Detection of AI-generated Content in Scientific Abstracts

Ernesto Luis Estevanell-Valladares
University of Havana
University of Alicante
ernesto.estevanell@ua.es
ernesto.estevanell@matcom.uh.cu

Alicia Picazo-Izquierdo
University of Alicante
alicia.picazo@ua.es

Ruslan Mitkov
Lancaster University
University of Alicante
r.mitkov@lancaster.ac.uk

Abstract

The growing use of generative AI in academic writing raises urgent questions about authorship and the integrity of scientific communication. This study addresses the detection of AI-generated scientific abstracts by constructing a temporally anchored dataset of paired abstracts—each with a human-written version that contains scientific abstracts of works published before 2021 and a synthetic version generated using GPT-4.1. We evaluate three approaches to authorship classification: zero-shot large language models (LLMs), fine-tuned encoder-based transformers, and traditional machine learning classifiers. Results show that LLMs perform near chance level, while a LoRA-fine-tuned DistilBERT and a PassiveAggressive classifier achieve near-perfect performance. These findings suggest that shallow lexical or stylistic patterns still differentiate human and AI writing, and that supervised learning is key to capturing these signals.

1 Rationale

The proliferation of generative artificial intelligence (AI) models, particularly large language models (LLMs), has significantly reshaped content creation across domains (Kreps et al., 2022), including scientific writing. While these models offer powerful tools for drafting, summarising, and translating academic texts, their capacity to autonomously generate scientific abstracts raises ethical concerns regarding authorship, originality, and the integrity of scholarly communication. In the context of peer-reviewed publication, the need of distinguishing between human-written and AI-generated content is becoming increasingly pressing. Without reliable detection methods, academic institutions, publishers, and reviewers face the risk of unknowingly legitimising AI-generated content, undermining trust in the scholarly record. As such, there is an urgent need for robust tools capable of accurately identifying AI-generated scientific writing, particularly in the early, high-stakes stages of academic dissemination—namely, paper abstracts.

Several recent approaches have emerged to address this challenge. Tools such as OpenAI's AI Text Classifier and GPTZero have attempted to leverage statistical and linguistic features to differentiate AI from human writing, with varying levels of success. In parallel, research studies have investigated stylometric patterns, perplexity metrics, and discourse-level anomalies as potential indicators of synthetic text. However, most of these efforts suffer from limitations including small or general-domain datasets, lack of temporal anchoring (e.g., comparing texts written before the advent of LLMs), and insufficient validation on high-quality, domain-specific academic corpora. Consequently, there remains substantial room for advancement in this area.

Our study seeks to address the above gaps by constructing a temporally controlled and domain-specific corpus for AI writing detection in scientific abstracts. By compiling a set of abstracts published prior to 2021—before the rise of transformer-based language models—and juxtaposing them with a parallel set of abstracts generated by state-of-the-art LLMs for the same papers, we aim to compare different models to distinguish between human and AI-generated scientific writing. This approach not only ensures a clear temporal boundary between human-authored and synthetic texts but also contributes a novel, curated dataset to the field of natural language processing.

The remainder of this paper is structured as follows: Section 2 reviews related work on AI text detection and scientific authorship analysis. Section 3 details the construction of the human and synthetic abstract corpora. Section 4 outlines our model architecture and experimental setup, as well as the results obtained. Finally, Section 5 discusses

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 21–29, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-003

the implications of our findings and Section 6 includes conclusions as well as directions for future research.

2 Related Work

The growing use of generative artificial intelligence (GAI), particularly large language models (LLMs), in scientific writing has inspired a broad spectrum of academic research. Recent research explores customisation strategies, potential pitfalls, and the promising capabilities of these tools in scholarly contexts. This section covers the use of generative AI in scientific writing and highlights different state-of-the-art methods for detecting AI-generated content.

2.1 AI-generated scientific writing

Emerging research highlights the ways in which commercial AI systems are being adapted for scientific use. Some studies compare multiple AI chatbots to demonstrate performance across academic writing tasks, with GPT-4 scoring highest in quantitative assessments, though all models failed to produce original scientific contributions (Lozić and Štular, 2023). Similarly, Biondi-Zoccai et al. (2025) provide a detailed overview of AI tools tailored for manuscript drafting, refinement, and literature review. While tools like ChatGPT, Grammarly, and SciSpace Copilot are becoming increasingly embedded in academic workflows, the authors caution against their uncritical adoption. In a practical example, Babl and Babl (2023) test ChatGPT's capacity to generate a conference abstract from fictitious data. The output, despite minor hallucination in the references, was structurally sound and content-appropriate, raising concerns over undetectable AI involvement in academic submissions.

A major concern addressed in the literature is the issue of hallucinations—false or fabricated information produced by AI. (Athaluri et al., 2023) (2023) thoroughly examine this phenomenon in scientific writing, warning of its potential to mislead readers and reviewers and to contaminate academic discourse. Another critical analysis comes from Jenko et al. (2024), who evaluate AI-generated literature reviews in musculoskeletal radiology. The study reveals significant factual inaccuracies and shallow content, concluding that current AI tools cannot yet replace expert domain knowledge in scientific synthesis. These risks are echoed in (Biondi-Zoccai et al., 2025), who warn of AI's susceptibility to generating fraudulent datasets and paper mill content. Traditional plagiarism detectors are ineffective against this sophisticated output, calling for robust AI detection mechanisms.

Despite these issues, several sources underscore the potential benefits of AI-assisted writing. (Huang and Tan, 2023) (2023) describe how ChatGPT can improve review article composition by accelerating literature organisation, enhancing linguistic clarity, and assisting non-native English speakers. They argue that AI serves best as a co-authoring assistant—providing structural and linguistic support while the scientist retains control over content and critical interpretation.

2.2 Detection of AI-generated content

As large language models (LLMs) such as GPT-4o and DeepSeek become capable of producing highly coherent and human-like text across multiple domains and languages, researchers have responded by developing diverse strategies and platforms to identify machine-generated content. These approaches generally fall into three categories: traditional machine learning, transformer-based detection models, and zero-shot evaluations using state-of-the-art LLMs themselves.

Early efforts in AI text detection relied heavily on traditional machine learning models using surface-level linguistic features (Alghamdi et al., 2023); (Jawahar et al., 2020). These include metrics such as token diversity, sentence length distributions, part-of-speech frequencies, and syntactic patterns. Classifiers such as Support Vector Machines (SVMs), trained on engineered features extracted from labelled datasets, have demonstrated moderate success. However, with the rise of transformer-based architectures, detection strategies have increasingly moved toward fine-tuned pretrained language models. Fine-tuning models such as BERT and DeBERTa-v3 on domain-specific corpora, often with techniques like Low-Rank Adaptation (LoRA), have shown improved performance (Hans et al., 2024);(He et al., 2021). A third, more recent direction involves evaluating the ability of advanced LLMs to detect AI-generated content in a zero-shot setting (Papageorgiou et al., 2024); (Forment et al., 2025). This strategy leverages the generative model itself—such as GPT-4o-mini—to assess whether a given text appears AI-generated.

Benchmark datasets have played a crucial role in driving these developments. Notable resources include the AuTexTification corpus (used in IberLEF 2023 and 2024), GPT-2 Output Dataset, HC3 and HC3 Plus for chat-based detection (Su et al., 2024), and domain-specific sets like TweepFake (Fagni et al., 2021) and MGTBench (He et al., 2024). These corpora span a range of languages, modalities, and genres—offering fertile ground for cross-domain benchmarking.

Detection tasks have also become the focus of organised evaluation campaigns. Shared tasks such as the IberLEF AuTexTification challenge, the SemEval 2024 Task 8 on authorship verification, and upcoming initiatives at RANLP and COLING have galvanised research efforts by offering competitive benchmarks and standardised test sets. These tasks increasingly emphasise multilingualism and domain diversity, reflecting real-world challenges where generative AI is used in both high-resource and under-resourced linguistic settings.

All in all, current detection platforms rely on a spectrum of techniques, from transparent ML classifiers (Alghamdi et al., 2023) to opaque but powerful deep learning systems (Hashmi et al., 2024); (Mahmud et al., 2024). Despite incremental gains in accuracy, no approach currently guarantees robust, generalisable detection across domains, languages, and use cases. The limitations of zero-shot LLM detection and the rising fluency of AI outputs all point to the need for hybrid approaches and labelled datasets.

3 Dataset

This project focuses on the development of a structured, balanced, and semantically coherent dataset designed to support research on the automatic identification and classification of machine-generated versus human-written scientific abstracts. In order to evaluate this task with high fidelity and domain diversity, we compiled a dataset that not only spans a wide range of scientific disciplines but also ensures that each data point includes two corresponding versions of the same abstract: one written by a human and another generated by a machine.

The entire data pipeline—from initial collection to the final preparation of train and test sets—was carefully engineered to respect the semantic integrity of abstract pairs and the thematic proportionality of the dataset. This section outlines the key stages of that process, namely the dataset compilation via API scraping and metadata filtering, followed by a custom train-test split procedure that guarantees class balance, category proportionality, and the preservation of human-machine abstract pairs.

3.1 Original and generated abstracts

The human-written abstracts were collected leveraging the Semantic Scholar Graph API to retrieve metadata and abstracts for a wide range of scientific papers across multiple disciplines. The query process was domain-driven, using keywords and filters to target articles in areas such as medicine, physics, environmental science, engineering, computer science, chemistry, biology, and materials science.

For each query result, the script extracted several fields of interest, including the paper's title, abstract, year of publication, venue, DOI, unique paper ID, and URL. Additional metadata was collected when available through integrations with the Unpaywall and Crossref APIs, which were used to verify open-access status and ensure the retrievability of the original documents.

To maintain linguistic and disciplinary consistency, the script applied a series of filtering criteria. First, only abstracts written in English were retained, as determined using the langdetect library. A minimum abstract length threshold was enforced to guarantee sufficient content for accurate language detection. Second, the script discarded non-research content, such as editorials or metadata-only entries, and prioritised papers for which a PDF was accessible or openly licensed. A human curation and review process was also implemented to verify abstract consistency and validity.

Once cleaned and filtered, each abstract was stored along with its associated metadata in a structured format. These abstracts constitute the human-authored portion of the final dataset.

The machine-generated abstracts were produced using the model GPT-4.1. For each scientific article retrieved in the previous stage, the first 10 pages of the full-text document were used as input to the model. These pages were either extracted from the available PDFs or obtained through additional metadata queries and processing pipelines that reconstructed the document's main body content.

The GPT-4.1 model was prompted to generate an abstract that closely followed the conventions of scientific abstract writing: summarising the re-

Domain	Human	Machine
biology	13,057	12,512
business	7,047	6,763
chemistry	11,770	10,838
computer science	9,163	8,361
economics	6,410	6,943
education	14,503	10,028
engineering	12,044	11,140
environmental science	18,313	13,229
materials science	8,574	7,740
medicine	20,063	17,782
physics	28,910	26,160
sociology	7,413	6,249

Table 1: Total word count per category for human- and machine-written abstracts.

search problem, methodology, and key findings in a concise and coherent format. No abstract was generated unless a minimum threshold of source content was available (i.e., a full 10-page span or an equivalent amount of text). This ensured that the machine-generated abstract had sufficient context and detail to mirror the function and structure of the original human-written abstract.

All generated abstracts were paired with their corresponding human-written versions using the paper's title as a unique ID, and both versions shared the same category and metadata. This pairing process resulted in a clean and balanced dataset where each title appears exactly twice—once under the label human and once under the label machine.

The total word count analysis reveals consistent patterns across categories, with human-written abstracts generally containing slightly more words than their machine-generated counterparts. This trend is observed in nearly all disciplines, most notably in fields like medicine, physics, and environmental science, which show the highest overall word volumes. The discrepancy in length may reflect differences in content density, verbosity, or summarisation strategies between human authors and the language model.

3.2 Split with pair integrity

Once the full dataset of human–machine abstract pairs had been compiled and validated, the next step was to divide it into a training set and a test set, in a way that would enable reliable supervised learning and fair evaluation. This division was carried out with particular attention to three key

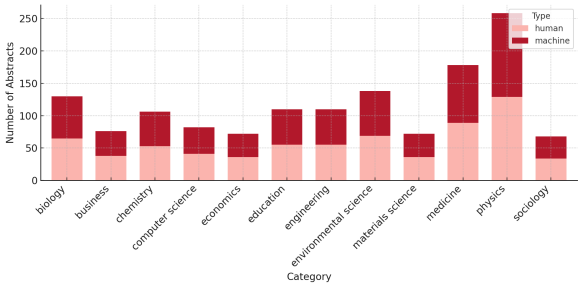

Figure 1: Train split

requirements: semantic pairing integrity, class balance, and thematic proportionality across scientific categories.

The core structural unit of the dataset is the abstract pair, consisting of one human-written and one machine-generated version of the same scientific paper. In order to prevent data leakage and preserve the semantic boundary between training and test samples, it was essential that these pairs remain intact during the split. That is, both the human and machine versions of a given abstract had to be assigned to the same subset—either training or test. Splitting the two across subsets would have introduced significant risk of semantic overlap, as both versions are derived from the same source paper and often convey similar core content.

To enforce this constraint, the split was performed at the level of the paper title, which uniquely identifies each pair. Only titles that appeared exactly twice in the dataset—once with each version—were eligible for inclusion. The total pool of such valid pairs was then randomly divided into training and test sets using an 80/20 stratified split, with stratification based on the category assigned to each paper. This ensured that the topical distribution of abstracts across disciplines (e.g., physics, medicine, computer science) remained proportionally balanced in both subsets.

After assigning titles to either the training or test set, all associated abstracts and metadata were recovered using the title as the join key. This approach guaranteed that the final training and test sets were (i) fully balanced in terms of class labels (human and machine); (ii) proportionally distributed across scientific categories (iii) free from any leakage or overlap of semantically equivalent texts.

Following the train–test split, a final validation step was performed to ensure the integrity of the abstract pair structure within each subset. This

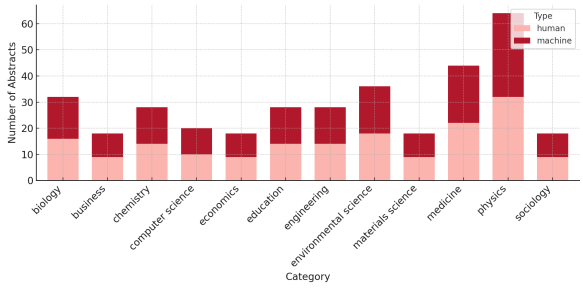

Figure 2: Test split

involved verifying that each paper ID appeared exactly once per version (i.e., once for the human-written abstract and once for the machine-generated one), and that both instances were assigned to the same subset.

This was achieved by counting the frequency of each ID in the training and test sets independently. The results of this verification confirmed that all pairs were preserved and correctly assigned, with no instances of cross-subset leakage or structural inconsistencies.

4 Experiments

This study investigates the capacity of computational models to classify scientific abstracts according to their authorship—human or machine-generated. The classification task was designed as a binary decision problem and explored through three complementary modelling approaches: (1) prompt-based classification using large language models (LLMs), (2) supervised fine-tuning of a transformer-based classifier with parameter-efficient adaptation, and (3) traditional machine learning pipelines based on bag-of-words representations.

4.1 Experimental setup

In this section we aim to explore which model and configuration performs best when classifying human vs. machine generated text. To this end, different setups have been explored and are detailed below.

Prompt-based LLM classification In the first setup, a suite of instruction-tuned large language models (LLMs) was used to perform zero-shot classification. Each model was prompted with a research abstract and asked to determine whether it had been written by a human or generated by a machine. A fixed prompt template was used for all models to ensure consistency and comparability across predictions.

- System prompt: You are a diligent assistant that labels research abstracts. Reply strictly with either 'human' or 'machine' and nothing else.
- User prompt: Classify the following abstract as written by a human or by a machine. Answer with only 'human' or 'machine'. Abstract: * Classification: *

No few-shot examples were provided, and no additional formatting was required from the model output beyond the binary label. Different models with different parameter configuration and size were used:

- OpenAI: GPT-4.1, o4-mini, GPT-4o-mini
- LLaMa 4: llama4-scout-instruct-basic, lama4-maverick-instruct-basic
- Qwen3: qwen3-30b-a3b, qwen3-235b-a22b
- DeepSeek: deepseek-r1-basic

Fine-tuned transformer with LoRA To complement zero-shot inference with supervised learning, we employed the AutoGOAL AutoML framework (Estevez-Velarde et al., 2020) to automatically explore and optimise deep learning pipelines based on transformer architectures. AutoGOAL was extended to include 44 pipeline variants across 13 transformer-based language models (introduced by Estevanell-Valladares et al., 2024), sourced from the Hugging Face model hub (Jain, 2022). These models included various fine-tuning strategies: full fine-tuning, partial fine-tuning (top-layer adaptation), and Low-Rank Adaptation (LoRA).

Training and evaluation were performed on a workstation equipped with an NVIDIA RTX 4090 GPU, allowing efficient gradient-based learning across configurations. Each pipeline was evaluated using 2-fold stratified cross-validation on the training set. The best-performing pipeline selected by AutoGOAL used LoRA fine-tuning over a DistilBERT base model.

Traditional machine learning baseline To establish a non-neural baseline, we also constructed and tuned a traditional machine learning pipeline built on sparse vector representations. The pipeline

consisted of a HashingVectorizer for text featurisation and a PassiveAggressiveClassifier for classification.

The HashingVectorizer was configured to use over two million features, binary encoding, and L1 normalisation, transforming text into a fixed-length sparse binary representation. The classifier was optimised with a high aggressiveness parameter (C=9.991) and evaluated using stratified validation on the training set.

4.2 Results

The performance of the large language models (LLMs) on the binary classification task—determining whether a scientific abstract was written by a human or generated by an AI—revealed a consistent trend: despite strong general-purpose capabilities, the models exhibited difficulty distinguishing between the two classes in a reliable manner.

Across all LLMs evaluated, F1 scores remained low, rarely exceeding 0.34. The best-performing model, Qwen3-235B, achieved an F1 score of 0.335, followed closely by GPT-4.1 and DeepSeek-R1, with scores of 0.333 and 0.332 respectively. Accuracy scores hovered near 49–50% for most models, suggesting that predictions were often close to chance level in aggregate, despite marginal gains in class-specific precision or recall. The confusion matrix in Figure 3 suggests that Qwen3-235B, which is the best LLM, almost always mistakes every machine-generated abstract for human writing.

This performance gap highlights a critical limitation of general-purpose LLMs when applied to subtle authorship attribution tasks involving highly similar content, such as pairs of human- and machine-written scientific abstracts derived from the same paper. The task appears to require more fine-grained discriminative capabilities than current zero-shot prompting strategies afford.

In contrast, the best-performing model emerged from a supervised approach using LoRA fine-tuning on top of the distilbert-base-multilingual-cased encoder. This configuration, discovered through AutoGOAL's AutoML pipeline search, achieved a markedly superior F1 score of 0.974, with equivalent levels of accuracy, precision, and recall. These results underscore the value of task-specific training, particularly when using parameter-efficient fine-tuning techniques like LoRA.

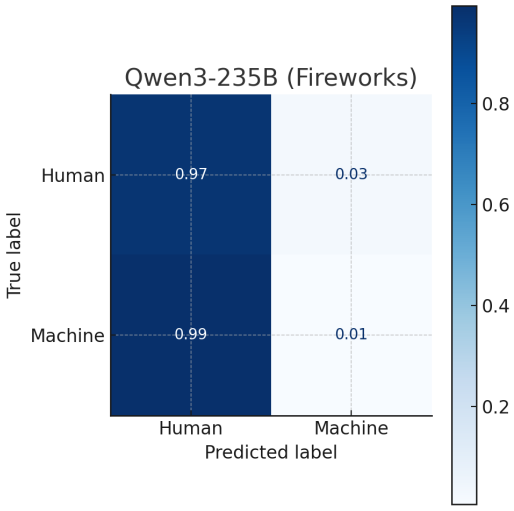

Figure 3: Confussion matrix for the best-performing LLM

Model	Acc	P	R	F1
LoRA DistilBERT	**0.974**	**0.974**	**0.974**	**0.974**
PassiveAggressive	0.972	0.972	0.972	0.972
Qwen3-235B	0.490	0.357	0.490	0.335
GPT-4.1	0.487	0.328	0.487	0.333
DeepSeek-R1	0.494	0.280	0.493	0.332

Table 2: Accuracy, precision, recall, and F1 score of the best-performing models across the classification approaches.

The fine-tuned encoder demonstrated consistent and robust performance across all metrics, correctly classifying nearly all abstracts in the test set. This outcome confirms that the classification signal—though subtle—can be captured by a discriminative model when exposed to labelled examples during training.

The traditional ML pipeline, consisting of a HashingVectorizer and a PassiveAggressiveClassifier, also performed strongly. With an F1 score of 0.972, it rivaled the fine-tuned transformer despite relying solely on sparse feature representations and linear decision boundaries. This result highlights that surface-level textual features may encode sufficient information to distinguish between human and machine authorship in abstracts, possibly due to differences in vocabulary frequency, sentence structure, or lexical density.

5 Discussion

The results of our experiments reveal a notable pattern in the performance of the classification

models: large language models (LLMs), including cutting-edge systems such as GPT-4.1 and Qwen3-235B, consistently performed near chance level in distinguishing between human- and machine-written scientific abstracts. In contrast, both the fine-tuned transformer model and the traditional classifier achieved near-perfect performance, with F1 scores of 0.974 and 0.972 respectively.

This sharp discrepancy raises several important questions about the nature of the detection task and the limitations of zero-shot LLM inference. The underwhelming results of the LLMs may stem from the zero-shot setup used in the experiments. Although LLMs have demonstrated broad competence in a range of generative and reasoning tasks, their performance in subtle classification settings—particularly without task-specific training—is often limited. In our case, the classification task relies on capturing fine-grained, often imperceptible linguistic differences between two texts that are topically identical and structurally similar. These nuances may not be readily detectable without additional context or calibration.

Another contributing factor is the in-domain similarity of the texts. Since both human- and machine-generated abstracts summarise the same research paper, they often share terminology, structure, and even phrasing. This results in minimal surface-level variation—precisely the kind of variation that LLMs may overlook in the absence of tailored prompting or fine-tuning.

Furthermore, LLMs are inherently generative, not discriminative. When repurposed for binary classification in a zero-shot setting, they rely heavily on probabilistic reasoning and internal priors, which may not be accurate for a highly specific detection task such as this. Their inability to identify stylistic markers of synthetic writing without explicit examples severely limits their utility in authorship verification.

The success of both the traditional PassiveAggressive classifier and the LoRA-fine-tuned DistilBERT suggests that authorship signals do exist in the data, but they are subtle and best captured by models with explicit supervision. The dataset shows a consistently higher word count in the generated versions by domain, which may have been a clear indicator for these models. There may be some lexical patterns such as "This paper/study/review presents/examines/provides..."

The PassiveAggressive classifier, leveraging a simple bag-of-words approach, likely benefits from capturing statistical regularities in vocabulary use, lexical density, or syntactic patterns that differ—perhaps subtly but consistently—between human and machine writers. These cues might include phrase redundancy, sentence-initial tokens, or unnatural repetition that are hard to detect perceptually but easily exploited by statistical models.

The DistilBERT model, fine-tuned via LoRA, excels likely because it is explicitly trained on the classification objective, allowing it to learn nuanced distinctions over multiple layers of abstraction. The results highlight the value of supervised discriminative learning even in tasks where the classes appear nearly indistinguishable to a human reader or an unadapted LLM.

These findings carry significant implications:

- The detection of AI authorship may not require deep semantic modelling, but rather benefits from the exploitation of shallow stylistic inconsistencies. This opens opportunities for lightweight, interpretable, and resource-efficient detection systems.
- Future detection strategies should consider ensemble approaches, combining the broad generalisation of LLMs with the precision of discriminative classifiers.

6 Conclusions and Future Work

This study investigated the detection of AI-generated content in scientific abstracts by evaluating a range of modelling strategies, including zero-shot prompting of large language models (LLMs), fine-tuned transformer encoders, and traditional machine learning classifiers. Surprisingly, the most advanced LLMs—including GPT-4.1 and Qwen3-235B—performed at near-chance levels in the binary classification task. In contrast, a lightweight encoder-based model fine-tuned with Low-Rank Adaptation (LoRA) and a traditional PassiveAggressive classifier achieved near-perfect classification accuracy.

These findings suggest that while LLMs excel at text generation and general reasoning, they are not well-suited for fine-grained authorship attribution in a zero-shot setting, especially when the candidate texts share substantial semantic overlap. On the other hand, task-specific supervised approaches—both neural and statistical—are capable of capturing subtle linguistic cues that differentiate human- and machine-generated writing.

Several limitations should be noted: (i) LLMs were tested exclusively in zero-shot mode, without prompt tuning, few-shot examples, or in-context learning strategies; (ii) all synthetic abstracts were produced by GPT-4.1, which may limit generalizability; (iii) the study focused exclusively on English-language abstracts.

Building on this limitations, several promising directions can be pursued:

- Explainable detection: Integrating explainability tools (e.g., SHAP, attention visualisation) into detection pipelines could reveal which linguistic features signal machine authorship and support trust in automated tools.
- Multilingual detection: Expanding the dataset and experiments to include other languages would allow evaluation of AI authorship detection across diverse linguistic and cultural contexts.
- Human-in-the-loop verification: Combining automated detection with expert judgment could yield hybrid frameworks that balance efficiency and reliability in academic publishing workflows.
- Comparison with abstracts from scientific papers published after gen-AI open-source tools, with the purpose of inferring whether automatic writing is being used in scientific writing.

Acknowledgements

This research has been partially funded by the University of Alicante, University of Havana, the Spanish Ministry of Science and Innovation, and the Generalitat Valenciana, through the "The limits and future of data-driven approaches: A comparative study of deep learning, knowledge-based and rule-based models and methods in Natural Language Processing" (CIDEXG/2023/12) project.

References

Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. 2023. A comprehensive survey on machine learning approaches for fake news detection.

Sai Anirudh Athaluri, Sandeep Varma Manthena, V S R Krishna Manoj Kesapragada, Vineel Yarlagadda, and Tirth Dave. 2023. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4).

Franz E. Babl and Maximilian P. Babl. 2023. Generative artificial intelligence: Can chatgpt write a quality abstract? *Emergency Medicine Australasia*, 35(5):809–811.

Giuseppe Biondi-Zoccai, Anna Cazzaro, Elisa Cobalchin, Diletta D'Auria, Giovanni Ardizzone, Salvatore Giordano, Ulvi Mirzoyev, Petar M. Seferovic, Gani Bajraktari, and Denisa Muraru. 2025. Artificial intelligence tools for scientific writing: The good, the bad and the ugly. *Top Italian Scientists Journal*, 2(1).

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Marc Forment, Juanan Pereira, Francisco García-Peñalvo, Maria Casañ, and Jose Cabré. 2025. Lamb: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards Interfaces*, 92:103940.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text.

Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Yamin, Subhan Ali, and Mohamed Abomhara. 2024. Advancing fake news detection: Hybrid deep learning with fasttext and explainable ai. *IEEE Access*, 12:44462 – 44480.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection.

Jingshan Huang and Ming Tan. 2023. The role of chatgpt in scientific communication: writing better scientific review articles. *American Journal of Cancer Research*, 13(4):1148–1154. Epub April 15, Published April 30.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nathan Jenko, Sisith Ariyaratne, Lee M. Jeys, Scott Evans, Krishna P. Iyengar, and Rajesh Botchu. 2024. An evaluation of ai generated literature reviews in musculoskeletal radiology. *The Surgeon*, 22(3):194–197.

Sarah Kreps, Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9:104–117.

Edisa Lozić and Benjamin Štular. 2023. Fluent but not factual: A comparative analysis of chatgpt and other ai chatbots' proficiency and originality in scientific writing for humanities. *Future Internet*, 15(10):336.

Tanjim Mahmud, Imran Hasan, Mohammad Tarek Aziz, Taohidur Rahman, Mohammad Shahadat Hossain, and Karl Andersson. 2024. Enhanced fake news detection through the fusion of deep learning and repeat vector representations. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 654–660.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8).

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2024. Hc3 plus: A semantic-invariant human chatgpt comparison corpus.