Can Large Language Models Identify Authorship?

Baixiang Huang[1] **Canyu Chen**[1] **Kai Shu**[2]
[1]Illinois Institute of Technology [2]Emory University
{bhuang15, cchen151}@hawk.iit.edu kai.shu@emory.edu

Abstract

The ability to accurately identify authorship is crucial for verifying content authenticity and mitigating misinformation. Large Language Models (LLMs) have demonstrated an exceptional capacity for reasoning and problem-solving. However, their potential in authorship analysis remains under-explored. Traditional studies have depended on hand-crafted stylistic features, whereas state-of-the-art approaches leverage text embeddings from pre-trained language models. These methods, which typically require fine-tuning on labeled data, often suffer from performance degradation in cross-domain applications and provide limited explainability. This work seeks to address three research questions: (1) Can LLMs perform zero-shot, end-to-end authorship verification effectively? (2) Are LLMs capable of accurately attributing authorship among multiple candidates authors (e.g., 10 and 20)? (3) Can LLMs provide explainability in authorship analysis, particularly through the role of linguistic features? Moreover, we investigate the integration of explicit linguistic features to guide LLMs in their reasoning processes. Our assessment demonstrates LLMs' proficiency in both tasks without the need for domain-specific fine-tuning, providing explanations into their decision making via a detailed analysis of linguistic features. This establishes a new benchmark for future research on LLM-based authorship analysis [1].

1 Introduction

Authorship analysis is the study of writing styles to determine the authorship of a piece of text, impacting areas from forensic investigation, such as distinguishing between murders and suicides (Chaski, 2005), to tracking terrorist threats (Winter, 2019; Cafiero and Camps, 2023). It addresses challenges in digital forensics and cybersecurity, including the fight against misinformation, impersonation, and cyber threats such as phishing and deceptive social media posts (Argamon, 2018; Shu et al., 2020; Stiff and Johansson, 2022a). Authorship analysis is essential for tracing cyber threats to their sources, combating plagiarism to uphold intellectual property rights (Stamatatos and Koppel, 2011), and identifying compromised accounts (Barbon et al., 2017). In addition, it helps link user accounts across social platforms (Shu et al., 2017; Sinnott and Wang, 2021) and detect fraudulent activities such as fake reviews (Ott et al., 2011).

Historically, authorship analysis relied on methods based on human expertise to distinguish between authors (Mosteller and Wallace, 1963). Later, a line of research known as stylometry emerged, which developed various features to quantify writing styles (Holmes, 1994). The evolution continued with the adoption of rule-based computational linguistic methods (Stamatatos, 2009). The development of statistical algorithms provides the capability to handle data with higher dimensions, enabling more expressive representations. These methods relied heavily on extensive text preprocessing and feature engineering (Bozkurt et al., 2007; Seroussi et al., 2014).

Compared to traditional statistical methods, deep learning techniques require less feature engineering. Among these techniques, pre-trained language models (PTMs) are widely used for representing authorship (Huang et al., 2024). These models, built predominantly on BERT-based architectures (Devlin et al., 2018) and contrastive learning paradigms, demonstrate efficacy in domain-specific applications. However, they fall short in cross-domain scenarios (Rivera-Soto et al., 2021). The performance of these methods also declines significantly with shorter query texts (Eder, 2015; Grieve et al., 2019) and limited data from the candidate authors. This reduction in performance limits their applicability in real-world situations,

[1]Code and data are publicly available at https://llm-authorship.github.io

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 445–460

where data scarcity and diversity are the norms. While some studies have attempted to overcome these challenges by applying text style transfer to learn content-independent style representations, they have not addressed the cross-domain issue effectively (Boenninghoff et al., 2019; Wegmann et al., 2022). These deep learning methods require extensive time and labeled data for training, are not effectively applicable across different data domains, and suffer from limited explainability.

Despite the rapid development of LLMs, there has been insufficient analysis and evaluation of their capabilities in authorship analysis (Huang et al., 2024). Some initial studies have utilized GPT-3 (Brown et al., 2020) for annotating data (Patel et al., 2023) before employing a T5 Encoder (Raffel et al., 2020) for learning representations of authorship. LLMs have demonstrated proficiency in zero-shot learning scenarios within domains lacking extensive resources (Kojima et al., 2022). However, their ability to grasp subtle nuances of language and extract critical features for authorship identification has not been extensively examined. Consequently, this paper aims to investigate the potential of LLMs for authorship identification by addressing the following research questions:

- **RQ1:** Can LLMs perform zero-shot, end-to-end authorship verification effectively?
- **RQ2:** Are LLMs capable of accurately attributing authorship among multiple candidates authors (e.g., 10 and 20)?
- **RQ3:** Can LLMs provide explainability in authorship analysis, particularly through the role of linguistic features?

We also propose a prompting technique named Linguistically Informed Prompting (LIP) to guide LLMs in identifying linguistic features that are used in practice by forensic linguists (Grant, 2022). This approach exploits the inherent linguistic knowledge embedded within LLMs, unleashing their potential to discern subtle stylistic nuances and linguistic patterns indicative of individual authorship. Figure 1 demonstrates the application of the LIP method in verifying authorship through linguistic feature analysis using GPT-4. It compares two texts from the Blog dataset (Schler et al., 2006). Analysis can provide specific linguistic evidence such as the use of informal language, punctuation patterns, and typographical errors.

Our empirical evaluation includes data with different genres and topics to validate the robustness and versatility of LLMs. The results demonstrate that LLMs can effectively perform zero-shot authorship verification and attribution, thereby obviating the need for fine-tuning. With the introduction of linguistic guidance, LLMs are further leveraged for authorship analysis, where our LIP technique sets a new benchmark for LLM-based authorship prediction. The key contributions of this work are summarized as follows:

- We conduct a comprehensive evaluation of LLMs in authorship attribution and verification tasks. Our results demonstrate that LLMs outperform existing BERT-based models in a zero-shot setting, showcasing their inherent stylometric knowledge essential for distinguishing authorship. This enables them to excel in authorship attribution and verification across low-resource domains without the need for domain-specific fine-tuning.
- We develop a pipeline for authorship analysis with LLMs, encompassing dataset preparation, baseline implementation, and evaluation. Our novel Linguistically Informed Prompting (LIP) technique guides LLMs to leverage linguistic features for accurate authorship analysis, enhancing their reasoning capabilities.
- Our end-to-end approach improves the explainability of authorship analysis. This approach elucidates the reasoning and evidence behind authorship predictions, shedding light on how various linguistic features influence these predictions. This contributes to a deeper understanding of the mechanisms behind LLM-based authorship identification.

2 Datasets

We choose two representative datasets to highlight the importance of user-generated content such as emails and social media posts. The first dataset is the Enron Email dataset, which consists of approximately half a million messages from senior Enron managers. This dataset offers insights into corporate communication, featuring long texts and a high variance in text-per-author (Klimt and Yang, 2004). The second dataset is the Blog Authorship Attribution corpus Schler et al. (2006). This corpus includes blog articles from 2004 and earlier, collected from Blogger.com, comprising over 680,000

Text 1: This is a typical PG-13 teen comedy, with a pretty original idea (to the best of my knowledge). The plot is good enough for this type of movie, but apart from the original concept, it seemed too slow and rarely very funny. The characters are believable, although a few come across as total dimwits, and the acting is good, mostly. The humor, while somewhat rare, is mostly pretty funny. There are some very entertaining court scenes near the end, which I enjoyed very much, as I'm a big fan of court dramas. The film is too slow, though, and if I had seen it at the cinema, I might have fallen asleep. I found it worth sitting through for the final court scenes, though, and I'd say that they are easily the funniest thing in the movie.

Text 2: I have to admit, I usually hate Stephen King films. I've seen The Shining, Carrie and It. I hate all of those. But this one was really good. It was scary and atmospheric, all the way through. The music was great, very moody and scary, the basic theme was pretty good, and the acting was, if not great, sufficient. The effects were mostly good, though I thought the moving corn, was, well, corny (no pun intended). It's interesting, all the way through, and has more likable main characters than your everyday horror movie. Both main characters are pretty well developed through the movie, and you actually care about them. This is a movie you should watch, at least once, if you like horror movies, even if you hate Stephen King (I do, and I loved the movie).

(a) Prompting with no guidance

Prompt: *Verify if two input texts were written by the same author*

Analysis: The input texts are not written by the same author. Text 1 is a positive review of a horror movie, while Text 2 is a review of a teen comedy.
Answer: False ❌

(b) Prompting with style guidance

Prompt: *Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content.*

Analysis: The writing styles of the two texts are quite different. Text 1 focuses on a PG-13 teen comedy movie, discussing the plot, characters, humor, and entertainment value. It uses phrases like 'too slow,' 'rarely very funny,' and 'worth sitting through for the final court scenes.' On the other hand, Text 2 discusses a Stephen King film, emphasizing its scariness, atmosphere, music, acting, and likable main characters. It uses phrases like 'scary and atmospheric,' 'likable main characters,' and 'if you like horror movies, even if you hate Stephen King.' The tone, vocabulary, and focus of the two texts are distinct, indicating different writing styles.
Answer: False ❌

(c) Prompting with grammar guidance

Prompt: *Verify if two input texts were written by the same author. Focus on grammatical styles indicative of authorship.*

Analysis: The analysis of the grammar style of the input texts suggests that they are not written by the same author. Text 1 uses more descriptive language and has longer sentences, while Text 2 uses simpler language and has shorter sentences. Additionally, Text 1 contains more punctuation marks and uses parentheses for clarification, while Text 2 does not. These differences in writing style indicate that the texts are likely written by different authors.
Answer: False ❌

(d) Linguistically Informed Prompting (LIP)

Prompt: *Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Reasoning based on linguistic features such as phrasal verbs, modal verbs, punctuation, rare words, affixes, quantities, humor, sarcasm, typographical errors, and misspellings.*

Analysis: Both texts exhibit a conversational and informal tone, with *frequent use of personal pronouns and subjective opinions*. They both use *a mix of simple and compound sentences*, and there is *a noticeable use of parenthetical statements* to add additional commentary. *Both texts also employ humor and sarcasm*, as seen in phrases like 'total dimwits' and 'corny (no pun intended)'. *The punctuation style is similar, with a liberal use of commas and occasional ellipses*. There are also *some typographical errors and informal language*, such as 'pretty original idea' and 'pretty well developed'. *Both texts use modal verbs* like 'might' and 'should', and there is *a consistent use of phrasal verbs and colloquial expressions*. Overall, the linguistic features and writing style are quite similar, suggesting that the texts were likely written by the same author.
Answer: True ✔

Figure 1: **A Comparison Between Linguistically Informed Prompting (LIP) and other Prompting Strategies for Authorship Verification**. "Analysis" and "Answer" are the output of prompting GPT-4. Only LIP strategy correctly identifies that the given two texts belong to the same author. Text colored in orange highlights the differences compared to vanilla prompting with no guidance. Text colored in blue indicates the linguistically informed reasoning process. Blue text represents the text referenced from the original documents.

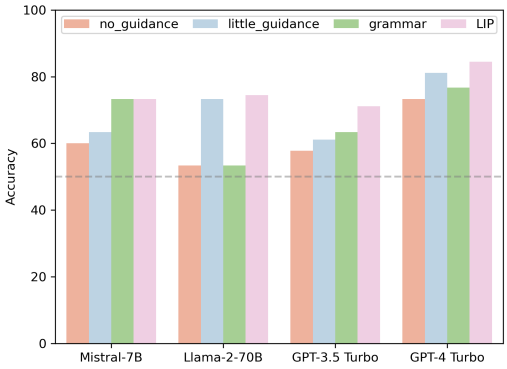


Figure 2: Authorship verification results in terms Accuracy (%) on the Blog dataset.

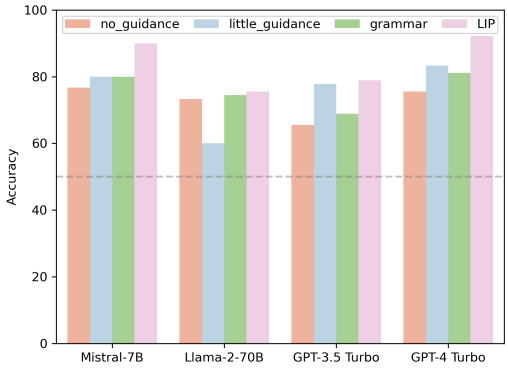


Figure 3: Authorship verification results in terms Accuracy (%) on the Email dataset.

posts from more than 19,000 authors, averaging 35 posts per author. The texts in this dataset are relatively short, with an average length of 79 tokens for the top five authors. For data preprocessing, we remove duplicate texts and authors with fewer than two texts and filter out non-English texts.

If the data formulation is not balanced, we observe that LLMs tend to predict any two given texts are written by different authors. LLMs are trained on varied datasets with many authors, which makes them better at detecting differences in writing style than similarities. The lack of multiple works by the same author in training data may prevents LLMs from learning individual authorial nuances. For our experiments, we organize and sub-sample these datasets for authorship attribution and authorship verification tasks separately. For the authorship verification task, we ensure a balanced distribution of positive and negative cases, meaning half of the texts are from the same author, and the other half are written by different authors. We sample 30 pairs of texts and conduct experiments three times for each dataset. We also make sure that all the authors and texts are unique. For the authorship attribution task, we ensure that the classes and authors are balanced: in each of the three repetitions, every query text is written by a different author, and all candidate authors are unique with only one correct author who also wrote the query text. We randomly select 10 or 20 different texts from different authors for every repetition. All texts in our sampled subsets are unique. We have saved the sampled data and used the same subsets across all baselines. The preprocessing code and sampled data are available on our GitHub repository.

3 Authorship Verification (RQ1)

Authorship verification involves assessing whether a single candidate author is the correct author of a query text. This process can be formulated as a one-class classification problem (Koppel et al., 2007). An important variant of authorship verification is to predict whether two or more given pieces of text were written by the same author or not (Koppel et al., 2012). The complexity of the author verification problem is amplified when it involves only a pair of documents for comparison. This scenario substantially limits the amount of reference material available for analysis. In this work, we address the problem of determining if two texts were authored by the same author.

The difference between "LIP", "`no_guidance`", "`little_guidance`", and "`grammar_guidance`", lies in their specificity and focus. "`no_guidance`" is the prompt that only offers a basic task description. "`little_guidance`" narrows this down by emphasizing writing style over content or topic differences, while "`grammar_guidance`" further specifies a focus on grammatical style that reflects authorship. "LIP" provides a specific list of linguistic features to analyze, such as phrasal verbs, modal verbs, punctuation, and typographical errors, ensuring a comprehensive and focused analysis. Such linguistic guidance allows for a more nuanced assessment, particularly effective in cases where an author writes on varied topics across different domain, thereby making "LIP" superior in authorship verification. Figure 6 in Appendix F shows the prompts we used for this task.

Comparing the performance of LLMs with the BERT model in our zero-shot setting is not straight-

		10 candidate authors			20 candidate authors		
Model	Prompt	Weighted F1	Macro F1	Micro F1	Weighted F1	Macro F1	Micro F1
TF-IDF		11.89	11.89	20.00	1.20	1.20	5.00
BERT		42.22	42.22	50.00	27.50	27.50	33.33
RoBERTa		34.44	34.44	43.33	23.33	23.33	28.33
ELECTRA		34.67	34.67	40.00	10.55	10.55	13.33
DeBERTa		38.89	38.89	46.67	19.09	19.09	23.33
GPT-3.5 Turbo	no_guidance	16.67	15.15	20.00	24.50	18.69	27.50
	little_guidance	27.22	24.75	33.33	37.83	27.94	37.50
	grammar	31.85	31.85	40.00	29.50	23.06	32.50
	LIP	30.56	27.78	40.00	33.33	25.20	37.50
GPT-4 Turbo	no_guidance	36.67	33.33	36.67	37.50	32.20	35.00
	little_guidance	36.67	33.33	36.67	40.83	37.12	40.00
	grammar	58.89	53.54	60.00	59.84	49.86	57.50
	LIP	**84.45**	**79.40**	**86.67**	**60.50**	**55.00**	**62.50**

Table 1: Authorship attribution results with 10 and 20 candidate authors in terms of Weighted F1(%), Macro F1(%), and Micro F1(%) on the Blog dataset.

		10 candidate authors			20 candidate authors		
Model	Prompt	Weighted F1	Macro F1	Micro F1	Weighted F1	Macro F1	Micro F1
TF-IDF		15.25	15.25	26.67	5.95	5.95	8.33
BERT		41.67	41.67	46.67	38.89	38.89	45.00
RoBERTa		36.67	36.67	43.33	22.06	22.06	26.67
ELECTRA		28.89	28.89	36.67	22.78	22.78	30.00
DeBERTa		27.22	27.22	33.33	21.67	21.67	25.00
GPT-3.5 Turbo	no_guidance	23.33	21.21	23.33	20.00	16.92	20.00
	little_guidance	28.89	26.26	30.00	26.67	24.24	26.67
	grammar	18.89	17.17	20.00	20.00	17.93	20.00
	LIP	22.22	20.20	23.33	28.89	26.26	30.00
GPT-4 Turbo	no_guidance	73.33	66.67	73.33	67.22	57.66	70.00
	little_guidance	71.67	65.15	73.33	73.33	66.67	73.33
	grammar	80.00	77.98	83.33	73.89	67.76	76.67
	LIP	**88.89**	**86.87**	**90.00**	**77.22**	**73.33**	**80.00**

Table 2: Authorship attribution results with 10 and 20 candidate authors in terms of Weighted F1(%), Macro F1(%), and Micro F1(%) on the Email dataset.

forward. This is because we instruct LLMs to directly output their final answer, as illustrated in the system instructions from Figure 6 in Appendix. In contrast, traditional pre-trained models, such as BERT, require fine-tuning along with a prediction head (typically a trained machine learning classifier) to map hidden embeddings to the final output. Our experiments indicate that employing cosine similarity with BERT embeddings, which are predominantly distributed around 0.9, makes it challenging to distinguish authorship for verification purposes in this zero-shot setting. The experimental results for both TF-IDF and BERT are provided in Appendix D.

Evaluation metrics for this task include accuracy, precision, recall, and F1 Score. Accuracy is a fundamental metric that measures the proportion of correct predictions out of the total predictions made. Precision focuses on the proportion of true positive predictions within the pool of positive predictions, evaluating the model's ability to avoid false positives. Recall assesses the model's ability to identify all actual positives, reflecting its capability to minimize false negatives. The F1 Score harmonizes precision and recall by providing a single metric that balances both aspects through the calculation of their harmonic mean.

The experiment results are demonstrated in Figure 2 and 3. They provide a comparative analysis of the performance of LLMs in authorship verification tasks across two different datasets. Four models are evaluated with four different prompt settings. GPT-4 Turbo consistently outperforms the other models in both datasets, indicating its superior capability in understanding authorship. The LIP method generally yields the highest scores across all metrics for

most models. Across both datasets, performance metrics improve as the level of prompt guidance increases from no guidance to LIP. This trend underscores the importance of linguistic guidance in leveraging LLMs for authorship verification. The analysis of these figures reveals that the effectiveness of LLMs in authorship verification tasks can be significantly influenced by the type of prompt guidance provided.

4 Authorship Attribution (RQ2)

In this section, we present a comprehensive analysis of experiments conducted to evaluate the efficacy of our proposed models on the zero-shot authorship attribution task. We selected the Blog and Enron email datasets to ensure a robust assessment across different domains and genres. Figure 7 in Appendix F shows the prompts of this task, we utilize four prompt similar to the authorship verification task, with LIP being the most effective due to its linguistic guidance effect. The experiments were structured to compare the performance of LLMs not only against each other but also against established benchmarks in the field, such as TF-IDF and BERT-based models.

Authorship attribution, the task of determining the most likely author of a given text from a set of candidates, is commonly formulated as a multi-class, single-label text classification problem. Tables 1 and 2 provide a overview of the performance of various models. These models were evaluated across two different datasets (Blog and Email) with varying numbers of candidate authors (10 and 20).

Weighted F1, micro F1, and macro F1 are used as evaluation metrics. Weighted F1 gives an average F1 score weighted by class size. Micro F1 calculates the overall average F1 score, combining all classes, and is sensitive to class imbalance. Macro F1 computes the unweighted average of F1 scores across classes, treating each class equally, ideal for assessing minority class performance.

We also tested Llama 2 and Mistral 7B. However, input texts from the datasets we used for evaluating other LLMs are too long and exceed the context limit of Llama 2 because of their context length limitations (4k tokens for Llama 2 and 8k for Mistral, versus 16k for GPT-3.5 Turbo and 128k for GPT-4 Turbo). Therefore the experiment on Mistral are shown in a separate table (Table 3).

Table 4 outlines the results of an ablation study focused on evaluating the impact of various linguistic features on the performance of LLMs. This study examines how the exclusion of specific features affects the models' abilities. Features such as affixes and quantities are crucial for GPT-4 Turbo, while misspellings hold more significance for GPT-3.5 Turbo. We use all of these linguistic features provided by forensic linguistics (Grant, 2022), and find that LLMs perform optimally when all of these features are provided, as shown in the LIP technique in Table 1, allowing LLMs to determine which features to utilize.

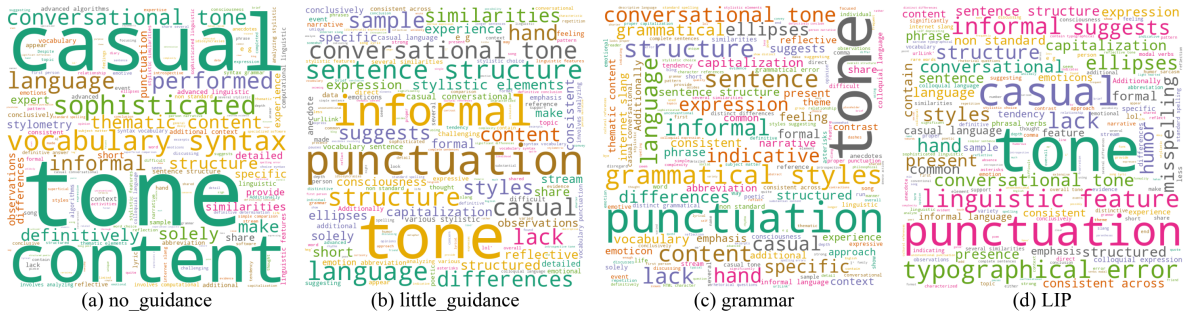
The experiment results highlight the superiority of GPT-4 Turbo over BERT-based language models and basic statistical approaches, such as TF-IDF. These advancements not only demonstrate significant improvements in scores but also show robustness against increased task complexity. The incorporation of linguistic guidance for LLMs markedly improves performance and generates more explainable authorship analysis. This progression emphasizes the importance of adopting LLMs for com-

Dataset	Prompt	Weighted F1	Macro F1	Micro F1
Blog	no_guidance	10.00	9.09	13.33
	little_guidance	6.89	6.26	10.00
	grammar	7.22	6.57	10.00
	LIP	10.56	9.90	13.33
Email	no_guidance	22.22	20.20	26.67
	little_guidance	22.45	20.40	26.67
	grammar	15.00	13.64	20.00
	LIP	29.44	28.53	33.33

Table 3: Mistral's performance on the authorship attribution task with 10 candidate authors.

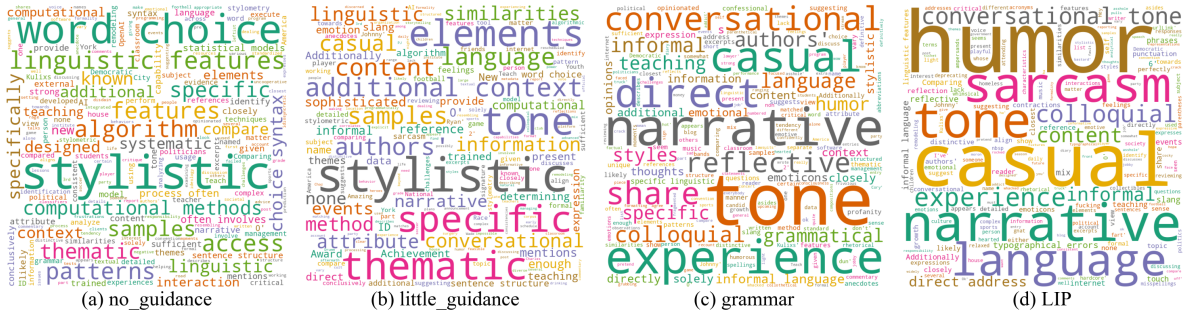
Model	Prompt	Weighted F1	Macro F1	Micro F1
GPT-3.5	phrasal verbs	22.67	20.61	30.00
	modal verbs	20.95	20.04	26.67
	punctuation	23.06	22.60	33.33
	rare words	26.00	23.64	33.33
	affixes	23.00	20.91	30.00
	quantities	19.44	18.23	30.00
	humor	23.22	21.66	33.33
	sarcasm	23.89	21.72	33.33
	typos	24.67	23.36	33.33
	misspellings	28.33	26.67	40.00
GPT-4	phrasal verbs	62.22	56.57	63.33
	modal verbs	56.67	51.52	56.67
	punctuation	71.11	67.27	73.33
	rare words	62.22	56.57	63.33
	affixes	75.56	71.32	76.67
	quantities	75.56	71.32	76.67
	humor	66.67	60.61	70.00
	sarcasm	72.22	65.66	73.33
	typos	55.56	50.51	56.67
	misspellings	46.67	42.42	46.67

Table 4: Ablation study on the impact of 10 linguistic features for the Blog dataset (with 10 candidate authors).



(a) no_guidance (b) little_guidance (c) grammar (d) LIP

Figure 4: Visualization of GPT-4's Authorship Verification Analysis on the Blog dataset with four Guidance Levels.



(a) no_guidance (b) little_guidance (c) grammar (d) LIP

Figure 5: Visualization of GPT-4's Authorship Attribution Analysis on the Blog dataset with four Guidance Levels.

plex authorship analysis tasks. Key observations are summarized as follows:

- Across both datasets and both sets of candidate authors, GPT-4 consistently outperforms traditional models and those using GPT-3.5 in all metrics. Specifically, linguistic-guided GPT-4 approaches show remarkable improvements, indicating the effectiveness of these advanced models in understanding and attributing authorship with higher precision. There is a clear trend of performance improvement when transitioning from traditional models and BERT-based models to LLMs. This suggests that more recent and sophisticated models are significantly better at handling the complexities of authorship attribution.

- The increase in the number of candidate authors from 10 to 20 generally results in a decline in performance across all models. However, the decline is less pronounced for LLMs, suggesting that they is better equipped to handle increased complexity and maintain higher levels of performance.

- Traditional pre-trained language models such as BERT, RoBERTa, DeBERTa, and ELECTRA exhibit considerable variability in their performance, suggesting they require domain-specific fine-tuning and need to be combined with other classifiers to perform this task effectively. TF-IDF shows the lowest performance across all metrics in both datasets.

5 Explainability (RQ3)

LLMs can provide explanations in natural language regarding their decision-making processes and provide linguistic evidence that references the original text, which is invaluable for verifying and understanding their conclusions. Our LIP method improves upon this by generating meaningful analyses of writing styles, which are instrumental in distinguishing between authors. We have made all the input texts and the corresponding authorship analyses, as produced by LLMs, accessible in our GitHub repository.

Figure 4 displays word clouds that represent the outputs from GPT-4 during the authorship verification task. These visualizations demonstrate the model's focus and specificity in analysis, where the density and size of terms indicate their prominence in the output. As guidance increases from

none to LIP, there is a clear shift from general and diverse terms to more specific linguistic features. The word cloud for LIP, being the most effective, underscores a thorough analysis by highlighting particular linguistic characteristics.

Similarly, the word clouds in Figure 5 illustrate that LLMs can offer in-depth explanations for authorship attribution tasks. The effectiveness and focus of these explanations can be significantly improved through explicit linguistic guidance, which directs the model to base its decisions on linguistic attributes used in practice (Grant, 2022). The word cloud of the LIP method prominently features terms such as "humor", "sarcasm", "casual", and "colloquial." This demonstrates that with LIP, the LLM is steered towards making decisions grounded in linguistic features, especially high-level and complex features such as humor and sarcasm. The specificity achieved through the LIP method highlights the model's ability to provide clear and focused explanations for its authorship decisions, offering a notable improvement over traditional methods that rely on hidden embeddings. The enhanced clarity in the LLM's outputs not only facilitates a better understanding of the decision-making process but also has the potential to increase the reliability of the authorship analysis process.

6 Related Work

In this section, we review the literature on traditional and contemporary methods of authorship analysis, as well as research on utilizing LLMs for authorship analysis and related tasks.

6.1 Authorship Analysis

The primary goal of authorship analysis is to analyze writing styles to determine authorship. It encompasses two main tasks: authorship attribution and verification. Authorship attribution, also known as authorship identification, aims to attribute a previously unseen text of unknown authorship to one of a set of known authors. Authorship verification involves determining whether a single candidate author wrote the query text by comparing text similarities (Koppel et al., 2007). This process requires establishing whether a query text was written by a specific author, compared to a set of their known works. Authorship attribution can be broken down into a series of authorship verification instances, focusing on measuring text similarity based on stylistic features. We specifically focus on closed-set authorship attribution, which deals with a predetermined, finite list of potential authors that always includes the true author of a query text. Authorship verification can also be seen as a specific case of authorship attribution, but with only one potential author.

Central to these tasks is the extraction of useful authorship features from textual data using natural language processing methods such as n-grams (Sharma et al., 2018), POS-tags (Sundararajan and Woodard, 2018), topic modeling (Seroussi et al., 2014), and Linguistic Inquiry and Word Count (LIWC) (Uchendu et al., 2020). More recently, the focus has shifted towards extracting embeddings from text, considering both content and style while often disregarding external contextual cues. These embeddings, serving as a numeric representation of a text segment, facilitate further analysis. When comparing a document embedding with another from the same author, the representation tends to orient toward the author's style rather than the document's content (Huertas-Tato et al., 2022).

Barlas and Stamatatos (2020) found that BERT models perform well when dealing with large vocabularies, outperforming multi-headed RNNs. Fabien et al. (2020) fine-tuned a BERT model for authorship attribution. They showed that incorporating stylometric and hybrid features into an ensemble model enhances its performance. Huertas-Tato et al. (2022) introduced a semi-supervised contrastive learning approach using a BERT-based model for cross-domain authorship attribution and profiling. Rivera-Soto et al. (2021) also explored cross-domain authorship representation learning through contrastive learning, revealing that neural authorship representations learned by deep learning models, such as Sentence-BERT (SBERT), are not universal. They concluded that topic diversity and the size of the training dataset are crucial for effective zero-shot cross-domain transfer. For instance, models trained on the Reddit comments (Baumgartner et al., 2020) exhibited significantly better transfer than those trained on the Amazon Reviews corpus (Ni et al., 2019) and the Fanfiction dataset (Bevendorff et al., 2020). Deep learning methods, despite their potential, require substantial training time and labeled data, offer limited generalization capabilities, and lack explainability. In contrast, our approach, which leverages the intrinsic linguistic knowledge and zero-shot reasoning abilities of LLMs, does not require fine-tuning and is effective in low-resource domains.

6.2 Large Language Models

Large Language Models (LLMs) excel at text generation, achieving a level of fluency and coherence that closely mimics human writing. Hence, numerous studies have focused on differentiating LLM-generated text from human-written text using various machine learning methods (Huang et al., 2024; Uchendu et al., 2020; Tang et al., 2023; Wu et al., 2023; Yang et al., 2023). In comparison, our research evaluates LLMs' capabilities in authorship verification and attribution, which are complex reasoning tasks. Unlike pre-trained language models (PTMs) that often require specific fine-tuning for different tasks, LLMs have an inherent capacity for reasoning and problem-solving. This is leveraged through instruction-based few-shot or zero-shot learning, allowing them to effectively conduct reasoning tasks with minimal examples (Brown et al., 2020; Kojima et al., 2022).

The application of LLMs in authorship analysis, particularly in authorship attribution and authorship verification, is rarely explored. Traditional methods have primarily used LLMs for auxiliary tasks, such as data extraction and annotation, rather than fully utilizing their capabilities (Patel et al., 2023). In contrast, our work is pioneering in exploring LLMs' end-to-end potential for authorship analysis tasks. This not only demonstrates the versatility and effectiveness of LLMs in complex linguistic tasks but also sets a new benchmark for future research in the field.

Moreover, this novel application of LLMs in authorship analysis aims to overcome the limitations of traditional methods, such as extensive feature engineering. Unlike BERT-based models, which require computationally expensive fine-tuning and large amounts of domain-specific data for optimal performance (Grieve et al., 2019), LLMs can generalize across various domains without any fine-tuning, addressing the issue of domain specificity (Barlas and Stamatatos, 2020). They are also capable of handling shorter texts, reducing the need for long inputs to derive meaningful representations (Eder, 2015). A key advantage of our LLM-based approach is its ability to provide understandable natural language explanations for its predictions, addressing the lack of transparency in traditional models' hidden text embeddings (Rivera-Soto et al., 2021). This improvement in explainability and versatility represents a significant advancement in overcoming the challenges related to data, domain specificity, text length requirements, and explainability faced by earlier methods.

7 Conclusion

This paper explores how to leverage LLMs for authorship analysis. Through comprehensive evaluation, it demonstrates that LLMs, equipped with the novel Linguistically Informed Prompting (LIP) technique, excel at identifying authorship without the need for domain-specific fine-tuning. By directly applying our end-to-end methods to authorship attribution and verification tasks, we aim to bypass the intermediate steps of feature extraction and manual annotation. This approach not only surpasses traditional and state-of-the-art methods in performance, especially in zero-shot and low-resource settings, but also enhances the explainability of authorship predictions by illuminating the role of linguistic features. The findings underscore the potential of LLMs to revolutionize authorship analysis, offering robust solutions for digital forensics, cybersecurity, and combating misinformation. This work paves the way for future research and applications in LLM-based authorship prediction.

8 Limitations

Scalability with Increasing Number of Authors The effectiveness of the method when the number of candidate authors increases is a major limitation. In real-world scenarios, especially in contexts like social media and large forums, the number of potential authors can be vast. If the model's performance degrades with more candidates, this restricts its utility in broader applications. Another potential limitation is the evaluation of machine-generated text for authorship analysis, particularly as machine-generated content becomes more common and sophisticated. Our method may not effectively distinguish between human-authored and machine-generated texts.

Explainability Although authorship analysis by LLMs offers a level of explainability through the linguistic features or insights highlighted during the analysis, the mechanistic interpretability of how these decisions are made at the neuronal level within the LLMs is not explored. This means that while we can observe the decisions that are made, the fundamental neural activities and interactions that lead to these decisions remain a black box. This lack of deeper explainability can be a drawback, particularly in critical applications where un-

derstanding the precise reasoning process is necessary for trust and verification.

9 Ethics Statement

The potential to reveal the identities of anonymous authors presents an ethical challenge. The paper discusses applications such as linking user accounts across platforms and identifying compromised accounts. These applications raise privacy concerns and ethical questions about surveillance and the profiling of individuals based on their writing style. The use of such methods must be carefully managed to protect individual privacy and adhere to ethical standards, particularly in sensitive areas such as journalism, political dissent, or corporate whistleblowing. Ensuring that authorship attribution methods are not used to undermine privacy rights or expose individuals to risks without their consent is crucial.

Acknowledgments

We thank our anonymous reviewers for their constructive feedback and recommendations. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04, NSF awards (SaTC-2241068, IIS-2339198, and POSE-2346158), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security and the National Science Foundation.

References

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE.

Shlomo Argamon. 2018. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law/Linguagem e Direito*, 5(2):7–37.

Sylvio Barbon, Rodrigo Augusto Igawa, and Bruno Bogaz Zarpelão. 2017. Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach. *Multimedia Tools and Applications*, 76:3213–3233.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*, pages 255–266. Springer.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Alimohammad Beigi, Zhen Tan, Nivedh Mudiam, Canyu Chen, Kai Shu, and Huan Liu. 2024. Model attribution in machine-generated disinformation: A domain generalization approach with supervised contrastive learning. *arXiv preprint arXiv:2407.21264*.

Nicole Mariah Sharon Belvisi, Naveed Muhammad, and Fernando Alonso-Fernandez. 2020. Forensic authorship analysis of microblogging texts using n-grams and stylometric features. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. 2020. Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 372–383. Springer.

Benedikt Boenninghoff, Robert M Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019. Similarity learning for authorship verification in social media. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2457–2461. IEEE.

Ilker Nadi Bozkurt, Ozgur Baghoglu, and Erkan Uyar. 2007. Authorship attribution. In *2007 22nd international symposium on computer and information sciences*, pages 1–5. IEEE.

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

Florian Cafiero and Jean-Baptiste Camps. 2023. Who could be behind qanon? authorship attribution with supervised machine-learning. *arXiv preprint arXiv:2303.02078*.

Carole E Chaski. 2005. Who's at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1):1–13.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, Xifeng Yan, William Yang Wang, Philip Torr, Dawn Song, and Kai Shu. 2024. Can editing llms inject harm? *arXiv preprint arXiv: 2407.20224*.

Canyu Chen and Kai Shu. 2024a. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Canyu Chen and Kai Shu. 2024b. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *ArXiv preprint*, abs/2211.05289.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maciej Eder. 2015. Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Tim Grant. 2022. *The Idea of Progress in forensic authorship analysis*. Cambridge University Press.

Jack Grieve, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. 2019. Attributing the bixby letter using n-gram tracing. *Digital Scholarship in the Humanities*, 34(3):493–512.

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 542–556.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

David I Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28:87–106.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *arXiv preprint arXiv: 2408.08946*.

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. Part: Pre-trained authorship representation transformer. *arXiv preprint arXiv:2209.15373*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*, pages 217–226. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. 2012. The "fundamental problem" of authorship attribution. *English Studies*, 93(3):284–291.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(6).

Moshe Koppel, Jonathan Schler, and Eran Messeri. 2008. Authorship attribution in law enforcement scenarios. *NATO Security Through Science Series D-Information and Communication Security*, 15:111.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. Plagiarism detection without reference collections. In *Advances in Data Analysis: Proceedings of the 30 th Annual Conference of the Gesellschaft für Klassifikation eV, Freie Universität Berlin, March 8–10, 2006*, pages 359–366. Springer.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ajay Patel, Delip Rao, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting llms. *arXiv preprint arXiv:2305.12696*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Abhay Sharma, Ananya Nandan, and Reetika Ralhan. 2018. An investigation of supervised learning methods for authorship attribution in short hinglish texts using char & word n-grams. *arXiv preprint arXiv:1812.10281*.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: Concepts, methods, and recent advancements. *Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities*, pages 1–19.

Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter*, 18(2):5–17.

Richard Sinnott and Zijian Wang. 2021. Linking user accounts across social media platforms. In *2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT'21)*, pages 18–27.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos and Moshe Koppel. 2011. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation*, 45:1–4.

Harald Stiff and Fredrik Johansson. 2022a. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

Harald Stiff and Fredrik Johansson. 2022b. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395.

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. *arXiv preprint arXiv:2204.04907*.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Jana Winter. 2019. Exclusive: Fbi document warns conspiracy theories are a new domestic terrorism threat. *Yahoo News*, 1.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023. A survey on llm-gernerated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*.

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*.

A Impact Statement

Authorship verification and attribution play an essential role in various applications such as combating misinformation (Shu et al., 2020; Chen and Shu, 2024a,b; Chen et al., 2022, 2024; Hanley and Durumeric, 2024; Stiff and Johansson, 2022b; Beigi et al., 2024), protecting intellectual property rights (Meyer zu Eissen et al., 2007; Stamatatos and Koppel, 2011), identifying fraudulent activities (Ott et al., 2011; Afroz et al., 2012), tracking terrorist threats (Winter, 2019; Cafiero and Camps, 2023), and aiding general criminal investigations (Koppel et al., 2008; Argamon, 2018; Belvisi et al., 2020).

B Future Work

The advent of LLMs has complicated the problem of authorship attribution since it is increasingly challenging to distinguish between LLM-generated and human-written texts (Huang et al., 2024). The hardness of differentiating the content produced by humans and machines potentially undermines the integrity of authorship, threatens the credibility of digital content and endangers safety of online space (Solaiman et al., 2023; Vidgen et al., 2024). More effort is desired to protect human authorship from the threat of LLM-generated content.

C Experiment Setup

The baselines used in this paper include: TF-IDF, pre-trained language models like BERT (`bert-base-uncased`) (Devlin et al., 2018), RoBERTa (`roberta-base`) (Liu et al., 2019), DeBERTa (`microsoft/deberta-base`) (He et al., 2020), and ELECTRA (`google/electra-base-discriminator`) (Clark et al., 2020), alongside LLMs represented by GPT-3.5 Turbo (`1106-preview`) and GPT-4 Turbo (`1106-preview`). We use GPT-3.5 Turbo (`1106-preview`) and GPT-4 Turbo (`1106-preview`) through the Microsoft Azure OpenAI API, setting the temperature to 0 for all our experiments. We conducted both authorship verification and attribution experiments three times and calculated the average score. We use py3langid [2] to filter out non-English texts. For running the quantized versions of Llama 2 (Llama-2-70B-chat-GPTQ) (Frantar et al., 2023) and Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023), we utilize an NVIDIA RTX A6000 with 48 GB of GPU memory. Both models are configured with the temperature set to 0 and top_p set to 1.

Dataset	Name	Accuracy	Precision	Recall	F1
Blog	TF-IDF	53.33	100.00	6.67	12.50
	BERT	50.00	50.00	100.00	66.67
Email	TF-IDF	73.33	100.00	46.67	63.64
	BERT	50.00	50.00	100.00	66.67

Table 5: Authorship Verification results on the Blog and the Email Dataset for BERT and TF-IDF.

D Additional Results

A challenge in evaluating zero-shot authorship verification is comparing our approach with conventional models, which often rely on trained classifiers for classification tasks. To ensure a fair comparison, we adapt these models to fit within a zero-shot framework. To establish a comparison, we consider null accuracy, which is 50% in a perfectly balanced dataset. Our experiments suggest that using cosine similarity scores of BERT embeddings are mostly distributed around 0.9 We use a threshold of 0.5, where above 0.5 means the same authorship, and vice versa.

The results shown in Table 5 mean that the BERT model exhibits a tendency to classify each pair of texts as having been authored by the same individual, resulting in a notably high recall rate. In contrast, the TF-IDF approach is characterized by high precision paired with low recall. This indicates that the model predominantly identifies pairs as being written by different authors.

E Scientific Artifacts

We use open-source scientific artifacts in this work, including pandas (Wes McKinney, 2010), pytorch (Paszke et al., 2019), HuggingFace transformers (Wolf et al., 2020), sklearn (Pedregosa et al., 2011), and NumPy (Harris et al., 2020).

F Prompt Design

This section provides details about the prompt we used for authorship verification (Figure 6) and attribution tasks (Figure 7). Including the system and user instructions for four levels of prompt designs including "LIP", `"no_guidance"`, `"little_guidance"`, and `"grammar_guidance"`.

[2] https://github.com/adbar/py3langid

AUTHORSHIP VERIFICATION:

System instruction: Respond with a JSON object including two key elements:
"analysis": Reasoning behind your answer.
"answer": A boolean (True/False) answer.

Prompting with no guidance: Verify if two input texts were written by the same author. Input text 1: <text 1>, text 2: <text 2>

Prompting with style guidance: Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Input text 1: <text 1>, text 2: <text 2>

Prompting with grammar guidance: Verify if two input texts were written by the same author. Focus on grammatical styles indicative of authorship. Input text 1: <text 1>, text 2: <text 2>

Linguistically Informed Prompting (LIP): Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Reasoning based on linguistic features such as phrasal verbs, modal verbs, punctuation, rare words, affixes, quantities, humor, sarcasm, typographical errors, and misspellings. Input text 1: <text 1>, text 2: <text 2>

Figure 6: **Prompt Design for the Authorship Verification Task**.

AUTHORSHIP ATTRIBUTION:

System instruction: Respond with a JSON object including two key elements:
"analysis": Reasoning behind your answer.
"answer": The query text's author ID.

Prompting with no guidance: Given a set of texts with known authors and a query text, determine the author of the query text. Input query text: <query text>; Texts from potential authors: <candidate texts>

Prompting with style guidance: Given a set of texts with known authors and a query text, determine the author of the query text. Do not consider topic differences. Input query text: <query text>; Texts from potential authors: <candidate texts>

Prompting with grammar guidance: Given a set of texts with known authors and a query text, determine the author of the query text. Focus on grammatical styles. Input query text: <query text>; Texts from potential authors: <candidate texts>

Linguistically Informed Prompting (LIP): Given a set of texts with known authors and a query text, determine the author of the query text. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Focus on linguistic features such as phrasal verbs, modal verbs, punctuation, rare words, affixes, quantities, humor, sarcasm, typographical errors, and misspellings. Input query text: <query text>; Texts from potential authors: <candidate texts>

Figure 7: **Prompt Design for the Authorship Attribution Task**. "query text" is the text whose authorship needs to be identified. "candidate texts" are a collection of texts written by each potential author, which is a JSON object formatted with author IDs as keys and values containing the texts written by them.