A Multi-Metric Analysis Between Human and AI Writing

Dinh-Thien-Loc Nguyen[1,2], Thien-Lac Quach[1,2][0009−0004−6310−5149], Hong-Tan-Tai Nguyen[1,2], Ton-Minh-Ky Tran[1,2][0009−0006−5027−2516], and Viet-Tham Huynh[1,2][0000−0002−8537−1331]

[1] Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{24125093,24125092,24125078,24125102}@apcs.fitus.edu.vn
hvlab@selab.hcmus.edu.vn

Abstract. The rapid advancement of Large Language Model (LLMs) has rendered the differentiation between human-written and AI-generated text increasingly challenging. In this paper, we propose a zero-shot, black-box, multi-metric analysis approach to distinguish these two text types by their linguistic features. We introduce six core metrics—lexical diversity, nominalization density, modal and epistemic rate, clause complexity, passive voice ratio, and sentence-to-sentence cosine similarity—to capture the nuanced differences in writing styles. Utilizing a comprehensive dataset of text samples from both human authors and various LLMs, we designed a pipeline to compute these metrics and analyze their distributions. Our findings reveal distinct linguistic fingerprints that can effectively differentiate human writing from AI-generated content Notably, human texts often exhibit significantly higher lexical diversity and passive voice frequency, while AI-generated texts prefer nominalizations and display higher semantic coherence. Moreover, humans tend to leverage specific, concrete vocabulary, whereas AI outputs are characterized by abstract and generalized terms. From these insights, we discuss the potential applications of our findings in developing a linguistic framework leveraging the strengths of both human and AI writing styles, particularly in educational contexts to assist English as a Foreign Language (EFL) learners. Our study lays the groundwork for future research in both detection methodologies and language learning applications.

Keywords: AI text detection · Multi-metric analysis · Linguistic features

1 Introduction

1.1 Problem statement

Differentiating human-written and artificial-intelligence-generated (AI-generated) text has been an increasingly important area of research due to concerns regarding the latter's societal and ethical impact. These include, but are not limited to,

the abuse of Large Language Models (LLMs) to disseminate misleading information [8], biases caused by the misrepresentation of marginalized demographics [12], or model collapse resulting from using model-generated data as training input [13]. Nonetheless, in the context of language learning, AI-generated text can be used to refine and further improve L2 students' writing [23]. By combining the best features from both domains of text, it is possible to create a "scaffolding" from which L2 learners' writing can be enhanced. Our research aims to differentiate both domains of text and highlight features from each in order to craft such a "scaffold" to assist learners who are new to English in their studies of the language.

However, commercial-tool- and human-based detection remains inadequate, often performing only marginally above chance, especially when faced against output of newer models [2, 14, 11]. On the other hand, more sophisticated, automated, detection methods have been researched and tested with notable results [22, 5, 1, 16]. Wu et al. (2024)[18] provides a comprehensive overview of these methods by categorizing them into three main types: watermarking, statistics-based, and neural-based; in each category, methods are further according to whether they have access to the models' internal workings (white-box methods) or not (black-box methods).

Among these methods, statistics-based, black-box ones are of interest to us given their high effectiveness [16, 22, 18], relative to their replicability, compared to the remaining categories. By processing the corpora given that can either be generated by human or AI, our goal is to develop metrics that consist of a structured feature set extracted from each piece of text, including quantitative linguistic indicators such as lexical diversity scores and nominalization counts. We also offer a comparative analysis that summarizes how these linguistic metrics differ between the human and AI-generated corpora, laying the groundwork for creating hybrid methods that leverage both human and AI strengths in English writing performance.

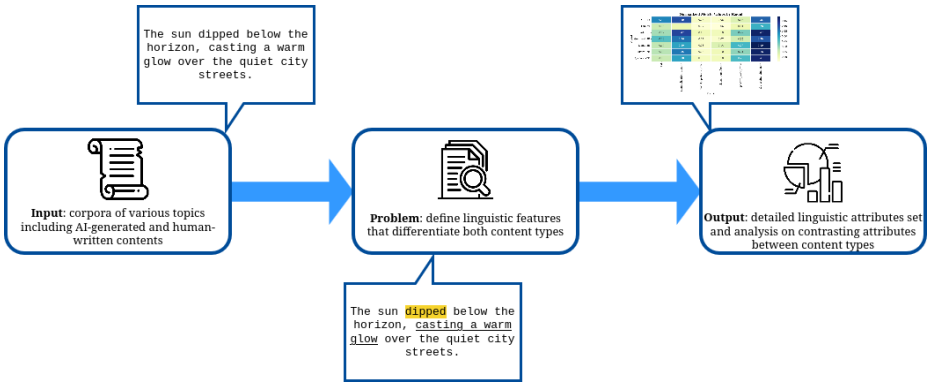


Fig. 1: The input and output of the problem

Applications and Challenges. By addressing the problem statement, we reinforce the overarching goal of differentiating between human-generated and AI-generated texts. It stands to reason that we encounter a few challenges in achieving these goals:

- Some linguistic features may not be consistent across various topics (e.g. academic writing tends to utilize more nominalization than casual writing).
- The proficiency and knowledge of writers may differ widely and thus cannot represent human writing as a whole.

Research scope Our scope includes texts that are fully AI-generated or fully human-written. To reduce sampling bias, we compare linguistic features only across texts matched in genre, prompt, and length.

1.2 Contributions

Our contributions lie in three folds:

- We develop a unified pipeline that integrates multiple linguistic metrics, extending prior studies that relied on limited or domain-specific indicators.
- We conduct a cross-genre comparison of human and AI-generated writing, revealing consistent linguistic differences across multi-domain texts.
- We provide empirical evidence and a feature-based foundation for future hybrid analytical methods that can combine human and AI strengths in writing assessment and support educational applications.

2 Related Work

Statistics-based analysis has been defined as one of the main approaches to detecting LLM-generated texts [17], with numerous studies producing highly accurate and consistent results [18, 7, 3]. These methods rely on analyzing statistical patterns in the output text that inspire our approach in the paper, such as logits, n-gram, or similarity scores, to highlight unique patterns in LLM-generated text. Among these, the black-box approaches, as they do not require access to the internal workings of LLM models, are of interest due to their simplicity and effectiveness, as well as their ability to generalize across different LLM architectures.

Recent literature has explored a variety of black-box methods to differentiate between human and LLM-generated content. A straightforward zero-shot black-box statistical analysis that compares the amount of grammatical corrections made by LLMs on a given text to determine its origin was proposed [18]. Another study used the logs-odd ratio to conclude the vocabulary range and collocation usage patterns of LLMs and humans [22]. Other works have focused on n-gram probability divergence similarity, text regeneration, and sentence-based cosine similarity [19, 21, 9].

Differences in quality between AI-generated and human-written texts have also been extensively discussed [15, 4]. Further studies have examined linguistic differences between human-written and AI-generated text using multi-metric analysis analyzed low-level textual features (e.g. sentence count, average word length) and high-level linguistic features (e.g. lexical and syntactic complexity) respectively [20, 6].

Zhu et al. (2025)[23] investigated the efficacy of an AI-powered corpus platform in improving EFL argumentative writing. Their study, which employed a quasi-experimental design, demonstrated that integrating Generative AI with authentic corpus data led to statistically significant improvements in learners' overall writing performance over time. Specifically, the research highlighted substantial gains in language-related dimensions, such as "Language Use" and "Grammar," as the tool enabled students to identify and assimilate advanced n-grams and rhetorical patterns. By characterizing the specific qualitative shifts—such as enhanced lexico-grammatical accuracy and structural coherence—resulting from AI scaffolding , this work provides a critical baseline for defining the linguistic features of "AI-assisted" writing, distinct from both unassisted human output and fully AI-generated text.

3 Approach

This research employs a zero-shot, black-box method to distinguish between human and AI-generated text. Unlike methods that require access to internal model weights or extensive supervised training, our approach utilizes a multi-metric statistical analysis of the output text itself. By analyzing specific linguistic patterns, such as lexical diversity, syntactic depth, and semantic consistency, we establish a robust detection baseline that operates without direct insight into the generative model's architecture. This zero-shot framework leverages six core linguistic metrics to fingerprint the statistical divergences between human writing and the output of Large Language Models (LLMs).

3.1 Dataset and Data Engineering

The research utilizes the dataset described in "Comprehensive Dataset for Human vs. AI Generated Text Detection"[10], which provides a massive scale of over 58,000 text samples. The dataset pairs high-fidelity human writing from *The New York Times* newspaper against outputs from six distinct Large Language Models: Gemma-2-9b, Mistral-7B, Qwen-2-72B, LLaMA-8B, Yi-Large, and GPT-4o. To ensure the statistical markers hold true across varied contexts, we categorize the data into major categories (e.g., Technology, Politics).

3.2 3.2. Key Metrics

We engineered six specific metrics to serve as the core of our zero-shot detection capability. These metrics quantify the linguistic differences between human and AI text.

1. **Measure of Textual Lexical Diversity (MTLD):** This metric assesses vocabulary richness based on the observation that humans employ varied, context-dependent vocabulary, whereas LLMs often default to repetitive, high-probability tokens.

$$\text{Score} = \frac{\text{Total Words}}{\text{Factors}}$$

2. **Nominalization Density:** Nominalization (converting verbs/adjectives into nouns) is a hallmark of formal human academic writing. Our algorithm improves upon simple suffix counting by verifying that the lemma differs from the surface form to avoid false positives.

$$D_{nom} = \frac{\text{Count}(\text{suffix} \in \{\text{-tion, -ment, -ness, -ity}\}) \times 1000}{\text{Total Words}}$$

3. **Modal & Epistemic Rate:** This metric targets "hedging" language, as humans frequently use epistemic markers (e.g., "might", "however") to convey uncertainty or nuance, unlike the declarative style of AI. Target tokens include *might, may, could, perhaps, possible, unlikely, however, although,* and *but*.

$$R_{modal} = \frac{\text{Count}(\text{Target Tokens}) \times 100}{\text{Total Words}}$$

4. **Clause Complexity:** To capture human "burstiness," we analyze syntactic depth. Using dependency parsing, we compute the maximum depth from the root verb to the furthest leaf node in the syntactic tree, differentiating deep human structures from flatter AI outputs.
5. **Passive Voice Ratio:** Stylistic preferences vary by genre; specifically, scientific human writing often leverages the passive voice, while AI models generally default to the active voice. The system flags sentences containing tokens tagged as `nsubjpass` (nominal subject passive) combined with `auxpass`.

$$\text{Ratio} = \frac{\text{Count}(\text{Passive Sentences})}{\text{Total Sentences}}$$

6. **Sentence-to-Sentence (S2S) Cosine Similarity:** AI models prioritize coherence, which often leads to high semantic overlap between sequential sentences, whereas humans frequently introduce new ideas that lower adjacent similarity. We employ Sentence-Transformers to measure this semantic flow.

$$S_{sim} = \cos(\boldsymbol{v}_n, \boldsymbol{v}_{n+1}) = \frac{\boldsymbol{v}_n \cdot \boldsymbol{v}_{n+1}}{||\boldsymbol{v}_n||\,||\boldsymbol{v}_{n+1}||}$$

3.3 Lexical Analysis

To capture lexical signatures that may not be reflected in syntactic metrics, we integrated a Python reimplementation of the IRAL Log–Odds analysis[22] that identifies tokens disproportionately associated with AI or human text. The implemented statistic uses additive smoothing:

$$\text{LogOdds}(w) = \ln\left(\frac{\text{Freq}(w)_{\text{AI}} + 0.5}{\text{Freq}(w)_{\text{Human}} + 0.5}\right),$$

and we compute an approximate standard error to obtain a z-score for each token. Tokens with $|z| > 1.96$ are reported as statistically significant (two-tailed test, $\alpha \approx 0.05$). Positive log-odds indicate AI association; negative values indicate human association.

- **Implementation:** The analysis is implemented in `src/iral/iral_lexical.py` and batched by `src/iral/iral_orchestrator.py`. Plots and visual summaries are produced by `src/iral/iral_plots.py`.
- **Preprocessing:** Tokenization and lemmatization use the shared `spaCy` pipeline; minimal normalization (lowercasing, punctuation stripping, lemma alignment) is applied so results reflect lexical choice rather than formatting artifacts.
- **Smoothing & cutoffs:** We use additive 0.5 smoothing to avoid zero counts; results are accompanied by frequency cutoffs and recommended minimum per-class counts to reduce spurious signals from extremely rare tokens.
- **Outputs:** The pipeline emits a scored CSV per model with columns: `token, lemma, freq_ai, freq_human, logodds, z, p`. It also emits ranked "giveaway" lists (tokens with $|z| > 1.96$) and log-odds bar charts.
- **Interpretation:** High absolute log-odds highlight candidate lexical signatures but must be qualitatively reviewed (domain terms vs. model idiosyncrasies). We report both effect size (log-odds) and statistical significance (z) to prioritize robust signals.

Reproducibility & robustness: The analysis is deterministic given cached parsed inputs (Parquet) produced by `parse_and_cache.py`; all seeds, frequency cutoffs, and smoothing constants are recorded in `metrics_config.yaml` to enable exact reruns. Recommended robustness checks include applying minimum corpus frequency filters and inspecting collocations to separate topical from model-specific effects.

Example (paper): Reported token lists show model-specific high log-odds (AI) and editorial/genre tokens with negative log-odds (Human), supporting claims about lexical divergence that complement syntactic and semantic metrics.

4 Experiments

4.1 Datasets

- **Dataset Name and Source:** The study utilizes the "Comprehensive Dataset for Human vs. AI Generated Text Detection," curated and introduced by Roy et al. (2025)[10] in the context of the Defactify 4 workshop.
- **Key Statistics:** The dataset comprises over 58,000 text samples distributed across 16 distinct topic shards, including major categories like Technology (1,459 samples) and Politics (1,187 samples). The data is divided into two

primary classes: Human (sourced from *The New York Times*) and AI, with the latter further subdivided into outputs from six distinct Large Language Models: Gemma-2-9b, Mistral-7B, Qwen-2-72B, LLaMA-8B, Yi-Large, and GPT-4o.

- **Preprocessing and Cleaning:** To address noise inherent in scraped data, a custom sanitization module (`src.ingest`) was employed. This included regex filtration to remove PDF artifacts (e.g., page headers, citations), structural cleaning to strip Markdown symbols, and strict column alignment to ensure valid paired statistical testing between human stories and model outputs. Additionally, rows with insufficient text length or clearly non-narrative content were automatically rejected.
- **Justification:** This dataset is uniquely appropriate for evaluating the proposed linguistic metrics because it provides a "human baseline" of high-quality, professionally edited journalism rather than unverified web text. The diversity of the six AI models ensures the model-independency of the method through being tested against a wide range of generation architectures, while the rigorous topic sharding validates that the proposed linguistic features remain robust across diverse semantic contexts.

4.2 Evaluation Protocol

- **Evaluation Metrics:** Since this study focuses on feature engineering and validation rather than binary classification accuracy, the primary metrics for success are statistical significance and effect size. We utilize Welch's t-test to calculate p-values for the divergence between human and AI distributions, and Cohen's d to quantify the effect size. An effect size of $d > 0.8$ is established as the threshold for a "Large Effect," indicating a metric's strong predictive capability.
- **Data Splitting Strategy:** The dataset is not split into traditional train/test sets for model tuning but is instead stratified into 16 distinct topic shards (e.g., Technology, Politics, Environment). Statistical validation is performed within each shard to ensure the metrics hold true across diverse semantic contexts, with results aggregated to verify robustness against topic-specific bias.
- **Baselines:** The "human baseline" is established using high-fidelity journalism from *The New York Times*. The methodological baseline for comparison is the word-count-based approach proposed by Herbold et al. (2023)[6]. Our method is evaluated on its ability to capture semantic and syntactic nuances (e.g., "burstiness," passive voice) that simple lexical counting fails to detect.
- **Task-Specific Conventions:** To ensure statistical validity, evaluation follows a strict paired testing protocol. Each AI-generated text sample is paired with its corresponding human-written source on the same prompt/topic. Any unaligned rows or samples with formatting artifacts that could skew complexity scores are excluded prior to calculation.

4.3 Experimental Settings

- **Architecture Settings and Hyperparameters:** The analysis is orchestrated via a modular Python pipeline. Key hyperparameters for the linguistic metrics include a Type-Token Ratio (TTR) threshold of 0.72 for the MTLD calculation and a Z-score threshold of > 1.96 for the Log-Odds Ratio analysis to identify significant lexical signatures.
- **Algorithms and Computation:** Linguistic parsing relies on the spaCy dependency parser for syntactic tree depth and passive voice detection. Semantic analysis utilizes the Sentence-Transformers library to generate embeddings for the Sentence-to-Sentence (S2S) Cosine Similarity metric.
- **Hardware and Software Environment:** The pipeline is built on a Python 3 tech stack. Data management is handled via DuckDB and cached in Parquet format to allow efficient sharded processing of the 58,000+ samples. Visualization and automated reporting are generated using matplotlib.
- **Data Preprocessing Pipeline:** Raw data ingestion is managed by a custom 'ingest' module that performs rigorous sanitization. This includes regex-based filtration to remove PDF scraping artifacts (e.g., "Page X", "More about..."), stripping of Markdown formatting (e.g., bolding, headers) to isolate pure prose, and automated rejection of rows containing non-narrative content such as copyright notices. We also manually verified cleanliness of the data by selecting only entries with lengthy, representative human text and discarding those that are of insufficient length (after processing) or are irrelevant to the topic. We manually selected the 25 texts we deemed the highest quality in each topic, then removed artifacts that the pipeline missed, including separated words joined together without spacing, placeholder text for hyperlinks, or extra information about blogs or authors.

4.4 Results and Discussion

Quantitative Analysis of Linguistic Features The quantitative analysis reveals distinct "linguistic fingerprints" that differentiate human writing from AI-generated text across the six proposed metrics. Figure 2 presents the normalized metric values for the Health shard, serving as a representative sample of model performance.

The heatmap highlights a fundamental divergence in lexical diversity. Human writing achieves an MTLD score of 0.61, significantly outperforming the most sophisticated model, GPT-4o (0.41), and smaller models like Qwen-2-72B (0.26). This confirms that human authors utilize a far richer and more context-dependent vocabulary than LLMs, which tend to converge on high-probability tokens. Conversely, nominalization density—a marker of academic formality—is markedly higher in GPT-4o (0.45) compared to humans (0.20), suggesting that while models can mimic formal tone, they often over-index on complex noun phrases to achieve it.

To visualize the holistic "shape" of these linguistic profiles, Figure 3 maps the six metrics onto a radar chart.

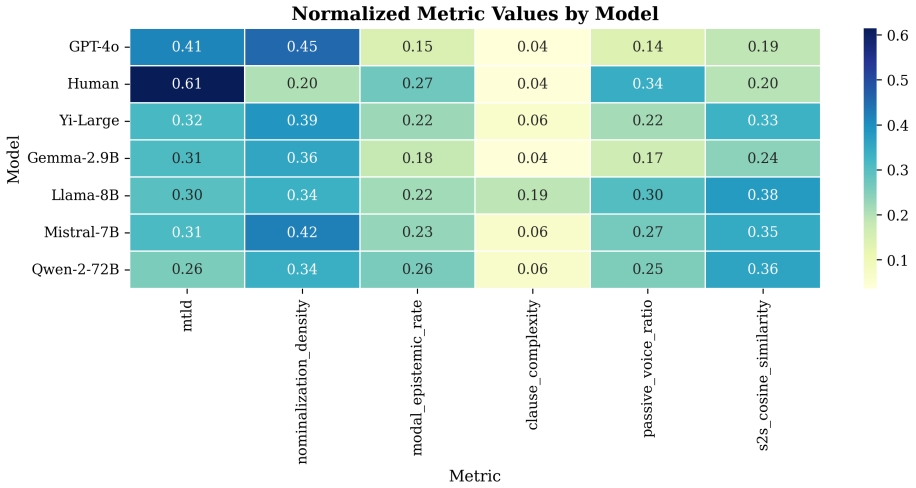


Fig. 2: Normalized Metric Values by Model (Topic: Health). Human writing exhibits significantly higher lexical diversity (MTLD) compared to all AI models, while AI models consistently show higher semantic similarity.

The radar chart demonstrates that no single model perfectly overlaps with the human baseline. The human profile (blue line) is distinctively spiked towards MTLD and clause complexity, whereas AI models cluster towards the bottom-right quadrant, favoring high Nominalization Density and S2S Cosine Similarity. Llama-8B (purple) stands out as an outlier with an extreme skew towards S2S Similarity, indicating highly repetitive and "safe" generation patterns.

Semantic Coherence and Burstiness A key finding of this study is the difference in semantic flow, quantified by the Sentence-to-Sentence (S2S) Cosine Similarity metric. As shown in Figure 4, human writing exhibits a lower mean similarity with high variance, reflecting the human tendency to make "semantic jumps"—introducing new ideas or changing topics abruptly (burstiness).

In contrast, models like Llama-8B show a tightly peaked distribution at a high similarity value (>0.4), confirming that these models prioritize coherence to the point of redundancy. GPT-4o, however, shows a broader distribution that more closely resembles the human shape, though its mean similarity remains higher. This suggests that advanced models are beginning to learn "burstiness," making them harder to detect via simple coherence metrics alone.

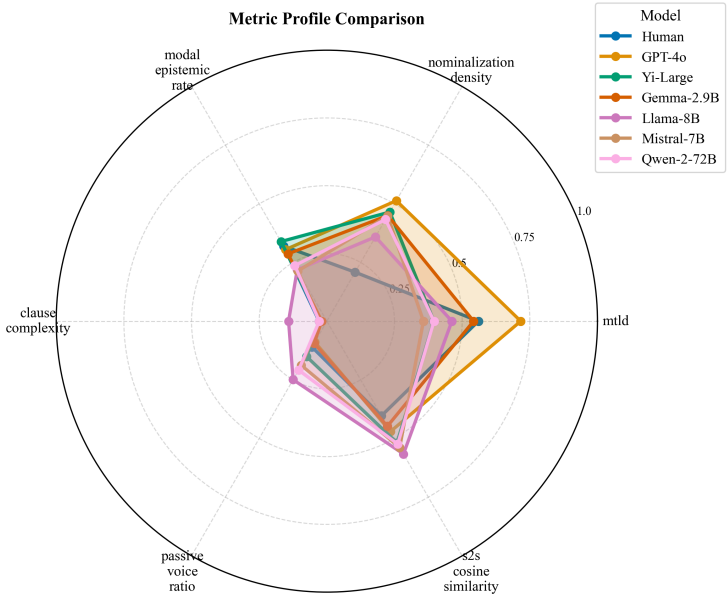


Fig. 3: Metric Profile Comparison (Topic: Health). The human profile (blue) is characterized by a sharp spike in lexical diversity (MTLD) and low semantic similarity, whereas models like Llama-8B (purple) are defined by extreme coherence and repetition.

Qualitative Lexical Divergence Beyond statistical metrics, a qualitative analysis of vocabulary usage reveals a stark contrast in semantic grounding. Figures 5a and 5b compare the most frequent unique tokens in GPT-4o output versus human stories.

The human word cloud is dominated by concrete, proper nouns and specific entities ("York," "Times," "Houston," "Earthquake"), reflecting a grounding in real-world events and facts. Conversely, the GPT-4o word cloud is filled with abstract, generalized concepts ("Community," "Cultural," "Journey," "Experience," "Vibrant"). This "hallucination of vagueness" is a critical qualitative differentiator, showing that AI models often simulate depth through high-level abstractions, whereas human narratives are built on specific, verifiable details.

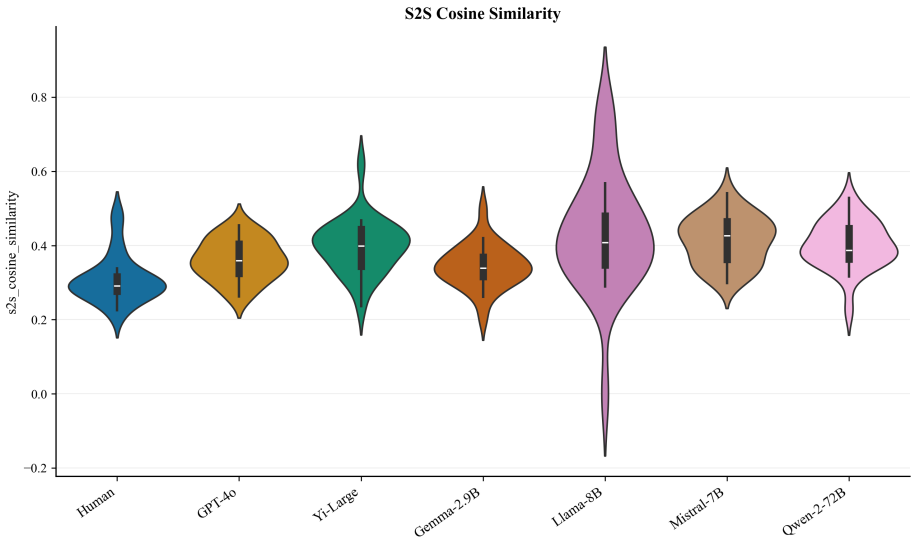


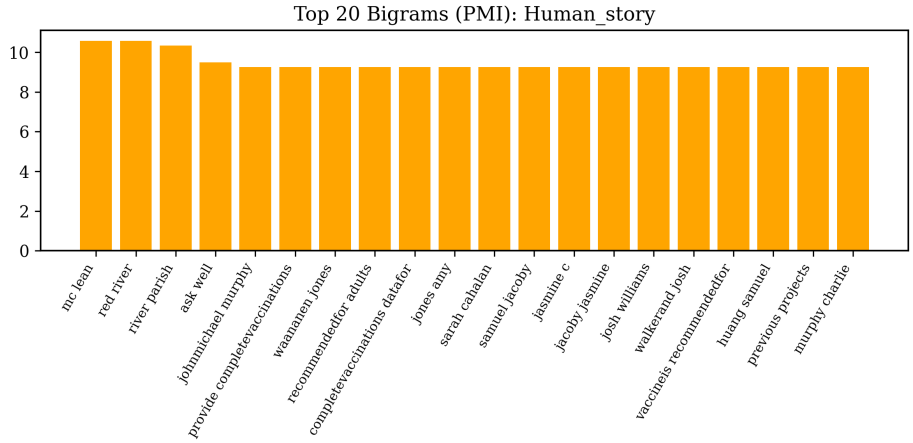
Fig. 4: Distribution of S2S Cosine Similarity (Topic: Health). Human writing shows a broader, lower distribution, indicating dynamic semantic flow. In contrast, Llama-8B exhibits a tight, high-value distribution, reflecting extreme repetitiveness.

(a) GPT-4o Word Cloud (b) Human Story Word Cloud

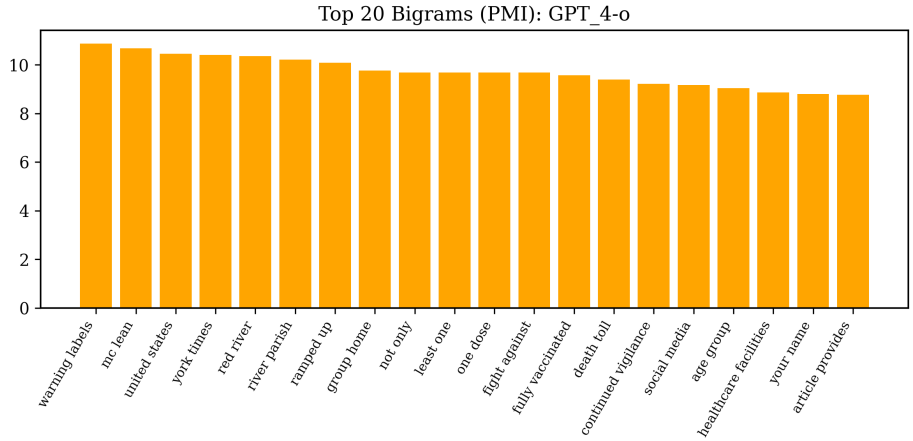
Fig. 5: Lexical Comparison. GPT-4o relies on abstract, "fluff" concepts (e.g., "Art", "Community", "Journey"), while human writing is grounded in specific entities and concrete nouns (e.g., "York", "Times", "Houston", "Dancing").

Collocation Analysis) To further investigate the semantic differences, we analyzed the top 20 bigrams, which highlights strongly associated word pairs rather than just frequent ones. Figure 6 illustrates the divergence between human and AI-generated text structure.

The human bigrams (Figure 6a) are dominated by unique, specific proper nouns and entities such as "Sarah Cahalan," "Samuel Jacoby," "River Parish," and "Waananen Jones". This strongly supports the hypothesis that human narratives are "grounded" in specific people, places, and singular events.



(a) Top 20 Bigrams (Human)



(b) Top 20 Bigrams (GPT-4o)

Fig. 6: PMI-ranked Bigram Analysis. Human writing is characterized by highly specific proper names and entities (e.g., "Sarah Cahalan", "Samuel Jacoby"), whereas GPT-4o output contains generic functional phrases ("not only", "social media") and meta-instructional artifacts ("your name", "article provides").

In contrast, the GPT-4o bigrams (Figure 6b) reveal two distinct markers of artificiality. First, the vocabulary is markedly more generic, relying on broad concepts like "United States," "social media," "age group," and "healthcare facilities" rather than specific names. Second, and most critically, the model outputs meta-instructional artifacts such as "your name" and "article provides". These hallucinations likely stem from the model's training on templates or its

attempt to follow the structure of a generic article prompt, a type of error entirely absent in the edited human journalism baseline.

Discussion of Model-Specific Performance The results highlight a "hierarchy of mimicry." GPT-4o emerges as the most sophisticated mimic, achieving a Clause Complexity score (0.04) identical to the human baseline (0.04) in the Health shard. However, it betrays its artificiality through excessive Nominalization Density (0.45 vs. Human 0.20). Smaller models like Llama-8B fail more transparently, characterized by low lexical diversity and extreme semantic repetitiveness.

Comparison Experiments This subsection reports direct comparisons with existing approaches:

- Performance comparison with classical baselines and recent state-of-the-art methods.
- Analysis of cases where the proposed method excels.
- Discussion of scenarios where performance is comparable or lower.
- Positioning of the proposed method within the broader research landscape. There should be multiple tables **??** and figures **??** for demonstration and clarity.

Ablation Study This subsection evaluates the contribution of each component of the proposed method:

- Removal or modification of individual modules or strategies.
- Measurement of performance change to quantify each component's importance.
- Justification of design decisions based on empirical evidence.

5 Conclusion

Our study investigated the distinction between human-generated essays and AI-generated essays. Our results showed that our hypotheses were correct; humans tend to utilize significantly less nominalization and passive voice than AI. Moreover, these suggest the attributes relating to readability that distinguish between human and different machine models generated texts, as well as a blueprint for an AI texts detection software. Our future works include a more detailed dissection of detection metrics and an improvement over the current model's accuracy.

References

[1] Guangsheng Bao et al. "Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature". In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=Bpcgcr8E8Z.

[2] Ahmed M. Elkhatat, Khaled Elsaid, and Saeed Almeer. "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text". In: *International Journal for Educational Integrity* 19.1 (Sept. 2023). ISSN: 1833-2595. DOI: 10.1007/s40979-023-00140-5.

[3] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. "GLTR: Statistical Detection and Visualization of Generated Text". In: *CoRR* abs/1906.04043 (2019). DOI: 10.48550/arxiv.1906.04043. arXiv: 1906.04043. URL: http://arxiv.org/abs/1906.04043.

[4] Ahmed Abdeen Hamed and Xindong Wu. "Detection of ChatGPT fake science with the xFakeSci learning algorithm". In: *Scientific Reports* 14.1 (July 2024). ISSN: 2045-2322. DOI: 10.1038/s41598-024-66784-6.

[5] Abhimanyu Hans et al. "Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text". In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=axl3FAkpik.

[6] Steffen Herbold et al. "A large-scale comparison of human-written versus ChatGPT-generated essays". In: *Scientific Reports* 13.1 (Oct. 2023). ISSN: 2045-2322. DOI: 10.1038/s41598-023-45644-9.

[7] Eric Mitchell et al. "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature". In: *CoRR* abs/2301.11305 (2023). DOI: 10.48550/ARXIV.2301.11305. arXiv: 2301.11305.

[8] Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. "Threat Scenarios and Best Practices to Detect Neural Fake News". In: *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*. Ed. by Nicoletta Calzolari et al. International Committee on Computational Linguistics, 2022, pp. 1233–1249. URL: https://aclanthology.org/2022.coling-1.106.

[9] Ali Quidwai, Chunhui Li, and Parijat Dube. "Beyond Black Box AI generated Plagiarism Detection: From Sentence to Document Level". In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, 2023, pp. 727–735. DOI: 10.18653/v1/2023.bea-1.58.

[10] Rajarshi Roy et al. "A Comprehensive Dataset for Human vs. AI Generated Text Detection". In: *CoRR* abs/2510.22874 (2025). DOI: 10.48550/ARXIV.2510.22874. arXiv: 2510.22874.

[11] Areg Mikael Sarvazyan et al. "Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains". In: *CoRR* abs/2309.11285 (2023). DOI: 10.48550/ARXIV.2309.11285. arXiv: 2309.11285.

[12] Emily Sheng et al. "Societal Biases in Language Generation: Progress and Challenges". In: *CoRR* abs/2105.04054 (2021). DOI: 10.48550/arxiv.2105.04054. arXiv: 2105.04054. URL: https://arxiv.org/abs/2105.04054.

[13] Ilia Shumailov et al. "AI models collapse when trained on recursively generated data". In: *Nat.* 631.8022 (2024), pp. 755–759. DOI: 10.1038/S41586-024-07566-Y.

[14] Mayank Soni and Vincent Wade. "Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms". In: *CoRR* abs/2303.17650 (2023). DOI: 10.48550/ARXIV.2303.17650. arXiv: 2303.17650.

[15] Karin Tengler and Gerhard Brandhofer. "Exploring the difference and quality of AI-generated versus human-written texts". In: *Discover Education* 4.1 (May 2025). ISSN: 2731-5525. DOI: 10.1007/s44217-025-00529-z.

[16] Vivek Verma et al. "Ghostbuster: Detecting Text Ghostwritten by Large Language Models". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*. Ed. by Kevin Duh, Helena Gómez-Adorno, and Steven Bethard. Association for Computational Linguistics, 2024, pp. 1702–1717. DOI: 10.18653/V1/2024.NAACL-LONG.95.

[17] Junchao Wu et al. "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions". In: *Comput. Linguistics* 51.1 (2025), pp. 275–338. DOI: 10.1162/COLI_A_00549.

[18] Junchao Wu et al. "Who Wrote This? The Key to Zero-Shot LLM-Generated Text Detection Is GECScore". In: *CoRR* abs/2405.04286 (2024). DOI: 10.48550/ARXIV.2405.04286. arXiv: 2405.04286.

[19] Xianjun Yang et al. "DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text". In: (May 2023). DOI: 10.48550/ARXIV.2305.17359. arXiv: 2305.17359 [cs.CL].

[20] Hatice Yildiz Durak, Figen Eğin, and Aytuğ Onan. "A Comparison of Human-Written Versus AI-Generated Text in Discussions at Educational Settings: Investigating Features for ChatGPT, Gemini and BingAI". In: *European Journal of Education* 60.1 (Jan. 2025). ISSN: 1465-3435. DOI: 10.1111/ejed.70014.

[21] Xiao Yu et al. "LLM Paternity Test: Generated Text Detection with LLM Genetic Inheritance". In: (May 2023). DOI: 10.48550/ARXIV.2305.12519. arXiv: 2305.12519 [cs.CL].

[22] Mengxuan Zhang and Peter Crosthwaite. "More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers". In: *International Review of Applied Linguistics in Language Teaching* (Jan. 2025). ISSN: 1613-4141. DOI: 10.1515/iral-2024-0196.

[23] Lihang Zhu et al. "Enhancing EFL argumentative writing through an AI-powered corpus: impact on learner writing proficiency". In: *Computer Assisted Language Learning* (2025). Published online: 16 Dec 2025. DOI: 10.1080/09588221.2025.2599152. URL: https://doi.org/10.1080/09588221.2025.2599152.