Open-World Authorship Attribution

Xinhao Tan, Songhua Liu, Cong Xia, Kunjun Li, Xinchao Wang[*]
National University of Singapore
{e0774267, songhua.liu, e0985916, kunjun}@u.nus.edu, xinchao@nus.edu.sg

Abstract

Recent years have witnessed rapid advancements in Large Language Models (LLMs). Nevertheless, it remains unclear *whether state-of-the-art LLMs can infer the author of an anonymous research paper solely from the text, without any additional information*. To investigate this novel challenge, which we define as *Open-World Authorship Attribution*, we introduce a benchmark comprising thousands of research papers across various fields to quantitatively assess model capabilities. Then, at the core of this paper, we tailor a two-stage framework to tackle this problem: candidate selection and authorship decision. Specifically, in the first stage, LLMs are prompted to generate multi-level key information, which are then used to identify potential candidates through Internet searches. In the second stage, we introduce key perspectives to guide LLMs in determining the most likely author from these candidates. Extensive experiments on our benchmark demonstrate the effectiveness of the proposed approach, achieving 60.7% and 44.3% accuracy in the two stages, respectively. We will release our benchmark and source codes to facilitate future research in this field.

1 Introduction

The advancement of Generative Artificial Intelligence (AI) and Large Language Models (LLMs) has revolutionized numerous fields due to their remarkable capabilities in Natural Language Processing (NLP) tasks (Hagos et al., 2024; Naveed et al., 2024; Cui et al., 2024; Zhang et al., 2024). Despite their widespread applications, their potential for authorship attribution—the task of identifying an author from anonymous text—remains largely unexplored. In this paper, we investigate an intriguing question: *Can state-of-the-art LLMs infer the author of an anonymous research paper without any additional information?* This problem is both practical and ambitious. On one hand, accurately attributing authorship in academic research is crucial for maintaining integrity, recognizing contributions, and detecting plagiarism or ghostwriting. On the other hand, directly applying modern LLMs to this task is challenging, as the relevant information is often dispersed across Internet-scale data, which makes it infeasible for these models to process efficiently.

In this paper, we define this challenging task as *Open-World Authorship Attribution*. Since no existing benchmark evaluates LLM performance in this area, we construct a dataset comprising thousands of academic papers from various research fields. Building on insights from this data, we propose a novel two-stage framework to address the task, including *Candidate Selection* and *Authorship Decision*.

Specifically, in the candidate-selection stage, we leverage LLMs to generate multi-level key representations of a target paper, which are then utilized to search the Internet for relevant authors and their publications at multiple levels of specificity. The retrieved authors along with authors in the citation list form our candidate pool. In the authorship-decision stage, LLMs assess potential authorship in the candidate pool by evaluating the anonymous text against multiple guidelines. Finally, a holistic decision is made to determine the most probable author, serving as the final output.

We conduct experiments using multiple state-of-the-art LLMs, including both open-source and closed-source models.

Extensive evaluation validate the effectiveness and the superiority of the proposed solution. Specifically, our approach achieves 60.7% accuracy of candidate selection and 44.3% accuracy of authorship decision. The contribution of this work is summarized as below:

- We are the first to define, study, and bench-

[*] Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 17744–17758
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

mark the task of open-world authorship attribution to the best of our knowledge.

- By leveraging impressive capacity of recent LLMs, we devise a novel two-stage pipeline, including candidate selection and authorship decision, to tackle this challenge.
- Extensive evaluations showcase the potential of modern LLMs and our proposed solution for open-world authorship attribution. We will release the dataset, prompts, and codes to support future research in this field.

2 Related Work

2.1 Large Language Models (LLMs)

LLMs have demonstrated remarkable capability in solving various Natural Language Processing (NLP) tasks, such as mathematical reasoning, and text summarization (Xuanfan Ni, 2024; Desta Haileselassie Hagos, 2024). The unique characteristic of LLMs lies in utilizing a unified paradigm without additional training to address various tasks (Qin et al., 2024).

Language modelling: Language modelling as the core to current LLMs has developed from the traditional statistical methods like n-gram (Sharma et al., 2018a) models to Neural Network language models. The transformer language models with self-attention mechanisms further lay the foundation for the current rapid development of LLMs (Vaswani et al., 2017). The introduction of the revolutionized transformer helped the development of the GPT-1 transformer-decoder structure and Bert's transformer-encoder structure.

LLMs Tuning: Tuning techniques have evolved alongside the development of LLMs. Tuning consists of full-parameter and partial-parameter tuning. Due to computational constraints, research has focused on Parameter-Efficient Fine-Tuning (PEFT), including prompt tuning, Adapter-Tuning, and LoRA. In-context learning, a form of prompt learning, enables adaptation without parameter updates by providing example-based prompts.

Instruction tuning is also the current focus. The purpose is to transform NLP tasks with natural language instruction which improves the performance of LLMs in zero-shot learning. Chain-of-Thought (Wei et al., 2023) is another reasoning strategy to resolve the issue of low performance in arithmetic reasoning, normal inference and symbol inference.

2.2 AI-generated texts Detection with LLMs

The widespread accessibility of generative models has led to a proliferation of AI-generated texts across the internet. Several detection approaches have been developed to detect LLM-generated works to address the issue of authenticity (Sun et al., 2025): (1) Training-based method adopt classifiers like Support Vector Machines (SVMs) or fine-tuned pre-trained language models like RoBERTa and T5(Yang et al., 2023; Tang et al., 2023). (2) Zero-shot Detection method directly uses the inherent properties embedded in LLMs (Yang et al., 2023). (3) Watermarking-based Detection like Inference-time watermarking (Tang et al., 2023) embeds unique patterns into text during generation by manipulating decoding processes, while post-hoc watermarking retroactively modifies generated text using rule-based or neural techniques to ensure traceability (Tang et al., 2023).

2.3 Authorship Identification

Several studies have already researched the authorship identification capabilities of LLMs, highlighting the importance of authorship attribution in forensic investigations, cybersecurity, and tackling misinformation (Huang et al., 2024; Wen et al., 2024; Huang et al., 2025).

Traditionally, authorship attribution and verification focus on analyzing writing styles to measure similarities and make authorial decisions. Early methods employed natural language processing (NLP) techniques, such as n-grams (Sharma et al., 2018a), part-of-speech (POS) tags (Sundararajan and Woodard, 2018), and Linguistic Inquiry and Word Count (LIWC) (Uchendu et al., 2020). These handcrafted features are designed to quantify stylistic patterns, including vocabulary richness, syntactic complexity, and semantic focus, for effective analysis (Huang et al., 2024, 2025).

More recently, with advancements in deep learning, text embeddings have become a prominent tool in authorship attribution. Text embeddings represent textual data as vectorized numerical representations, enabling models to encode both semantic and stylistic nuances. (Kumarage and Liu, 2023) emphasizes the potential of leveraging large pre-trained language models (LLMs) like BERT and GPT to generate embeddings that capture deeper stylistic and contextual patterns, thus applying them to authorship attribution tasks. Another significant change in authorship attribution is the in-

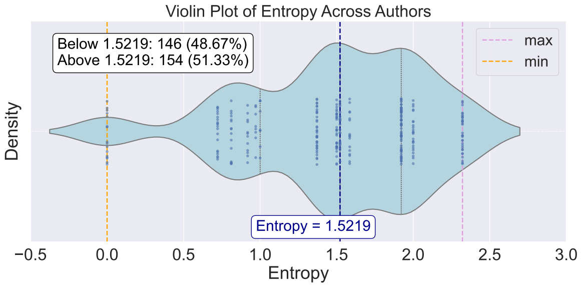


Figure 1: Violin plot illustrating the distribution of information entropy among 300 authors. An information entropy value of 1.5219 indicates that, among the five collected articles for a given author, two articles share the same topic group, another two belong to a different topic group, and the remaining article falls into a separate group

tegration of contrastive learning techniques into embedding-based methods(Patel et al., 2023).

Some methods developed the prompt pipeline for authorship identification, leveraging the inherent stylistic and linguistic extraction capabilities of LLMs (Huang et al., 2024; Wen et al., 2024; Huang et al., 2025). The results demonstrate the ability of LLMs to capture nuanced stylistic features without explicit feature engineering. However, limitations of the study are noticeable, such as dependency on pre-collected candidate authors which hinders its application in large-scale candidate pools. In most cases, the number of candidate authors is fewer than 50, making the approach impractical for real-world applications. Their focus on stylometric feature analysis and prioritizing explainability in the authorship decision-making limits its efficiency.

3 Methods

In this section, we elaborate on the proposed benchmark and two-stage approach for open-world authorship attribution. Sec. 3.1 introduces the data sources and the construction of the benchmark. Secs. 3.2 and 3.3 describe the main pipelines of the two stages: candidate selection and authorship decision, respectively, in the proposed solution. In the first stage, candidate papers are retrieved from the Internet using LLM-generated keywords. In the second stage, LLMs determine the most likely author from these candidates. Fig. 3 provides an overview of the entire streamline.

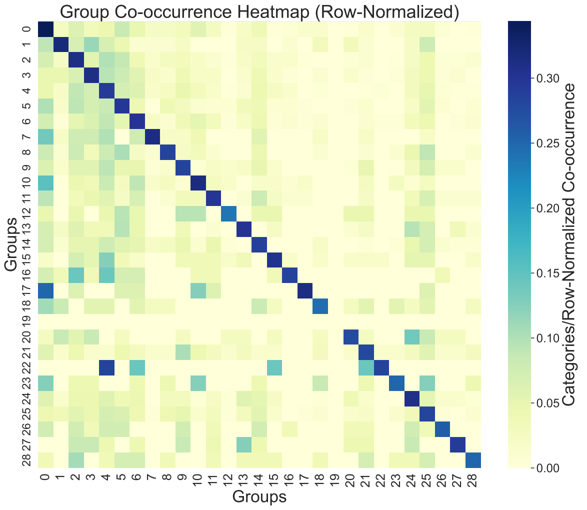


Figure 2: This heatmap shows the extent to which authors choose the same topics across their publications. Each cell represents the co-occurrence strength between topics for the same author, with darker shades indicating a higher likelihood of an author selecting the same topic in their papers. Each number in axises indicates different topic groups 6

.

3.1 Data Curation

Considering there is no off-the-shelf benchmark for the task of open-world authorship attribution, we construct a dataset in this work. Specifically, To ensure diversity, we select papers from CVPR 2024, spanning 30 subfields in computer vision. We also curated 50 additional samples from non-CVPR fields, including topics in Mathematics, Quantitative Finance, Physics, and Economics. The test results for these non-CVPR samples are provided in Appendix 5. Moreover, to guarantee sufficient online reference materials for candidate selection, we filter out authors with fewer than five first-author papers. This results in a dataset comprising 300 authors and 1,500 papers. More details are provided in Sec. 4.1.

3.2 Candidate Selection

The core challenge of open-world authorship attribution lies in handling Internet-scale data, which significantly exceeds the processing capabilities of LLMs. Therefore, identifying common patterns among papers by the same author is crucial for narrowing down potential candidates from such a vast data source.

During our investigation and collection of the dataset, we observe that many authors' published papers demonstrate a correlation in research topics.

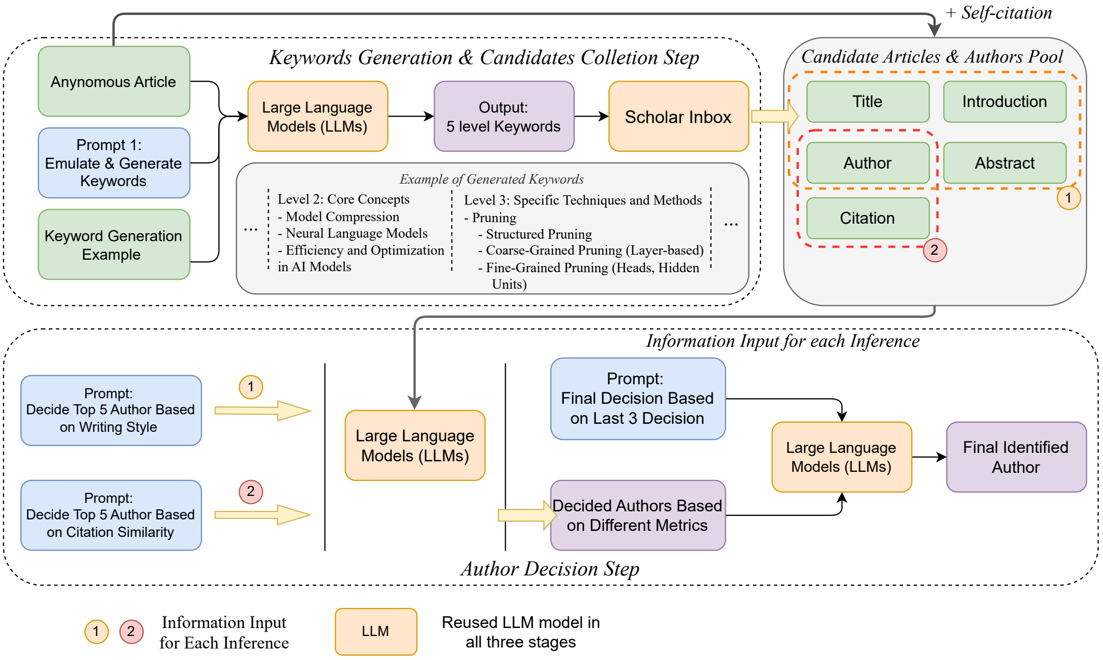

+ Self-citation
Keywords Generation & Candidates Colletion Step
Anynomous Article
Prompt 1: Emulate & Generate Keywords
Keyword Generation Example
Large Language Models (LLMs)
Output: 5 level Keywords
Scholar Inbox
Example of Generated Keywords
Level 2: Core Concepts
- Model Compression
- Neural Language Models
- Efficiency and Optimization in AI Models
Level 3: Specific Techniques and Methods
- Pruning
- Structured Pruning
- Coarse-Grained Pruning (Layer-based)
- Fine-Grained Pruning (Heads, Hidden Units)
...
Candidate Articles & Authors Pool
Title
Introduction
Author
Abstract
Citation
1
2
Information Input for each Inference
Prompt: Decide Top 5 Author Based on Writing Style
Prompt: Decide Top 5 Author Based on Citation Similarity
1
2
Large Language Models (LLMs)
Prompt: Final Decision Based on Last 3 Decision
Decided Authors Based on Different Metrics
Large Language Models (LLMs)
Final Identified Author
Author Decision Step
1 2 Information Input for Each Inference
LLM
Reused LLM model in all three stages

Figure 3: The whole process of the proposed automatic end-to-end authorship identification consists of 2 stages. The first stage will contribute to the collection of relevant candidates. The second stage performs 2 rounds of inference decision based on different metrics with different input information from candidate pool. The final decision is based on the 2 inferences.

To analyze the relationship between authors and research topics, we leverage Information Entropy and perform a statistical analysis. Specifically, for each author, we evaluate the randomness and diversity of the involved research fields via:

$$H(T) = -\sum_{i=1}^{N} p(T_i) \log_2 p(T_i), \qquad (1)$$

where $p(T_i)$ represents the probability of an author's papers belonging to topic T_i. Higher entropy suggests greater diversity in research areas, while lower entropy indicates topic consistency, aiding in author identification.

In the violin plot shown in Fig. 1, we observe that only a small subset of authors exhibit high entropy and diversity across research fields. In Fig. 2, we also visualize the extent to which authors in a given field also conduct research in other fields. Strong diagonal activations suggest a high likelihood of topic overlap within an author's publications.

Based on the correlation of topics between authors' different papers above, we can rely on this to search for the author. Therefore, we need some keywords which can summarize the topics from anonymous text and be utilized for searching the relevant articles. This is where LLMs can help in our candidate selection stage - keyword generation. These candidate articles will be used further in the next stage of decision-making.

Keywords Generation. The keywords generated need to accurately capture the contents of the anonymous input for the effective searching of the relevant articles. Therefore, we decided to generate the relevant keywords in hierarchies to describe the anonymous content. The different levels should range from general (level 1) to specific (level 5). In this way, we can search from the most specific to the most general to get the relevant articles as our candidates prepare for the next stage of decision-making.

Few-shot Prompting. If we ask LLMs to generate the keywords directly in hierarchies, the output may be in different formats and the quality of generation is not guaranteed by the simple instruction of "Generate 5-level keywords". Few-shot prompting is utilized for LLMs to demonstrate how the response should look. We will manually create an example 6 and use it as an example within the

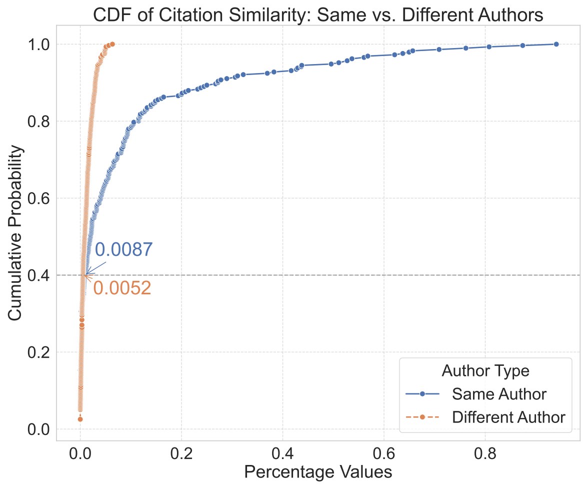


Figure 4: This figure presents the cumulative density plot (CDF) of citation similarity as it increases. The varying rate of cumulative percentage growth indicates that articles within the same topic tend to exhibit higher citation similarity.

Algorithm 1 Open World Authorship Attribution

```
Input: Anonymous article $x$
Prompts: $metrics = \{style, citation\}$
Output: Final attributed author
Step 1: Keyword Generation
# Extract representative keywords using LLMs:
$keywords \leftarrow LLMs(x, Prompts, Example)$
Step 2: Candidates Collection
for each $level$ in $keywords$ do
    # Use web scraping to collect potential authors and articles:
    $CandidatePool \mathrel{+}= WebScraping(level)$
end for
Step 3: Iterative Filtering by Metrics
for $i = 1$ to $|metrics|$ do
    # Use LLMs to rank authors based on current metric:
    $TopAuthors \leftarrow LLMs(x, CPool, metrics_i)$
    # Append with the top-ranked authors:
    $CandidatePool \leftarrow TopAuthors$
end for
Step 4: Final Attribution
#determine the most likely author from $TopAuthors$:
$FinalAuthor \leftarrow LLMs(x, TopAuthors)$
Return: $FinalAuthor$
```

prompt. This will maximize the possibility of effective keyword generation which follows our proposed method.

Candidates Search & Collect. Different levels of the keywords will be sent to the search engine to collect the potential candidates. The search engine we choose is Scholar Inbox which has a semantic section to input the different level keywords. In each search result of the level keywords, we decided to collect the first 20 papers with their respective titles, authors and abstracts. If the searched article provides an arxiv link, we will try to retrieve the introduction and citation to help the stage 2 decision-making.

Appending Self-citation. Most authors tend to cite their previous works, especially when focused on a specific research area. This characteristic increases the likelihood of successfully including the true author in the candidate pool, which serves as the foundation for Stage 2. To leverage this, we incorporate self-citation into our candidate selection process, and their metadata is retrieved through web scraping to enhance the effectiveness of our decision-making.

3.3 Authorship Decision

LLMs also show their great capability in analyzing large-scale data. Therefore In this stage, after creating our potential candidate pool, we need to enable LLMs to decide the most possible author. However, direct instruction like "Please decide the most possible author from the candidate pool" is not detailed enough for LLMs to accurately decide the most possible author. Therefore, we establish different metrics to guide LLMs in identifying the most probable author step by step. Chain-of-Thought (CoT) (Wei et al., 2023) prompting is a technique designed to enhance the reasoning capabilities of large language models (LLMs) by guiding them to generate intermediate reasoning steps before arriving at a final answer. This approach mirrors the human thought process, and by simulating it, LLMs can achieve more accurate analysis and decision-making outcomes.

For each metric, we ask the LLMs to list the top 5 authors that most match the metrics. In the last step, we input all the decision results from its decision and let LLMs decide the most possible one holistically. In this way, we can identify the most possible author. LLMs will perform 3 rounds of most possible author inference based on contents, writing styles, and citation similarities. The full prompting can be referred to in Appendix 7.

3.3.1 Contents

As mentioned in the above section, due to the similarities of authors chosen topic in their published paper, content is the important metric to determine the actual author (Halvani and Graner, 2021; Potha and Stamatatos, 2019). the specific metrics within the contents is preferred topics and Domain-Specific term used in the writing. We will input author names, titles, abstracts and introduction (not every paper have) from the candidates pool for LLMs inference.

3.3.2 Writing Style

Relying on stylometry is the traditional way of authorship attribution. Evolving from the human skills in identifying (Argamon et al., 2009), computational methods gradually become the main trend in the analysis of authors' unique linguistic features in stylometry methods(Lagutina et al., 2019; Neal et al., 2017). Machine learning methods with LLMs further advance the computational methods with their powerful ability to extract features (Boenninghoff et al., 2019; Kojima et al., 2022).

In our proposed method, we also utilize writing style as an important factor in identifying the author from our collected candidate pool. Different people have different habits or underlying characteristics in writing. Some typical metrics are repetition Patterns in words and phrases (Sharma et al., 2018b), sentence complexity, paragraph structure, sentence length and variation.

To demonstrate the importance of writing style in differentiating authors, we select three paragraphs from three articles—two written by the same author and one by a different author(Appendix 7). To eliminate the influence of topic variation, all three articles cover the same subject. We then prompt GPT-4o to analyze the texts and determine which two paragraphs are authored by the same individual. GPT-4o demonstrates its ability to make this distinction based on factors such as writing tone, repetition patterns, sentence complexity, and paragraph structure.

In this metric, we have the same input as content-based inference which is title, author, abstract and introduction.

3.3.3 Citation Similarity

As mentioned above, the topics for an author's published papers are largely overlapped. This may also indicate that the literature review - citations of an author tend to be similar. In other words, the author has a preference for some citations and they may prefer to reuse these citations in their other works. Based on this assumption we establish the third metric which is the citation similarities. In this stage, we only utilize the author names, articles titles with the corresponding citations as the input for the LLMs.

We conduct a citation similarity analysis on our collected dataset, calculating the similarity between the citations of test articles and titles authored by the same or different authors. The plotted cumulative distribution function (CDF) Figure 4 illustrates the distribution of citation similarity under the same or different authors. Approximately 40% of the articles exhibit similar citation match probabilities between the same author and different authors. However, the remaining 60% show a clear distinction in citation similarity. In general, articles written by the same author tend to have higher citation similarity values.

4 Experiments

4.1 Experimental Setup

Models. We choose nowadays popular LLMs to conduct the test. We download the open source Meta-Llama-3.1-8B-Instruct (Grattafiori and et al., 2024) for the initial test with both abstract-only and abstract-plus-introduction as input information input. We also use the recent GPT-4o-mini (OpenAI et al., 2024) to conduct the test. GPT-4o-mini was accessed and tested via API requests.

Dataset. Due to a lack of academic paper datasets available online, We self-collected our dataset for our testing. Our dataset includes 300 authors, with 5 papers selected for each author. To ensure relevance and keep the writing style of the author, we only selected the paper where the author is listed as the first author or second author of the paper. For every author, the first paper we collected is ensured to be the most recent published from the author (simultaneously ensuring the author is the first author), which can minimize the possible bias that the paper is included as the pretraining data of the popular LLMs. To facilitate the extraction of relevant information such as authors, titles, abstracts, introductions or citations, we collect the paper link in the sample. Additionally, we assume that the authors' papers are published on the arxiv.org website. Hence, all the paper links are arxiv links. For every test, we use every author's first paper as the anonymous text input. The rest 4 papers are used for other analysis.

Implementations. Meta-Llama-3.1-8B-Instruct model was downloaded and deployed on 4-10 RTX 4090 GPUs, with the max new token set to 2000 to guarantee complete output. GPT-4o-mini was utilized through API request. The maximum input tokens for GPT-4o-mini is about 200K. During the searching process, if we collect the same article as our anonymous test input, we will ignore it as

Table 1: Test Results of Two Stages: Candidate Selection and Authorship Decision. Extra test results from non-CVPR datasets are shown in Appendix 5

Model	Stage 1 (%)	Stage 2 (%)						
		Content		Writing Style		Citation Similarity		Final Accuracy
		Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	(%)
Llama-3.1-8B	60.7	19.0	47.0	21.3	21.3	52.0	57.0	31.3
GPT-4o-mini	56.3	27.3	52.3	36.3	61.3	49.3	66.0	41.3

Table 2: Classification Report of Different ML Models

Methods	Model	Writing Style		Citation		Final Score (%)
		Top1	Top5	Top1	Top5	
LIP (Huang et al., 2024)	Llama-3.1-8B	10.0	18.3	-		8.7
	GPT-4o-mini	6.7	17.7	-		9.3
	DeepSeek-V3	4.0	21.7	-		10.3
AIDBench (Wen et al., 2024)	Llama-3.1-8B	9.7	18.0	-		10.0
	GPT-4o-mini	11.3	21.7	-		11.3
	DeepSeek-V3	6.3	23.0	-		11.67
Our Method	Llama-3.1-8B	8.7	16.0	12.3	20.0	11.6
	GPT-4o-mini	18.0	29.7	15.0	26.7	17.7
	DeepSeek-V3	9.3	30.7	7.7	28.7	18.7

the assumption of our method is that the test paper should not appear on the website.

Evaluations. Our evaluations are divided into 2 parts, the first part is to examine the effectiveness of searching and collecting the true author from the crawling based on the generated keywords from LLMs. The second part is to examine the ability of the LLMs to identify the true author from the candidates pool. If the first stage fails to collect the true author, we add the same authors' other papers from our dataset to the candidate pool and allow the LLMs to reattempt the stage 2 test.

Baseline. We identify related baselines (Wen et al., 2024; Huang et al., 2024) that also employ prompt-based methods, closely aligning with our approach. Both methods emphasize writing style analysis within their prompts. Accordingly, we adopt these baselines in the authorship decision stage and apply them to the same candidate sets produced by our candidate selection method. This setup ensures a fair comparison under consistent input conditions.

Additionally, we conducted an overall score evaluation, as our proposed method is an automated, end-to-end process. This test measures the percentage of cases where the correct author is successfully selected and identified throughout the entire pipeline.

Multi-Conversation Handling. The input token limits of OpenAI API requests are 128,000 to 200,000 tokens. When the input token number exceeds the token limitations due to additional information such as citation and introduction, or a large number of papers collected in the candidate pool, we implement a hierarchical batch processing approach. The candidate pool is divided into equal-sized batches, with each containing almost same number of candidates. We first identify the top 5 candidates from each batch independently. These preliminary selections are recorded and subsequently aggregated to form a consolidated candidate pool. We then extract the complete information profiles for these candidates from the original dataset, enabling a comprehensive final evaluation. In this way, we construct small candidate pool based on each metrics and then make the holistic

Table 3: Ablation experiment by using different input and different prompts. The experiment was conducted using GPT-4o-mini.

INPUT	METRICS FOR PROMPTING			FINAL ACCURACY (%)
	CONTENT	WRITING STYLE	CITATION SIMILARITY	
ABSTRACT	√	√		26.0
ABSTRACT+INTRO+CITATION	√		√	39.7
ABSTRACT+INTRO+CITATION	√	√	√	41.3
ABSTRACT+INTRO+CITATION		√	√	44.3

Table 4: Evaluating the accuracy of searching the correct author's other published articles in our stage 1. The keywords are generated by Llama-3.1-8B.

KEYWORD INSTRUCTION	STAGE-1 ACC. (%)
NO INSTRUCTONS	7.3
FROM GENERAL TO SPECIFIC	12.0
WITH LEVELS	15.7
SELF-CITATION	55.3
WITH LEVELS + SELF-CITATION	60.7

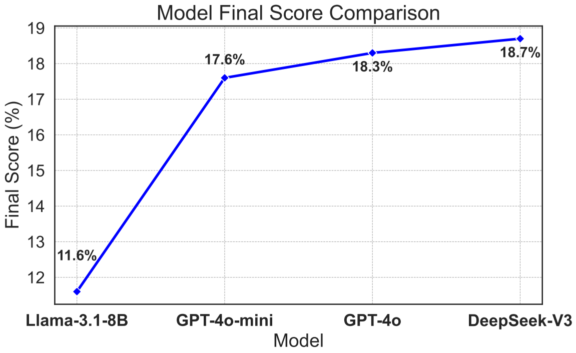


Figure 5: The final Score comparison across different models using our methods.

decision of the final author.

4.2 Results

First, the experiment results of Stage 1 and Stage 2 are summarized in Table 1. In Stage 1, we utilize our hierarchical levels to guide the models in emulating the keyword generation process. Each level serves as a basis for retrieving candidate authors, which are then used in Stage 2 for decision-making. Additionally, we incorporate citation information from the anonymous articles into our candidate pool to enhance selection accuracy. Using keywords generated by Llama-3.1-8B, we achieved an accuracy of 59.3%. Keywords emulated and generated by GPT-4o-mini are slightly less effective, yielding an accuracy of 56.3%.

In Stage 2, we collect information about the candidate authors and input this data into LLMs for reference-based evaluation using three key metrics: content similarity, writing style, and citation similarity. Across all three metrics, Top-5 accuracy consistently exceeded 50%. Among these metrics, writing style outperform content similarity in distinguishing authors. However, citation similarity achieves the highest accuracy, with Top-1 accuracy reaching 52% for Llama-3.1-8B and 49.3% for GPT-4o-mini. Finally, by integrating these three metrics, our final decision accuracies are 31.3% for Llama-3.1-8B and 41.3% for GPT-4o-mini.

We also experiment to evaluate the overall accuracy by combining Stage 1 and Stage 2 (Table 2). The final score indicates the probability of correctly identifying the author from searching to authorship-decision, Our method achieves the best results using GPT-4o-mini (17.6%), outperforming the baseline models LIP (Huang et al., 2024) and AIDBench (Wen et al., 2024) in both Llama and GPT models.

To validate the effectiveness of our proposed hierarchical keyword levels for candidate collection, we conduct ablation experiments on keyword analysis in Table 4. The results demonstrate that using different prompts leads to varying search performance. When applying our proposed levels, we achieve an accuracy rate of 15.7%. Furthermore, when combined with self-citation, the overall accuracy increases significantly to 60.7%.

We also investigate the impact of incorporating the introduction and citation in author attribution in Table 3, finding that it significantly improves the LLMs' ability to identify the correct author, with accuracy increasing from 26% to approximately

40%. When prompting the model to perform inference based on content, writing style, and citation similarity, the results are slightly lower than the accuracy achieved using only writing style and citation similarity (44.3%).

Finally, in Figure 5, we achieve the highest overall score of 18.3% using the latest GPT-4o model.

5 Conclusion

In this paper, we introduce the first benchmark and dedicated solution for *Open-World Authorship Attribution*. Leveraging recent advancements in LLMs, we propose a two-stage pipeline: candidate selection and authorship decision. In the first stage, multi-levels keywords extracted from the target paper are used to search the Internet. The retrieved results, combined with citation lists, form a pool of potential candidates. In the second stage, LLMs infer authorship based on writing style and citation similarity from these candidates. Extensive experiments demonstrate the effectiveness and superiority of our approach over multiple potential baseline methods.

6 Limitations

Table & Figure Features. Another distinguishing feature in authorship attribution is the unique preferences authors showcase in structuring and designing their tables and figures. These characteristics manifest in various ways, including the choice of color palettes, where some authors consistently favor specific hues or grayscale representations. Differences also emerge in plotting styles, such as the use of bar charts, scatter plots, heatmaps, or line graphs, along with variations in grid usage, axis formatting, and legend placement. Labeling and annotation preferences also contribute to stylistic distinctions, as authors may differ in font choices, caption positioning, and the inclusion of callout markers. Additionally, the structuring of tables varies, with some researchers favoring detailed grid layouts while others opt for minimalistic designs with selective use of horizontal and vertical lines. Another notable characteristic is the numbering and referencing approach, with some authors preferring “Figure 1” while others use “Fig. 1,” along with variations in how they cross-reference visual elements within the text. In future work, we aim to systematically analyze and quantify these stylistic preferences, leveraging feature extraction techniques and deep learning models to explore how visual elements can enhance authorship attribution accuracy.

Large input. Since our method follows an open-world authorship attribution approach with an end-to-end pipeline, it requires collecting a substantial amount of information as input for the LLMs. This often exceeds the maximum token limit of many models, which results in extra strategies to handle multi-turn conversations, as these models do not have built-in memory functions.

7 Future Work

Source Code. The code will be made available to facilitate reproduction of our results.

Low Final Score. The methodology we present, along with our dataset, provides valuable resources for further advancement in this field. Our work establishes an important foundation and benchmark upon which future research can build more robust and accurate solutions. However, we acknowledge that the final accuracy scores still have a large room for improvement to support real-world high-stakes applications like plagiarism detection or forensic analysis. We will continuously find new methods to improve the result.

More LLMs for Testing. Currently, we use LLaMA, the GPT series, and DeepSeek as our primary models for testing. Other LLMs, such as Claude, may not support large input token lengths like GPT, making inference cumbersome. We will continue to monitor the availability of other LLMs and plan to include them in future experiments.

Impact of Author Position. In this paper, we curate the dataset using the first and second authors as the label. Our primary assumption is that first and second authors typically contribute most significantly to the writing style and content of papers, making them more suitable for authorship attribution. We will investigate how different authorship positions as well as co-authorship affect the performance in the future.

Search Engine & Collection of Papers. Currently, our candidate pool is generated solely through Scholar Inbox, which offers convenient access to paper searches. In future work, we plan to develop additional agents to support more search engines.

For the choice of retrieving the first 20 papers, it is based on practical considerations, which provide a good balance between search time and search accuracy, allowing us to identify the right author with a reasonable trade-off. In the future, we will analyze how different search engines, ranking algorithms, and retrieval counts affect the overall performance of our method.

More Baselines. Although there are still many baselines that could be discussed, some remain impractical for our specific setting. For example CAVE (Ramnath et al., 2025) is designed for pairwise authorship verification between two texts. which makes it impractical for identification with 100 candidates. The Bayesian Approach method (Hu et al., 2024) requires retrieving the logits of each token predicted by open-source LLMs, making it computationally expensive in both runtime and memory for our setting. Due to the lack of publicly available source code for reproduction and the complexity and intricate details of the method, it is challenging for us to accurately implement the method. We will continue to explore and incorporate new methods into our comparisons to further enrich our dataset.

Acknowledgements

This project is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002).

References

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

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

Jing Cui, Yishi Xu, Zhewei Huang, Shuchang Zhou, Jianbin Jiao, and Junge Zhang. 2024. Recent advances in attack and defense approaches of large language models. *Preprint*, arXiv:2409.03274.

Danda B. Rawat Desta Haileselassie Hagos, Rick Battle. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *arXiv preprint arXiv:2407.14962*.

Aaron Grattafiori and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *Preprint*, arXiv:2407.14962.

Oren Halvani and Lukas Graner. 2021. Posnoise: An effective countermeasure against topic biases in authorship analysis. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–12.

Zhengmian Hu, Tong Zheng, and Heng Huang. 2024. A bayesian approach to harnessing the power of llms in authorship attribution. *Preprint*, arXiv:2410.21716.

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

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *Preprint*, arXiv:2408.08946.

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.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. *arXiv preprint arXiv:2308.07305*. Computation and Language (cs.CL); Artificial Intelligence (cs.AI).

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P. G. Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

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

OpenAI, :, Aaron Hurst, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting llms. *Preprint*, arXiv:2305.12696.

Nektaria Potha and Efstathios Stamatatos. 2019. Improving author verification based on topic modeling. *Journal of the Association for Information Science and Technology*, 70(10):1074–1088.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Large language models meet nlp: A survey. *Preprint*, arXiv:2405.12819.

Sahana Ramnath, Kartik Pandey, Elizabeth Boschee, and Xiang Ren. 2025. Cave: Controllable authorship verification explanations. *Preprint*, arXiv:2406.16672.

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

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

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. *Preprint*, arXiv:2502.12150.

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

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*. Accessed: January 7, 2025.

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.

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, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zichen Wen, Dadi Guo, and Huishuai Zhang. 2024. Aidbench: A benchmark for evaluating the authorship identification capability of large language models. *arXiv preprint arXiv:2411.13226*.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared llama: Accelerating language model pre-training via structured pruning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Presented at ICLR 2024.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. *arXiv preprint*, arXiv:2204.00408. Accepted to ACL 2022. The code and models are available at https://doi.org/10.48550/arXiv.2204.00408.

Piji Li Xuanfan Ni. 2024. A systematic evaluation of large language models for natural language generation tasks. *arXiv preprint arXiv:2405.10251*.

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*. Accessed: January 7, 2025.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Preprint*, arXiv:2401.13601.

A Appendix

A.1 Ethical Discussion

The application of this method may lead to unintended consequences, such as identifying authors during the Open Review stage of anonymous submissions, which would constitute an inappropriate and unethical use of our approach. Misuse of this technique in peer review processes could compromise the integrity of double-blind evaluation systems, introducing bias in scholarly assessments.

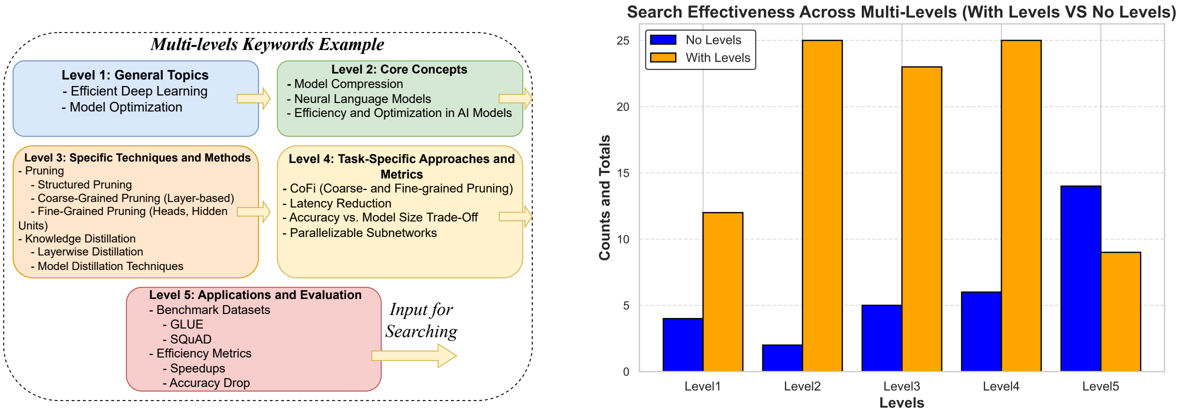
To mitigate such risks, we strongly advocate for the ethical application of our method in domains where it can serve as a tool for transparency, accountability, and integrity. These include plagiarism detection, where it can help identify unauthorized reproduction of content; authenticity verification, which ensures the legitimacy of texts to detect spam and fraudulent writing; and forensic linguistic analysis, where attribution techniques contribute to research integrity.

Furthermore, we emphasize the importance of responsible deployment, encouraging institutions, publishers, and AI practitioners to implement strict ethical guidelines when leveraging authorship attribution technologies. By doing so, we can ensure that such methods are used only in contexts that promote fairness, trust, and the credibility of research and publishing.

A.2 Cost Discussion

During testing, the primary cost is associated with API access to OpenAI models. GPT-4o-mini and GPT-4o are priced at 0.15 and 2.50 USD per million input tokens, respectively. Due to the large input size generated by our candidate pools, each test round involving 300 authors incurs an estimated cost of 5 USD when using GPT-4o-mini and 50 USD per round when using GPT-4o.

A.3 Effectiveness Analysis of Multi-Level Keywords Retrieval



Multi-levels Keywords Example

Level 1: General Topics
- Efficient Deep Learning
- Model Optimization

Level 2: Core Concepts
- Model Compression
- Neural Language Models
- Efficiency and Optimization in AI Models

Level 3: Specific Techniques and Methods
- Pruning
  - Structured Pruning
  - Coarse-Grained Pruning (Layer-based)
  - Fine-Grained Pruning (Heads, Hidden Units)
- Knowledge Distillation
  - Layerwise Distillation
  - Model Distillation Techniques

Level 4: Task-Specific Approaches and Metrics
- CoFi (Coarse- and Fine-grained Pruning)
- Latency Reduction
- Accuracy vs. Model Size Trade-Off
- Parallelizable Subnetworks

Level 5: Applications and Evaluation
- Benchmark Datasets
  - GLUE
  - SQuAD
- Efficiency Metrics
  - Speedups
  - Accuracy Drop

Input for Searching

Figure 6: Our proposed Multi-Levels keywords serve as the foundation for searching and collecting candidate authors. These keywords are used as example prompts for LLMs to generate emulated results. The Keywords Level Analysis evaluates the effectiveness of each multi-levels in the search process, comparing the results against those obtained without any guiding instructions.

A.4 Expand the Dataset

Here, we expand our dataset to include authors from other fields beyond computer vision. We collect data from 54 additional authors: 20 from mathematics, 10 from quantitative finance, and 24 from economics. We conduct evaluation using GPT-4o-mini across these new domains, and the results shown below demonstrate the effectiveness of our method beyond the field of computer vision.

Table 5: GPT-4o-mini results of 54 additional authors in fields of mathematics, finance, physics and economics.

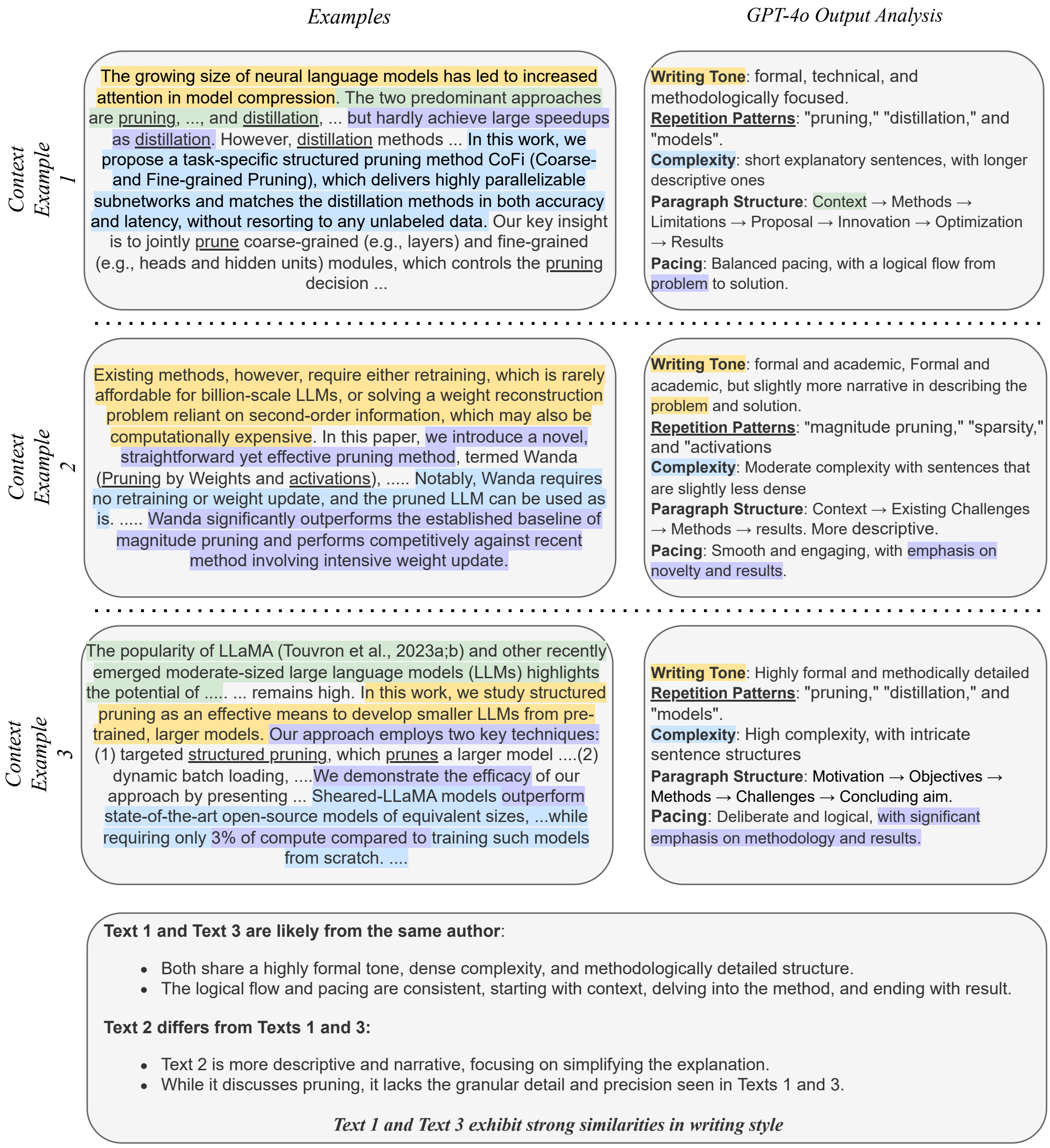
STEP1 Accuracy(%)	STEP2 Top 5 Acc(%)	STEP2 Top 1 Acc(%)
61.11	90.90	21.67

Table 6: Categories & Groups in Heatmap

Groups	Categories	Groups	Categories
1	3D Vision and Reconstruction	16	Natural Language Processing (NLP)
2	Medical Imaging and Diagnostics	17	Scene Understanding and Parsing
3	Image Processing and Enhancement	18	Tracking and Re-Identification
4	Video Understanding and Generation	19	Federated and Distributed Learning
5	Vision-Language Models	20	AI Ethics and Explainability
6	Generative Models and Techniques	21	Physics and Scientific Topics
7	Object Detection and Recognition	22	Data Representation and Augmentation
8	Semantic and Instance Segmentation	23	Audio and Speech Processing
9	Adversarial Techniques and Robustness	24	Emotion and Human-Centric Applications
10	Optimization and Efficiency	25	Novel Applications and Emerging Topics
11	Robotics and Navigation	26	Large Language Models (LLM)
12	Graph Neural Networks and Hyperbolic Models	27	Neural Architecture Optimization
13	Multimodal and Hybrid Models	28	Other
14	Scientific Modeling and Mathematics	29	Deep Learning and Foundational Models
15	Self-Supervised and Semi-Supervised Learning		

Table 7: Full Instruction & Prompting used in decision stage based on 3 different metrics: Content, Writing Style and Citation Similarities.

Metrics	Instruction as Input
Content	I will provide the information of the anonymous article's title, abstract, dataset, introduction or extra information, please remember them. Then, Please choose the top 5 possible articles' author(s) among all the candidates with their corresponding information. In this time, Decide the author(s) based on the Content like topics covered. You need to evaluate based on metrics focused on contents includes:(a) Preferred Topics: Common themes or subjects frequently addressed by the author. (b)Domain-Specific Terms: Use of jargon or technical language tied to the author's expertise. Now i will start to give you the list of candidates for you to decide!
Writing Style	Please choose the top 5 most possible articles' author(s) among all the candidates with their corresponding information. In this time Decide the author based on the writing style. Metrics for evaluation include: (a) Writing Tone: Formal, casual, emotional, or neutral tone in the text; (b) Repetition Patterns: Tendency to repeat certain ideas, phrases, or structures; (c) Complexity: Use of compound or complex sentences, and overall readability level; (d) Paragraph Structure: Length and organization of paragraphs; (e) Vocabulary Usage: Word choices, diversity, and domain-specific terms; (f) Punctuation Patterns: Frequency and style of punctuation usage; (g) Sentence Length and Variation: Average length and variability of sentences; (h) Personal Pronouns and Voice: Usage of pronouns and active/passive voice; (i) Lexical Density: Ratio of content words to function words; (j) Rhythm and Flow: Natural sentence progression and rhythm.
Citation Similarity	Please choose the top 5 most possible articles' author(s) among all the candidates with their corresponding information. In this time Decide the author based on the citations. Different papers with the same author tend to share similarities in references. Therefore, please refer to References and Sources: Citation patterns, including the types of resources cited (e.g., scholarly papers, blogs).



The growing size of neural language models has led to increased attention in model compression. The two predominant approaches are pruning, ..., and distillation, ... but hardly achieve large speedups as distillation. However, distillation methods ... In this work, we propose a task-specific structured pruning method CoFi (Coarse- and Fine-grained Pruning), which delivers highly parallelizable subnetworks and matches the distillation methods in both accuracy and latency, without resorting to any unlabeled data. Our key insight is to jointly prune coarse-grained (e.g., layers) and fine-grained (e.g., heads and hidden units) modules, which controls the pruning decision ...

Writing Tone: formal, technical, and methodologically focused.
Repetition Patterns: "pruning," "distillation," and "models".
Complexity: short explanatory sentences, with longer descriptive ones
Paragraph Structure: Context → Methods → Limitations → Proposal → Innovation → Optimization → Results
Pacing: Balanced pacing, with a logical flow from problem to solution.

Existing methods, however, require either retraining, which is rarely affordable for billion-scale LLMs, or solving a weight reconstruction problem reliant on second-order information, which may also be computationally expensive. In this paper, we introduce a novel, straightforward yet effective pruning method, termed Wanda (Pruning by Weights and activations), ... Notably, Wanda requires no retraining or weight update, and the pruned LLM can be used as is. ..... Wanda significantly outperforms the established baseline of magnitude pruning and performs competitively against recent method involving intensive weight update.

Writing Tone: formal and academic, Formal and academic, but slightly more narrative in describing the problem and solution.
Repetition Patterns: "magnitude pruning," "sparsity," and "activations
Complexity: Moderate complexity with sentences that are slightly less dense
Paragraph Structure: Context → Existing Challenges → Methods → results. More descriptive.
Pacing: Smooth and engaging, with emphasis on novelty and results.

The popularity of LLaMA (Touvron et al., 2023a;b) and other recently emerged moderate-sized large language models (LLMs) highlights the potential of .... remains high. In this work, we study structured pruning as an effective means to develop smaller LLMs from pre-trained, larger models. Our approach employs two key techniques: (1) targeted structured pruning, which prunes a larger model ....(2) dynamic batch loading, ...We demonstrate the efficacy of our approach by presenting ... Sheared-LLaMA models outperform state-of-the-art open-source models of equivalent sizes, ...while requiring only 3% of compute compared to training such models from scratch. ....

Writing Tone: Highly formal and methodically detailed
Repetition Patterns: "pruning," "distillation," and "models".
Complexity: High complexity, with intricate sentence structures
Paragraph Structure: Motivation → Objectives → Methods → Challenges → Concluding aim.
Pacing: Deliberate and logical, with significant emphasis on methodology and results.

Text 1 and Text 3 are likely from the same author:

- Both share a highly formal tone, dense complexity, and methodologically detailed structure.
- The logical flow and pacing are consistent, starting with context, delving into the method, and ending with result.

Text 2 differs from Texts 1 and 3:

- Text 2 is more descriptive and narrative, focusing on simplifying the explanation.
- While it discusses pruning, it lacks the granular detail and precision seen in Texts 1 and 3.

Text 1 and Text 3 exhibit strong similarities in writing style

Figure 7: The GPT-4o analysis was conducted on three selected texts. To minimize the influence of topic variation on the LLM's ability to determine authorship, all three texts were chosen to focus on the same topics: model pruning and large language models (LLMs). The first and third text are from the same author (Xia et al., 2022, 2024). Therefore the analysis shown LLMs' ability to distinguish and identify authors based on writing style.

Table 8: Full Example of Different Writing Styles. Example 1 (Xia et al., 2022), Example 2(Sun et al., 2023), Example 3(Xia et al., 2024).

Class	Abstract
Example 1	The growing size of neural language models has led to increased attention in model compression. The two predominant approaches are pruning, which gradually removes weights from a pre-trained model, and distillation, which trains a smaller compact model to match a larger one. Pruning methods can significantly reduce the model size but hardly achieve large speedups as distillation. However, distillation methods require large amounts of unlabeled data and are expensive to train. In this work, we propose a task-specific structured pruning method CoFi (Coarse- and Fine-grained Pruning), which delivers highly parallelizable subnetworks and matches the distillation methods in both accuracy and latency, without resorting to any unlabeled data. Our key insight is to jointly prune coarse-grained (e.g., layers) and fine-grained (e.g., heads and hidden units) modules, which controls the pruning decision of each parameter with masks of different granularity. We also devise a layerwise distillation strategy to transfer knowledge from unpruned to pruned models during optimization. Our experiments on GLUE and SQuAD datasets show that CoFi yields models with over 10x speedups with a small accuracy drop, showing its effectiveness and efficiency compared to previous pruning and distillation approaches.
Example 2	As their size increases, Large Language Models (LLMs) are natural candidates for network pruning methods: approaches that drop a subset of network weights while striving to preserve performance. Existing methods, however, require either retraining, which is rarely affordable for billion-scale LLMs, or solving a weight reconstruction problem reliant on second-order information, which may also be computationally expensive. In this paper, we introduce a novel, straightforward yet effective pruning method, termed Wanda (Pruning by Weights and Activations), designed to induce sparsity in pretrained LLMs. Motivated by the recent observation of emergent large magnitude features in LLMs, our approach prunes weights with the smallest magnitudes multiplied by the corresponding input activations, on a per-output basis. Notably, Wanda requires no retraining or weight update, and the pruned LLM can be used as is. We conduct a thorough evaluation of our method Wanda on LLaMA and LLaMA-2 across various language benchmarks. Wanda significantly outperforms the established baseline of magnitude pruning and performs competitively against recent methods involving intensive weight update.
Example 3	The popularity of LLaMA (Touvron et al., 2023a;b) and other recently emerged moderate-sized large language models (LLMs) highlights the potential of building smaller yet powerful LLMs. Regardless, the cost of training such models from scratch on trillions of tokens remains high. In this work, we study structured pruning as an effective means to develop smaller LLMs from pre-trained, larger models. Our approach employs two key techniques: (1) targeted structured pruning, which prunes a larger model to a specified target shape by removing layers, heads, and intermediate and hidden dimensions in an end-to-end manner, and (2) dynamic batch loading, which dynamically updates the composition of sampled data in each training batch based on varying losses across different domains. We demonstrate the efficacy of our approach by presenting the Sheared-LLaMA series, pruning the LLaMA2-7B model down to 1.3B and 2.7B parameters. Sheared-LLaMA models outperform state-of-the-art open-source models of equivalent sizes, such as Pythia, INCITE, and OpenLLaMA models, on a wide range of downstream and instruction tuning evaluations, while requiring only 3% of compute compared to training such models from scratch. This work provides compelling evidence that leveraging existing LLMs with structured pruning is a far more cost-effective approach for building smaller LLMs.