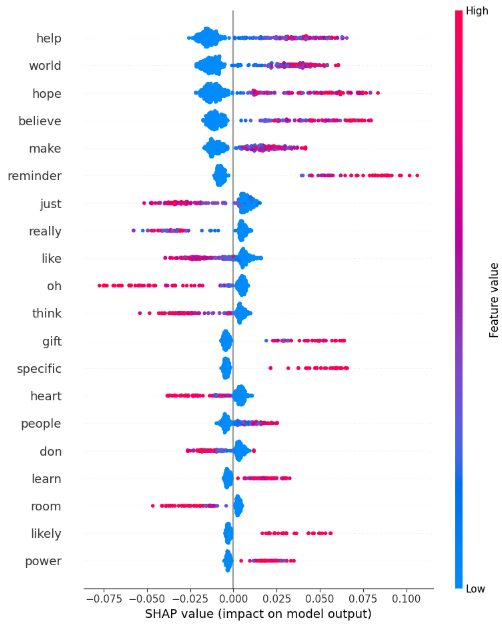
A.2 Final model details

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**

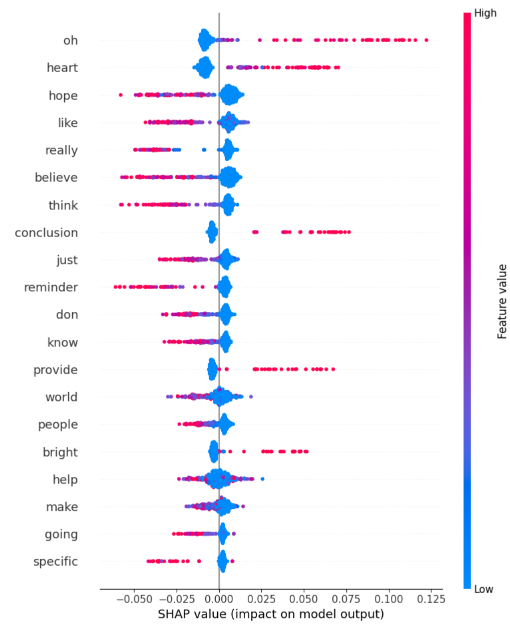
Table 6: Multiclass performance (Human / GPT / BARD) without genre for TF-IDF, Perplexity, RoBERTa, and the meta-classifier.

A.3 Poems Generated by ChatGPT-4o

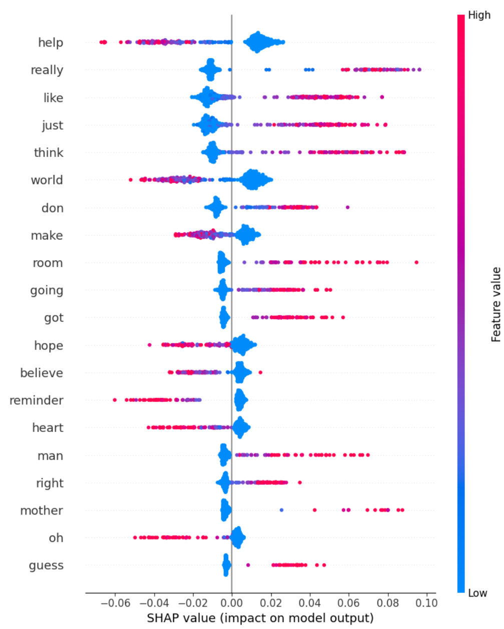
Prompt. `Give me two poems: one that feels like it was written by a human, and one that feels like it was generated by an AI.`



(a) BARD



(b) GPT



(c) Human

Figure 9: SHAP summary plots for each class in the multiclass setting.

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.
```

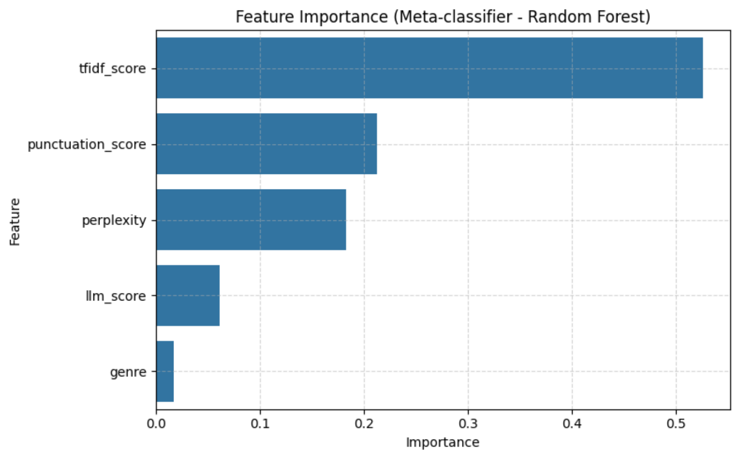


Figure 10: Feature importance from the final Random Forest classifier (binary setting).

```
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.
```