A Bayesian Approach to Harnessing the Power of LLMs in Authorship Attribution

Zhengmian Hu[1,2*], **Tong Zheng**[1*], **Heng Huang**[1]

[1]Department of Computer Science, University of Maryland, College Park, MD 20742
[2]Adobe Research

huzhengmian@gmail.com, zhengtong12356@gmail.com, heng@umd.edu,

Abstract

Authorship attribution aims to identify the origin or author of a document. Traditional approaches have heavily relied on manual features and fail to capture long-range correlations, limiting their effectiveness. Recent advancements leverage text embeddings from pre-trained language models, which require significant fine-tuning on labeled data, posing challenges in data dependency and limited interpretability. Large Language Models (LLMs), with their deep reasoning capabilities and ability to maintain long-range textual associations, offer a promising alternative. This study explores the potential of pre-trained LLMs in one-shot authorship attribution, specifically utilizing Bayesian approaches and probability outputs of LLMs. Our methodology calculates the probability that a text entails previous writings of an author, reflecting a more nuanced understanding of authorship. By utilizing only pre-trained models such as Llama-3-70B, our results on the IMDb and blog datasets show an impressive 85% accuracy in one-shot authorship classification across ten authors. Our findings set new baselines for one-shot authorship analysis using LLMs and expand the application scope of these models in forensic linguistics. This work also includes extensive ablation studies to validate our approach.

1 Introduction

Authorship attribution, the process of identifying the origin or author of a document, has been a longstanding challenge in forensic linguistics. It has numerous applications, including detecting plagiarism (Alzahrani et al., 2011) and attribution of historical text (Silva et al., 2023). As the digital age progresses, the need for reliable methods to determine authorship has become increasingly important, especially in the context of combating misinformation spread through social media and conducting forensic analysis. The ability to attribute authorship can also lead to challenges around privacy and anonymity (Juola et al., 2008).

The field traces its roots back to the early 19th century (Mechti and Almansour, 2021), with early studies focusing on stylistic features and human expert analysis (Mosteller and Wallace, 1963). Traditional methods often relied on stylometry, which quantifies writing styles (Holmes, 1994), and rule-based computational linguistic methods (Stamatatos, 2009) to deduce authorship. Later, statistical algorithms incorporating extensive text preprocessing and feature engineering (Bozkurt et al., 2007; Seroussi et al., 2014) were introduced to improve accuracy. However, these methods often struggled with capturing long-range dependencies in text and require careful setup of specific thresholds for various indicators, which can be challenging to select effectively. They also involve designing complex, high-quality features, which can be costly and time-consuming.

The advent of deep learning has transformed the landscape of authorship attribution by turning the problem into a multi-class classification challenge, allowing for the capture of more features and addressing more complex scenarios effectively (Ruder et al., 2016; Ge et al., 2016; Shrestha et al., 2017; Zhang et al., 2018). However, these neural network (NN) models often lack interpretability and struggle with generalization in cases of limited samples.

Despite advancements, the field still faces significant challenges. Obtaining large, balanced datasets that represent multiple authors fairly is difficult, and as the number of authors increases, the accuracy of machine learning models tends to decrease.

On the other hand, language models, central to modern NLP applications, define the probability of distributions of words or sequences of words and have traditionally been used to predict and generate plausible language. Yet, for a long time, these

*These authors contributed equally to this work.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13216–13227
November 12-16, 2024

models, including high-bias models like bag-of-words and n-gram models, struggled to fit the true probability distributions of natural language. Deep learning's rapid development has enabled orders of magnitude scaling up of computing and data, facilitating the use of more complex models such as Random Forests (Breiman, 2001), character-level CNNs (Zafar et al., 2020), Recurrent Neural Networks (Bagnall, 2015), and Transformer (Vaswani et al., 2017).

The recent rapid evolution of Large Language Models (LLMs) has dramatically improved the ability to fit natural language distributions. Trained on massive corpora exceeding 1 trillion tokens, these models have become highly capable of handling a wide range of linguistic tasks, including understanding, generation, and meaningful dialogue (Liang et al., 2022; Bubeck et al., 2023; Zhang et al., 2023a, 2024). They can also explain complex concepts and capture subtle nuances of language. They have been extensively applied in various applications such as chatbots, writing assistants, information retrieval, and translation services. More impressively, LLMs have expanded their utility to novel tasks without additional training, simply through the use of prompts and in-context learning (Brown et al., 2020). This unique ability motivates researchers to adapt LLMs to an even broader range of tasks and topics including reasoning (Wei et al., 2022), theory of mind (Kosinski, 2023) and medical scenario (Singhal et al., 2023).

Interestingly, language models have also been explored for authorship attribution (Agun and Yilmazel, 2017; Le and Mikolov, 2014; McCallum, 1999). Recently, research has utilized LLMs for question answering (QA) tasks within the application of authorship verification and authorship attribution (Huang et al., 2024), though these have primarily been tested in small-scale settings. Other approaches have attempted to leverage model embeddings and fine-tuning for authorship attribution, such as using GAN-BERT (Silva et al., 2023) and BERTAA (Fabien et al., 2020). However, these techniques often face challenges with scalability and need retraining when updating candidate authors. Moreover, they require relatively large dataset and multiple epochs of fine-tuning to converge. Given the challenges with current approaches, a natural question arises: *How can we harness LLMs for more effective authorship attribution?*

Two aspects of evidence provide insights to answer the above questions. First, recent studies on LLMs have shown that these models possess hallucination problems (Ji et al., 2023). More interestingly, the outputs of LLMs given prompts may disagree with their internal thinking (Liu et al., 2023). Therefore, it is advisable not to rely solely on direct sampling result from LLMs. Second, the training objective of LLMs is to maximize the likelihood of the next token given all previous tokens. This indicates that probability may be a potential indicator for attributing texts to authors.

Language models are essentially probabilistic models, but we find the probabilistic nature of LLMs and their potential for authorship identification remains underexploited. Our study seeks to bridge this gap. Specifically, we explore the capability of LLMs to perform one-shot authorship attribution among multiple candidates.

We propose a novel approach based on a Bayesian framework that utilizes the probability outputs from LLMs. By deriving text-level log probabilities from token-level log probabilities, we establish a reliable measure of likelihood that a query text was written by a specific author given example texts from each candidate author. We also design suitable prompts to enhance the accuracy of these log probabilities. By calculating the posterior probability of authorship, we can infer the most likely author of a document (Figure 1). Due to the pivotal role of log probability in our algorithm, we coined our approach the "Logprob method."

Our new method has three main advantages:

- No Need for Fine-Tuning: Our approach aligns the classification task with the pretraining objective, both focusing on computing entailment probability. This avoids any objective mismatch introduced by fine-tuning. Moreover, our method leverages the inherent capabilities of pretrained LLMs and avoids knowledge forgetting that often occurs during fine-tuning.

- Speed and Efficiency: This approach requires only a single forward pass through the model for each author, making it significantly faster and more cost-effective compared to normal question-answering method of language models which involves sampling a sequence of tokens as answer, with one forward pass for each token generated.

- No Need for Manual Feature Engineering: The pre-training on diverse data enables LLMs to au-

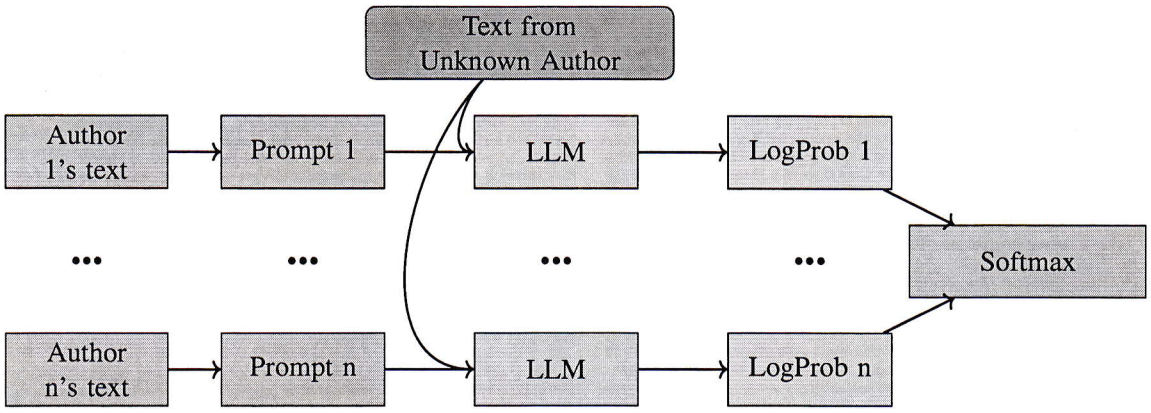

Figure 1: Illustration of bayesian authorship attribution using LLM.

tomatically capture and utilize subtle nuances in language, thus eliminating the need for manually designing complex features, which can be costly and time-consuming.

By applying this technique, we have achieved state-of-the-art results in one-shot learning on the IMDb and blog datasets, demonstrating an impressive 85% accuracy across ten authors. This advancement establishes a new baseline for one-shot authorship analysis and illustrates the robust potential of LLMs in forensic linguistics.

2 Method

Our approach to authorship attribution is based on a Bayesian framework. Given a document whose authorship is unknown, our objective is to identify the most probable author from a set using the capabilities of Large Language Models (LLMs).

We consider a scenario where we have a set of authors $\mathcal{A} = \{a_1, \ldots, a_n\}$ and a set of all possible texts $\mathcal{E}$. Given an authorship attribution problem, where each author a_i has written a set of texts $t_{i,1}, t_{i,2}, \ldots, t_{i,m_i} \in \mathcal{E}$, we denote the collection of known texts of an author a_i as $\boldsymbol{t}(a_i) = (t_{i,1}, t_{i,2}, \ldots, t_{i,m_i})$. For an unknown text $u \in \mathcal{E}$, we aim to determine the most likely author from the set $\mathcal{A}$.

To estimate the author of text u, we use a Bayesian framework where the probability that u was written by author a_i is given by:

$$P(a_i|u) = \frac{P(u|a_i)P(a_i)}{P(u)}. \tag{1}$$

Here, $P(a_i)$ is the prior probability of each author, assumed to be equal unless stated otherwise, making the problem focus primarily on estimating $P(u|a_i)$.

Assuming that each author a_i has a unique writing style represented by a probability distribution $P(\cdot|a_i)$, texts written by a_i are samples from this distribution. To estimate $P(u|a_i)$, we consider the independence assumption: texts by the same author are independently and identically distributed (i.i.d.). Thus, the unknown text u is also presumed to be drawn from $P(\cdot|a_i)$ for some author a_i and is independent of other texts from that author.

Notice that although texts are independent under the i.i.d. assumption when conditioned on a particular author, there exists a correlation between the unknown text u and the set of known texts $\boldsymbol{t}(a)$ in the absence of knowledge about the author. This correlation can be exploited to deduce the most likely author of u using the known texts.

Specifically, we have

$$\begin{aligned} P(u|\boldsymbol{t}(a_i)) &= \sum_{a_j \in \mathcal{A}} P(u, a_j|\boldsymbol{t}(a_i)) \\ &= \sum_{a_j \in \mathcal{A}} P(u|a_j, \boldsymbol{t}(a_i))P(a_j|\boldsymbol{t}(a_i)) \\ &= \sum_{a_j \in \mathcal{A}} P(u|a_j)P(a_j|\boldsymbol{t}(a_i)), \end{aligned} \tag{2}$$

where the last equality uses the i.i.d. assumption, meaning that when conditioned on a specific author a_j, u is independent of other texts.

We then introduce the "sufficient training set" assumption, where:

$$P(a_j|\boldsymbol{t}(a_i)) = \begin{cases} 1 & a_i = a_j \\ 0 & a_i \neq a_j. \end{cases} \tag{3}$$

This implies that the training set is sufficiently

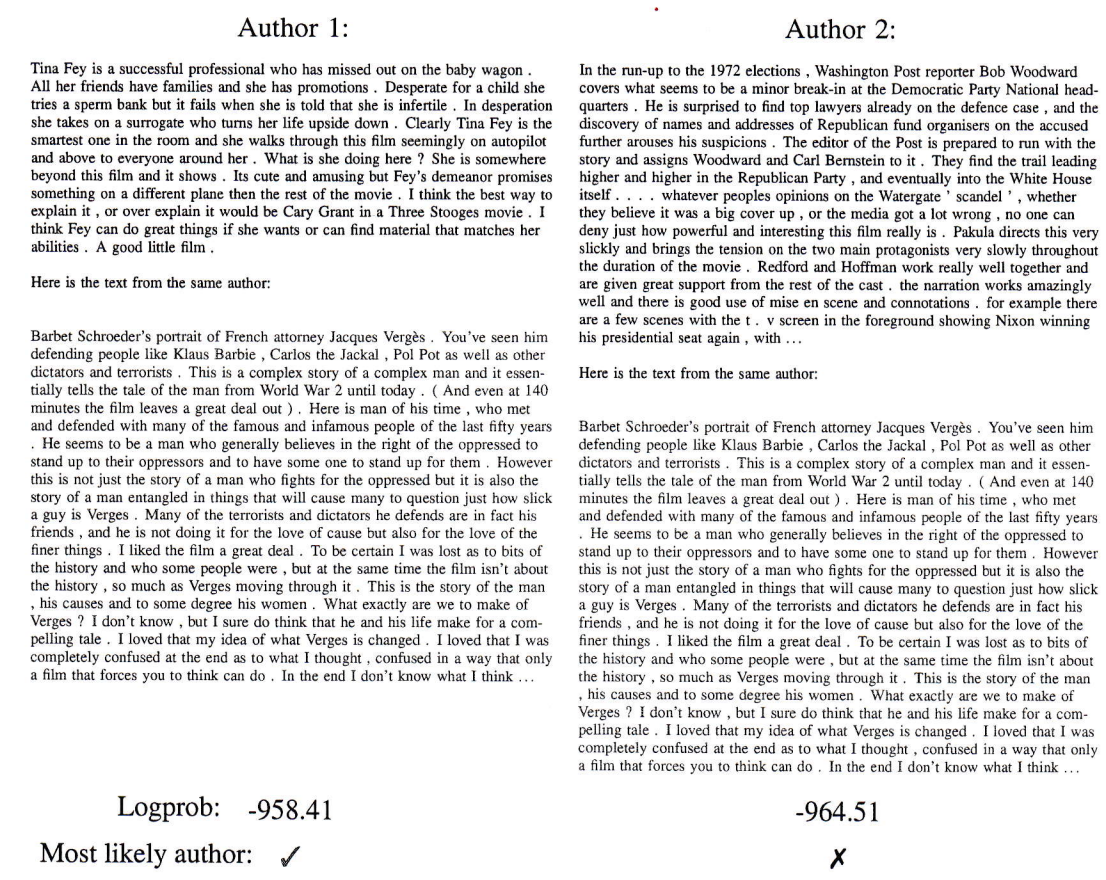

Figure 2: Example of prompt construction and authorship attribution based on log probabilities. The logprob is computed on the orange part, which represents the text from unknown author.

comprehensive to unambiguously differentiate authors, leading to:

$$P(u|\boldsymbol{t}(a_i)) = P(u|a_j), \tag{4}$$

where a_j is the assumed true author of text u.

We use Large Language Models (LLMs) to estimate $P(u|\boldsymbol{t}(a_i))$, which represents the probability that a new text u was written by the author of a given set of texts $\boldsymbol{t}(a_i)$.

The probability nature of language models means that they typically calculate the probability of a token or a sequence of tokens given prior context. For a vocabulary set Σ, the input to a language model might be a sequence of tokens $x_1, \ldots, x_m \in \Sigma$, and the model's output would be the probability distribution $P_{\text{LLM}}(\cdot|x_1, \ldots, x_m)$, typically stored in logarithmic scale for numerical stability.

When using an autoregressive language model, we can measure not only the probability of the next token but also the probability of a subsequent sequence of tokens. For instance, if we have a prompt consisting of tokens $x_1, \ldots, x_m \in \Sigma$, and we want to measure the probability of a sequence $y_1, \ldots, y_s \in \Sigma$, we calculate:

$$\begin{aligned} &P_{\text{LLM}}(y_1, \ldots, y_s|x_1, \ldots, x_m) \\ =&\prod_{i=1}^{s} P_{\text{LLM}}(y_i|x_1, \ldots, x_m, y_1, \ldots, y_{i-1}). \end{aligned} \tag{5}$$

To estimate $P(u|\boldsymbol{t}(a_i))$ for authorship attribution, we define:

$$\begin{aligned} &P(u|\boldsymbol{t}(a_i)) \\ =&P_{\text{LLM}}(u|\,\text{prompt\_construction}(\boldsymbol{t}(a_i))). \end{aligned} \tag{6}$$

The prompt construction can vary, providing flexibility in how we use the model to estimate probabilities. Our method involves constructing a prompt steering the LLM uses to predict the likelihood that the unknown text was written by the same author (Figure 2).

In summary, our approach is straightforward and simple. By leveraging the capabilities of Large Language Models, we calculate the likelihood that an unknown text originates from a known author based on existing samples of their writing. This

probability assessment allows us to identify the most likely author from a set without the need for fine-tuning or feature engineering.

3 Experimental Setups

3.1 Models & Baselines

Models We selected two widely-used LLM families: 1) LLaMA family, which includes LLaMA-2 (Touvron et al., 2023), LLaMA-3, CodeLLaMA (Roziere et al., 2023), available in various parameter sizes and configurations, with some models specifically fine-tuned for dialogue use cases; 2) the GPT family (Brown et al., 2020), featuring GPT-3.5-Turbo and GPT-4-Turbo (Achiam et al., 2023), where we specifically used versions gpt-4-turbo-2024-04-09 and gpt-3.5-turbo-0125. The LLaMA family models were deployed using the vLLM framework (Kwon et al., 2023) if used for Logprob method and are deployed on Azure if used for question-answering. Apart from Table 1, all ablation studies of Logprob method uses LLaMA-3-70B model.

Baselines We chose two types of baselines for comparison. 1) embedding-based methods such as BertAA (Fabien et al., 2020) and GAN-BERT (Silva et al., 2023), which require training or fine-tuning, 2) LLM-based methods such as those described in (Huang et al., 2024), which utilize LLMs for authorship attribution tasks through a question-answering (QA) approach.

3.2 Evaluations

Datasets We evaluated our method on two widely used author attribution datasets: 1) IMDB62 dataset, a truncated version of IMDB dataset (Seroussi et al., 2014) and 2) Blog Dataset (Schler et al., 2006). IMDB62 dataset comprises 62k movie reviews from 62 authors, with each author contributing 1000 samples. Additionally, it also provides some extra information such as the rating score. The Blog dataset, contains 681k blog comments, each with an assigned authorID. Besides the raw text and authorID, each entry includes extra information such as gender and age. Both datasets are accessible via HuggingFace.

Benchmark Construction Unlike fixed author sets used in many previous studies, we constructed a random author set for each test to minimize variance. By default, unless specified otherwise, each experiment in our experiments involved a 10-author one-shot setting, and we conducted 100 tests for each experiment to reduce variance. Each test involved the following steps: 1) Ten candidate authors were randomly selected. 2) For each author, one (or n for n-shot) article was randomly selected as the training set. 3) One author was randomly selected from the ten candidates as the test author. 4) One article not in the training set was randomly selected from the test author's articles as the test set (with size of 1). 5) We run the authorship attribution algorithm to classify the test article into 10 categories.

Our evaluation pipeline can avoid potential biases from fixed author sets and better measure the efficacy of LLMs in authorship attribution tasks. We also share our pipeline for fair evaluations of future related works.

Notably, aforementioned pipeline is suitable for non-training based methods like ours and QA approaches. However, for training-based methods such as embedding approaches, each train-test split is followed by a retraining, demanding significant computational resources. Therefore, in this work, we directly cited scores from the original papers.

Evaluation Metrics We adopt three metrics: top-1, top-2 and top-5 accuracies. Specifically, top k accuracy is computed as follows:

$$\text{Top } k \text{ Accuracy} = \frac{\text{Num}^{\text{k}}_{\text{correct}}}{\text{Num}_{\text{all}}}, \qquad (7)$$

where $\text{Num}^{\text{k}}_{\text{correct}}$ represents the number of tests where the actual author is among the top k predictions, and Num_{all} represents the total number of tests.

4 Experiments

Firstly, we evaluate different methods for author attribution in Section 4.1, noting that our Logprob method significantly outperformed QA-based methods in accuracy and stability across datasets. Then, we study the impact of increasing candidate numbers on performance in Section 4.2, where our method maintained high accuracy despite a larger pool of candidates. Next, in Section 4.3, we analyze prompt sensitivity, concluding that while prompt use is crucial, variations in prompt design did not significantly affect the performance. Further, in Section 4.4, we explore bias in author attribution and in Section 4.5, we measure performance variations across different subgroups. Finally, in

Section 4.6, we compared the efficiency of different author attribution methods.

4.1 Author Attribution Performance

Table 1 shows the main results for different methods on the IMDB62 and Blog datasets concerning authorship attribution capabilities. We make the following observations:

- **LLMs with QA-based methods cannot perform author attribution tasks effectively.** For example, GPT-4-Turbo can only achieve a top-1 accuracy of 34% on the IMDB62 dataset and 62% on the Blog dataset. Notably, there are two interesting phenomena: 1) GPT-4-Turbo and GPT-3.5-Turbo exhibit inconsistent higher accuracy across different datasets, highlighting inherent instability in the prompt-based approach. 2) Older LLMs with smaller context window lengths are unable to perform author attribution due to the prompt exceeding the context window. These phenomena indicate that QA methods are not a good option for enabling LLMs to conduct author attribution tasks effectively.

- **Our Logprob method helps LLMs perform author attribution tasks more effectively.** With LLaMA-3-70B, we achieved top-1 accuracy of 85%, and both top-2 and top-5 accuracies were even higher. This suggests that LLMs equipped with our method can effectively narrow down large candidate sets. Additionally, two another things worth noting are that 1) LLMs with the Logprob method exhibit more stable performance across both tasks, something QA methods struggle with, and 2) LLMs with Logprob can conduct authorship attribution tasks with lower requirements for context window length. For instance, LLaMA-2-70B-Chat with the Logprob method can handle authorship attribution, whereas the same model with a QA approach fails when the collective text of 10 authors exceeds the context window length. These findings highlight the superiority of our Logprob method.

- **Training-free method can achieve comparable or even superior performance to training-based methods.** The Blog dataset showed higher top-1 accuracy with LLaMA + Logprob compared to GAN-BERT and BertAA. While the IMDB62 dataset exhibited lower performance relative to embedding-based methods, it is important to note that Logprob achieves this as a one-shot method, whereas embedding-based approaches require much more data for training to converge. This demonstrates that Logprob can more effectively capture the nuances necessary for authorship attribution.

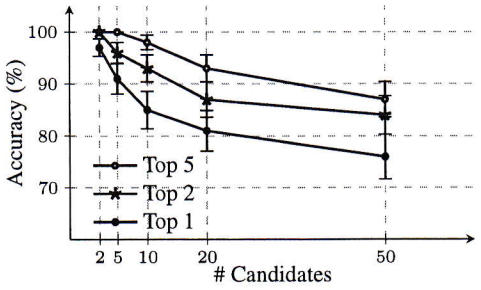

Figure 3: Accuracy vs. number of candidates.

4.2 Performance vs. Number of Candidates

One of the challenges in authorship attribution is the difficulty in correctly identifying the author as the number of candidates increases, which generally leads to decreased accuracy. Figure 3 shows the author attribution performance across different candidate counts on the IMDB62 dataset. We made the following observations:

- First, performance indeed decreases as the number of candidates increases.

- Second, across all settings, all metrics maintain relatively high scores. For example, in the setting with 50 candidates, our method achieved 76% top-1 accuracy, 84% top-2 accuracy, and 87% top-5 accuracy.

- Third, top-2 and top-5 accuracies are more stable compared to top-1 accuracy. The model may not always place the correct author at the top, but it often includes the correct author within the top few predictions. This attribute is also crucial as it allows the narrowing down of a large pool of candidates to a smaller subset of likely candidates.

4.3 Analysis of Prompt Sensitivity

Our method relies on suitable prompt as in Figure 2. Here, we discuss the sensitivity of our accuracy to different prompt constructions in Table 2. We made the following observations:

- **Using prompts is essential for enhancing the accuracy of our method (#1 vs. #2).** This phenomenon is aligned with previous studies

Method	Model	IMDB62 Dataset					BLOG Dataset				
		#Candidate	n-Shot	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.	#Candidate	n-Shot	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.
LogProb	LLaMA-2-7B	10	1	80.0 ± 4.0	88.0 ± 3.3	97.0 ± 1.7	10	1	79.0 ± 4.1	84.0 ± 3.7	**98.0 ± 1.4**
	LLaMA-2-7B-Chat	10	1	68.0 ± 4.7	80.0 ± 4.0	88.0 ± 3.3	10	1	69.0 ± 4.6	78.0 ± 4.1	89.0 ± 3.1
	LLaMA-2-13B	10	1	84.0 ± 3.7	88.0 ± 3.3	**100.0 ± 0.0**	10	1	81.0 ± 3.9	86.0 ± 3.5	94.0 ± 2.4
	LLaMA-2-70B	10	1	**88.0 ± 3.3**	**94.0 ± 2.4**	**99.0 ± 1.0**	10	1	**88.0 ± 3.3**	**90.0 ± 3.0**	95.0 ± 2.2
	LLaMA-2-70B-Chat	10	1	79.0 ± 4.1	85.0 ± 3.6	95.0 ± 2.2	10	1	83.0 ± 3.8	85.0 ± 3.6	**97.0 ± 1.7**
	Code-LLaMA-7B	10	1	71.0 ± 4.5	84.0 ± 3.7	96.0 ± 2.0	10	1	78.0 ± 4.1	84.0 ± 3.7	94.0 ± 2.4
	Code-LLaMA-13B	10	1	70.0 ± 4.6	84.0 ± 3.7	98.0 ± 1.4	10	1	77.0 ± 4.2	85.0 ± 3.6	92.0 ± 2.7
	Code-LLaMA-34B	10	1	75.0 ± 4.3	84.0 ± 3.7	98.0 ± 1.4	10	1	78.0 ± 4.1	83.0 ± 3.8	94.0 ± 2.4
	LLaMA-3-8B	10	1	82.0 ± 3.8	89.0 ± 3.1	98.0 ± 1.4	10	1	84.0 ± 3.7	**89.0 ± 3.1**	95.0 ± 2.2
	LLaMA-3-8B-Instruct	10	1	69.0 ± 4.6	77.0 ± 4.2	90.0 ± 3.0	10	1	68.0 ± 4.7	77.0 ± 4.2	90.0 ± 3.0
	LLaMA-3-70B	10	1	**85.0 ± 3.6**	**93.0 ± 2.6**	98.0 ± 1.4	10	1	82.0 ± 3.8	**88.0 ± 3.3**	95.0 ± 2.2
	LLaMA-3-70B-Instruct	10	1	79.0 ± 4.1	89.0 ± 3.1	**99.0 ± 1.0**	10	1	79.0 ± 4.1	**87.0 ± 3.4**	**96.0 ± 2.0**
QA	LLaMA-2-70B-Chat	10	1	Failed	-	-	10	1	Failed	-	-
	LLaMA-3-70B-Instruct	10	1	31.0 ± 4.6	-	-	10	1	22.0 ± 4.1	-	-
	GPT-3.5-Turbo	10	1	69.0 ± 4.6	-	-	10	1	47.0 ± 5.0	-	-
	GPT-4-Turbo	10	1	34.0 ± 4.7	-	-	10	1	62.0 ± 4.9	-	-
Other Baseline	GAN-BERT	20	80	96.0	-	-	20	80	40.0	-	-
	BertAA	62	80	93.0	-	-	10	80	65.0	-	-

Table 1: Author attribution results on IMDB62 and Blog dataset. Prompt construction for QA method is in consistent with Huang et al. (2024).

Prompt 1: Here is the text from the same author:
Prompt 2: Analyze the writing styles of the input texts, disregarding the differences in topic and content. Here is the text from the same author:
Prompt 3: Focus on grammatical styles indicative of authorship. Here is the text from the same author:
Prompt 4: 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. Here is the text from the same author:

#	Prompting	Top 1 Accuracy	Top 2 Accuracy	Top 5 Accuracy
1	<Example Text> + <Query Text>	70.0 ± 4.6	81.0 ± 3.9	92.0 ± 2.7
2	<Example Text> + <Prompt 1> + <Query Text>	85.0 ± 3.6	92.0 ± 2.7	99.0 ± 1.0
3	<Example Text> + <Prompt 2> + <Query Text>	83.0 ± 3.8	87.0 ± 3.4	100.0 ± 0.0
4	<Example Text> + <Prompt 3> + <Query Text>	86.0 ± 3.5	90.0 ± 3.0	100.0 ± 0.0
5	<Example Text> + <Prompt 4> + <Query Text>	87.0 ± 3.4	90.0 ± 3.0	99.0 ± 1.0

Table 2: Author attribution performance vs. different prompting choices on IMDB62 dataset.

Gender	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.
Both	84.0 ± 1.6	90.8 ± 1.3	95.8 ± 1.0
Male	81.4 ± 2.5	88.6 ± 2.1	95.4 ± 1.4
Female	**86.3 ± 2.1**	**92.8 ± 1.6**	**96.2 ± 1.2**

Table 3: Gender bias in author attribution performance.

Gender	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.
Male	77.0 ± 4.2	82.0 ± 3.8	92.0 ± 2.7
Female	**89.0 ± 3.1**	**91.0 ± 2.9**	**95.0 ± 2.2**

Table 4: Author attribution performance in each gender subgroup.

(Wei et al., 2022) that have demonstrated that prompting is beneficial for unlocking the full potential of LLMs.

- **There is no statistically significant evidence to suggest that specific prompt designs impact performance significantly (#2 vs. #3 vs. #4 vs. #5).** The results show very close performance metrics across different prompt constructions.

Discussions Prompting sensitivity (Sclar et al., 2023) is a widely acknowledged property in the generation process of LLMs. This also has motivated a trend of research on prompting engineering (Zhang et al., 2023b; Guo et al., 2024) as different promptings can lead to completely different performance. However, our method appears to be relatively insensitive to the choice of prompt, which makes our method more robust, maintaining high performance and stability across various settings.

4.4 Bias Analysis

An algorithm trained on an entire dataset may exhibit different accuracy levels across different subgroups during testing (Chouldechova and G'Sell, 2017; Pastor et al., 2021). This section discusses such bias issues and measures how the algorithm's accuracy varies for different subgroups.

Interval	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.
$[1-2]$	82.0 ± 3.8	89.0 ± 3.1	96.0 ± 2.0
$[3-4]$	87.0 ± 3.4	94.0 ± 2.4	99.0 ± 1.0
$[5-6]$	$\mathbf{90.0 \pm 3.0}$	$\mathbf{96.0 \pm 2.0}$	$\mathbf{100.0 \pm 0.0}$
$[7-8]$	88.0 ± 3.3	92.0 ± 2.7	97.0 ± 1.7
$[9-10]$	89.0 ± 3.1	93.0 ± 2.6	96.0 ± 2.0

(a) performance in each rating subgroup.

Age	Top 1 Acc.	Top 2 Acc.	Top 5 Acc.
$[13-17]$	$\mathbf{90.0 \pm 3.0}$	$\mathbf{94.0 \pm 2.4}$	$\mathbf{99.0 \pm 1.0}$
$[18-34]$	84.0 ± 3.7	89.0 ± 3.1	95.0 ± 2.2
$[35-44]$	80.0 ± 4.0	87.0 ± 3.4	94.0 ± 2.4
$[45-48]$	81.0 ± 3.9	85.0 ± 3.6	95.0 ± 2.2

(b) performance in each age subgroup.

Table 5: Author attribution performance in each rating subgroup and age subgroup.

#	Foundation Models	Deployment Resource	Method	Inference Time (s)	Accuracy
1	LLama-3-70B	8 × A6000 (VLLM)	Logprob	462.1	85.0 ± 3.6
2	GPT-4-Turbo	OpenAI	QA	663.1	34.0 ± 4.7
3	LLama-3-70B-Instruct	Azure	QA	2065.6	31.0 ± 4.6

Table 6: Efficiency analysis between prompt-based method and logprob-based method on Blog dataset.

Influence of Gender We conduct 500 tests which consists of 237 tests for blogs written by male authors and 263 tests for blogs written by female authors and show their accuracy of authorship attribution separately in Table 3. The results indicate that authorship attribution for blogs written by female authors exhibits higher accuracy. This suggests that female-authored blogs might contain more distinct personal styles, making it easier to infer the author.

4.5 Subgroup Analysis

When considering authorship attribution restricted to specific subgroups, the task can either become simpler or more difficult. Certain subgroups may express personal styles more distinctly, making authorship attribution easier, while others may be more homogeneous, making it more challenging. Here, we consider three subgroup factors: gender, age, and rating, to analyze the performance under each group.

Subgroup by Gender As shown in Table 4, we evaluated the performance of authorship attribution within different gender subgroups in the Blog dataset. We observed that authorship attribution performed better within the female subgroup, consistent with findings in Section 4.4, suggesting female-authored blogs possess more distinctive personal styles.

Subgroup by Rating Table 5 (a) shows the performance of authorship attribution across different rating ranges in the IMDb review dataset. Overall, we can see that rating does influence performance, with review in the $[5-6]$ rating range easier to attribute. Despite such difference, our method consistently obtains good performance across all subgroups.

Subgroup of Age Table 5 (b) shows the performance of authorship attribution across different age ranges of bloggers in the Blog dataset. We observed that age significantly influences performance. The youngest age group $[13-17]$ exhibited the highest top-1 accuracy at 90%, while accuracy decreased with increasing author age. This suggests that younger authors tend to have more distinct opinions and identifiable writing styles. Despite performance differences, our method maintained relatively overall high performance, with the lowest accuracy still surpassing that of GPT-4-Turbo with QA method.

4.6 Efficiency Analysis

Table 6 shows the efficiency comparison of different methods on the imdb dataset. Our Logprob method operates with notably lower runtime compared to QA methods. This is primarily due to the Logprob method requiring only a single forward pass through the LLM for each author to estimate the log probabilities. In contrast, QA methods generally need multiple iterations of token generations to form a response, which increases computation time substantially. In the mean time, our method achieves an accuracy of up to 85%, surpassing QA method based on GPT-4-Turbo in both efficiency and accuracy.

In summary, our method proves to be effective and efficient in performing authorship attribution across various datasets and setups.

5 Conclusion

In this paper, we study the problem of authorship attribution. We demonstrate the effectiveness of utilizing pre-trained Large Language Models (LLMs) for one-shot author attribution. Our Bayesian approach leverages the probabilistic nature of language models like Llama-3 to infer authorship. Our method does not require fine-tuning, therefore reduces computational overhead and data requirements. Our experiments validate that our method is more effective and efficient compared to existing techniques.

6 Limitations

The main limitations arise due to the dependence on LLMs.

Our method relies heavily on the capabilities of LLMs, and the performance of our approach is highly affected by the size and training objectives of the LLMs. As shown in Table 1, models that are only pre-trained rather than fine-tuned for dialogue or code task performs better.

While larger models generally perform better, they also entail higher costs, posing scalability and accessibility challenges for broader applications.

Another limitation is due to training data of LLMs. If the training data lacks diversity or fails to include certain writing styles, the model may not fully capture the intricacies of an author's style, potentially leading to misclassifications. This limitation underscores the importance of using diverse and comprehensive training datasets.

Furthermore, any biases present in the training data can also be absorbed by the model. These biases will influence the performance of our authorship attribution method.

On the broader societal level, the potential for misuse of this technology is a significant concern. The challenge of regulating and overseeing the use of such powerful tools is still not fully addressed.

Lastly, while our approach avoids the need for extensive retraining or fine-tuning, which is an advantage in many cases, this also means that our method might not adapt well to scenarios where lots of training data and computation is available, which justifies more complex and computationally intensive methods.

Acknowledgments

ZH, TZ and HH were partially supported by NSF IIS 2347592, 2347604, 2348159, 2348169, DBI 2405416, CCF 2348306, CNS 2347617.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hayri Volkan Agun and Ozgur Yilmazel. 2017. Document embedding approach for efficient authorship attribution. In *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pages 194–198. IEEE.

Salha M Alzahrani, Naomie Salim, and Ajith Abraham. 2011. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2):133–149.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Alexandra Chouldechova and Max G'Sell. 2017. Fairer and more accurate, but for whom? *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Maël Fabien, Esau 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, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Zhenhao Ge, Yufang Sun, and Mark Smith. 2016. Authorship attribution using a neural network language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

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

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *arXiv preprint arXiv:2403.08213*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797, Singapore. Association for Computational Linguistics.

Andrew Kachites McCallum. 1999. Multi-label text classification with a mixture model trained by em. In *AAAI'99 workshop on text learning*.

Seif Mechti and Fahad Almansour. 2021. An orderly survey on author attribution methods: From stylistic features to machine learning models. *Int. J. Adv. Res. Eng. Technol*, 12:528–538.

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.

Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Identifying biased subgroups in ranking and classification. *Measures and Best Practices for Responsible AI at KDD 2021*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*.

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.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

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

Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers*, pages 669–674.

Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov. 2023. Authorship attribution of late 19th century novels using gan-bert. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 310–320.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sarim Zafar, Muhammad Usman Sarwar, Saeed Salem, and Muhammad Zubair Malik. 2020. Language and obfuscation oblivious source code authorship attribution. *IEEE Access*, 8:197581–197596.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Ethical Considerations

Our method using LLMs for authorship attribution brings several ethical considerations that must be addressed to ensure responsible and fair use of the technology.

Privacy and Anonymity The capacity of LLMs to attribute authorship with high accuracy can lead to ethical challenges regarding privacy and anonymity. Individuals who wish to remain anonymous or protect their identity could be compromised if authorship attribution tools are misused. Therefore, it is crucial to establish strict guidelines and ethical standards on the use of such technologies to prevent breaches of privacy.

Potential for Abuse Despite multiple beneficial applications, the misuse potential of authorship attribution tools is significant. Risks include the use of this technology to suppress free speech or to endanger personal safety by identifying individuals in contexts where anonymity is crucial for safety. Addressing these risks requires robust governance to prevent misuse and to ensure that the technology is used ethically and responsibly.

Bias Issue The performance of authorship attribution methods can vary across different demographics, leading to potential biases. It is important to continually assess and correct these biases to ensure fairness in the application of this technology.

Misclassification Issue Given the high stakes involved, especially in forensic contexts, the accuracy of authorship attribution is important. Misclassifications can have serious consequences, including wrongful accusations or legal implications. It is essential for authorship attribution methods to be reliable and for their limitations to be transparently communicated to users.

B Broader Impact

Our study of authorship attribution using LLMs contributes to advancements in various domains:

Forensic Linguistics Our research contributes to the field of forensic linguistics by providing tools that can solve crimes involving anonymous or disputed texts. This can be particularly useful for law enforcement and legal professionals who need to gather evidence and make more informed decisions.

Intellectual Property Protection Our method can serve as a powerful tool in identifying the authors of texts, which can help protect intellectual property rights and resolve disputes in copyright.

Historical Text Attribution In literary and historical studies, determining the authorship of texts can provide insights into their origins and contexts, enhancing our understanding and interpretation.

Enhanced Content Management Media and content companies can use this technology to manage content more effectively by accurately attributing authorship to various contributors.

Educational Applications In educational settings, our method can help prevent plagiarism and promote academic integrity. It can also serve as

a teaching tool to help students understand and appreciate stylistic differences between authors.

While our method holds promise across multiple applications, it is crucial to deploy it with caution. Ensuring that the technology is used responsibly and ethically will be key to maximizing its benefits while minimizing potential harm.