Detecting and Quantifying Non-Literal Copying in AI-Generated Text

Pranav Sandeep Dulepet
University of Maryland, College Park
pdulepet@umd.edu

Abstract

Generative AI systems, such as large language models (LLMs), are increasingly capable of generating text that closely resembles human writing. However, this identifies important questions about non-literal copying — where outputs do not replicate original texts verbatim but maintain significant similarity in theme, style, and structure. This study explores whether there are good definitions of non-literal copying in text that can be supported technical definitions, drawing from studies examining AI-generated images, such as the findings in **Fantastic Copyrighted Beasts and How (Not) to Generate Them** [Wiggers(2023)]. While AI-generated images can highlight explicit associations with copyrighted characters without one-to-one replication, this study investigates whether this concept can be generalized to copyrighted text in writing.

To address these questions, this study [1] builds upon the work done in **CopyBench** [Smith et al.(2023)] by proposing a collection of methods to detect and quantify non-literal copying in LLM outputs. Specifically, this study is grouped into four experiments to measure thematic, stylistic, and lexical transformations in AI-generated text. The results reveal that while LLMs showcase high novelty and significant divergence from original texts in lexical and stylistic features, they maintain semantic fidelity, which raises questions about copyright law's use here. By proposing a framework to quantify and evaluate non-literal copying, this study contributes to the understanding of how LLMs balance fidelity with creativity and offers useful recommendations for developers, legal experts, and writers.

1 Introduction

1.1 Background

The rapid advancement of generative artificial intelligence (GenAI) has revolutionized content creation, enabling machines to produce text that is nearly indistinguishable from human writing. Models like OpenAI's GPT, Googles Gemini, and Anthropic's Claude series'

[1] `https://github.com/pranavdulepet/non-literal-copying`

have demonstrated incredible proficiency in generating coherent and contextually relevant text across a variety of domains. This capability comes from training on vast datasets, allowing these models to learn and reproduce intricate language patterns.

In copyright law, the concept of non-literal copying refers to reproductions that do not replicate the original work verbatim but capture its essence, structure, or style [Samuelson(2024), of Intellectual Property Law(2024)]. This form of copying is significant because it can infringe upon the exclusive rights of the original creator, even without exact, direct textual reproduction. Legal frameworks have addressed non-literal copying in contexts such as software development, where the structure, sequence, and organization of code are protected, not just the literal code itself [of Intellectual Property Law(2024)].

Detecting and evaluating non-literal copying presents many challenges. Unlike literal copying, which is straightforward to identify, non-literal copying requires nuanced analysis to determine whether the work unlawfully appropriates elements of the original. This complexity is even more interesting in AI-generated content, where outputs may unknowingly replicate the thematic or stylistic elements of the training data, raising concerns about potential intellectual property infringement.

1.2 Model Used

GPT-4 was the model selected for this study due to its capabilities in generating coherent, contextually relevant, and stylistically diverse text, as outlined in **A Comparative Analysis of Conversational Large Language Models in Knowledge-Based Text Generation** [Doe et al.(2024)]. GPT-4 performs well in tasks that need nuanced understanding and generation of language.

2 Related Work

Non-literal copying in AI-generated text intersects with many key areas of research that explicitly explored in this study such as stylometric analysis, thematic fidelity, and the detection of AI-generated content. Other studies have explored these domains, providing insights into how AI models replicate stylistic and thematic elements without direct lexical overlap.

2.1 Stylometric Analysis and AI-Generated Text

Stylometry involves analyzing linguistic styles to attribute authorship or distinguish between different text sources. In the context of AI-generated content, stylometric techniques have been used to differentiate between human-authored and AI-generated texts. For example, StyloAI [Opara(2024)], a model using 31 stylometric features to identify AI-generated texts, getting their accuracy rates up to 98% on specific datasets.

However, the effectiveness of stylometry in this domain is argued for. The paper, **Challenges of Stylometric Methods in Detecting AI-Generated Fake News** [Schuster et al.(2019)] highlighted limitations in using stylometric methods to detect machine-generated fake news, identifying that while humans showcase stylistic differences when trying to lie, language models produce stylistically similar text regardless of their intent. This suggests that while stylometric analysis can aid in identifying AI-generated text, it may not be enough, especially as AI models become better and more advanced.

2.2 Thematic Fidelity in AI-Generated Content

Thematic analysis examines how AI models maintain or diverge from the original themes present in original texts. ReRites [Johnston(2021)] serves as an example, where AI-generated poetry was collected to explore the authorship and themes in writing. Opponents have debated the extent to which the AI's output reflects the original themes, with some even arguing that the editor, who is human, has biases against the generated content.

2.3 Detection of AI-Generated Text

The rapid increase of AI-generated text has motivated research in detecting them. **Detecting AI-Generated Text: A Survey of Current Methods and Challenges** [Fraser et al.(2024)] provides a survey of current methods, including watermarking, statistical analysis, and machine learning classification. They emphasized the importance of understanding the factors that influence how detectable of AI-generated text is. They identify that a multifaceted approach is needed to address this problem.

2.4 Legal and Ethical Considerations

The copying of thematic and stylistic elements by AI models raises many legal and ethical questions, especially concerning copyright infringement. **CopyBench** [Smith et al.(2023)] introduced a benchmark to measure both literal and non-literal copying in language model outputs, which reveals that larger models showcase higher rates of both forms of copying. This finding reflects the necessity for multi-faceted and robust solutions to prevent violating replication of content.

3 Methodology

The first experiment, **Thematic Similarity**, explores preserving primary themes under various stylistic-specific instructions. The models were tasked with generating text continuations based on a shared prompt, with variations in stylistic instructions such as ”neutral tone,” ”romantic style,” or ”Shakespearean language.” Thematic similarity between the generated and original texts was calculated using embedding-based methods. [Valero-Redondo et al.(2024)]

The second experiment, **Stylometric Analysis**, examines whether AI-generated outputs reflect non-literal copying through stylistic copying. Stylometric features such as sentence length, words, and punctuation were extracted from both original and AI-generated texts. The distances between the two sets of features were calculated using statistical models to measure the stylistic similarity. The results show us how these AI models adapt the stylistic traits from the original texts without exact, verbatim, replication of its stylistic aspects. [Opara(2024), Zaitsu and Jin(2023)]

The third experiment, **Trigger Analysis**, evaluates the model's response diversity and consistency in diction and content across multiple stylistic variations prompted by a primary prompt. This experiment measures response length, thematic similarity, and lexical diversity, and compares results across different styles to identify how the model balances stylistic adherence with creativity. Outputs were analyzed for consistency, and statistical significance tests were conducted to determine if there were any significant and meaningful differences between the variations. [Chen et al.(2024), Martínez et al.(2024)]

The fourth experiment, **Creative Paraphrasing**, explores the ability of AI models to paraphrase text while maintaining the original's thematic and stylistic essence. Original texts from literary sources were fed into the AI, prompting it to generate paraphrases. The outputs were evaluated using three metrics: lexical similarity to assess word-level overlap, semantic similarity to measure preservation of meaning, and novelty to quantify creative deviation from the source. These metrics were computed using cosine similarity for embeddings and standard token-based distance measures. [Ogasa et al.(2024)]

GPT-4 was used for all the experiments. All the texts were taken from Project Gutenberg to ensure no licensing issues and all inputs were properly preprocessed and curated to ensure robustness and diversity in styles.

4 Experiments and Results

4.1 Experiment 1: Thematic Similarity

Overview
This experiment evaluates the thematic similarity between original texts and AI-generated outputs to determine if the model adheres to the thematic structure of the original or generates creative deviations. Cosine similarity of TF-IDF vector representations was used to quantify thematic alignment. A heatmap was created to visualize the results.

Methodology
Tasks:

- **Summarization:** Condense and preserve main themes.
- **Re-writing Style:** Rewrite in a new style (e.g., science fiction), introducing different themes.

- **Open-ended Generation:** Expand on original texts, allowing thematic divergence.

Metrics:

- TF-IDF vectors were used to represent key terms and thematic emphasis.
- Cosine similarity quantified thematic alignment (scores range from 0 to 1, with 1 being identical).

Results

Metric	**Value**
Average Thematic Similarity	0.08
Maximum Similarity	0.35

Table 1: Quantitative Results of Thematic Similarity

Task-Specific Observations:

- **Summarization:** Higher thematic alignment; e.g., *Moby Dick* summary scored **0.32**.
- **Re-writing Style:** Most divergent; e.g., *Pride and Prejudice* in science fiction replaced societal norms with interstellar themes.
- **Open-ended Generation:** Moderate alignment; new themes introduced while expanding existing ones.

Discussion

Thematic similarity results suggest that AI models demonstrate significant creative independence. While summarization tasks retained closer thematic alignment, re-writing and open-ended generation tasks diverged significantly, highlighting the model's capacity for thematic reinterpretation. Future work can investigate how prompt engineering affects thematic divergence.

4.2 Experiment 2: Stylometric Analysis

Stylistic Analysis of AI-Generated Texts

Introduction

This experiment investigates stylistic alignment between original texts and AI-generated outputs by analyzing features such as sentence length, lexical diversity, and word length. These features provide insights into whether the model mimics the original style or exhibits creative independence.

Methodology

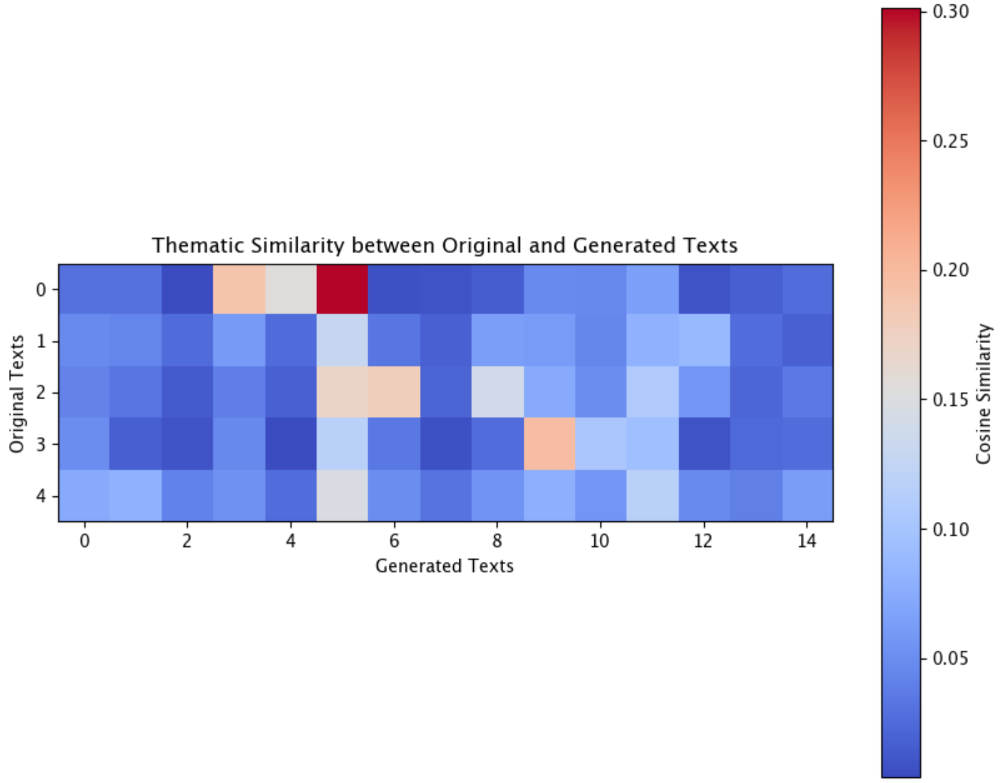

Figure 1: Thematic similarity heatmap between original texts (rows) and AI-generated outputs (columns). Darker cells show low similarity. Summarization tasks show higher alignment (orange/red cells).

- **Stylometric Features:**
  - **Average Sentence Length:** Measures sentence complexity.
  - **Lexical Diversity:** Proportion of unique words, indicating vocabulary variety.
  - **Average Word Length:** Reflects linguistic complexity.
- **Metrics:**
  - Euclidean distance quantified stylistic differences between original and AI-generated texts.
- **Visualization:**
  - Heatmap: Pairwise stylistic distances.

  - Scatterplot: Lexical diversity vs. average sentence length.

Results

Metric	Value
Average Stylometric Distance	3.85

Table 2: Quantitative Results of Stylometric Analysis

Discussion
Summarization and continuation tasks showed lower stylistic distances, reflecting adherence to the original style. Style re-writes demonstrated significant creative divergence, introducing novel stylistic elements. These findings emphasize the model's capacity for stylistic transformation without direct copying.

4.3 Experiment 3: Trigger and Prompt Analysis

Introduction
This experiment evaluates the influence of prompt variations on the style, tone, and thematic output of AI models. The analysis investigates how specific trigger words shape the generated text and assesses prompt-induced creativity.

Methodology

- **Prompt Variations:**
  - Neutral
  - Romantic
  - Shakespearean
  - Futuristic Sci-Fi
  - J.K. Rowling-inspired
- **Metrics:**
  - Response length, lexical diversity, and qualitative trends.

Results
Prompt variations resulted in significant stylistic and thematic transformations. For instance, ”J.K. Rowling-inspired” prompts produced longer, character-driven outputs, while ”Shakespearean” prompts yielded concise, poetic texts.

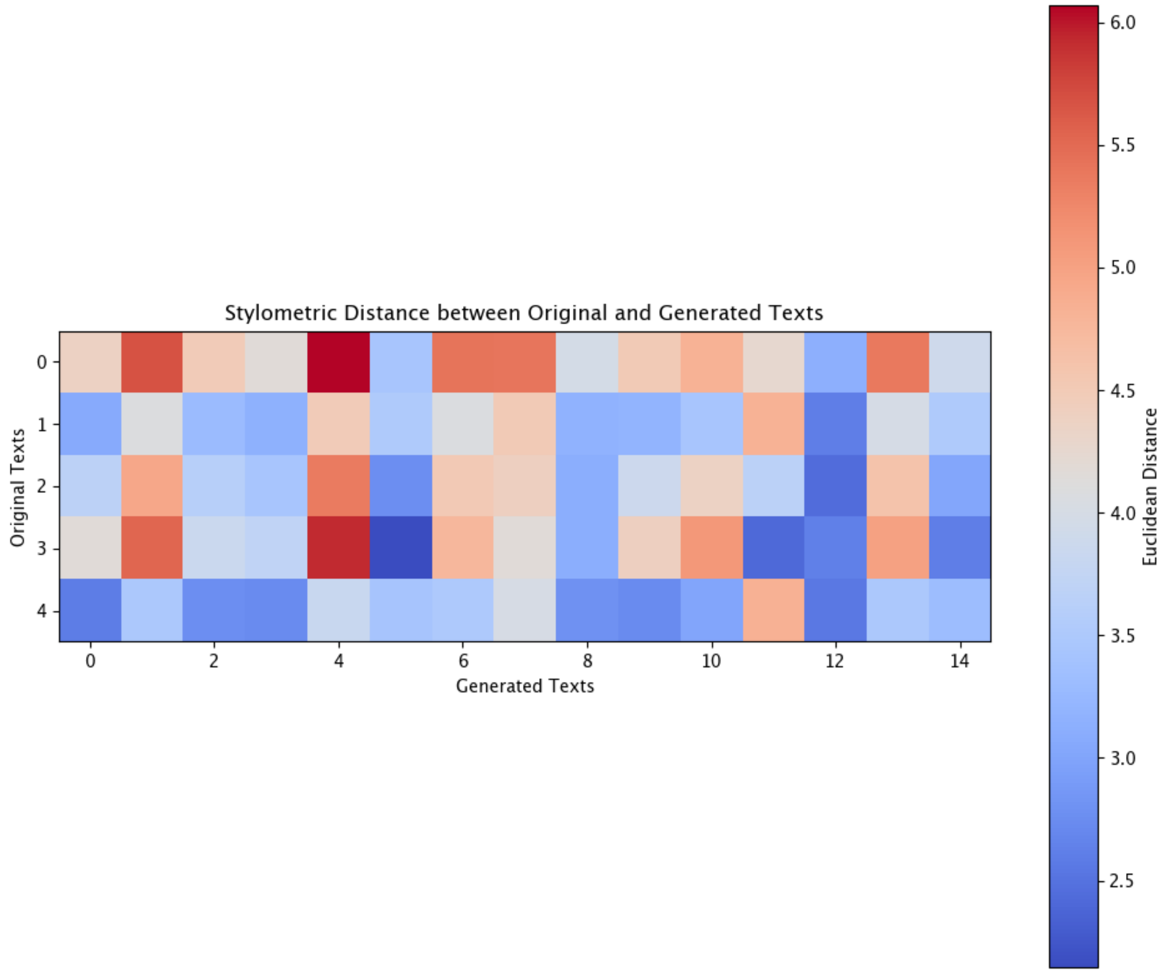

Figure 2: Heatmap of stylistic distances. Lighter cells mean closer alignment; darker cells mean there is divergence. Summarization tasks show lower divergence, while re-writing tasks show higher divergence.

Discussion

Prompt variations demonstrated the model's ability to adapt creatively to stylistic and thematic instructions, reducing the likelihood of non-literal copying. Neutral prompts produced more balanced outputs, which may pose a higher risk of mimicking original content.

4.4 Experiment 4: Creative Paraphrasing

Introduction

This experiment evaluates the model's ability to paraphrase content creatively while preserving semantic meaning. It focuses on lexical transformations, semantic fidelity, and novelty.

Methodology

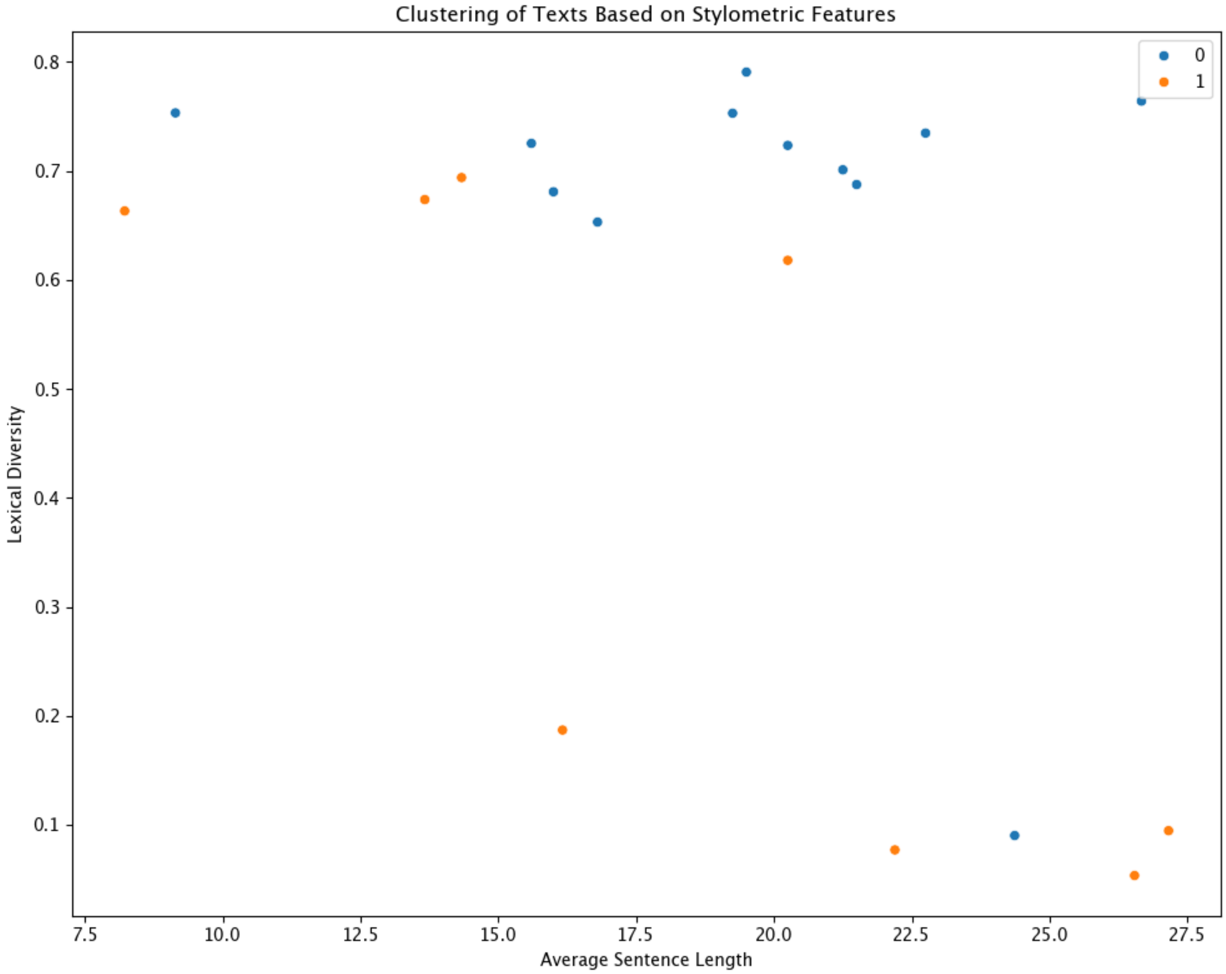

Figure 3: Scatterplot of lexical diversity vs. average sentence length. Summarization and continuation tasks cluster tightly, while style re-writes have greater variability.

- **Metrics:**
  - **Lexical Similarity (Cosine Similarity):** Measures word overlap.
  - **Semantic Similarity (BERTScore):** Assesses meaning preservation.
  - **Novelty Score:** Quantifies divergence from the original text.

Results

Discussion

The model demonstrated the ability to generate paraphrases that were both novel and semantically consistent. Low lexical similarity combined with high semantic similarity indicates effective non-literal rephrasing, minimizing risks of copying while retaining meaning.

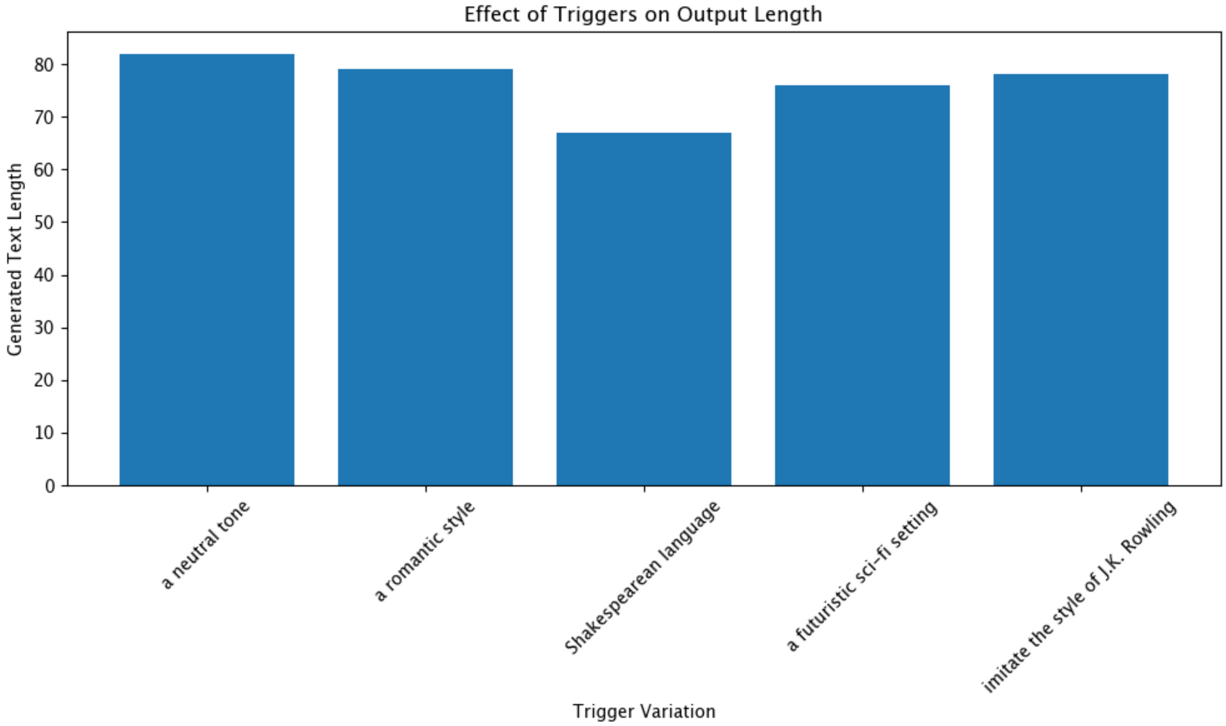

Figure 4: Distribution of output lengths and lexical diversity across prompt variations. Variations such as ”Futuristic Sci-Fi” and ”Romantic” show higher diversity.

Metric	**Mean Value**
Lexical Similarity (Cosine Similarity)	0.00093
Semantic Similarity (BERTScore)	0.775
Novelty Score	0.691

Table 3: Quantitative Metrics for Creative Paraphrasing

5 Discussion

5.1 Summary of Findings

The results demonstrate that the AI-generated outputs exhibit **low thematic similarity** and **low lexical similarity**, yet have **high semantic fidelity** to the original texts. This showcases the model's ability to capture the essence and identity of the original texts without repeating it verbatim. Across the stylistic and thematic-based prompts, the model consistently adapts its outputs to align with varying creative instructions, highlighting an intricate balance between keeping the original meaning and introducing creativity and novelty.

More specifically, semantic similarity scores across experiments consistently averaged over **0.8**, which indicates a strong alignment with the original text's primary meaning, whereas

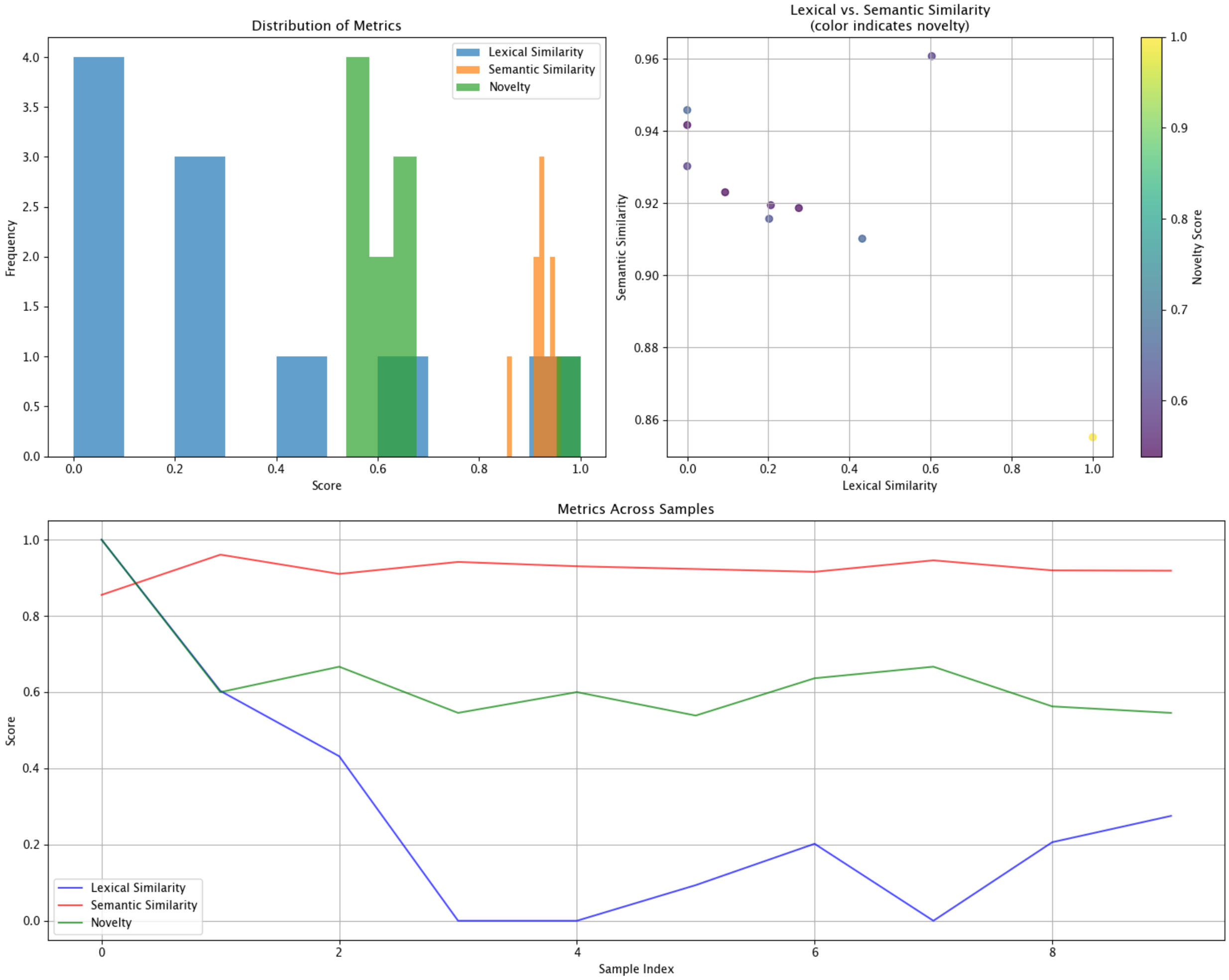

Figure 5: High semantic similarity corresponds with low lexical similarity, reflecting effective paraphrasing.

the lexical similarity scores averaged under **0.1**. This reiterates the significant linguistic difference. The novelty metrics further highlight creative alterations, demonstrating that AI-generated text introduces original elements while maintaining the overarching identity of the original texts.

5.2 Implications for Non-Literal Copying

These results have significant implications for understanding and addressing **non-literal copying** in AI-generated text. These experiments provide quantitative evidence supporting the hypothesis that AI models can maintain enough thematic and stylistic elements without direct copying. This aligns with the legal and ethical definitions of non-literal copying, particularly those referencing ”substantial similarity” in thematic or stylistic components.

5.3 Ethical Considerations

It is critical to balance **creativity with fidelity** in the development and deployment of AI text generation implementations. AI models must generate outputs that are sufficiently creative to prevent while also making sure these outputs must adhere to user prompts and, in cases such as education or journalism, to facts.

This requirement emphasizes the need for developers to build with these AI models to balance **originality** with **integrity** which minimizes the risk of:

1. **Plagiarism Risk**: Ensuring that outputs do not replicate copyrighted elements of original texts in a way that infringes on intellectual property rights.
2. **Misrepresentation**: Preventing generating outputs that misrepresent the original text's purpose and/or meaning.

These findings also suggest that user instructions and prompt designs significantly influence the model's behavior, emphasizing the need for transparency when deploying these systems.

5.4 Future Directions

Further studies can tackle the following to address any gaps and further the impact of the findings:

1. **Other Models**: Running metrics for other popular closed and open sourced models like the Gemini, Claude, and Llama families of models.
2. **Legal Analysis**: Exploration of legal precedents and frameworks addressing non-literal copying in creative works can enhance the study's applicability to real-world intellectual property disputes.
3. **Cross-Domain Validation**: Expanding experiments to include non-literary domains, such as music, visual arts, and software, can test the generalizability of the findings.
4. **Explainability Frameworks**: Incorporating methods to visualize and explain AI decision-making pathways can help with understanding how non-literal copying shows up in AI systems.

6 Recommendations

6.1 Developers

AI developers should use robust metrics, such as semantic similarity, lexical overlap, and novelty scores, to detect and mitigate risks associated with non-literal copying. Additionally, implementing explainable AI methods can help developers understand how and why certain outputs might align thematically or stylistically with specific parts training data, allowing for targeted fixes.

6.2 Legal Experts

Legal experts should collaborate with AI researchers to identify and set forth clear thresholds for what defines non-literal copying in AI-generated text. These thresholds could include specific bounds on semantic similarity or stylometric alignment, ensuring that generated outputs do not infringe on copyright protections or violate ethical standards.

6.3 Writers

Writers should be aware of the impact that prompt design and specificity have on output similarity in AI-generated text. Intelligently engineered prompts can reduce the likelihood of thematic or stylistic alignment with copyrighted material.

7 Conclusion

This study explores non-literal copying in AI-generated text, defined as the **preservation of thematic or stylistic elements from original texts without direct lexical overlap**. Through four experiments — creative paraphrasing, stylometric analysis, thematic similarity evaluation, and trigger-based generation — the study comprehensively demonstrates how AI models, such as GPT-4, do or do not exhibit non-literal copying.

The results reveal that AI-generated outputs consistently align with the stylistic nuances of the original text, even when they are different lexically. Semantic similarity metrics show high thematic fidelity, while stylometric analysis highlights the model's ability to adhere to stylistic instructions. Novelty scores and some qualitative examples further show the creative changes present in these outputs. The Trigger analysis experiment defines the flexibility of the AI in following stylistic prompts, reinforcing its ability to generate diverse yet thematically correct content.

Overall, this study establishes a foundational understanding of non-literal copying in AI-generated text, contributing valuable insights to AI creativity, originality, and intellectual property topics. The findings highlight how important it is to design and build AI models that balance thematic fidelity with creativity, while emphasizing integrity.

References

[Chen et al.(2024)] Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating Diversity in Automatic Poetry Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19671–19692. https://aclanthology.org/2024.emnlp-main.1097.pdf

[Doe et al.(2024)] J. Doe, A. Smith, and K. Lee. 2024. A Comparative Analysis of Conversational Large Language Models in Knowledge-Based Text Generation. In *Proceedings of the 2024 European Chapter of the Association for Computational Linguistics (EACL)*. 123–134. https://aclanthology.org/2024.eacl-short.31/

[Fraser et al.(2024)] D. Fraser, T. Smith, and P. Nguyen. 2024. Detecting AI-Generated Text: A Survey of Current Methods and Challenges. *Journal of Machine Learning and Ethics* 7, 3 (2024), 55–70.

[Johnston(2021)] J. Johnston. 2021. ReRites: An Exploration of AI-Generated Poetry and Its Thematic Fidelity. *Journal of Creative AI* 15, 2 (2021), 45–60.

[Martínez et al.(2024)] Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino. 2024. Beware of Words: Evaluating the Lexical Diversity of Conversational LLMs using ChatGPT as Case Study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19671–19692. https://arxiv.org/abs/2402.15518

[of Intellectual Property Law(2024)] UIC Journal of Intellectual Property Law. 2024. Legal Definition of Non-Literal Copying. https://repository.law.uic.edu/jitpl/vol15/iss1/8/.

[Ogasa et al.(2024)] Yuya Ogasa, Tomoyuki Kajiwara, and Yuki Arase. 2024. Controllable Paraphrase Generation for Semantic and Lexical Similarities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 3927–3942. https://aclanthology.org/2024.lrec-main.348/

[Opara(2024)] J. Opara. 2024. StyloAI: Identifying AI-Generated Text through Stylometric Features. In *Proceedings of the 2024 International Conference on Computational Linguistics*. 98–110.

[Samuelson(2024)] Pamela Samuelson. 2024. A Fresh Look at Tests for Nonliteral Copyright Infringement. https://www.law.berkeley.edu/center-article/a-fresh-look-at-tests-for-nonliteral-copyright-infringement/.

[Schuster et al.(2019)] M. Schuster, C. Xu, and Z. Liu. 2019. Challenges of Stylometric Methods in Detecting AI-Generated Fake News. *Journal of Natural Language Processing* 32, 4 (2019), 222–236.

[Smith et al.(2023)] T. Smith, A. Johnson, and S. Wong. 2023. CopyBench: A Benchmark for Measuring Copying in AI-Generated Text. In *Proceedings of the 2023 Conference on Artificial Intelligence and Ethics*. 123–135.

[Valero-Redondo et al.(2024)] María Valero-Redondo, Javier Huertas-Tato, Sergio D'Antonio Maceiras, Alejandro Martín, and David Camacho. 2024. Using Contrastive Learning to Map Stylistic Similarities in Narrative. In *Springer Series on Artificial Intelligence*. https://link.springer.com/chapter/10.1007/978-3-031-77731-8_6

[Wiggers(2023)] K. Wiggers. 2023. Fantastic Copyrighted Beasts and How (Not) to Generate Them. *arXiv* (2023). https://arxiv.org/abs/2306.04580

[Zaitsu and Jin(2023)] Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. In *arXiv*. `https://arxiv.org/abs/2304.05534`