M-DAIGT: A Shared Task on Multi-Domain Detection of AI-Generated Text

Salima Lamsiyah[1], Saad Ezzini[2], Abdelkader El Mahdaouy[3], Hamza Alami[4], Abdessamad Benlahbib[4], Samir El Amrany[1], Salmane Chafik[3], Hicham Hammouchi[1]

[1]University of Luxembourg, Luxembourg
[2]King Fahd University of Petroleum and Minerals, Saudi Arabia
[3]Mohammed VI Polytechnic University, Morocco
[4]Sidi Mohamed Ben Abdellah University, Morocco

Abstract

The generation of highly fluent text by Large Language Models (LLMs) poses a significant challenge to information integrity and academic research. In this paper, we introduce the Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task, which focuses on detecting AI-generated text across multiple domains, particularly in news articles and academic writing. M-DAIGT comprises two binary classification subtasks: News Article Detection (NAD) (Subtask 1) and Academic Writing Detection (AWD) (Subtask 2). To support this task, we developed and released a new large-scale benchmark dataset of 30,000 samples, balanced between human-written and AI-generated texts. The AI-generated content was produced using a variety of modern LLMs (e.g., GPT-4, Claude) and diverse prompting strategies. A total of 46 unique teams registered for the shared task, of which four teams submitted final results. All four teams participated in both Subtask 1 and Subtask 2. We describe the methods employed by these participating teams and briefly discuss future directions for M-DAIGT.

1 Introduction

The recent advancements in large language models have created a paradigm shift in content generation (Naveed et al., 2023; Chang et al., 2024). These models offer numerous opportunities to improve a wide range of applications, including academic research and journalism (Chung et al., 2023). However, their powerful capabilities also raise critical concerns regarding the integrity of the information ecosystem (Wu et al., 2025). In journalism, the potential for large-scale automated generation of misinformation and fake news represents a serious societal threat, with AI-generated articles already appearing on both mainstream and disinformation websites (Wu et al., 2025; Ali et al., 2025). In academia, LLMs challenge the fundamental principles of academic honesty (Bittle and El-Gayar, 2025), and the accessibility of these tools has made it easier for students to generate ghostwritten assignments, contributing to a noticeable rise in academic misconduct (Bittle and El-Gayar, 2025; Go et al., 2025). Research indicates that a significant number of students acknowledge using such tools for their coursework, making it increasingly difficult to distinguish between appropriate academic support and plagiarism (Kovari, 2025).

Distinguishing AI-generated text from human writing is a non-trivial scientific challenge. Modern LLMs produce text that is grammatically correct, stylistically coherent, and often factually plausible, making it difficult to differentiate from human output (Brown et al., 2020; Urlana et al., 2024; Mitchell et al., 2023). Empirical studies have shown that humans, including experienced educators with high confidence in their judgment, perform only marginally better than random chance when attempting to distinguish AI-generated text from human-written content (Urlana et al., 2024). Moreover, recent detection approaches, such as entropy-based statistical methods (Shen et al., 2023), syntactic pattern analysis (Tassopoulou et al., 2021), and neural classifiers (Ippolito et al., 2020; Li et al., 2025), show promise yet remain vulnerable to paraphrasing and prompt variation (Rivera Soto et al., 2025; Kirchenbauer et al., 2023). The field is effectively locked in a technological "arms race": as detection tools improve, so do generative models and the methods used to evade them, including paraphrase attacks and text "humanizers" (Wu et al., 2025; Sadasivan et al., 2023).

Therefore, this rapidly evolving landscape underscores the need for ongoing research and rigorous evaluation methods for AI content detection. The motivation for advancing detection methodologies extends beyond a reactive approach aimed solely at identifying academic dishonesty. Rather, it serves as a proactive strategy to preserve the integrity of the digital information ecosystem. One key con-

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 1–9, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-001

cern is the phenomenon of recursive degradation, where future language models may be trained on vast amounts of unlabeled AI-generated text collected from the internet. This process risks diminishing the quality, originality, and diversity of training data, potentially leading to a degradation of model performance over time (Wang et al., 2024b). Given that news articles and academic publications constitute essential sources of high-quality training data, maintaining their authenticity is crucial for ensuring the long-term robustness, reliability, and generalization capabilities of future AI systems.

To address some of these challenges and to further encourage work on AI-generated text detection, we organized the **Multi-Domain Detection of AI-Generated Text (M-DAIGT)** shared task. M-DAIGT focuses on two domains where the authenticity of text is particularly vital: news articles and academic writing. Specifically, the task is structured into two binary classification subtasks:

- **Subtask 1. News Article Detection (NAD):** Classifying news content as human-written or AI-generated.
- **Subtask 2. Academic Writing Detection (AWD):** Classifying academic texts as human-written or AI-generated.

The key contributions of this work are as follows: (1) the creation and public release of a large and diverse dataset of 30,000 samples specifically designed for AI-generated text detection in the domains of news and academia, featuring outputs from models like GPT-4 and Claude using varied prompts (Wang et al., 2024b); and (2) a comprehensive analysis of participating systems that range from statistical methods to transformer-based detectors (Li et al., 2025; Kuznetsov et al., 2025), offering insights into the current state-of-the-art and highlighting key challenges for future research.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI-generated text detection. Section 3 presents the dataset creation process and evaluation metrics. Section 4 presents the baseline and participant models, along with the evaluation methodology and results. Finally, Sections 5 and 5 conclude the paper and discuss the limitations of the shared task.

2 Related Work

AI-Generated Text Detection Methods. The detection of AI-generated text is a rapidly evolving research domain, with increasing attention due to the widespread development of large language models (Wu et al., 2025). Several methods have been proposed for AI-generated text detection, which can be broadly classified into statistics-based methods, neural-based methods, watermarking, and the use of LLMs as detectors.

Statistics-based approaches aim to exploit intrinsic differences in linguistic features between human and machine-generated texts. Early efforts, such as those Shen et al. (2023) and Tassopoulou et al. (2021), leveraged entropy measures and n-gram frequency analysis to differentiate between text origins. Krishna et al. (2022) utilized sentence repetition patterns, noting that LLMs often assign high probability to repetitive content. DetectGPT (Mitchell et al., 2023) proposed a perturbation-based method to identify whether text lies in negatively curved regions of the log-likelihood landscape.

Neural-based methods dominate recent advances in AI-generated text detection due to their high accuracy and adaptability. Early methods adopt fine-tuned models like BERT (Devlin et al., 2019) and RoBERTa (Solaiman et al., 2019). Furthermore, Ippolito et al. (2020) demonstrated that training on outputs generated using diverse decoding strategies (e.g., top-k sampling, nucleus sampling, temperature control) significantly improves detection robustness. Recently, Li et al. (2025) proposed IRON, a robust adversarial training framework that improves resilience against attacks designed to evade detection systems. Jiao et al. (2025) introduced M-RangeDetector, which enhances model generalization via multi-range attention masks. Similarly, Kuznetsov et al. (2025) provided feature-level interpretability through sparse autoencoders, offering insights into which patterns distinguish AI and human text. Tong et al. (2025) combined reinforced sampling with LLM augmentation for improved fake news detection, while Ali et al. (2025) extended neural classifiers to low-resource languages, specifically addressing Urdu fake news detection. These efforts reflect a growing focus on robustness, explainability, and multilingual applicability in neural detection research.

Watermarking-based approaches offer proactive detection capabilities by embedding or identifying implicit signals in generated text. Early methods include synonym replacement, lexical substitution (Li et al., 2023; Sadasivan et al., 2023), and soft watermarking using curated token lists

(Kirchenbauer et al., 2023). Hidden-space watermarking approaches (Zhao et al., 2023) manipulate token-level probability vectors to introduce tamper-resistant signatures. Some methods, like Bhattacharjee and Liu (2024), aim to exploit surface-level word randomness as a trigger for detection. Recently, Rivera Soto et al. (2025) proposed Paraphrase Inversion, a novel technique to counter paraphrase attacks that aim to remove watermark signals by recovering semantic intent. This approach highlights the challenges posed by adversaries seeking to bypass detection through surface-level text alterations. While many watermarking techniques rely on controlled generation, this method contributes a defensive post-processing solution that does not depend on direct access to generation mechanisms.

Lastly, LLMs themselves are increasingly used as detectors. Tools such as GPTZero[1], ZeroGPT[2], and OpenAI's[3] AI text classifier exemplify this trend. Sadasivan et al. (2023) proposed a zero-shot framework using clustering to differentiate between watermarked and unwatermarked text. Wang et al. (2024b) proposed M4, a comprehensive black-box framework for machine-generated text detection that operates across multiple generators, domains, and languages. Their approach focuses on generalization under realistic, diverse conditions by evaluating detectors on unseen generators and multilingual datasets, setting a new benchmark for robust and scalable AI text detection. More recently, Su et al. (2025) introduced HACo-Det, which focuses on fine-grained detection of human-AI coauthored text, a challenging scenario due to subtle stylistic blending. Go et al. (2025) proposed XDAC, a detection and attribution framework using explainable AI for Korean-language content. Li and Wan (2025) examined how authorial intent and role influence AI-text detectability, emphasizing the social and cognitive dimensions of authorship. These works represent a trend toward leveraging LLMs as meta-models that interpret, explain, and critique textual content.

Benchmark Datasets and Shared Tasks. Standardized evaluation frameworks through both benchmark datasets and shared tasks play a crucial role in advancing AI-generated text detection by providing standardized, diverse, and challenging evaluation settings. A variety of benchmarks have been proposed to test generalization across languages, domains, modalities, and attack scenarios. The MultiSocial dataset (Macko et al., 2025) supports multilingual detection on social media content, while XDAC (Go et al., 2025) introduces explainable detection and attribution for LLM-generated news comments in Korean. Double Entendre (Frohmann et al., 2025) expands detection tasks beyond pure text through a multimodal benchmark focused on audio-based AI-generated lyrics. To assess robustness under adversarial conditions, IRON (Li et al., 2025) incorporates adversarially perturbed examples, and Stress-Testing (Pedrotti et al., 2025) manipulates LLM writing styles to mislead detectors. In parallel, feature-level datasets (Kuznetsov et al., 2025) offer interpretable benchmarks using sparse autoencoders, while M4GT-Bench (Wang et al., 2024b) evaluates black-box detectors across multiple generators, domains, and languages, which is critical for real-world deployment. Additional public resources, such as the *AI-and-Human-Generated-Text* dataset available on Hugging Face[4] and the *GPT-generated Text Detection: Benchmark Dataset and Tensor-based Detection Method* (Qazi et al., 2024), further enrich the landscape of available datasets.

Complementing benchmark datasets, shared tasks have emerged as key drivers of progress in AI-generated text detection by offering standardized, competitive, and collaborative evaluation platforms. The SemEval-2024 Task 8 (Wang et al., 2024a) was specifically designed to evaluate detection systems under multimodal, multidomain, and multilingual settings in a black-box scenario, challenging participants to detect text generated by unseen language models across diverse languages and content types. The task highlighted major real-world concerns such as domain shift, lack of training-time transparency, and linguistic variability. Among the participating systems, TrustAI (Urlana et al., 2024) provided a comprehensive analysis of multi-domain machine-generated text detection techniques, implementing various strategies across statistical, neural, and ensemble approaches. Their findings underscored the importance of domain-specific fine-tuning and robust feature extraction in black-box detection contexts. In parallel, the 1st Workshop on GenAI Content De-

[1] https://gptzero.me/

[2] https://www.zerogpt.com/

[3] https://platform.openai.com/ai-text-classifier

[4] https://huggingface.co/datasets/Ateeqq/AI-and-Human-Generated-Text

tection (GenAIDetect) (Alam et al., 2025), held at COLING 2025, provided a dedicated forum for advancing research on generative content detection. It addressed key challenges such as multilingual robustness, adversarial evasion, and watermarking techniques, fostering discussion around emerging benchmarks and methodological innovation. Building on prior efforts, M-DAIGT focused on AI-generated text detection in two critical domains: news journalism and academic writing. It features two binary classification subtasks, News Article Detection (NAD) and Academic Writing Detection (AWD), supported by a newly released dataset.

3 Datasets and Evaluation Metrics

3.1 Datasets Collection

To support the M-DAIGT shared task, we curated a dataset tailored to evaluate systems on detecting AI-generated news and academic texts.

3.1.1 News dataset:

We gathered 7,000 manually written news articles from the CNN Daily News website, covering more than 40 categories. To create the AI-generated counterparts, we used the titles extracted from these human-written articles as input prompts. Multiple language models were employed to generate news content, including LLaMA3.2-3B-Instruct, Qwen2.5-3B-Instruct, Mistral-7B-Instruct-v2.0, and various models from the GPT family (GPT-4o, GPT-3.5, GPT-4o-mini). Each model was prompted using the standardized prompt shown in Listing 1, with the role definition randomly selected at runtime to encourage stylistic diversity in the generated outputs.

3.1.2 Academic texts dataset:

We collected 7,000 abstracts from published papers on ArXiv, covering a range of categories. To minimize the likelihood of including AI-generated content, only papers published before 2019 were selected. For each human-written abstract, we extracted the corresponding paper title and used it as a prompt to generate an AI-written counterpart. The same models described earlier were employed for this task. Each model was prompted using one of the two prompts shown in Listing 2,

In conclusion, we compiled balanced datasets of manually written and AI-generated texts for both news articles and academic abstracts, totaling 14,000 examples per task. Each dataset was

Listing 1: Prompt's Key Components for Generating News Articles

```
-- Each time this prompt is used, a role is randomly selected to influence the assistant writing style.

-- Randomly select one of the following journalist roles:

Role Definition:
    - "You are an expert journalist."
    - "You are a professional news writer with a focus on clear, unbiased reporting."
    - "You are a friendly and engaging journalist, writing in an easy-to-understand style."
    - "You are an opinion writer, focusing on offering personal insights on current news."

-- Instructions:

Generate a news article of approximately '{article_length}'-words on the following topic: '{Title}'

Write only the article content. Do not include a title or any additional commentary.
```

Listing 2: Prompts for Generating Scientific Abstracts

```
-- Prompt 1:

You are a researcher working on a research paper. Your English proficiency level is '{english_proficiency}'.
Your task is to write a well-structured abstract of approximately 250 words for your research paper in response to the given topic: '{title}'.
Ensure your abstract is clear and concise, following the standard format: 'background', 'objective', 'methodology', 'key findings', and 'conclusion'.
The response should contain only the abstract text, without titles or introductory phrases.

-- Prompt 2:

Generate a 250-word abstract for work with the given topic: '{title}'.
Describe the 'results obtained', the 'problem' the work attempts to solve, and the 'key ideas' and 'methodology' in a formal academic and scientific writing voice.
Use the first plural person form. Use active voice.
Please provide only the abstract text, excluding any titles or introductory phrases.
```

randomly divided into 10,000 samples for training, 2,000 for development, and 2,000 for testing, providing a robust foundation for evaluating models performance across different stages.

3.2 Evaluation Metrics

The performance of the participating systems in both the News Article Detection (NAD) and Academic Writing Detection (AWD) subtasks was evaluated based on standard classification metrics. The official ranking of the teams was determined by the F1-score. The primary metrics used for evaluation were:

- **Accuracy:** The proportion of correctly classified instances.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of a model's performance.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall:** The ratio of correctly predicted positive observations to all observations in the actual class.

In addition to these primary metrics, a secondary analysis was planned to assess model robustness across different text lengths, writing styles, topic domains, and the various generation models used to create the dataset.

4 Shared Task Teams & Results

In this section, we present the shared task baseline models, participating systems descriptions, and their obtained results.

4.1 Baselines

We evaluate three simple baselines on both subtasks:

- **ARBERTv2:** A transformer-based model pretrained on large-scale Arabic text (Abdul-Mageed et al., 2021), fine-tuned on each task (5 epochs, learning rate 2 × 10^{-5}).
- **LogReg (char 2–5):** Logistic Regression using character-level n-grams (2–5) with TF–IDF features, designed to capture fine-grained morphological patterns.
- **LogReg (word 1–2):** Logistic Regression using word-level n-grams (1–2) with TF–IDF features, providing simple but effective word co-occurrence representations.

Model	**P**	**R**	**F_1**	**Supp.**
human				
ARBERTv2	0.9979	0.9410	0.9686	1000
LogReg (char 2–5)	0.9791	0.9820	0.9805	1000
LogReg (word 1–2)	0.9679	0.9940	0.9808	1000
machine				
ARBERTv2	0.9442	0.9980	0.9703	1000
LogReg (char 2–5)	0.9819	0.9790	0.9805	1000
LogReg (word 1–2)	0.9938	0.9670	0.9802	1000
accuracy			0.9695	2000
macro avg	0.9710	0.9695	0.9695	2000
weighted avg	0.9710	0.9695	0.9695	2000

Table 1: Task 1 (NAD) development set.

Model	**P**	**R**	**F_1**	**Supp.**
human				
ARBERTv2	0.9946	0.9240	0.9580	1000
LogReg (char 2–5)	0.9759	0.9700	0.9729	1000
LogReg (word 1–2)	0.9529	0.9920	0.9721	1000
machine				
ARBERTv2	0.9290	0.9950	0.9609	1000
LogReg (char 2–5)	0.9702	0.9760	0.9731	1000
LogReg (word 1–2)	0.9917	0.9510	0.9709	1000
accuracy			0.9595	2000
macro avg	0.9618	0.9595	0.9594	2000
weighted avg	0.9618	0.9595	0.9594	2000

Table 2: Task 1 (NAD) test set.

Tables 1–4 report the full per-class metrics on the development and test splits.

On the news domain sub-task, both logistic-regression baselines outperform ARBERTv2, achieving 98.05 F_1 on dev (vs. 96.95) and 97.15 F_1 on test (vs. 95.95), indicating that simple n-gram features capture domain-specific style cues very effectively.

For the academic domain sub-task, ARBERTv2 reaches near-perfect performance (99.85 F_1 dev, 99.75 F_1 test), slightly outperforming the n-gram baselines. These results set a high bar for future participants: transformer fine-tuning excels on formal academic text, while lightweight n-gram classifiers remain surprisingly competitive, especially on news.

4.2 Participants Systems

Four teams submitted system description papers, and their approaches are summarized as follows.

Zain et al. team explored three different architectures: a fine-tuned RoBERTa-base model, a TF-IDF based system with a Linear SVM classifier, and an experimental system named Candace that

Model	P	R	F_1	Supp.
human				
ARBERTv2	0.9980	0.9990	0.9985	1000
LogReg (char 2–5)	0.9950	0.9990	0.9970	1000
LogReg (word 1–2)	0.9881	0.9970	0.9925	1000
machine				
ARBERTv2	0.9990	0.9980	0.9985	1000
LogReg (char 2–5)	0.9990	0.9950	0.9970	1000
LogReg (word 1–2)	0.9970	0.9880	0.9925	1000
accuracy			0.9985	2000
macro avg	0.9985	0.9985	0.9985	2000
weighted avg	0.9985	0.9985	0.9985	2000

Table 3: Task 2 (AWD) development set.

Model	P	R	F_1	Supp.
human				
ARBERTv2	1.0000	0.9950	0.9975	1000
LogReg (char 2–5)	0.9950	0.9980	0.9965	1000
LogReg (word 1–2)	0.9920	0.9960	0.9940	1000
machine				
ARBERTv2	0.9950	1.0000	0.9975	1000
LogReg (char 2–5)	0.9980	0.9950	0.9965	1000
LogReg (word 1–2)	0.9960	0.9920	0.9940	1000
accuracy			0.9975	2000
macro avg	0.9975	0.9975	0.9975	2000
weighted avg	0.9975	0.9975	0.9975	2000

Table 4: Task 2 (AWD) test set.

used probabilistic features from multiple Llama-3.2 models (Zain et al., 2025). Their final submission was based on the fine-tuned **RoBERTa-base** model, which yielded the highest performance on the development sets.

IntegrityAI team proposed a multimodal architecture that combines textual features from a pre-trained **ELECTRA** model with four handcrafted stylometric features: word count, average sentence length, vocabulary richness (TTR), and average word length (IntegrityAI, 2025). For the news subtask, they also employed a pseudo-labeling technique to augment their training data.

Hamada Nayel team focused on classical machine learning algorithms, submitting a system based on a **Linear Support Vector Machine (SVM)** classifier with **TF-IDF** features (Ashraf et al., 2025). Their approach emphasized efficiency and interpretability, demonstrating that traditional methods can achieve competitive performance without the need for resource-intensive deep learning models.

CNLP-NITS-PP team developed a hybrid model that fine-tuned a **DeBERTa-base** transformer and augmented it with nine auxiliary stylometric features, such as Unique Word Count, Stop Word Count, and Type-Token Ratio (Yadagiri et al., 2025). The contextual embedding from DeBERTa was concatenated with the feature vector before being passed to a final classification layer.

4.3 Results

The official results for both subtasks are presented in Tables 5 and 6. All participating teams achieved exceptionally high scores, indicating the high quality of the submitted systems.

Team	F1	Acc.	Prec.	Rec.
Zain et al.	1.000	1.000	1.000	1.000
IntegrityAI	0.996	0.996	0.996	0.996
Hamada Nayel	0.990	0.990	0.980	0.990
CNLP-NITS-PP	0.898	0.898	0.898	0.898

Table 5: Official results for Subtask 1 (NAD).

Team	F1	Acc.	Prec.	Rec.
Zain et al.	1.000	1.000	1.000	1.000
CNLP-NITS-PP	1.000	1.000	1.000	1.000
IntegrityAI	0.999	0.999	0.999	0.999

Table 6: Official results for Subtask 2 (AWD). The team Hamada Nayel focused their paper on Subtask 1.

In Subtask 1 (NAD), the top-performing systems were all based on transformer architectures. The winning system from **Zain et al.**, a fine-tuned RoBERTa model, achieved a perfect F1-score of 1.000. The **IntegrityAI** team, using ELECTRA with stylometric features, also achieved a near-perfect score of 0.996. Notably, the classical SVM-based system from **Hamada Nayel** secured the third rank with an F1-score of 0.990, outperforming one of the transformer-based systems and demonstrating the viability of simpler models.

In Subtask 2 (AWD), the performance was even higher across the board, with two teams, **Zain et al.** (RoBERTa) and **CNLP-NITS-PP** (DeBERTa + features), achieving perfect scores. The **IntegrityAI** system was just behind with an F1-score of 0.999. The near-perfect results from all teams suggest that detecting AI-generated text in the academic writing domain, at least with the data provided, was a less challenging task compared to the news domain. The structured and formal nature of academic abstracts may provide more distinct signals for classifiers to distinguish between human and machine-generated content. The general trend indicates that while fine-tuned transformers are dominant, augmenting them with stylometric features is a popular and effective strategy.

5 Conclusion

The M-DAIGT shared task aimed to advance the detection of AI-generated text in the critical domains of news and academic writing. The results demonstrate the remarkable effectiveness of current state-of-the-art models, with participating systems achieving near-perfect to perfect scores on both subtasks. The primary findings indicate that fine-tuned transformer models, such as RoBERTa, ELECTRA, and DeBERTa, are highly proficient at this task. Furthermore, the integration of stylometric features proved to be a valuable strategy for several teams, suggesting that a hybrid approach combining deep contextual understanding with traditional linguistic analysis is robust. The strong performance of a classical TF-IDF+SVM model in the news subtask also highlights that resource-efficient methods remain highly competitive. Overall, this shared task provides a valuable benchmark and dataset for the community, confirming the strength of existing methods while also pointing to the nuanced challenges posed by different domains.

Limitations

Despite the success of the shared task, several limitations should be acknowledged. First, the dataset, while diverse in its use of generator models and prompts, represents a static snapshot of LLM capabilities. The rapid evolution of generative models means that detectors trained on this data may not generalize well to text produced by future, more sophisticated LLMs. Second, the task was framed as a binary classification problem (human vs. AI), which does not capture the increasingly common scenario of human-AI collaborative writing, where text is partially generated and then edited by a human. Detecting such mixed-authorship content remains a significant open challenge. Third, the task did not explicitly evaluate the robustness of systems against adversarial attacks, such as paraphrasing or "humanization" techniques designed to evade detection. The exceptionally high scores, particularly in the academic subtask, might also indicate that the detection task within our dataset's parameters was not sufficiently challenging to fully differentiate the capabilities of the top systems. Finally, our study was confined to the English language, and the findings may not be directly applicable to other languages with different linguistic structures. Future iterations of this shared task could address these limitations by incorporating more recent LLMs, including co-authored text, introducing adversarial evaluation tracks, and expanding to multilingual contexts.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros, editors. 2025. *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*. International Conference on Computational Linguistics, Abu Dhabi, UAE.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in Urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428, Vienna, Austria. Association for Computational Linguistics.

Nsrin Ashraf, Mariam Labib, and Hamada Nayel. 2025. Inside the box: A streamlined model for AI-generated news article detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21.

Kyle Bittle and Omar El-Gayar. 2025. Generative ai and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4):296.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Frohmann, Gabriel Meseguer-Brocal, Markus Schedl, and Elena V. Epure. 2025. Double entendre: Robust audio-based AI-generated lyrics detection via multi-view fusion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1914–1926, Vienna, Austria. Association for Computational Linguistics.

Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. 2025. XDAC: XAI-driven detection and attribution of LLM-generated news comments in Korean. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22728–22750, Vienna, Austria. Association for Computational Linguistics.

IntegrityAI. 2025. A multimodal transformer-based approach for cross-domain detection of machine-generated text. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. 2025. M-RangeDetector: Enhancing generalization in machine-generated text detection through multi-range attention masks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8971–8983, Vienna, Austria. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Attila Kovari. 2025. Ethical use of chatgpt in education—best practices to combat ai-induced plagiarism. In *Frontiers in Education*, volume 9, page 1465703. Frontiers Media SA.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics.

Jiatao Li and Xiaojun Wan. 2025. Who writes what: Unveiling the impact of author roles on AI-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26620–26658, Vienna, Austria. Association for Computational Linguistics.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. 2025. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3091–3113, Vienna, Austria. Association for Computational Linguistics.

Zongjie Li, Chaozheng Wang, Shuai Wang, and Cuiyun Gao. 2023. Protecting intellectual property of large language model-based code generation apis via watermarks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2336–2350.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Andrea Pedrotti, Michele Papucci, Cristiano Ciaccio, Alessio Miaschi, Giovanni Puccetti, Felice Dell'Orletta, and Andrea Esuli. 2025. Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3010–3031, Vienna, Austria. Association for Computational Linguistics.

Zubair Qazi, William Shiao, and Evangelos E Papalexakis. 2024. Gpt-generated text detection: Benchmark dataset and tensor-based detection method. In *Companion Proceedings of the ACM Web Conference 2024*, pages 842–846.

Rafael Alberto Rivera Soto, Barry Y. Chen, and Nicholas Andrews. 2025. Mitigating paraphrase attacks on machine-text detection via paraphrase inversion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4421–4433, Vienna, Austria. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge, Xing Yang, and Yanghe Feng. 2023. Textdefense: Adversarial text detection based on word importance entropy. *arXiv preprint arXiv:2302.05892*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22015–22036, Vienna, Austria. Association for Computational Linguistics.

Vasiliki Tassopoulou, George Retsinas, and Petros Maragos. 2021. Enhancing handwritten text recognition with n-gram sequence decomposition and multitask learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10555–10560. IEEE.

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290, Vienna, Austria. Association for Computational Linguistics.

Ashok Urlana, Aditya Saibewar, Bala Mallikarjunarao Garlapati, Charaka Vinayak Kumar, Ajeet Singh, and Srinivasa Rao Chalamala. 2024. TrustAI at SemEval-2024 task 8: A comprehensive analysis of multi-domain machine generated text detection techniques. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 927–934, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Annepaka Yadagiri, L. D. M. S. Sai Teja, Partha Pakray, and Chukhu Chunka. 2025. AI-generated text detection using DeBERTa with auxiliary stylometric features. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Ali Zain, Sareem Farooqui, and Muhammad Rafi. 2025. A multi-strategy approach for AI-generated text detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.