A Ship of Theseus:
Curious Cases of Paraphrasing in LLM-Generated Texts

Anonymous ACL submission

Abstract

In the realm of text manipulation and linguistic transformation, the question of authorship has always been a subject of fascination and philosophical inquiry. Much like the **Ship of Theseus paradox**, which ponders whether a ship remains the same when each of its original planks is replaced, our research delves into an intriguing question: *Does a text retain its original authorship when it undergoes numerous paraphrasing iterations?* Specifically, since Large Language Models (LLMs) have demonstrated remarkable proficiency in both the generation of original content and the modification of human-authored texts, a pivotal question emerges concerning the determination of authorship in instances where LLMs or similar paraphrasing tools are employed to rephrase the text - *whether authorship should be attributed to the original human author or the AI-powered tool.* Therefore, we embark on a philosophical voyage through the seas of language and authorship to unravel this intricate puzzle. Using a computational approach, we discover that the diminishing performance in text classification models with each successive paraphrasing iteration is closely associated with the extent of deviation from the original author's style, thus provoking a reconsideration of the current notion of authorship.

1 Introduction

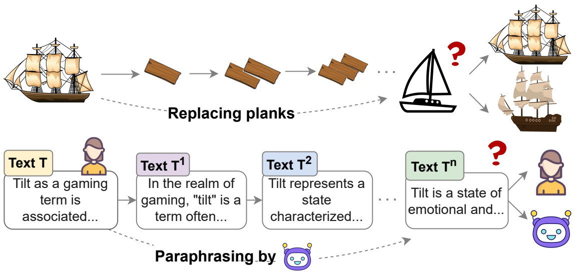
Replacing planks
Text T
Tilt as a gaming term is associated...
Text T1
In the realm of gaming, "tilt" is a term often...
Text T2
Tilt represents a state characterized...
Text Tn
Tilt is a state of emotional and...
Paraphrasing by

Figure 1: Ship of Theseus paradox in text paraphrasing scenario: who should be considered the author of T^n?

The Ship of Theseus paradox is a philosophical thought experiment (Scaltsas, 1980) that questions the concept of originality and change over time. The paradox begins with the premise that a ship, called the **Ship of Theseus**, gradually has all its planks replaced over time with new, identical planks. The paradox then poses the question: *Is the fully modified ship, with none of its original parts remaining, still the Ship of Theseus, or is it an entirely different ship?* Just like the Ship of Theseus, our study involves the successive transformation of text through paraphrasing as illustrated in Figure 1. Each paraphrase iteration can be seen as a replacement of linguistic "planks." We aim to explore whether, like the Ship of Theseus, the essence of the original authorship remains intact or whether it morphs into something entirely new.

Paraphrasing involves rewriting texts to convey the same meaning while employing different words or sentence structures (Bhagat and Hovy, 2013). Although paraphrasing has long been employed to enhance writing, it has been the subject of ongoing ethical and plagiarism-related debates (Prentice and Kinden, 2018; Roe and Perkins, 2022). Nevertheless, paraphrasing has always been considered a tool to aid in rewriting content rather than generating entirely original material. However, recent advancements in LLMs have altered this paradigm as they can function as paraphrasers while also autonomously generating original content without explicit prompts. As illustrated in the examples in Figure 2, a situation will arise in contemporary times where paraphrasing a text (T^0) using an LLM to produce the paraphrased version (T^1) might closely resemble the text (G) independently generated by the LLM on the same subject matter. Consequently, this situation prompts inquiries about the authorship of text T^1, akin to the philosophical dilemma posed by the **Ship of Theseus**.

Two contrasting perspectives on this matter are evident within the existing literature (Figure 2).

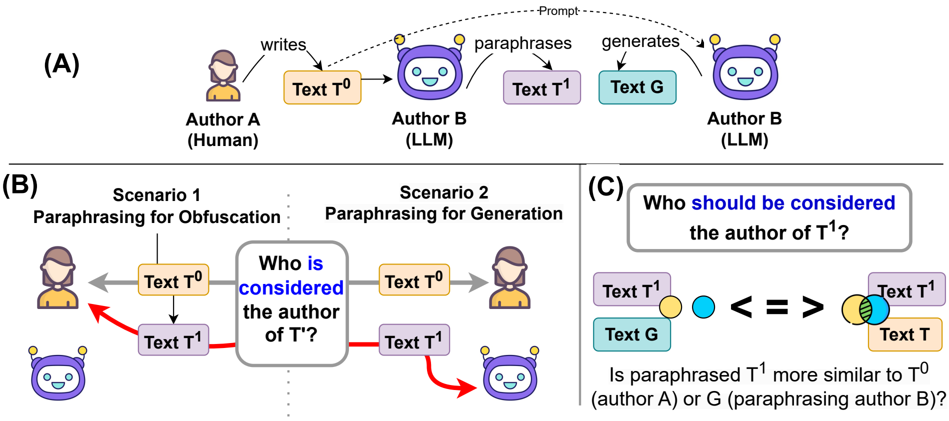
(A)
writes
Prompt
Text T0
paraphrases
Text T1
generates
Text G
Author A
(Human)
Author B
(LLM)
Author B
(LLM)
(B)
Scenario 1
Paraphrasing for Obfuscation
Text T0
Text T1
Who is considered the author of T'?
Scenario 2
Paraphrasing for Generation
Text T0
Text T1
(C)
Who should be considered the author of T1?
Text T1
Text G
< = >
Text T1
Text T
Is paraphrased T1 more similar to T0
(author A) or G (paraphrasing author B)?

Figure 2: (A) indicates how LLMs can paraphrase as well as generate texts, (B) portrays the two alternative scenarios regarding authorship, and (C) shows how authorship should be determined.

Paraphrasing has often been employed as a text obfuscation or perturbation method (Potthast et al., 2016; Bevendorff et al., 2019, 2020). In line with this perspective, several studies (Krishna et al., 2023; Sadasivan et al., 2023; Hu et al., 2023) argue that the weakness of a text classifier or AI-text detector is evident if it fails to attribute a paraphrased text to its original source precisely. Thus, these studies assume that **authorship remains the same after paraphrasing**. Conversely, paraphrasing can also serve as a text-generation technique. A growing number of recent studies (Yu et al., 2023; Zhang et al., 2023; Lucas et al., 2023) utilize LLMs to rewrite human-generated texts through paraphrasing to create AI-generated datasets. Consequently, these studies presume that **authorship changes after paraphrasing**.

While authorship identification is a well-established discipline in text classification (Neal et al., 2017), it has garnered renewed interest with the advent of LLMs that mimic human-like text generation across diverse contexts (Uchendu et al., 2023; Tripto et al., 2023). The attribution of authorship to a text is fundamentally influenced by two key factors: **content**, denoting the subject matter or what the text pertains to, and **style**, reflecting the way of expression (Sari et al., 2018). In light of this, our research investigates the extent of content and stylistic alterations resulting from successive iterations of paraphrasing sourced from original texts and its implications for text classifier performance. Our motivation stems from the plausible scenarios facilitated by contemporary LLMs, as illustrated in Figure 2. To investigate this, we employ LLMs as paraphrasers alongside other paraphrasing models (PMs) like Pegasus (Zhang et al., 2020), operating at the sentence level, and Dipper (Krishna et al., 2023), capable of whole-text paraphrasing while preserving contextual coherence and offering control over lexical diversity. Our comprehensive analysis encompasses various text sources, including human-authored content and texts generated by six LLMs in seven distinct datasets.

Our study stands apart from other research in authorship analysis, paraphrasing detection, AI-text detection, or style analysis. The major contribution of our paper is as follows:

- We aim to offer a solid resolution to the counter-intuitive assumptions surrounding paraphrasing and authorship, employing a comprehensive computational perspective supported by philosophical theory.
- We identify the difference among paraphrasers regarding their effect on authorship.
- We create a paraphrased corpora[1] consisting of seven sources (with humans), seven datasets, and four paraphrasers.

2 Related Work

Our study extends prior research examining authorship from various perspectives, including style and content. Notably, Sari et al. (2018) found that content-based features are more effective for datasets with high topical variance, while datasets with lower variance benefit more from style-based features. Several assessments and benchmarks on stylistic analysis have aimed to identify and infer style across different domains. The XSLUE

[1] Available at https://anonymous.4open.science/r/Ship_of_theseus_paraphrased_copus-B4B5

benchmark (Kang and Hovy, 2021) comprehensively evaluates sentence-level cross-style language understanding in 15 different styles. Additionally, the STEL framework (Wegmann and Nguyen, 2021; Wegmann et al., 2022) introduces four specific assessments measuring the stylistic content of authorship representations: formality, simplicity, contraction usage, and number substitution preference. Recent research has also explored the learning of authorship representations (Boenninghoff et al., 2019; Hay et al., 2020) in diverse cross-domain settings. For instance, Rivera-Soto et al. (2021) introduced the concept of universal authorship representations with a recent extension (Wang et al., 2023) to validate their capacity to capture stylistic features. However, it is essential to note that these studies primarily focus on performing classification tasks related to authorship in various setups. Our task distinguishes itself by delving into the established concept of ground truth concerning authorship in paraphrasing in the era of LLMs.

Another body of related research revolves around paraphrasing detection and plagiarism. These studies aim to determine whether a pair of texts constitutes a paraphrased version of one another (Becker et al., 2023). Paraphrasing detection stands as a critical challenge within the domain of plagiarism identification (Chowdhury and Bhattacharyya, 2018). It is also a subject of inquiry in evaluating a proposed model's capacity for addressing natural language understanding tasks (Wang et al., 2018). Previous research encompasses both human-generated (Seraj et al., 2015; Dong et al., 2021) and machine-generated (Foltýnek et al., 2020; Wahle et al., 2022a) paraphrased versions. A recent investigation by Wahle et al. (2022b) suggests that machine-generated paraphrases bear greater similarity to the original source text than human-generated paraphrases. This phenomenon resurges the discussion: if an LLM, such as ChatGPT, is employed to paraphrase a text, should ChatGPT be regarded as the author? Therefore, our study seeks to explore the connection between authorship and paraphrasing by bridging the gap among these distinct lines of research.

3 Methodology

The classical authorship attribution problem aims to determine the author (A) of a given text T^0 from a set of candidate authors, typically treated as a multi-class classification task. However, when text T^0 is paraphrased into T^1 by an LLM (B), who is also a potential author in the candidate set, it raises a question of what should be considered the ground truth. The **traditional** perspective designates the original author A as the author of T^1, while an **alternative** perspective assigns LLM B as the author. Text T^1 may substantially diverge from T^0 in style and content, potentially more similar to text G, independently generated by LLM B on the same subject. A similar scenario is also applicable to human vs. AI text detection problems. Therefore, our methodology (Figure 3) focuses on assessing the classifier's performance, considering both perspectives and exploring how variations in style and content account for the observed differences.

Dataset development: We built our dataset from the benchmark by Li et al. (2023), which features text generated by various LLMs using the same prompt, specifically instructing LLMs to continue generating text based on the first 30 words of the original human-written text. This choice enables us to explore a realistic scenario where multiple authors have written on the same subject. Given the remarkable similarity between recent LLM-generated (AI) text and human text, it is not meaningful to classify authors at the single-sentence level (Yang et al., 2023). Therefore, we selected seven datasets with paragraph-level texts from Li et al. (2023) and included one LLM from each language model family. Table 1 provides a concise overview of these datasets, the selected authors, and the paraphrasers (LLM or PM).

Each dataset is divided into a 50:50 split, allocating half for training classifiers and constructing style models and the other half for paraphrasing evaluations. To prevent the classifiers from being exclusively trained on the text's topic or content, we ensured that the train and test portions contain identical split (based on originating source) from all authors. We paraphrased each text in test portion three times, sequentially, i.e. the original text T is paraphrased once to obtain text T^1, which is then paraphrased again to get T^2 and then once again to generate T^3.

Author style model: While specific contrastive learning-based techniques (Wegmann and Nguyen, 2021; Wang et al., 2023) aim to discern context-independent style embeddings, we opted not to employ them as our style model. Firstly, these techniques operate as black-box embeddings, making it challenging to comprehend their inner workings

Dataset with Sample Size	Authors: Organization	Paraphraser
Xsum (Narayan et al., 2018): 956 news articles in various topics	**Human**: original source of writings	**ChatGPT**: we utilize the prompt "*paraphrase the following text. keep similar length*" to paraphrase any given text. We set the max length as the allowed max length and keep the other parameters as default.
TLDR: 766 articles collected from daily tech newsletter [2]	**ChatGPT** (*gpt 3.5 turbo*): OpenAI	
SCI_GEN (Moosavi et al., 2021) 944 abstracts of scientific articles	**PaLM2** (*text-bison@001*) (Anil et al., 2023): Google [3]	**PaLM2**: similar technique. PaLM2 often generates text with formatting that we remove to keep the plain text only
CMV (Tan et al., 2016): total 514 statements from r/ChangeMyView SubReddit	**LLaMA**-65B (Touvron et al., 2023): Meta	**Dipper** (Krishna et al., 2023): can paraphrase the whole text by controlling output diversity. We consider lexical_ diversity(*lex*) = 60, order_diversity(*order*)=60 as the default **dipper(moderate)** setting. We perform ablation with **dipper(high)** settings as *lex=100,order=100* and **dipper(low)** setting as *lex=20,order=20.*
WP (Fan et al., 2018): 942 stories based on prompts from r/WritingPrompts SubReddit	GLM-130B (Zeng et al., 2022): **Tshinghua**	
ELI5 (Fan et al., 2018): 954 answers from r/ExplainLikeIam5 SubReddit	**BLOOM**-7B1 (Scao et al., 2022): BigScience	**Pegasus** (Zhang et al., 2020): a sentence-wise paraphraser. We paraphrase all sentences in a text as default setting. We perform ablation study with **pegasus(slight)** variation that paraphrases random 25% sentences in a text.
YELP (Zhang et al., 2015): 986 reviews from yelp dataset	GPT-NeoX-20B (Black et al., 2022): **EleutherAI**	

Table 1: Summary of datasets, authors, and paraphrasers. ChatGPT and PaLM2 serve as both candidate authors for text generation and paraphrasers as well. Sample size indicates the original human writings that were considered for each dataset. For instance, **xsum** dataset will contain approximately 956×50%(test split)×7(authors)×4(paraphrasers)×3(times paraphrasing) ≈ 40K samples.

and ensure explainability (Angelov et al., 2021). Additionally, our aim to validate whether the drop in classification performance can be attributed to changes in style, employing a more accessible perspective. Hence, we utilized a feature-based approach, incorporating features from LIWC (Pennebaker et al., 2001), and WritePrints (Abbasi and Chen, 2008), to construct our style model for each author in individual datasets.

We validated our style model using statistical tests, confirming that, for each author, the Mahalanobis distance (McLachlan, 1999) between the style model of the original test samples and the style model from the reserved training samples of that author is lower than the distances from the training style models of other authors in that dataset. The performance of the style model as a classifier also validates its effectiveness in classifying text in its original state (details in Appendix A.2).

Content similarity: We employ *text-embedding-ada-002* from OpenAI to assess the deviation of paraphrased text from the original text's content. It is known for its high performance in the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2022) and can take lengthy texts as input (up to 8191 tokens) compared to others. Our analysis also establishes its correlation with pairwise BERT (Zhang et al., 2019) and BLUE (Papineni et al., 2002) scores between original and paraphrased texts.

4 Experimental Results

We primarily focus on evaluating the impact of paraphrasing in the context of authorship attribution, which translates into a seven-class classification problem. Our objective is not to devise new text classification methods but to investigate how ground truth influences their performance and its correlation with changes in style and content. We employ established text classification methods, including **Finetuned BERT** *(bert-base-cased)* as a representative of the finetuned language model (LM), our style model with XGBoost (Chen and Guestrin, 2016) classifier as a representation of **stylometry**, **GPT-who** (Venkatraman et al., 2023) for information density-based multi-class classification, and **TF-IDF** with logistic regression for classic n-gram-based analysis. We also explore the human vs. AI text detection scenario, a binary classification problem, using different finetuned and zero shot approaches.

Authorship attribution results: Table 2 presents the impact of different paraphrasing iterations on classifier performance in the **traditional perspective** of ground truth. A notable performance drop is observed after the first paraphrased version (from

[2] https://huggingface.co/datasets/JulesBelveze/tldr_news

[3] Original FLAN-T5 samples in (Li et al., 2023) were substantially smaller compared to other LLMs, so we replace them by generating text from PaLM2 with same prompt

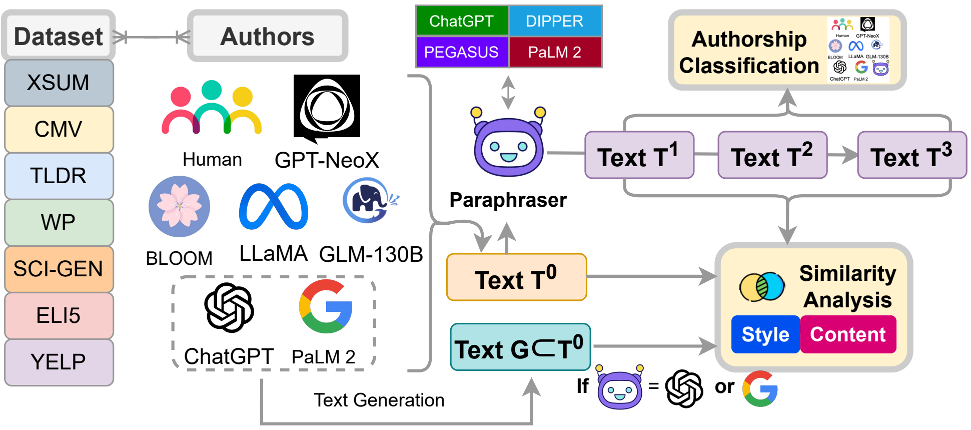

Figure 3: The original benchmark from Li et al. (2023) has several datasets with samples from different sources (human and LLMs) in each dataset. Original texts (T^0) are paraphrased sequentially three times, utilizing diverse paraphrasers (LLMs or PMs). We assess classifier performance in each iteration and measure their resemblance to T^0. For LLM paraphrasers, we additionally evaluate their similarity with text G, generated by the respective LLM.

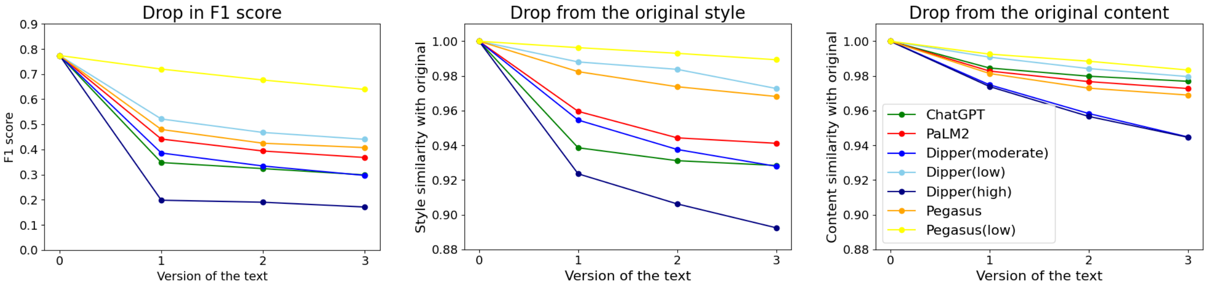

Figure 4: Comparison of classification performance (avg. macro F1 score for Fine-tuned BERT), style deviation from the original, and content shift in successive paraphrased versions across various paraphrasing methods (averaged across all datasets and sources).

Text T^n		BERT	Stylometry	GPT-who	TF-IDF
Original	**0**	**0.77**	0.71	0.62	0.61
ChatGPT	**1**	0.33 (↓57.1%)	0.32 (↓54.9%)	0.28 (↓54.8%)	**0.35** (↓42.6%)
	2	0.31 (↓6.1%)	0.29 (↓9.4%)	0.27 (↓3.6%)	**0.33** (↓5.7%)
	3	0.29 (↓6.5%)	0.27 (↓6.9%)	0.25 (↓7.4%)	**0.32** (↓3.0%)
PaLM2	**1**	**0.44** (↓42.9%)	0.43 (↓39.4%)	0.35 (↓43.5%)	0.43 (↓29.5%)
	2	0.39 (↓11.4%)	0.38 (↓11.6%)	0.32 (↓8.6%)	**0.4** (↓7.0%)
	3	0.37 (↓5.1%)	0.37 (↓2.6%)	0.3 (↓6.3%)	**0.39** (↓2.5%)
Dipper	**1**	0.38 (↓50.6%)	0.35 (↓50.7%)	0.33 (↓46.8%)	**0.44** (↓27.9%)
	2	0.33 (↓13.2%)	0.31 (↓11.4%)	0.28 (↓15.2%)	**0.38** (↓13.6%)
	3	0.29 (↓12.1%)	0.29 (↓6.5%)	0.25 (↓10.7%)	**0.35** (↓7.9%)
Pegasus	**1**	**0.55** (↓28.6%)	0.49 (↓31.0%)	0.44 (↓29.0%)	0.49 (↓19.7%)
	2	**0.49** (↓10.9%)	0.46 (↓6.1%)	0.4 (↓9.1%)	0.45 (↓8.2%)
	3	**0.47** (↓4.1%)	0.42 (↓8.7%)	0.38 (↓5.0%)	0.42 (↓6.7%)

Table 2: Performance (avg. of macro f1 score across datasets) of classifiers for various paraphrasers across different versions (**traditional** perspective). ↓ denotes performance drop from the previous version T^{n-1}.

Scenario		xsum	tldr	sci_gen	cmv	wp	eli5	yelp
Original		0.72	0.74	0.77	0.82	0.81	0.78	0.79
Chatgpt	**traditional**	0.29	0.35	0.43	0.32	0.35	0.23	0.33
	alternative	0.65	0.67	0.69	0.75	0.67	0.74	0.72
PaLM2	**traditional**	0.45	0.46	0.55	0.38	0.45	0.34	0.38
	alternative	0.62	0.66	0.68	0.71	0.7	0.71	0.71

Table 3: Performance comparison of traditional and alternative perspective with original version (T^0) of datasets for finetuned BERT after LLM paraphrasing.

original T^0 to T^1), with subsequent iterations causing marginal decreases in all cases. ChatGPT paraphrasers exhibit the most substantial performance drop, while Pegasus has the slightest effect. Additionally, it is interesting to note the performance variation among classifiers. BERT, which achieves the highest performance in original datasets (T^0), is the most affected by paraphrasing, followed by stylometry. In contrast, TF-IDF, initially the lowest performer in the original dataset, exhibits the highest F1 score when dealing with paraphrased text, except for Pegasus. This suggests that paraphrasing primarily impacts style while retaining similar content, making it challenging for classifiers to attribute samples accurately. Figure 5 delves deeper into the causes of performance drops and misclassifications of authors after paraphrasing.

Table 3 shows the results of authorship attribution for the **alternative perspective**, where authorship is considered to change after paraphrasing. Interestingly, the classification model demonstrates a markedly higher performance under this alternative perspective than the traditional viewpoint. While adopting the alternative perspective as the ground truth in all scenarios may seem appealing, it will not be a universally applicable approach, as we elaborate in the subsequent section.

Style and content similarity drop: For each paraphrase, we examine how the paraphrased text deviates from the original text across two crucial dimensions: **content** and **style**. Figure 4 illustrates the reduction in style and content similarity between the original text and its paraphrased versions, with the performance drop. We note that style deviates more substantially than content, with a substantial drop after the first paraphrasing and marginal changes in subsequent iterations, resembling the F1 score trend for all paraphrasers. A Pearson correlation test confirmed the statistically significant relation (p value<0.05) between the decline in F1 scores and a drop in style similarity. Figure 6 shows the individual breakdown of style drop for different authors and datasets.

Dataset	**Scenario**	**BERT** (finetuned)	**Detect-GPT**	**GPT-Zero**	**RoBERTa** (zero-shot)	**Long-Former**
xsum	original	**0.97**	0.66	0.67	0.3	0.9
	traditional	**0.83**	0.66	0.66	0.36	0.68
	alternative	0.6	**0.67**	**0.67**	0.33	0.36
eli5	original	**0.96**	0.67	0.71	0.34	0.86
	traditional	0.51	0.66	0.66	0.39	**0.69**
	alternative	**0.88**	0.67	0.75	0.33	0.4

Table 4: Performance of AI text detectors along different authorship perspectives after ChatGPT paraphrasing.

Figure 7 illustrates that for specific authors, following successive paraphrasing, a substantial portion of LLM paraphrased text can become more similar to the style of the LLM.

AI text detection results: Using four zero-shot AI text detectors (DetectGPT Mitchell et al., 2023, GPTZero Tian, 2023, RoBERTa, and LongFormer Li et al., 2023) and fine-tuned BERT, we assess current AI text detectors in three scenarios. As expected, fine-tuned BERT surpasses all zero-shot approaches. LongFormer exhibits the best performance among the zero-shot methods, likely due to their fine-tuning dataset's overlap with ours. Our results suggest that for formal writings, AI text detectors perform better if evaluated under the traditional setting while informal writing can be more easily detected if authorship is assumed to be with the paraphraser (Detailed results in Table 4).

5 Discussion

5.1 Major Findings

We discuss the significant findings from results that contribute essential insights to our subsequent discussion regarding paraphrasing and authorship.

Style deviates a lot more than content after paraphrasing: Our observation highlights a substantial divergence in style compared to content when paraphrasing is applied. PMs and LLMs seek to preserve the original semantics of the text while paraphrasing, and this deviation from style is why classifiers fail to correctly attribute the paraphrased versions.

LLM paraphrasers deviate the style to the LLM style model: LLM paraphrasers tend to align the paraphrased text's style with the LLM's style model (Figure 10). The paraphrased text also often exhibits greater stylistic similarity to the LLM's original text than the actual author (Figure 7). This explains the misclassification of paraphrased texts as the corresponding LLM label (Figure 5).

Subsequent paraphrasing differs for LLM and PM paraphrasers: Our findings reveal a notable performance drop and style deviation following the first paraphrasing. Though, subsequently we note a mere 3%-4% average performance decrease from the second paraphrase onward, signifying less impact. In contrast, PM paraphrasers display a consistent decline in style compared to LLM paraphrasers in subsequent versions (T^1 to T^2 and T^2 to T^3).

The choice of paraphraser impacts performance and style deviation: ChatGPT is a stronger paraphraser than PaLM2 as per both performance drop and style deviation. Additionally, our observations highlight the significance of lexical diversity, as evidenced by the variations in Dipper's low, moderate, and high versions.

Performance varies across datasets and sources: The impact of paraphrasing is milder for formal writing datasets like **xsum**, **tldr**, and **scigen** versus informal ones, such as **eli5**, **cmv**, and **yelp**.

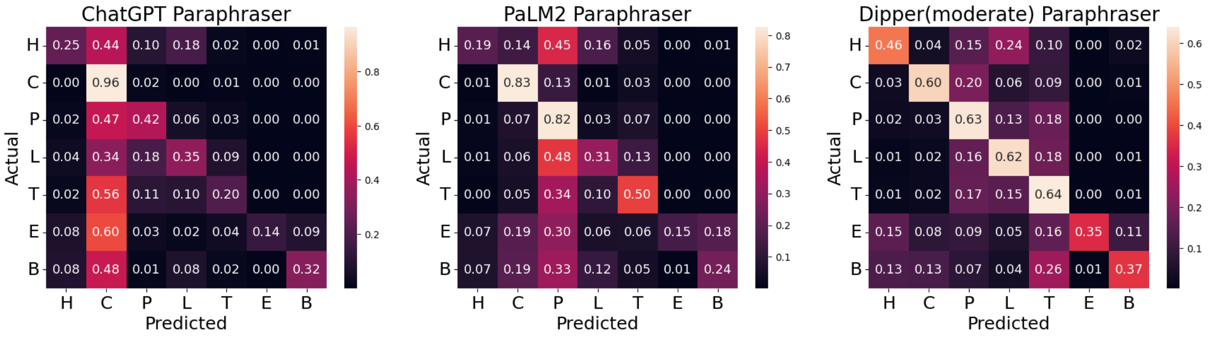

Figure 5: Confusion matrix for the Fine-tuned BERT classifier after the first version of paraphrasing (H: Human, C: ChatGPT, P:PaLM2, L: LLAMA, T: Tsinghua, E:Eleuther-AI, B: Bloom). In the case of LLM paraphrasers (ChatGPT and PaLM2), paraphrased samples are predominantly misclassified as corresponding LLMs, whereas for other PMs, such as Dipper (and Pegasus also), misclassifications are distributed more uniformly.

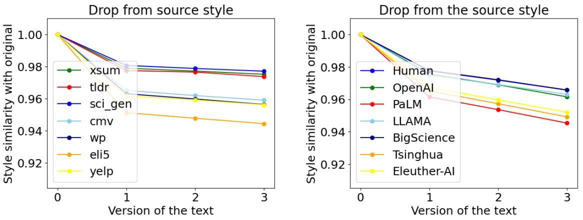

Figure 6: How style similarity drops from original text for different authors and different datasets.

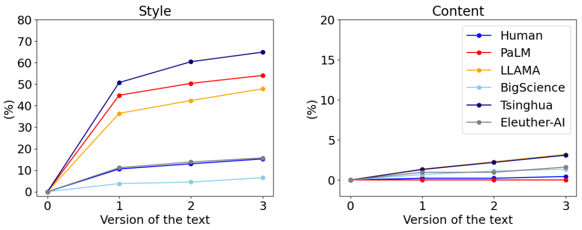

Figure 7: Percentages of ChatGPT paraphrased text (T^n) more similar to ChatGPT-generated original text (G) than the original authors' text (T^0) in style/content.

Formal writings, such as scientific abstracts and news articles, often exhibit a consistent style across sources. In contrast, informal writings, such as Reddit comments or reviews, have greater style diversity and variance. Thus, paraphrasing substantially alters the style of informal text. Also, Human and ChatGPT-generated texts maintain their original style through paraphrasing iterations, while PaLM2 and Tsinghua texts undergo substantial stylistic changes after paraphrasing.

5.2 Philosophical Perspective of Authorship

The original Ship of Theseus paradox that motivates our research remains a topic of philosophical debate without universal consensus (Pickup, 2016). Table 5 summarizes potential solution scenarios, drawing parallels to authorship scenarios and supporting theories. The notion of authorship is multifaceted and context-sensitive. Whether LLM paraphrasing should alter authorship depends on the use cases discussed below.

When "content" of text is more important: For presentations of original ideas like scientific articles, core **content** and ideas have the utmost importance. Thus, LLM paraphrasing should not alter authorship, which should remain with the original content creator, aligning with the **idea-expression dichotomy** (Samuels, 1988). For example, **current ACL policy** mentions that using tools that only assist with language, like Grammarly or spell checkers, *does not need to be disclosed*. However, the stylistic influence of LLMs like ChatGPT could raise flags with AI text detection tools, which authors should consider.

When "style" of text is more important: Expecting detectors to identify heavily paraphrased text as the source poses challenges, as LLM-paraphrased text exhibits the LLM's style. Our alternative ground truth findings further showcase this. Thus, substantial paraphrasing should change authorship, aligning with the **death of the author theory** (Barthes, 2016), and its utility as a perturbation method remains debatable. However, the ongoing cat-and-mouse game between LLMs and AI text detectors necessitates an authorship preservation metric- surpassing its threshold denotes authorship change.

When both "content" and "style" are important: Maintaining the author's unique tone and style is crucial when both the originality of content and creative expression are paramount. The

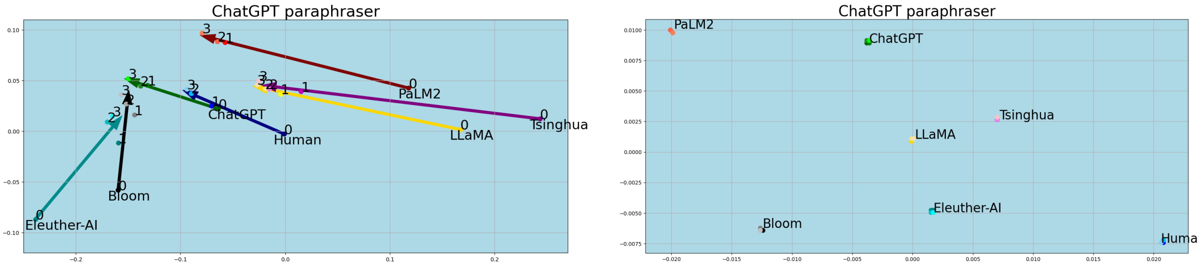

Figure 8: PCA visualization of author style (left) and content (right) shifts in **cmv** dataset for ChatGPT paraphraser. While style shifts substantially, the content of paraphrased versions remains same (details in Figure 10, Appendix).

Ship scenario	Supporting Philosophical Theory	Authorship Scenario	Supporting Writing Theory	Applicability
The original ship with *planks replaced* should be considered original	**Bundle theory** (Pike, 1967): Identity of an object is tied to the persistent of the specific characteristics that it present. As long as the ship's bundle of properties/characteristics remain unchanged, it is the same ship.	Paraphrasing **should not change** authorship	**The idea-expression dichotomy** (Samuels, 1988): Authorship is based on the author's unique expression of ideas, concepts, or thoughts. If a paraphraser modifies the expression while preserving the core ideas, the original authorship may still be attributed to the original author.	Where **content** of the text matters most, such as copyright & plagiarism of scientific articles
The new ship *formed with original planks* should be considered original	**Identity through continuity theory** (Wiggins et al., 1967): An object retains its identity if there is a continuous chain of physical connections between its various stages. Since the original planks were used to construct the new ship, it creates a direct continuity between the original ship and the new one.	Paraphrasing **should change** authorship	**The death of the author theory** (Barthes, 2016): Once a text is created, it takes on a life of its own and becomes independent of the author's intention or identity. So, the paraphrasing tool can be considered the author because it is the one actively transforming the original text into a new version.	Where **style** or probability distribution of text matters most, such as detection of AI vs human text
Both ships exist simultaneously	**Dual identity theory** (Brown, 2005): The ship will be the same in terms of its historical identity (refers to the object's history, narrative, or the sequence of events), while it's physical identity (related to the object's material composition and current state) changes due to the replacement of its planks.	Authorship as collaborative endeavor as shown in (Stillinger, 1991) & (Chen, 2011)	**Distributed authorship** (Ascott, 2005): Authorship is no longer an individual process but is instead shared among multiple entities who contribute to the creation, editing, and dissemination of a text. So, due recognition is extended to both the original author and the paraphraser.	If the use of LLM is normalized as writing improvement tool similar to the widespread integration of grammar correction tools (Ferris, 2004)

Table 5: Different Ship of Theseus "solution" and corresponding authorship scenario in case of paraphrasing

widespread use of ChatGPT or other LLMs in modifying text raises concerns about authenticity, as exemplified by ChatGPT's impact on Clarkesworld, leading to a submission suspension[4]. While employing LLMs for minor proofreading is acceptable, authors should strive to produce original content and preserve their distinctive style in a significant portion of their writing. Thus, authorship determination should rely on the flow of ideas and their articulation within the text.

Paraphrasing as AI text generation method? Paraphrasing serves as a common technique for data augmentation (Beddiar et al., 2021; Okur et al., 2022; Sharma et al., 2022; Li et al., 2022; Macko et al., 2023), particularly valuable in low-resource settings with limited data availability. Our work demonstrates that paraphrasing with an LLM like ChatGPT can align the style with that of the specific LLM. We approximated a *fixed* style for the LLM in a *specific dataset* based on samples generated with minimal prompts. However, this style is modifiable through prompting and varies in other datasets. Therefore, if LLM paraphrasing is employed for AI text generation, attribution as the author should only occur when a substantial portion of the text is independently generated, not derived from the original prompt.

6 Conclusion

In light of the increasing mainstream popularity of LLMs, this study explores the diverse notions of authorship regarding paraphrasing, inspired by the philosophical Ship of Theseus paradox. Our findings suggest that authorship should be task-dependent, and we substantiate our empirical results with theoretical and philosophical perspectives. Given the increasing prevalence of LLMs in generating and enhancing text, our research can provide a sound basis for addressing plagiarism and copyright disputes in the future involving LLMs.

[4] https://shorturl.at/gvN25

Limitations

While our study offers a comprehensive computational and philosophical exploration of the paraphrasing and authorship scenario, it is essential to acknowledge its limitations. When utilizing LLMs as paraphrasers, we employ their default settings with a generic prompt, neglecting specifically tailored instructions. LLM-paraphrased text's style can vary if instructed to generate in a particular tone or style.

A limitation of our study is its restriction to the English language. To explore how LLMs and paraphrasing tools in other languages deviate from the source style, further research and expertise in those languages are required. Furthermore, as we have shown that LLMs can shift the style to conform with the LLM style distribution, it raises the question of whether the reverse is feasible. Can humans paraphrase LLM-generated text to render it with a human-like style? This intriguing avenue warrants additional investigation, and it is critical to include the perspectives of human specialists, including linguists and computational experts, on these ambivalent concepts about authorship present in the current scenario.

Ethics Statement

This research was conducted with careful consideration of ethical principles. The tasks of this paper involve paraphrasing existing datasets from humans and LLMs, adhering to their licenses. When paraphrasing text samples, we verified for the addition of Protectable Personal Information (PII) by the paraphraser.

The potential societal impacts of this work, both positive and negative, were contemplated. On the positive side, this research aims to spur thoughtful discussion around emerging issues of authorship attribution and ownership in the age of large language models. On the negative side, the techniques presented could be misused to misattribute authorship or obfuscate plagiarism intentionally. However, promoting awareness of these capabilities will enable more informed policy decisions rather than attempts at prohibition, which are unlikely to succeed.

While observational, this study was conducted ethically and does not directly recommend for or against any particular applications of paraphrasing technology. The authors hope the insights will inform ongoing debates among scholars and policymakers about AI writing assistants' proper and fair usage. Any future research building upon these findings should continue to consider the ethical implications of how text authorship is assigned, quantified, and detected.

Acknowledgements

Removed for double-blind reviewing.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Roy Ascott. 2005. Distance makes the art grow further: Distributed authorship and telematic textuality in la plissure du texte. In Annmarie Chandler and Norie Neumark, editors, *At a Distance: Precursors to Art and Activism on the Internet*, pages 282–297. MIT Press, Cambridge, Massachusetts.

Roland Barthes. 2016. The death of the author. In *Readings in the Theory of Religion*, pages 141–145. Routledge.

Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2023. Paraphrase detection: Human vs. machine content. *arXiv preprint arXiv:2303.13989*.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108.

Janek Bevendorff, Tobias Wenzel, Martin Potthast, Matthias Hagen, and Benno Stein. 2020. On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification. *IT-Information Technology*, 62(2):99–115.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE.

Christopher Brown. 2005. *Aquinas and the ship of Theseus: solving puzzles about material objects*. A&C Black.

Shun-ling Chen. 2011. Collaborative authorship: From folklore to the wikiborg. *U. Ill. JL Tech. & Pol'y*, page 131.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Hussain A Chowdhury and Dhruba K Bhattacharyya. 2018. Plagiarism: Taxonomy, tools and detection techniques. *arXiv preprint arXiv:1801.06323*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. pages 8440–8451.

William Jay Conover. 1999. *Practical nonparametric statistics*, volume 350. john wiley & sons.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Dana R Ferris. 2004. The "grammar correction" debate in l2 writing: Where are we, and where do we go from here?(and what do we do in the meantime…?). *Journal of second language writing*, 13(1):49–62.

Tomáš Foltýnek, Terry Ruas, Philipp Scharpf, Norman Meuschke, Moritz Schubotz, William Grosky, and Bela Gipp. 2020. Detecting machine-obfuscated plagiarism. In *International conference on information*, pages 816–827. Springer.

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. MGTBench: Benchmarking Machine-Generated Text Detection. *CoRR abs/2303.14822*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *arXiv preprint arXiv:2307.03838*.

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-style language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

Goeffrey J McLachlan. 1999. Mahalanobis distance. *Resonance*, 4(6):20–26.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume PMLR 202, pages 24950–24962.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Learning to reason for text generation from scientific tables. *arXiv preprint arXiv:2104.08296*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Shashi Narayan, Shay Cohen, and Maria Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4114–4125, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Martin Pickup. 2016. A situationalist solution to the ship of theseus puzzle. *Erkenntnis*, 81:973–992.

Nelson Pike. 1967. Hume's bundle theory of the self: A limited defense. *American Philosophical Quarterly*, 4(2):159–165.

Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. *CLEF (Working Notes)*, pages 716–749.

Felicity M Prentice and Clare E Kinden. 2018. Paraphrasing tools, language translation tools and plagiarism: an exploratory study. *International Journal for Educational Integrity*, 14(1):1–16.

Rafael A Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919.

Jasper Roe and Mike Perkins. 2022. What are automated paraphrasing tools and how do we address them? a review of a growing threat to academic integrity. *International Journal for Educational Integrity*, 18(1):15.

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

Edward Samuels. 1988. The idea-expression dichotomy in copyright law. *Tenn. L. Rev.*, 56:321.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th international conference on computational linguistics*, pages 343–353.

Theodore Scaltsas. 1980. The ship of theseus. *Analysis*, 40(3):152–157.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390.

Saket Sharma, Aviral Joshi, Namrata Mukhija, Yiyun Zhao, Hanoz Bhathena, Prateek Singh, Sashank Santhanam, and Pritam Biswas. 2022. Systematic review of effect of data augmentation using paraphrasing on named entity recognition. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Jack Stillinger. 1991. *Multiple authorship and the myth of solitary genius*. Oxford University Press, USA.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Edward Tian. 2023. Gptzero. Online; accessed 23-Mar-2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Nafis Irtiza Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. Hansen: Human and ai spoken text benchmark for authorship analysis. In *Findings of Conf. on Empirical Methods in Natural Language Processing (EMNLP-Findings)*.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv e-prints*, pages arXiv–2209.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. *arXiv preprint arXiv:2310.06202*.

Jan Philip Wahle, Terry Ruas, Tomáš Foltỳnek, Norman Meuschke, and Bela Gipp. 2022a. Identifying machine-paraphrased plagiarism. In *International Conference on Information*, pages 393–413. Springer.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022b. How large language models are transforming machine-paraphrased plagiarism. *arXiv preprint arXiv:2210.03568*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *arXiv preprint arXiv:2308.11490*.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268.

David Wiggins et al. 1967. *Identity and spatio-temporal continuity*. Blackwell Oxford.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *arXiv preprint arXiv:2304.12008*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Methodological Details

This section delves into a nuanced analysis of our methodology, focusing mainly on the datasets and author style models.

A.1 Dataset examples

For a clearer understanding of multiple authors, paraphrasers, and iterations, Table 6 provides an example from our paraphrased corpus. As discussed in Section 3, we use identical samples from all authors as training samples to ensure fair training and style model creation, mitigating any potential bias from the content of the texts and making the classification task more challenging and realistic in settings where multiple authors have writings on the same topic.

While primarily used for author attribution tasks to validate both traditional and alternative perspectives of ground truths, we also leverage a subset of our datasets to address the human vs. AI text detection problem as follows.

- **Normal:** In a normal scenario, without paraphrasing, we designate T^0(human) as **human** and T^0(ChatGPT) as **AI** text for each dataset.

- **Traditional:** In the traditional setting, where paraphrasing maintains authorship, we designate ChatGPT paraphrased (after the first iteration) of the original human text, T^1(human) as **human** text, and similarly T^1(ChatGPT) as **AI** text.

- **Alternative:** In the alternative scenario, where paraphrasing alters authorship, we label ChatGPT paraphrased (after the first iteration) of the original human text, T^1(human), as **AI** text, while the original version of the human text, T^0(human), is designated as **human** text.

A.2 Validity of author style model

Section 3 explains our motivation for opting for feature-based methods to approximate a style model for any author. We substantiate our choice both statistically and based on classification performance. It is essential to note that our style model is approximated individually for each dataset. Therefore, the human style model in xsum, for instance, differs from the human style model in the cmv dataset.

Statistical significance test: We define the style model derived from each author's original test samples (T^0) as its baseline and assess its deviation from the styles in T^1, T^2, and T^3, respectively. We employ the remaining training portion for each author to validate its applicability. A robust style model should yield similar results between the style models from the training and original test samples (T^0). Mathematically, for authors A and B with respective training and test style models, their distances should adhere to the following properties:

$$\begin{aligned} |\text{train}(A) - \text{test}(A)| &< |\text{train}(B) - \text{test}(A)| \\ |\text{train}(B) - \text{test}(B)| &< |\text{train}(A) - \text{test}(B)| \end{aligned}$$

In simpler terms, an author's training and test style models should exhibit high similarity when considered as distributions. We validate this behavior using Mahalanobis distance (MD) (McLachlan, 1999), measuring the distance between a point and a distribution. For instance, to validate $|\text{train}(A) - \text{test}(A)| < |\text{train}(B) - \text{test}(A)|$, we calculate the Mahalanobis distance $MD(x, \text{train}(A))$ and $MD(x, \text{train}(B))$ for each point $x \in test(A)$. We utilize a one-sample Wilcoxon signed-rank test (Conover, 1999) to demonstrate that $MD(x, \text{train}(A)) < MD(x, \text{train}(B))$ is statistically significant for any given author A and B. The results consistently yield a p-value < 0.001 across all datasets and authors. Figure 9 illustrates that the distribution of train and test samples for any specific author appears similar in the 2D space, validating that the style model from T^0 should approximate the style of that particular author.

Classification performance: We employ the style model as a stylometry measure for author attribution. Despite its simplicity, it achieves the second-best performance in the original version (T^0), with a slight decrease compared to the best-performing Fine-tuned BERT. The ablation study (Section C) demonstrates that utilizing LIWC and WritePrint yields better results than considering them individually. Future work will focus on identifying feature importance to have a more nuanced understanding of style and how paraphrasing impacts it.

B Experimental Details

This section details our authorship attribution and AI text detection methods, encompassing a preprocessing step for compatibility. We eliminated

Author	Original (T^0)	ChatGPT paraphrased (T^1)	ChatGPT paraphrased (T^2)
Human	🟩 GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly novel domain, a task which GANs are inherently unable to do. It is also desirable to have a level of control so that there is a degree of artistic direction rather than purely curation of random results.	🟩 GANs have the ability to produce realistic images based on the data they were trained on. However, individuals who wish to use GANs for creative purposes often desire to generate images from completely new domains, which GANs are incapable of doing naturally. Additionally, it is ... S=0.983 = S′=0.983 ǀ C=0.972 > C′=0.932	🟩 GANs can generate realistic images using the provided data, but they cannot naturally create images from different domains. People who want to use GANs for creative purposes often want to generate images from new domains and have some control over the output... S=0.965 < S′=0.973 ǀ C=0.985 > C′=0.924
OpenAI	🟥 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel source, without having to manually gather and label training data. In recent years, a technique called StyleGAN has gained popularity as a way to generate novel images....	🟩 GANs have the ability to produce realistic images based on their training data. However, those who desire to use them for creative purposes often seek to generate pictures from a completely new source, without the need to manually collect and categorize data for training.... S=S′=0.994 ǀ C=C′=0.989	🟩 GANs are capable of generating realistic images based on their training data. However, individuals interested in using GANs for *creative purposes often want to generate* images from a completely new source without the need to manually collect and categorize data for training... S=S′=0.992 ǀ C=C′=0.984
PaLM2	🟩 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel domain. One way to achieve this is to use CLIP, a large language model, to provide the text prompt for the GAN....	🟩 GANs have the ability to produce realistic images based on the data they were trained on. However, individuals seeking to utilize GANs for creative purposes may desire to create images from completely new domains. To accomplish this, one method involves utilizing CLIP, a language model... S=0.9811 < S′=0.9814 ǀ C=0.987 > C′=0.885	🟥 GANs can generate realistic images using trained data, but those who want to use GANs for creative purposes may want to create images from different domains. To do this, one approach is to use CLIP, a large language model, to provide text prompts for the GAN. CLIP can generate ... S=0.974 < S′=0.979 ǀ C=0.964 > C′=0.867
LLaMA	🟩 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel domain, and this is challenging due to the narrow focus of GANs. I will discuss the phenomenon of mode collapse, the difficulty of training a GAN to generate a diverse set of images...	🟩 GANs have the ability to produce realistic images that resemble those in their training dataset. However, when using GANs for creative purposes, there is a desire to generate images that belong to a completely new domain. This poses a challenge as GANs tend to have a limited focus. In this discussion, I will explore the concept of ... S=0.971 < S′=0.980 ǀ C=0.977 > C′ = 0.882	🟥 GANs possess the capability to create authentic-looking images that resemble those found in their training dataset. However, when GANs are employed for creative purposes, there is a desire to generate images that belong to an entirely new domain. This presents a challenge since GANs tend to have a narrow focus. In this discourse ... S= 0.966 < S′=0.982 ǀ C=0.977 > C′=0.880
Tsinghua	🟩 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel domain. Unfortunately, it is often difficult to find a suitable domain for training a GAN in this manner. As a result, the image generation quality is often not satisfactory. In this paper, we propose a novel approach ...	🟥 GANs have the ability to create realistic images that belong to the same domain as their training data. However, individuals looking to use GANs for creative purposes often desire to generate images from a completely new domain. Unfortunately, finding a suitable training dataset for this purpose is often challenging, resulting in ... S=0.898 < S′=0.976 ǀ C=0.952 > C′ = 0.891	🟥 GANs possess the capability to generate realistic images within the same domain as their training data. However, when it comes to utilizing GANs for creative purposes, individuals often desire the generation of images from an entirely novel domain. Unfortunately, finding an appropriate training dataset for this ... S=0.884 < S′=0.980 ǀ C=0.946 > C′ = 0.817
Eleuther -AI	🟩 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel perspective or with new aesthetic qualities that are not present in existing photos [1,2]. In this paper we propose an approach which allows us control over "where" input ...	🟥 In this paper, we introduce a method that provides us with the ability to manipulate the input noise in GAN-generated images. This manipulation is achieved through the use of attention maps generated by self-organizing networks (SOM). Our approach allows for the generation of images from ... S=0.975 < S′=0.978 ǀ C=0.956 > C′ = 0.839	🟥 This paper presents a technique that enables us to control the input noise in images generated by GANs. By utilizing attention maps created by self-organizing networks (SOM), we are able to manipulate the noise and generate images with distinct viewpoints and artistic qualities, surpassing S=0.974 < S′=0.981 ǀ C=0.952 > C′ = 0.829
Big-Science	🟩 *GANs can generate photo-realistic images from the domain of their training data. However, those wanting to use them for creative purposes often want to generate imagery from a truly* novel viewpoint. Our paper describes an approach based on multi-view learning that enables one-to-many style transfer when generating artistic photographs using untrained DNN ...	🟥 GANs have the ability to generate realistic images based on the training data they receive. However, for creative purposes, it is often desirable to generate images from unique viewpoints. Our research paper presents a method that utilizes multi-view learning to enable one-to-many style ... S=0.981 = S′=0.981 ǀ C=0.981 > C′ = 0.901	🟥 GANs possess the capability to generate lifelike images based on the training data they receive. However, when it comes to artistic purposes, there is often a desire to produce images from unique perspectives. Our research paper introduces a method that utilizes multi-view learning to enable ... S=0.980 < S′=0.981 ǀ C=0.971 > C′ =0.875

Table 6: Example of our paraphrased corpus (**sci_gen** dataset). The original text (T^0) from each author was paraphrased subsequently by **ChatGPT** to generate T^1, T^2, ... The *italic part* of T^0 was the **prompt** for generating from other authors (LLMs). **S** and **S′** identify the style similarity with the original text version (T^0) from the corresponding author and **ChatGPT**, respectively. Similarly, **C** and **C′** show the content similarity. Green cells identify that it was correctly predicted as the respective author (Finetuned BERT), whereas red shows misclassifications. We observe **C>C′**, whereas **S** is mostly less than **S′** and decreases from the previous iteration.

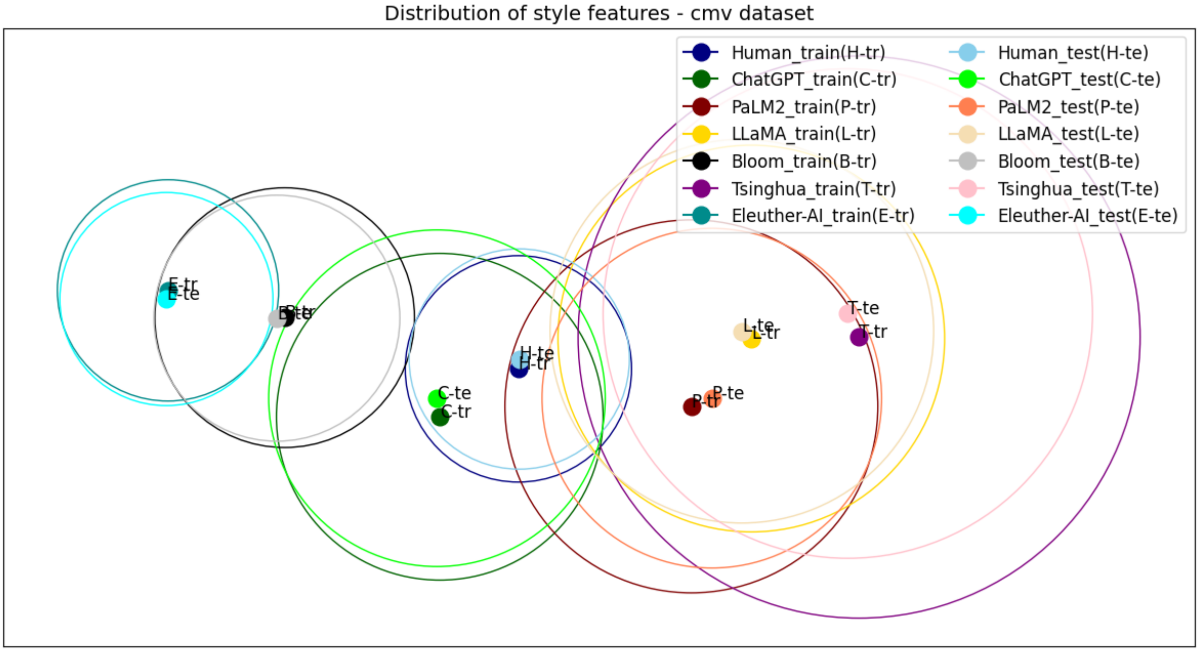

Figure 9: PCA visualization of the style features for train samples and original (T^0) test samples for different authors. The point represents the mean of the distribution, and the circle approximates the distribution (containing 90% of all samples).

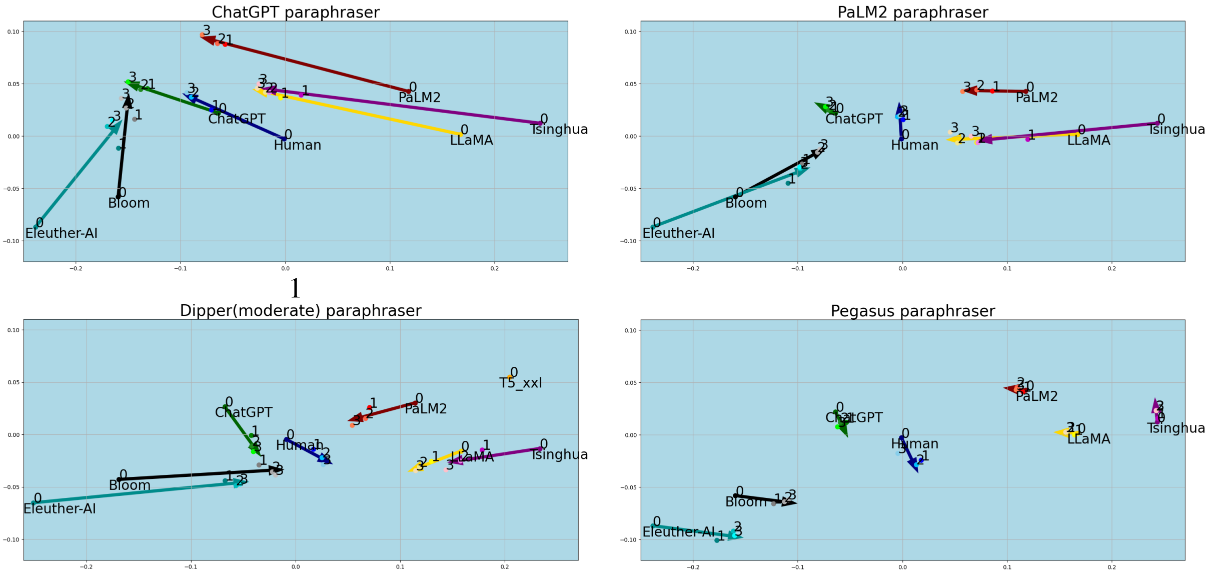

Figure 10: PCA visualization of author style shifts in **cmv** dataset using various paraphrasers. Points represent the center of samples for each author and version (0-original, 1,2,3-subsequent paraphrased). Arrows indicate style shifts. ChatGPT significantly alters the style of all authors, centralizing them around ChatGPT's style; PaLM2 exhibits a similar though less pronounced behavior. Pegasus induces minimal changes, while Dipper, despite a substantial shift, diverges from the style of its training LM, T5-xxl. However, we do not observe any content shift from the original texts for all paraphrasers, as depicted in Figure 8 for ChatGPT.

samples falling below-specified thresholds (100 for authorship attribution and 200 for AI text detection). To mitigate randomness, we also conducted the experiments five times for each text classification task, reporting the average in all tables. Figures 10 and 11 show how the style and content shift for different. paraphrasers

B.1 Authorship attribution methods

Since our authorship attribution is a seven-class text classification problem, we rely on the supervised/finetuned method for classification.

Method	xsum	tldr	sci_gen	cmv	wp	eli5	yelp
BERT(best model)	0.72	0.74	0.77	0.82	0.81	0.78	0.79
Style model	0.67	0.63	0.7	0.8	0.76	0.71	0.72
WritePrints only	0.64 (↓4.5%)	0.61 (↓3.2%)	0.68 (↓2.9%)	0.78 (↓2.5%)	0.75 (↓1.3%)	0.7 (↓1.4%)	0.7 (↓2.8%)
LIWC only	0.52 (↓22.4%)	0.48 (↓23.8%)	0.57 (↓18.6%)	0.67 (↓16.2%)	0.64 (↓15.8%)	0.56 (↓21.1%)	0.57 (↓20.8%)

Table 7: Ablation study for performance of style model in authorship attribution. ↓ denotes performance drop if a specific component is removed, compared to style model.

Finetuned BERT: As fine-tuned language models have been state of the art in text classification tasks, we fine-tune BERT (*bert-base-cased*) on each dataset training set and evaluate it on the test set.

Stylometry: We employ our style model that combines LIWC (Pennebaker et al., 2001) and WritePrint (Abbasi and Chen, 2008) features with an XGBoost classifier as the stylometry method. LIWC analyzes text using over 60 categories representing a range of social, cognitive, and affective processes. WritePrint extracts lexical and syntactic features, encompassing char, word, letter, bigram, trigram, vocabulary richness, pos-tags, punctuation, and function words. In total, we utilize 623 features.

GPT-who: GPT-who (Venkatraman et al., 2023), a psycho-linguistically-aware multi-class domain-agnostic statistical-based detector, utilizes UID-based features to capture a unique statistical signature. Initially designed for AI text detection, we repurpose it for our multi-class settings. The UID features are generated through inference from GPT-2, and an Logistic Regression (LR) model is trained on the dataset.

TF-IDF: We employ character n-grams (n=2 to 5) represented by TF-IDF scores in conjunction with an LR classifier. While n-grams excel in traditional authorship attribution tasks (Tyo et al., 2022; Tripto et al., 2023), their performance is comparatively lower in our dataset since all authors have training samples on similar topics.

B.2 AI text detection methods:

While fine-tuning language models enhances AI text detection performance on specific datasets (He et al., 2023), depending solely on this approach is not a comprehensive solution, given the rapid growth of LLMs and their generated texts. Therefore, we restrict ourselves to one fine-tuned method and incorporate mostly zero-shot/statistical detectors in our experiments.

Finetuned BERT: Like authorship attribution, we finetune our BERT model for two classes (human and AI) and evaluate performance on the test sets.

DetectGPT: DetecGPT (Mitchell et al., 2023) is a zero-shot AI text classifier that generates perturbed samples from the original text and calculates their probabilities under the model parameters. We utilize T5-3b as the mask-filling model and generate 50 samples as perturbed examples.

GPT-Zero GPT-Zero (Tian, 2023) employs perplexity to gauge the complexity of the text and Burstiness to assess variations in sentences, determining whether the text is AI-generated.

RoBERTa-large: Initially developed as the GPT-2 output detector, this model was created through fine-tuning a RoBERTa large model using the outputs of the 1.5B-parameter GPT-2 model (Conneau et al., 2020).

LongFormer: Longformer (Beltagy et al., 2020), a modified Transformer architecture, overcomes the limitations of traditional transformer models by efficiently handling more than 512 tokens. It employs an attention pattern scaling linearly with sequence length, facilitating the processing of longer documents. The Longformer used in our study (Li et al., 2023) is based on a comprehensive dataset comprising 447,674 human-written and machine-generated texts.

C Ablation Study

For ablation study, we have conducted several experiments supporting our decisions and/or findings.

Paraphrasing more than three times While the Ship of Theseus undergoes numerous modifications before posing the paradox, our experimental constraints lead us to limit paraphrasing iterations to three for most findings. Beyond the initial iterations, we observe minimal shifts in performance

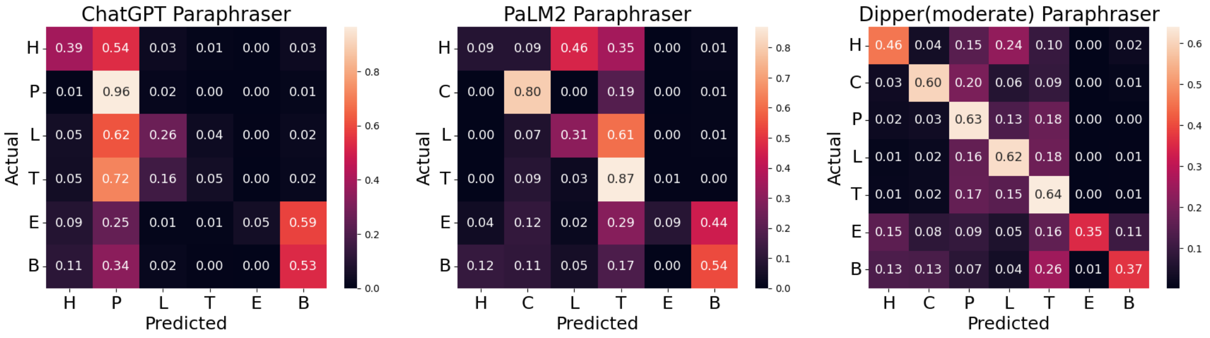

Figure 11: Confusion matrix for the Fine-tuned BERT classifier when the respective paraphraser LLM (ChatGPT or PaLM2) is left out from the training. (H: Human, C: ChatGPT, P: PaLM2, L: LLAMA, T: Tsinghua, E: Eleuther-AI, B: Bloom). Even in this scenario, misclassifications are aligned to another LLM (although different from the paraphrasing LLM) compared to the Dipper paraphraser.

Dataset	**Paraphraser**	**Text** $T^0 \Rightarrow T^1 \Rightarrow T^2 \Rightarrow T^3 \Rightarrow T^4 \Rightarrow T^5 \Rightarrow T^6 \Rightarrow T^7$
xsum	**ChatGPT**	0.72 ⇒ 0.26 ⇒ 0.24 ⇒ 0.22 ⇒ 0.21 ⇒ 0.2 ⇒ 0.21 ⇒ 0.18
	Dipper	0.7 ⇒ 0.27 ⇒ 0.27 ⇒ 0.24 ⇒ 0.22 ⇒ 0.19 ⇒ 0.2 ⇒ 0.16
cmv	**ChatGPT**	0.79 ⇒ 0.3 ⇒ 0.24 ⇒ 0.24 ⇒ 0.24 ⇒ 0.23 ⇒ 0.22 ⇒ 0.22
	Dipper	0.76 ⇒ 0.45 ⇒ 0.41 ⇒ 0.36 ⇒ 0.32 ⇒ 0.28 ⇒ 0.27 ⇒ 0.27
sci_gen	**ChatGPT**	0.74 ⇒ 0.39 ⇒ 0.33 ⇒ 0.3 ⇒ 0.31 ⇒ 0.29 ⇒ 0.31 ⇒ 0.32
	Dipper	0.73 ⇒ 0.37 ⇒ 0.22 ⇒ 0.16 ⇒ 0.18 ⇒ 0.16 ⇒ 0.11 ⇒ 0.15

Table 8: Performance of Finetuned BERT classifier up to seven paraphrasing iterations (traditional perspective).

and style/content changes. To further investigate, we conducted an ablation study by paraphrasing a subset of our datasets up to seven times. Table 8 presents the performance of the Finetuned BERT classifier (best model) after seven paraphrasing iterations by two paraphrasers (ChatGPT as an LLM paraphraser and Dipper as a PM paraphraser). The results support our choice of three iterations in experiments as sufficient, as the drop in classification performance becomes negligible for the later versions. Notably, Dipper's paraphrasing leads to a more rapid performance decline than ChatGPT.

Style model without all components In Table 7, a comparison of using different style models for authorship attribution is provided. It shows that the used combined style model is more suitable than using just the existing WritePrints or LIWC features.

Misclassifications when paraphrasing LLM is not an author While Dipper paraphraser causes slightly more performance drops and style shifts compared to ChatGPT and PaLM2 paraphrasers, its misclassifications exhibit a more uniform distribution across all classes (Figure 5) , in contrast to LLM paraphrasers. This phenomenon may be attributed to the absence of a PM-specific class label. To address this issue, we excluded ChatGPT/PaLM2 from classifier training and examined the distribution of ChatGPT/PaLM2-generated texts among other classes after classification. Figure 11 presents such authorship attribution results in the form of a confusion matrix (in comparison to Figure 5). Despite this exclusion, LLM paraphrasers still converge the style to a specific LLM, albeit different from the paraphrasing LLM, as it is excluded from the possible authors.