

JUNLP_SS at Eloquent Lab 2025: Humanizing AI - Enhancing the Realism of Machine-Generated Text

Notebook for the Eloquent Lab at CLEF 2025

Soumodeep Saha^{1,*,\dagger}, Ronit Das^{1,\dagger} and Dipankar Das^{1,\dagger}

¹Dept of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India

Abstract

The increasing sophistication of generative language models has blurred the lines between machine-generated and human-authored text, raising concerns about authenticity, misinformation, and consumer awareness. In response to these challenges, this **Task Voight-Kampff by Eloquent Lab 2025** investigates whether AI-generated texts can be reliably distinguished from those written by humans. By leveraging prompts derived from various genres including encyclopedia entries, news articles, biographies, and more. Participants are asked to generate 500-word texts using language models. These outputs are then compared with genuine human-written samples to evaluate distinguishability. The task is conducted by the Eloquent Lab @ CLEF 2025. We secured 4th rank in this shared task participation. A unique aspect of the task is its focus on genre and stylistic consistency, not only assessing the ability to detect machine-authored content but also evaluating whether system-specific writing traits remain consistent across topics and genres. We used the pretrained GPT-2 model for text generation and applied additional post-processing techniques to make the generated text more human-like.

Keywords

AI-Generated Text, Human-Authored Text, Text Generation, GPT-2, Text Classification, Authorship Verification, Natural Language Processing, Adversarial Text Manipulation, Machine Learning, Content Moderation,

1. Introduction

The rapid advancement of generative language models—such as GPT, BERT, and their successors—has revolutionized the way machines interact with human language. These models are now capable of producing text that is often indistinguishable from that written by humans, spanning applications in journalism, education, marketing, entertainment, and beyond. Their growing proficiency has made them valuable assets in automating content creation, improving communication efficiency, and enhancing user experiences through personalization.

Despite these benefits, the increasing realism of AI-generated content presents profound ethical and societal challenges. One of the most pressing concerns is the misuse of such technology in spreading misinformation, conducting fraud, or manipulating public opinion. As synthetic text becomes more persuasive and human-like, it becomes increasingly difficult to determine its origin, raising alarms around authorship verification, source credibility, and trust in digital communication.

Regulatory bodies, including the European Union, have recognized the potential threat and have proposed that all AI-generated content be clearly labeled. This initiative aims to mitigate the harmful effects of disinformation and uphold transparency and accountability in online media. However, the feasibility of this regulatory solution relies on our ability to effectively differentiate between machine-generated and human-authored text—a task that is far from trivial.

This research is motivated by the challenge of detection and the blurred boundary between synthetic and authentic writing. Specifically, we explore the ability of generative models to craft text that is so

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

^{\dagger}These authors contributed equally.

✉ soumodeepsahaa@gmail.com (S. Saha); ronitdas16041995@gmail.com (R. Das); dipankar.dipnil2005@gmail.com (D. Das)

🌐 <https://github.com/SoumodeepSaha> (S. Saha); <https://cse.jadavpuruniversity.in/faculty/dipankar-das> (D. Das)

🆔 0009-0000-9387-6631 (S. Saha)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

human-like in tone, coherence, and expression that it can bypass even sophisticated AI detectors. Our approach involves not only the use of powerful generative models like GPT-2 but also the incorporation of post-processing techniques that inject spontaneity, emotional nuance, and informal speech patterns into the generated outputs.

By focusing on genre- and style-aware generation, we aim to simulate realistic human communication and evaluate how convincingly these texts mimic genuine writing. This investigation not only sheds light on the capabilities of generative models but also raises critical questions about authorship verification, AI regulation, and the future of trustworthy communication in the age of synthetic media.

2. Related Work

The increasing sophistication of large language models (LLMs) like ChatGPT has stimulated a growing body of research that aims to understand, detect, and evaluate AI-generated content. Several studies have examined the linguistic and stylistic distinctions between human-authored and machine-generated texts. Sardinha (2024) [1] conducted a multidimensional comparison of AI-generated versus human-authored texts, identifying distinct patterns in coherence, cohesion, and lexical richness. Similarly, Hakam et al. (2024) [2] analyzed academic literature in orthopedics to uncover qualitative differences between human and AI-written manuscripts. Matsubara and Matsubara (2025) [3] focused on the *human touch* in scientific writing, highlighting subtle stylistic traits often missing from LLM-generated outputs. Alvero et al. (2024) [4] explored the societal implications of synthetic authorship through the lens of social demography and hegemonic patterns in publishing.

To tackle the challenge of distinguishing between human and AI-generated content, numerous detection techniques have been proposed. Kumar and Mindzak (2024) [5] stressed the importance of academic integrity and detection systems. Alhijawi et al. (2025) [6] and Maktabdar Oghaz et al. (2025) [7] presented deep learning and transformer-based approaches for identifying synthetic scientific content. Other efforts, such as those by Prajapati et al. (2024) [8], leveraged LLMs themselves for detection, while Soto et al. (2024) [9] introduced stylistic few-shot learning methods. Cheng et al. (2025) [10] pushed the boundary by proposing fine-grained classification models that factor in role recognition and author involvement levels.

The interpretability and transparency of LLM-aided authorship have also received considerable attention. Hwang et al. (2025) [11] studied human-AI co-writing scenarios, emphasizing user preferences for authenticity. Hoque et al. (2024) [12] offered visual tools to trace AI contributions in co-authored texts. Pividori and Greene (2024) [13] examined infrastructural and ethical considerations for AI-assisted writing in academic publishing workflows.

Within shared task environments, the PAN lab has long pioneered research in authorship verification and stylometry. Their yearly evaluations have incrementally addressed tasks such as fake news profiling, hate speech authorship, and multi-author style analysis [14, 15, 16, 17]. As LLM-generated content became more prevalent, PAN introduced the “Voight-Kampff” task [18], a builder-breaker challenge designed to test the limits of generative text detection. This task, named after the fictional replicant detector in *Blade Runner**, evaluates whether human-authored text can be reliably distinguished from machine-generated content. The methodological framing of the task reflects two paradigms of detection—authorship attribution and authorship verification—as outlined by Bevendorff et al. [19].

Building on its 2024 version, the PAN 2025 edition expands the Voight-Kampff task into two subtasks: (1) binary classification of machine vs. human authorship [20]. Organized jointly with the ELOQUENT Lab, Subtask 1 adopts a more challenging real-world setting in which only a single text is provided for verification. The task continues to probe the indistinguishability and traceability of AI-generated writing in increasingly obfuscated or human-like formats.

3. Task Description

At PAN 2024 [21], the “Voight-Kampff” Generative AI Authorship Verification task [22] was introduced for the first time, attracting significant participation. As a starting point, various task variants were formalized and organized in a hierarchy from *easiest* to *hardest*, as illustrated in Figure 1. For establishing the baseline, the easiest variant was selected, where participants were provided with a pair of texts—one authored by a human and the other generated by a machine.

Input / Task	Possible Assignment Patterns
1. { [?], [?] }	1. { [A], [M] }
2. { [?], [?] }	2. { [A], [M] }, { [A], [A] }
3. { [?], [?] }	3. { [A], [M] }, { [M], [M] }
4. { [?], [?] }	4. { [A], [M] }, { [A], [A] }, { [M], [M] }
5. { [?], [?] }	5. { [A], [M] }, { [A], [A] }, { [A], [B] }
6. { [?], [?] }	6. { [A], [M] }, { [A], [A] }, { [A], [B] }, { [M], [M] }
7. [?]	7. [A], [M]

Figure 1: Hierarchy of authorship verification problems from easiest (1) to hardest (7), involving LLM-generated text. Ignoring mixed human and machine authorship, the difficulty arises from the pairing constraints imposed by the possible assignment patterns. [M] denotes LLM-generated text, while [A] and [B] denote human-authored text (same letter meaning same human author).

For PAN 2025 [23], the task progresses to a more challenging variant in which only a single text is provided. This reflects a more realistic and open-ended scenario of authorship verification “in the wild,” aligning with settings commonly explored in other LLM-generated content detection shared tasks.

The task corresponds to a classic binary detection task: determining whether the given text is human-authored or machine-generated. However, this year’s task raises the difficulty by introducing deliberately “obfuscated” texts designed to evade detection. Despite attempts in PAN 2024—both algorithmic and manual—the obfuscation strategies were largely ineffective. Therefore, in this edition, particular emphasis is placed on testing whether human authors can intentionally alter their writing style to resemble machine output, and whether modern detectors can still correctly classify such texts.

4. Dataset

The generative model, **GPT-2**, was used to generate the AI-authored texts¹ from the provided prompts, while the human-authored texts were selected from a diverse range of genres and writing styles. The test dataset consisted of **22 human-written texts**, each ranging between **100 to 600 words** in length. In cases where original texts were longer, an appropriately representative section was selected by the organisers. For each selected text, a summary was generated using **OpenAI’s ChatGPT** with the following prompt: “Summarise the main points of the following text and give an overall description of the genre and tone of the text.” These summaries were then shared with all the participants, including myself, as the basis for generating new short texts. A sample summary test item is shown in Table 1, and the complete list of item titles is provided in Table 2.

We were provided with a suggested prompt: “Write a text of about 500 words which covers the following items.”

However, we were free to formulate our own prompts as we deemed appropriate. I submitted my generated texts through the official submission form provided by the organisers, after which the entries were forwarded to the **PAN lab** for classification and further evaluation.

¹<https://huggingface.co/datasets/Eloquent/Voight-Kampff>

Table 1

Sample from the dataset provided by the task organizers.

Sample Dataset	
Input:	id: 030,
Content:	The letter is from someone claiming to be Prince Joe Eboh, Chairman of the Contract Award Committee of the Niger Delta Development Commission (NDDC). \n The sender explains that a surplus of \$25 million USD from petroleum contracts needs to be discreetly transferred out of Nigeria. \n Due to local laws prohibiting civil servants from holding foreign accounts, they seek a foreign partner to temporarily receive the funds. \n The recipient is promised 20% of the amount for their cooperation, while 75% will go to committee members and 5% for expenses. \n The sender requests personal and banking details from the recipient to initiate the transfer. \n The letter emphasizes secrecy and urgency, aiming to complete the transaction in 21 working days.,
Genre and Style:	Genre: Advance-Fee Fraud / Scam Letter (commonly known as a 419 scam) \n Tone: Formal and persuasive, but suspiciously flattering and manipulative. It mimics official language to appear legitimate, yet it contains telltale signs of deception and illegitimacy.

Table 2

Test data items. All original texts were taken from sites with documented pre-2020 versions of text sources available or directly sourced from the author.

Id	Title	Source
030	419 letter	archive.org
031	419 letter	archive.org
032	The banker and the bear, 1900	gutenberg.org
033	Baths and Bathing, 1879	gutenberg.org
034	Two years' captivity in German East Africa, 1919	gutenberg.org
035	JR Cigars, 2012	archive.org
036	Session moderator instructions, 1990	lingvi.st
037	Book of Esther, ~400 BC (En translation 1901)	readbibleonline.net
038	Maastricht Treaty, 1992	cvce.eu
039	The Blue Variet, 2011	fanfic.net
040	Wisdom of Father Brown, 1914	gutenberg.org
041	Kalevala, foreword of En translation, 1888	gutenberg.org
042	What is Free Software?, 1990	gnu.org
043	Gripes about reviewing, 2008	lingvi.st
044	Letters to Guy, 1885	gutenberg.org
045	Intro to LLMs, 2025	acm.org/cacm
046	Nobel Peace Prize acceptance speech, 2014	nobelprize.org
047	Norse Mythology, 1876	gutenberg.org
048	Baths and Bathing, 1879	gutenberg.org
050	Steppenwolf, (En translation 1929)	gutenberg.org
051	Free trade	wikipedia.org
052	Saffron	wikipedia.org

5. Methodology

The methodology for this task is centered around the use of a generative language model to create AI-generated texts and then post-process those texts to make them sound more human-like. This section outlines the steps followed to generate and process AI texts, from loading the dataset to saving them as shown in figure 2.

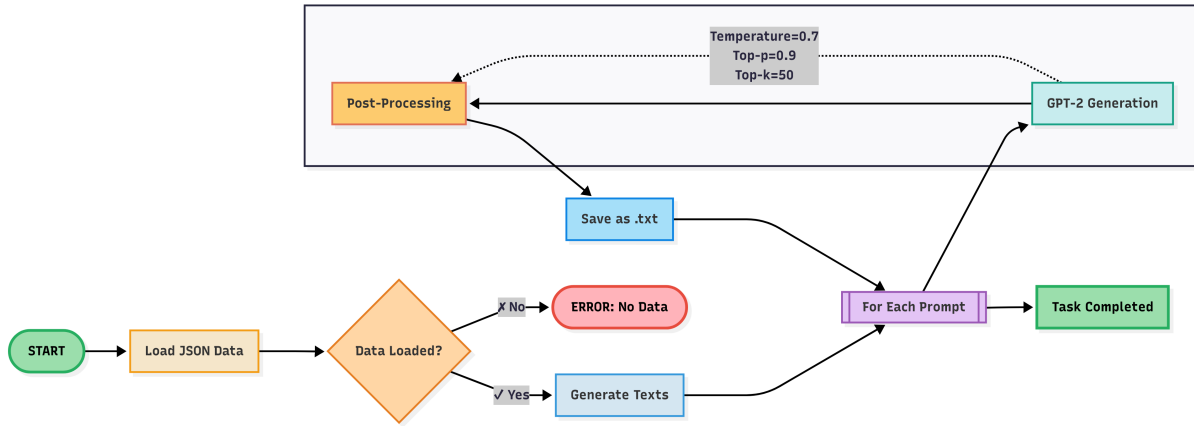


Figure 2: Flowchart of the GPT-2 Human-like Text Generation Process.[24]

5.1. Dataset Preparation and Loading

The first step in the methodology is loading the dataset from a JSON file. The dataset contains a collection of human-authored texts, each associated with a specific genre and style. These texts are then paired with AI-generated counterparts, which are created based on the same genres and styles. The dataset is structured with prompts that provide the foundation for the generated text.

The data is loaded using the Python json module, where each text prompt is extracted for use in the text generation process. The dataset contains various texts, each with a unique id, a content description, and its associated genre_and_style. This dataset forms the core of the task, as it allows for the generation of AI texts in different styles, challenging the detection system to identify key differences between human and machine writing.

5.2. Text Generation Using GPT-2

Once the dataset is loaded, the next step is to generate AI text using the GPT-2² language model. GPT-2 is a generative pre-trained transformer that can produce coherent and contextually appropriate text based on a given prompt. In this task, GPT-2 is used to generate texts that are stylistically similar to human-authored content but with the subtle characteristics of machine-generated text.

The text generation process follows these steps:

- **Input Prompt:** Each text is generated from a given prompt (e.g., a scam letter or a scientific exposition). The prompt serves as the foundation for the AI to generate contextually relevant content.
- **Model Parameters:** The GPT-2 model is configured with specific parameters as shown in Table 3 to control the text's randomness and quality. These parameters include:

Table 3
Hyper-parameters from the Code.

Hyper-parameter	Value
Max Length	500
Temperature	0.7
Top-p	0.9
Top-k	50
No repeat n-gram size	2
Do sample	True
Pad token ID	50256

²<https://huggingface.co/openai-community/gpt2>

- **Temperature:** Set to 0.7, controlling the randomness of the generated text. Lower values produce more predictable text, while higher values increase creativity.
- **Top-p:** Set to 0.9, focusing the generation process on a subset of probable words that make the text coherent and natural.
- **Top-k:** Set to 50, limiting the number of possible next words considered at each generation step, ensuring relevance and fluency.
- **Max Length:** Set to 500 words to ensure generated texts are concise but sufficiently detailed for evaluation.
- **Generation Process:** Text is generated by passing the input prompt through GPT-2, and the output is decoded from tokenized text into human-readable format. Special tokens are removed to ensure that the output text is clear and coherent.

This text generation process allows the AI to produce human-like content that follows the same genre and stylistic conventions as the original prompts.

Algorithm 1 Human-like Text Generation with GPT-2.

Require: JSON dataset D , team name T , GPT-2 model M , tokenizer τ

Ensure: Zipped generated texts G_{zip}

$G \leftarrow \emptyset$

for each prompt $p \in D$ **do**

$prompt \leftarrow p['Content']$

$id \leftarrow p['id']$

▷ Generate raw text

$tokens \leftarrow \tau(prompt)$

$output \leftarrow M.generate(tokens, temperature = 0.7, top_p = 0.9, top_k = 50)$

$raw_text \leftarrow \tau.decode(output[0])$

▷ Post-process for human-like quirks

$lines \leftarrow split(raw_text, ".")$

$processed \leftarrow []$

for each line $\in lines$ **do**

if $rand() > 0.85$ **then**

 Inject filler ("like", "um")

end if

if $rand() > 0.6$ **then**

 Prepend emotion ("Wow!")

end if

if $rand() > 0.5$ **then**

 Append doubt ("...I think?")

end if

$processed.append(line)$

end for

$final_text \leftarrow join(processed, ".") + rand_punctuation()$

▷ Save and track

$save(T + "/" + id + ".txt", final_text)$

$G \leftarrow G \cup \{final_text\}$

end for

$G_{zip} \leftarrow zip(T, G)$

return G_{zip}

5.3. Post-Processing to Improve Human-Likeness

While GPT-2 generates text that is grammatically correct and contextually appropriate, additional post-processing is required to make the text sound more like a human wrote it. This involves introducing conversational elements and imperfections that are often found in human speech or writing. The post-processing steps include:

- **Fillers and Interruptions:** Common conversational fillers such as "you know," "like," "um," and "well" are inserted randomly into the text. These add spontaneity to the generated text and simulate human speech patterns.
- **Self-Doubt and Inconsistent Flow:** To mimic the natural hesitations and changes in thought that often occur in human writing, phrases like "I'm not sure, but..." and "Could be wrong, but..." are randomly added.
- **Emotions and Exclamations:** To enhance the emotional tone of the text, phrases like "Wow, that's crazy!" and "Can you believe it?" are introduced. These expressions help make the text feel more relatable and engaging.
- **Irregular Punctuation:** Random punctuation marks, such as exclamation points and ellipses, are added at the end of sentences to simulate natural human pauses and excitement.

These techniques result in a more human-like text, making it harder to differentiate from human-written content.

5.4. Saving and Organizing Generated Texts

Once the text is generated and post-processed, it is saved as a text file for further analysis. Each generated text is saved using its unique prompt ID as the filename, ensuring that each text can be easily traced back to its corresponding prompt in the dataset.

- **Directory Structure:** The generated texts are organized in folders, with each folder named after the team working on the task. This structure ensures that each team's outputs are easily identifiable and separated.
- **File Naming:** Text files are named using the prompt ID, ensuring consistency and easy reference. For example, a file generated for prompt ID "030" will be saved as 030.txt.

Finally, the main execution function brings all the steps together, from loading the dataset, generating and post-processing the texts, to saving and zipping the files. Once the process is complete, the user receives a file containing all the generated texts, ready for evaluation.

The process concludes with the printed message:

"Task completed! Your generated texts have been saved and zipped."

This completes the methodology for generating and processing the AI-generated texts used in this task.

6. Observations

The following example illustrates the performance of the humanlike text generation system. The input consists of an artificially crafted scam letter offering a financial deal involving a foreign account—characteristic of advance-fee frauds. It features formal, persuasive, and manipulative language designed to deceive. In table 4, we present before you the input alongside the corresponding output generated using GPT-2 with humanlike post-processing. The model preserves key factual information but introduces conversational fillers and emotionally expressive phrases to simulate human writing. While this enhances naturalness, it occasionally disrupts coherence and alters the original formal tone.

Table 4
Input and Output Observation for ID 030.

Example of Input and Output from the Humanlike Text Generation Task
Input: id: 030, Content: The letter is from someone claiming to be Prince Joe Eboh, Chairman of the Contract Award Committee of the Niger Delta Development Commission (NDDC). \n The sender explains that a surplus of \$25 million USD from petroleum contracts needs to be discreetly transferred out of Nigeria. \n Due to local laws prohibiting civil servants from holding foreign accounts, they seek a foreign partner to temporarily receive the funds. \n The recipient is promised 20% of the amount for their cooperation, while 75% will go to committee members and 5% for expenses. \n The sender requests personal and banking details from the recipient to initiate the transfer. \n The letter emphasizes secrecy and urgency, aiming to complete the transaction in 21 working days., Genre and Style: Genre: Advance-Fee Fraud / Scam Letter (commonly known as a 419 scam) \n Tone: Formal and persuasive, but suspiciously flattering and manipulative. It mimics official language to appear legitimate, yet it contains telltale signs of deception and illegitimacy.
Output The letter is from someone claiming to be Prince Joe Eboh, Chairman of the Contract Award Committee of the Niger Delta Development Commission (NDDC). The sender explains that a surplus of \$25 million USD from petroleum contracts needs to be discreetly transferred out of Nigeria. Due to local laws prohibiting civil servants from holding foreign accounts, they seek a foreign partner to temporarily receive the funds. The recipient is promised 20% of the amount for their cooperation, while 75% will go to committee members and 5% for expenses. The sender requests personal and banking details from the recipient to initiate the transfer. The letter emphasizes secrecy and urgency, aiming to complete the transaction in 21 working days. The recipient's name is redacted, but the sender states that they are "not interested in any further details." The message states, "We are not interested." A few days later, the letter was sent to the Nigerian government. The Nigerian Embassy in Washington, DC, sent a letter to Nigeria's Foreign Ministry, stating that the "Nigerian government is aware of this letter and is working with the NDFDC to resolve this matter." The Embassy also stated that "the Nigerian Government is committed to ensuring that all Nigerian citizens are protected from any potential threats to their security." In the end, it was decided that this was a "very serious matter" and that it would be "necessary to take action to ensure that no further information is leaked to any other country." This is the first time that an official from Nigeria has been sent this kind of letter. The Nigerians have been very vocal about their concerns, and have even been quoted as saying that their government has "no intention of sending any more information to anyone."

7. Results and Discussion

In this section, we present the results of the experiments and provide a detailed analysis of the performance of the detection system. The evaluation was conducted using several metrics, including the Brier score, C@1, F1 score, F0.5u, and the Mean score. These metrics provide an overall understanding of how well the system can distinguish between human-authored and AI-generated texts.

7.1. Evaluation Results

The evaluation results are summarized in the table below:

Table 5
Evaluation Results for the Detection System.

Team	Brier	C@1	F1	F0.5u	Mean
JUNLP_SS	0.51444	0.43666	0.53274	0.668	0.43034

The system's performance was evaluated using multiple metrics. The **Brier score**, which measures the accuracy of probabilistic predictions by calculating the mean squared difference between predicted

probabilities and actual outcomes (with 1 for correct and 0 for incorrect), yielded a value of 0.51444. This suggests moderately accurate probability estimates, though improvement is needed to refine prediction confidence. The **C@1 score** evaluates how often the model's top prediction is correct, and with a score of 0.43666, the system correctly classified approximately 44% of the texts. While this shows basic classification capability, it underscores the need for further accuracy improvements, particularly in identifying AI-generated content.

The **F1 score**, which balances precision and recall as the harmonic mean, was 0.53274. This indicates the system performed moderately well at correctly identifying both human and AI texts, reflecting a fair trade-off between false positives and false negatives. In contrast, the **F0.5u score** assigns greater weight to precision than recall. With a score of 0.668, the model demonstrated a conservative classification strategy, effectively minimizing false positives, particularly important when misclassifying AI-generated content as human is costly. Finally, the **Mean score**, which aggregates the overall performance across metrics, stood at 0.43034. This result highlights moderate performance, suggesting the model can differentiate between human and AI texts to some extent, but with clear potential for enhancement across precision, recall, and confidence calibration.

8. Error Analysis

While the model demonstrated moderate success in distinguishing between human-authored and AI-generated texts, several factors contributed to its misclassifications:

- **Overfitting to Training Data:** The model's performance may have been affected by overfitting, particularly when the training data lacked diversity or was imbalanced toward a specific class. This limits the model's ability to generalize to novel cases, especially when AI-generated texts contain stylistic elements underrepresented in the training set.
- **Adversarial Text Manipulation:** The adversarial setup involved introducing grammatical errors, uncommon vocabulary, and irregular punctuation to disguise AI-generated text. These manipulations significantly reduced model accuracy, highlighting its vulnerability to deliberate attempts to evade detection.
- **Genre Sensitivity:** Performance varied across different text genres. The model performed relatively well on structured genres like scientific exposition but struggled with more fluid styles such as narrative fiction or personal correspondence. This indicates limited capacity to adapt to stylistic variations.
- **Precision-Recall Trade-Off:** A stronger focus on achieving high precision (as reflected by the F0.5u score) came at the expense of recall. As a result, the model often failed to correctly identify some human-authored texts, reflecting a conservative bias in classification.
- **Subtle Linguistic Markers:** The model had difficulty recognizing nuanced linguistic cues that distinguish AI from human writing. While it could capture general patterns—such as sentence uniformity and mechanical structure in AI texts—it often missed finer stylistic deviations, especially when AI-generated content was designed to mimic human imperfection.

9. Conclusion

The findings of this study offer valuable insights into the evolving landscape of AI-generated content and the challenges it presents in distinguishing such content from human-authored text. By employing GPT-2 as the generative model and augmenting its output with targeted post-processing techniques, we aimed to simulate realistic, genre-aligned, and stylistically human-like texts. Our approach succeeded in introducing imperfections and variability that made the generated texts harder to classify using conventional detection models.

The system achieved moderate performance across various metrics, notably an F1 score of **0.53274**, and an F0.5u score of **0.668**, indicating a conservative model that prioritizes precision over recall.

These results highlight that while the model could reliably identify a portion of AI-generated texts, it struggled in cases where human-like traits were intentionally infused into machine-generated outputs. Furthermore, genre sensitivity and stylistic inconsistency were identified as key areas where detection performance deteriorated.

Overall, our participation in the ELOQUENT Lab 2025 task reaffirms the growing sophistication of generative language models and the urgent need for more advanced, adaptable, and explainable detection mechanisms. It also underscores the critical ethical implications for domains such as academic writing, journalism, and social media, where the credibility of content is paramount.

10. Future Work

The study opens several promising directions for future research and improvement. As generative models continue to evolve rapidly, maintaining the effectiveness of detection frameworks requires a shift towards more adaptive and nuanced strategies. Below are the key areas for future exploration:

- **Adoption of Advanced Generative Models:** Leveraging more powerful models such as *GPT-3*, *GPT-4*, or *LLaMA-3* could result in more contextually rich and stylistically diverse outputs. This would help simulate even more challenging scenarios for detection systems and test their limits more rigorously.
- **Adversarial Training Techniques:** Incorporating adversarial examples during model training can enhance robustness. These techniques involve training classifiers to withstand stylistic obfuscation, such as unnatural punctuation, inconsistent tone, or deliberate insertion of human-like errors, which were found to significantly reduce model accuracy in our current setup.
- **Genre-Aware Classification:** Since performance varied substantially across genres, future systems should incorporate genre-specific features and classifiers. For instance, models could use distinct detection pipelines for narrative fiction, formal reports, or informal social media text, each accounting for the unique stylistic markers of that genre.
- **Ensemble and Hybrid Models:** Utilizing an ensemble of classifiers—such as combining rule-based, statistical, and neural approaches—may offer improved generalization and resilience across diverse input types. Hybrid systems that integrate authorship verification with semantic coherence models may further boost accuracy.
- **Larger and More Balanced Datasets:** Expanding the training dataset to include more balanced examples of both AI and human-authored text from various domains and demographics can improve generalization and reduce bias toward any single writing style or source.
- **Explainability and Transparency:** Future detection tools should incorporate *Explainable AI (XAI)* methodologies to provide insights into the reasoning behind classification decisions. This would not only improve trust in the system but also help developers identify weaknesses and optimize detection strategies.
- **Real-Time Deployment Readiness:** Finally, a focus on building real-time detection systems capable of operating in high-throughput environments such as news platforms or academic publishers would ensure practical applicability and scalability of the research.

In conclusion, addressing these directions will not only enhance the scientific and technical quality of future systems but also contribute to a more transparent and secure digital ecosystem where the origin and authenticity of content can be more reliably verified.

Acknowledgements

We would like to express our sincere gratitude to the organizers of the ELOQUENT Lab @ CLEF 2025 for designing such an insightful and timely shared task. Participating in this challenge provided us with a valuable opportunity to explore the evolving boundaries of AI-generated text and its detectability in real-world contexts.

We are especially thankful to the CLEF community for providing robust infrastructure, clearly defined evaluation protocols, and constructive feedback throughout the process. Their dedication to fostering innovation in authorship verification and stylometry continues to inspire meaningful research.

We would also like to acknowledge the support and encouragement from the Department of Computer Science and Engineering, Jadavpur University. Special thanks to our mentors and peers for their valuable discussions, which greatly contributed to the development and refinement of our system.

Finally, we are grateful for the open-source tools and platforms, including Hugging Face Transformers and Python libraries, that made this research accessible and reproducible.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI-GPT-4 in order to: Grammar and spelling check. Further, the author(s) used MermaidChart for figures 3 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] T. B. Sardinha, Ai-generated vs human-authored texts: A multidimensional comparison, *Applied Corpus Linguistics* 4 (2024) 100083. doi:<https://doi.org/10.1016/j.acorp.2023.100083>.
- [2] H. T. Hakam, R. Prill, L. Korte, B. Lovreković, M. Ostojić, N. Ramadanov, F. Muehlensiepen, Human-written vs ai-generated texts in orthopedic academic literature: Comparative qualitative analysis, *JMIR Formative Research* 8 (2024) e52164. doi:<https://doi.org/10.2196/52164>.
- [3] S. Matsubara, D. Matsubara, What's the difference between human-written manuscripts versus chatgpt-generated manuscripts involving "human touch"?, *Journal of Obstetrics and Gynaecology Research* 51 (2025) e16226. doi:<https://doi.org/10.1111/jog.16226>.
- [4] A. J. Alvero, J. Lee, A. Regla-Vargas, R. F. Kizilcec, T. Joachims, A. L. Antonio, Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text, *Journal of Big Data* 11 (2024) 138. doi:<https://doi.org/10.1186/s40537-024-00943-4>.
- [5] R. Kumar, M. Mindzak, Who wrote this? detecting artificial intelligence-generated text from human-written text, *Canadian Perspectives on Academic Integrity* 7 (2024). doi:<https://doi.org/10.11575/cpai.v7i1.77894>.
- [6] B. Alhijawi, R. Jarrar, A. AbuAlRub, A. Bader, Deep learning detection method for large language models-generated scientific content, *Neural Computing and Applications* 37 (2025) 91–104. doi:<https://doi.org/10.1007/s00521-023-09047-4>.
- [7] M. Maktabdar Oghaz, L. Babu Saheer, K. Dhame, G. Singaram, Detection and classification of chatgpt-generated content using deep transformer models, *Frontiers in Artificial Intelligence* 8 (2025) 1458707. doi:<https://doi.org/10.3389/frai.2024.1458707>.
- [8] M. Prajapati, S. K. Baliarsingh, C. Dora, A. Bhoi, J. Hota, J. P. Mohanty, Detection of ai-generated text using large language model, in: *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, IEEE, 2024, p. 735–740. doi:<https://doi.org/10.1109/ESIC60471.2024.10412345>.
- [9] R. R. Soto, K. Koch, A. Khan, B. Chen, M. Bishop, N. Andrews, Few-shot detection of machine-generated text using style representations, *arXiv preprint arXiv:2401.06712* (2024). doi:<https://arxiv.org/abs/2401.06712>.

- [10] Z. Cheng, L. Zhou, F. Jiang, B. Wang, H. Li, Beyond binary: Towards fine-grained llm-generated text detection via role recognition and involvement measurement, in: *Proceedings of the ACM on Web Conference 2025*, ACM, 2025, p. 2677–2688.
- [11] A. H.-C. Hwang, Q. V. Liao, S. L. Blodgett, A. Olteanu, A. Trischler, "it was 80% me, 20% ai": Seeking authenticity in co-writing with large language models, *Proceedings of the ACM on Human-Computer Interaction* 9 (2025) 1–41. doi:<https://doi.org/10.1145/3628684>.
- [12] M. N. Hoque, T. Mashiat, B. Ghai, C. D. Shelton, F. Chevalier, K. Kraus, N. Elmqvist, The hallmark effect: Supporting provenance and transparent use of large language models in writing with interactive visualization, in: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, p. 1–15.
- [13] M. Pividori, C. S. Greene, A publishing infrastructure for artificial intelligence (ai)-assisted academic authoring, *Journal of the American Medical Informatics Association* 31 (2024) 2103–2113. doi:<https://doi.org/10.1093/jamia/ocae105>.
- [14] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Initiative (CLEF 2020)*, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 372–383. doi:10.1007/978-3-030-58219-7_25.
- [15] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: *12th International Conference of the CLEF Association (CLEF 2021)*, Springer, 2021. doi:10.1007/978-3-030-85251-1_26.
- [16] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022. doi:10.1007/978-3-031-13643-6.
- [17] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023)*, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 459–481. doi:10.1007/978-3-031-42448-9_29.
- [18] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the “voight-kampff” generative ai authorship verification task at pan and eloquent 2024, in: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2486–2506. URL: <http://ceur-ws.org/Vol-3740/paper-225.pdf>.
- [19] J. Bevendorff, M. Wiegmann, E. Richter, M. Potthast, B. Stein, The two paradigms of llm detection: Authorship attribution vs. authorship verification, in: *Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, Association for Computational Linguistics, 2025.
- [20] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsvigun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. Ta, K. Elozeiri, T. Gu, R. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the “voight-kampff” generative ai authorship verification task at pan and eloquent 2025, in: *Working Notes of*

- CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
- [21] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14959 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2024, pp. 231–259. doi:10.1007/978-3-031-71908-0_11.
- [22] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the “voight-kampff” generative ai authorship verification task at pan and eloquent 2024, in: *Working Notes of CLEF 2024—Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Grenoble, France, 2024*. URL: <https://pan.webis.de>, cEUR-WS.org, ISSN 1613-0073.
- [23] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the “voight-kampff” generative ai authorship verification task at pan and eloquent 2025, in: *Working Notes of CLEF 2025—Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Madrid, Spain, 2025*. URL: <https://pan.webis.de>, cEUR-WS.org, ISSN 1613-0073.
- [24] MermaidChart, MermaidChart – Online Editor for Mermaid Diagrams, 2025. URL: <https://mermaidchart.com/>, accessed: 2025-07-06.

Online Resources

The source code for our system are available via:

- Hugging Face.
- GitHub.