Unraveling Interwoven Roles of Large Language Models in Authorship Privacy: Obfuscation, Mimicking, and Verification

Tuc Nguyen[1], **Yifan Hu**[2], **Thai Le**[1]
[1]Indiana University [2]Northeastern University
{tucnguye, tle}@iu.edu yif.hu@northeastern.edu

Abstract

Recent advancements in large language models (LLMs) have been fueled by large-scale training corpora drawn from diverse sources such as websites, news articles, and books. These datasets often contain *explicit* user information, such as person names, addresses, that LLMs may unintentionally reproduce in their generated outputs. Beyond such explicit content, LLMs can also leak identity-revealing cues through *implicit* signals such as distinctive writing styles, raising significant concerns about authorship privacy. There are three major automated tasks in authorship privacy, namely authorship obfuscation (AO), authorship mimicking (AM), and authorship verification (AV). Prior research has studied AO, AM, and AV independently. However, their interplays remain under-explored, which leaves a major research gap, especially in the era of LLMs, where they are profoundly shaping how we curate and share user-generated content, and the distinction between machine-generated and human-authored text is also increasingly blurred. This work then presents the first unified framework for analyzing the dynamic relationships among LLM-enabled AO, AM, and AV in the context of authorship privacy. We quantify how they interact with each other to transform human-authored text, examining effects at a single point in time and iteratively over time. We also examine the role of demographic metadata, such as gender, academic background, in modulating their performances, inter-task dynamics, and privacy risks. The code is available at https://github.com/nguyentuc/authorship_privacy.

1 Introduction

Recent advances in LLMs have been extraordinary, driven largely by the massive amounts of training data indiscriminately sourced from diverse online platforms such as websites, news outlets, and books (Brown, 2020; Le Scao et al., 2023; Touvron et al., 2023; Achiam et al., 2023). This training data often includes extensive writing contributions by the same authors, publicly shared across various platforms (Gao et al., 2020; Raffel et al., 2020). These sources frequently contain *explicit* user information, such as names, addresses, and phone numbers, which LLMs can inadvertently expose during their text generation process (Weidinger et al., 2021; Kim et al., 2024). Beyond explicit details, user identification can also be inferred from *implicit* information, such as their distinctive writing styles, that does not immediately give out the authors' identities. Research in human cognitive science and linguistics highlights that individual backgrounds significantly shape writing styles (Zheng et al., 2006; Cheng et al., 2023a; Deshpande et al., 2023; Xing et al., 2024; He et al., 2025), facilitating bidirectional inferences between implicit information (e.g., writing style) and explicit information (e.g., names, ages, or areas of expertise). Recent studies also reveal that text generated by LLMs can also capture human personality traits (Karra et al., 2022; Jiang et al., 2024a,b; Bang et al., 2024; An et al., 2024), and vice versa–i.e., explicit information about specific individuals or groups can be

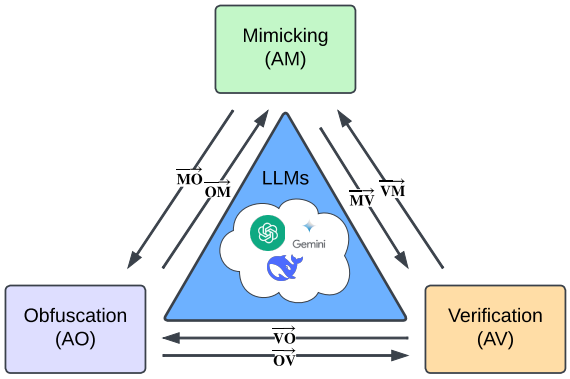

Figure 1: The authorship privacy interactive influence loop between LLMs, obfuscation, mimicking, and verification.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14915–14931
November 4-9, 2025

used by LLMs to produce personalized outputs or *mimic individuals' writing styles* (Chen et al., 2024; Salemi et al., 2024).

Although the authorship mimicking (AM) capabilities of LLMs–i.e., their ability to replicate an individual's writing style, are impressive, this capability could also enable malicious activities, such as impersonating public figures to spread misinformation or commit fraud (Deshpande et al., 2023; Jiang et al., 2024a). For instance, a fraudster could fine-tune an LLM on publicly available texts authored or spoken by a target victim (e.g., social media posts, interviews) and prompt LLMs to generate spam emails or persuasive messages that pretend to be delivered by the victim (Salewski et al., 2023). Contrasting with AM, authorship obfuscation (AO) (Uchendu et al., 2024) aims to conceal an author's identity by altering stylistic features of text while preserving its original meaning. By masking writing style before public dissemination (e.g., on social media), AO can help protect whistleblowers, such as writers or speakers, from potential anonymity exposure. In addition, authorship verification (AV) is the process of determining the author of a particular piece of writing. AV poses significant privacy risks by enabling the deanonymization of individuals through their writing style, which can facilitate surveillance, behavioral profiling, and misuse without informed consent.

While AO, AM, and AV have each been studied in isolation, their interactions within a unified framework remain underexplored or limited to only specific pairwise formulations, such as AV and AO in the context of LLM-generated text Uchendu et al. (2023). In addition, real-world scenarios often involve multiple rounds of text transformation, where content is repeatedly mimicked, obfuscated, and verified—either by different LLMs or within multi-turn dialogue settings where LLMs interact with one another (Duan et al., 2024). To address this gap, our study investigates three key scenarios in which LLMs play triple roles in authorship privacy (Fig. 1), analyzing their individual effects, interdependencies, and collaborative influences. Understanding the interplay among these capabilities is crucial for netizens in today's LLM era, where users may rely on LLMs to obfuscate their writing style, while others may utilize LLMs to recover or attribute the original authorship. **Our contributions** include: (1) the first unified framework for studying the bidirectional effects among AO, AM, and AV; (2) empirical findings revealing distinct task-specific strengths of various commercial LLMs; (3) detailed analysis showing how demographic and metadata influence these interactions. Our analysis shows that obfuscation tends to outperform mimicking in interactive settings, effectively disrupting authorial signals. However, mimicking can partially reverse obfuscation over successive cycles, gradually restoring aspects of the original writing style. Furthermore, models with stronger reasoning abilities (e.g., o3-mini, Deepseek) according to the benchmark [1], excel at verification and concealing authorial traits but are less effective at faithfully replicating an author's distinctive style.

2 Related Works

Beyond explicit metadata leakage such as names, social security numbers, LLMs' generations can also reflect *implicit and private authorship signals* such as writing style, tone, or rhetorical structure, many of which are uniquely identifiable to specific individuals (Zheng et al., 2006; Cheng et al., 2023a; Deshpande et al., 2023; Xing et al., 2024; He et al., 2025). Thus, these models may memorize and reproduce identifiable features of authorship through their generated texts, so-called AM, introducing interesting interwoven relationships with LLM-enabled AO and AV.

Authorship Obfuscation (AO) hides the original author's identity by altering stylistic cues without compromising semantic content. Recent methods include ALISON (Xing et al., 2024), which performs obfuscation by substituting stylistic sequences, and StyleRemix (Fisher et al., 2024), which utilizes AdapterMixup (Nguyen and Le, 2024) to train adapters for various stylistic dimensions and mix them. Different prompting-based approaches using LLMs have also been proposed (Hung et al., 2023; Pape et al., 2024).

Authorship Mimicking (AM) is the reverse of AO, aiming to generate text in the style of a specific author. LLMs excel in this task due to their few-shot and in-context learning capabilities, raising ethical concerns around impersonation, misinformation, and malicious use (Deshpande et al., 2023; Jiang et al., 2024a). Recent work has shown that LLMs can be fine-tuned or prompted to convincingly replicate individual writing styles from publicly available content (Salewski et al., 2023), making these capabilities intersect with privacy risks, such as

[1] https://www.vals.ai/benchmarks/math500-05-09-2025

when the LLMs leak memorized training examples (Carlini et al., 2023; Zhang et al., 2023).

Authorship Verification (AV) seeks to determine or confirm whether a given text was written by a particular author, based on linguistic cues or stylistic fingerprints (Huang et al., 2025). With the advancement of model size scaling laws, LLMs can now perform AV in few-shot settings (Hung et al., 2023; Huang et al., 2024).

Interdependency of AO, AM, and AV. Prior research has largely treated AO, AM, and AV in isolation. However, their **pairwise interactions**, especially under the influence of LLMs, remain underexplored and foundational to many practical scenarios. For instance, for AO–AV, users obfuscate their writing style to protect identity, while adversaries re-identify authorship, creating a privacy-versus-attribution dynamic; for AM–AV, attackers mimic a target author's style to deceive attribution models, challenging the robustness of verification systems; and for AO–AM, one can attempt to reconstruct authorial style from obfuscated text, testing the boundaries of stylistic recovery. Moreover, AO, AM, and AV can also form a closed loop in a **triplet-wise interaction**, reflecting how a text authorship changes under the influence of LLMs overtime. Our work is the first to address all pairwise and triplet-wise interdependencies of LLM-enabled AO, AM, and AV.

3 Research Questions and Formulation

3.1 Research Questions

We propose three **research questions (RQs)** to investigate both isolated and multi-level interdependencies among LLM-enabled authorship privacy tasks AO, AM, and AV, aiming to understand how individual and joint model behaviors influence the privacy and stylometry–i.e., writing styles, in complex authorship pipelines. Practical implications of our RQs are motivated in Appendix. A.1.

RQ1: How effectively can different LLMs perform AO, AM, AV *in isolation*, and which models are best suited for specific goals such as privacy preservation and stylistic imitation?

RQ2: How do LLM-enabled AO, AM, and AV influence one another to transform individuals' stylometries when used *in conjunction at one point in time*, including their pairwise and triplet interactions?

RQ3: How do LLM-enabled AO, AM, and AV influence one another to transform individuals' stylometries when used *in conjunction iteratively through time*?

To answer these RQs, we first formally define the evaluation of AO, AM, and AV of a target LLM $f(\cdot)$. For a given author a, let $\mathcal{D}_a=\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ represent a set of a's original written documents paired with their corresponding author labels. M_a denotes the metadata associated with author a, such as name, field of study. We define $C_a=\{M_a, \mathcal{D}_a\}$ as the context available to $f(\cdot)$. For example, $f^{AO}(x|C_a)$ denotes the output obfuscation text of LLM $f(\cdot)$ on the input text x given the context C_a. $d(\cdot)$ is a stylometric distance defined on the two sets of input texts.

3.2 Isolation - No Interdependency

We begin by formulating AO, AM, and AV in isolation to evaluate the standalone performance of a specific LLM f. This setting is the most common in prior work, where researchers aim to quantify how well an individual LLM performs on specific authorship privacy tasks (Hung et al., 2023; Huang et al., 2024; Fisher et al., 2024; Pape et al., 2024; Salewski et al., 2023).

AO. To evaluate the effectiveness of AO on an input text x, we compute the distance $d(\cdot)$ between the original authentic texts and the obfuscated one (Eq. 1). The larger the distance, the more divergent the obfuscated text becomes from the original, suggesting more effective obfuscation.

$$AO = d(f^{AO}(x|C_a),\ \mathcal{D}_a) \tag{1}$$

AM. We evaluate the effectiveness of AM on an input text x by computing the distance between the original texts and the mimicked text (Eq. 2). The smaller stylometric distance $d(\cdot)$, the more similar the mimicked text is to the original, suggesting more effective mimicking.

$$AM = d(f^{AM}(x|C_a),\ \mathcal{D}_a) \tag{2}$$

AV. We evaluate the effectiveness of AV on an input text x by comparing its binary predictive verification –i.e., whether the text was written by author a or by someone else (Eq. 3). The higher verification accuracy, the more effective $f(\cdot)$ is at correctly identifying the author's text.

$$AV = \mathbb{I}(f^{AV}(x|C_a) == a) \tag{3}$$

3.3 Pairwise Interdependency

Netizens are increasingly relying on LLMs to refine or disguise their writing through polishing, paraphrasing, or rephrasing, before sharing and publishing their content. These scenarios highlight a growing trend in which multiple LLMs are employed within a single pipeline: one model generates or modifies text, while another evaluates or attributes authorship. Consequently, the input to these models is not always original author-written text but may already have undergone AI-driven transformation (Uchendu et al., 2023). To better understand these interactions, we conduct *pairwise interdependency* evaluations that measure their bidirectional relationships–i.e., *how one LLM's capabilities influence the performance of others* (Fig. 1). To reflect the realistic scenario where the users prefer the best models for specific tasks, we designate a *"judge"* f_{judge} for each task, or the LLM that is selected based on its highest standalone performance in isolation (§ 3.2), for this evaluation.

Influence of Obfuscation. We factorize the influence of AO in the authorship pipeline into (1) how AO influences AM ($\overrightarrow{OM}$) and (2) how AO influences AV ($\overrightarrow{OV}$). For $\overrightarrow{OM}$, we first generate the obfuscated versions of an input text x, denoted x_{obf}, using various LLMs. Each of the obfuscated texts then serve as an input for the mimicking "judge" - a "ground-truth" LLM with the highest AM performance in isolation (§ 3.2), which attempts to reconstruct the original style of input x (Eq. 4). We compare the mimicked outputs to the original, authentic texts. The greater their stylistic divergence is, the more effective the obfuscated input, and hence the more influential the corresponding AO, and vice versa:

$$\overrightarrow{OM} = d(f^{\text{AM}}_{judge}(x|x_{obf}),\ \mathcal{D}_a) \tag{4}$$

For $\overrightarrow{OV}$, we pass the obfuscated texts x_{obf} to a verification "judge". We compute verification accuracy on the original input x given the obfuscated texts (Eq. 5). The lower the accuracy, the more effective the obfuscation is; otherwise, it suggests the author's style remains identifiable. This evaluation provides a practical measure of AO by testing whether others can still attribute the distorted writing to its original author. Such insights are particularly valuable in privacy-sensitive settings—e.g., anonymous investigative journalism or whistleblowing—where safeguarding the author's identity is paramount:

$$\overrightarrow{OV} = \mathbb{I}(f^{AV}_{judge}(x|x_{obf}) == a) \tag{5}$$

Influence of Mimicking. We factorize the influence of AM in the authorship pipeline into (1) how AM influences AO ($\overrightarrow{MO}$) and (2) how AM influences AV ($\overrightarrow{MV}$). For ($\overrightarrow{MO}$), we first generate mimicking versions of the input text x, denoted as x_{mimic}, using various LLMs. These mimicked texts then serve as the reference inputs for the obfuscation "judge". Then, we compare the resulting obfuscated outputs to the original, authentic texts (Eq. 6). Obfuscation style significantly diverging from the originals indicates that the mimicking was effective in replicating the author's writing style, and vice versa:

$$\overrightarrow{MO} = d(f^{AO}_{judge}(x|x_{mimic}),\ \mathcal{D}_a) \tag{6}$$

For $\overrightarrow{MV}$, we feed x_{mimic} into a verification "judge". We calculate the verification accuracy of the predictive author with x's original author a (Eq. 7). A high verification accuracy indicates that the mimicked text effectively replicates the original author's writing style, whereas a low accuracy suggests poor stylistic imitation:

$$\overrightarrow{MV} = \mathbb{I}(f^{AV}_{judge}(x|x_{mimic}) == a) \tag{7}$$

Influence of Verification. We factorize the influence of AV in the authorship pipeline into (1) how AV influences AO ($\overrightarrow{VO}$) and (2) how AV influences AM ($\overrightarrow{VM}$). In other words, AV acts as a filtering process to select only the texts verified as being authored by a as the input *contexts* for AO and AM. Intuitively, AV decides how pure or contaminated C_a is. To do this, we randomly sample n *noisy texts* or documents written by authors different from a, supposedly these are imposter samples. In both settings, we assess AV performance under two conditions: (1) *perfect* C_a*:* where all input context are genuine samples from the target author, and (2) *noisy* $\overline{C}_a$*:* where we introduce imposter samples from other authors that the model nonetheless classifies as the target author. Persistent positive classification of these imposter texts indicates weaker verification robustness. We then compute the distance of mimicking and obfuscation texts on the original input x, with the ground truth samples are all genuine and noisy (Eq. 8, Eq. 9).

$$\overrightarrow{VO} = d(f^{AO}_{judge}(x|C_a),\ f^{AO}_{judge}(x|\overline{C}_a)) \tag{8}$$

$$\overrightarrow{VM} = d(f^{AM}_{judge}(x|C_a),\ f^{AM}_{judge}(x|\overline{C}_a)) \tag{9}$$

Models	AO		AM		AV
	PPL (↑)	SIM (↓)	PPL (↓)	SIM (↑)	Acc (↑)
4o-mini	0.72	0.12	**0.65**	**0.13**	0.45
o3-mini	**2.71**	**0.10**	1.57	0.11	**0.89**
deepseek	1.08	0.11	1.86	0.12	0.74
gemini	0.31	0.12	1.00	**0.13**	0.39

Table 1: Isolation evaluation on AO, AM, and AV across different models. **Bold** and underline indicate each metric's best and second-best performance, respectively.

Dataset	# Exam	Avg length	Avg doc. length	Avg #sen. per doc.	# Authors
Speech	5,172	58.20	17.44	3.34	3
Quora	9,899	294.62	18.83	15.64	5
Essays	154	225.87	9.43	7.24	3

Table 2: Statistics of the evaluation datasets.

3.4 Triplet-wise Interdependency

While previous evaluations identify which models excel at individual tasks and how they are pairwise-interdependent, this section investigates *the authorship pipeline cycle as a whole* (Fig. 1)–i.e., how AO and AM alter verification accuracy and the linguistic distribution of original human texts. By orchestrating multiple LLMs, each deployed for its strongest capability, whether AO, AM or AV, we evaluate their collective impact on authorship privacy. This integrated perspective mirrors real-world workflows in which texts undergo successive AI-mediated transformations, from iterative edits in anonymous online forums to chained paraphrasing and verify in whistleblowing activities.

4 Experiment Setup

4.1 Models & Datasets

Models. We utilize the well-known commercial LLMs of varying presence of reasoning capability and origins: GPT-4o-mini (Achiam et al., 2023), GPT-o3-mini (Brown, 2020), Gemini-2.0 (Team et al., 2023), and Deepseek-v3 (Liu et al., 2024).

Datasets. We utilize three datasets: *Speech*: US Presidents' speeches from Fisher et al. (2024), *Quora*: Quora blog posts by diverse users with active online presence that we collect ourself; and *Essay:* writing essays from layperson (Li and Wan, 2025). These corpora vary in text length and author notoriety, descending from *Speech*, *Quora* and *Essay*. They also allow us to evaluate LLMs' performance on writing by both native and non-native English speakers. The dataset statistics are presented in Table 2. *Details of the datasets are provided in the Appendix. A.2.*

4.2 Prompt Construction

Following previous works such as LIP (Huang et al., 2024), we design prompts along four key dimensions: *Context*, *Task*, *Instruction*, and *Output* to characterize open-ended LLMs' behavior systematically (Cheng et al., 2023b). Specifically, we prompt LLMs to focus on writing style rather than differences in topic or content. *We provide details of the prompts design and ablation studies in the Appendix. A.3.*

4.3 Metrics

In our work, authorship privacy depends on identifying linguistic traits that are unique to individuals and can also help differentiate human-authored text from that generated by LLMs. Particularly, we examine how *4 key linguistic features* change *before and after* an authorship task AO, AM and AV is performed. Central to this is word distribution, quantified using TF-IDF similarity (denoted as **SIM**), which is also widely applied in detecting deepfake text by revealing unnatural or overly consistent vocabulary usage (Becker et al., 2023). Additionally, we evaluate language naturalness using perplexity (denoted as **PPL**) and also report the **KL** divergence over the distribution of text PPL scores. This metric is commonly employed to capture the natural writing patterns of individuals and to detect machine-generated text that may appear overly fluent or statistically optimized compared to genuine human writing. In our experimental setup, we conduct evaluations both with metadata C_a=$\{M_a, \mathcal{D}_a\}$ and without metadata C_a=$\{\mathcal{D}_a\}$. *Details of the metrics are provided in the Appendix. A.4.*

5 Experiment Results

5.1 Isolation Evaluation (RQ1)

Overall, *o3-mini* performs the best in AO and AV tasks, and *4o-mini* leads in faithful AM (Table 1). Particularly, *o3-mini* achieves the highest perplexity (2.71) and lowest similarity (0.10) in AO, indicating more distinct and less traceable outputs. For AM, *4o-mini* excels with the lowest PPL (0.65) and highest similarity (0.13), reflecting better stylistic imitation of the original texts. For AV, *o3-mini* identifies authorships with the highest accuracy (0.89).

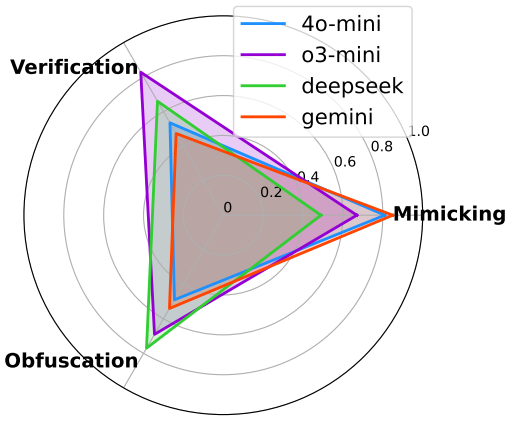

Figure 2: We present an overall pairwise interdependency evaluation of each LLM across the tasks of AO, AM, and AV. For each aspect, the final score is computed as the average across two "judge" evaluations to enable relative comparison.

	Models	Speech			Quora			Essay		
		KL	SIM	ACC	KL	SIM	ACC	KL	SIM	ACC
w. meta	*4o-mini*	0.14	0.08	0.71	1.21	0.19	0.67	1.99	0.18	0.59
	o3-mini	0.91	0.08	0.58	1.96	0.16	0.59	2.15	**0.15**	**0.51**
	gemini	0.23	0.09	0.66	1.82	0.18	0.69	1.83	**0.15**	0.61
	deepseek	**1.15**	**0.08**	**0.53**	**2.15**	**0.13**	**0.57**	**2.23**	**0.15**	**0.51**
w.o. meta	*4o-mini*	0.39	0.08	0.63	1.25	0.16	**0.62**	1.87	0.17	0.62
	o3-mini	1.41	0.07	0.59	1.84	0.16	0.63	2.01	**0.15**	**0.53**
	gemini	0.05	0.08	0.70	1.76	0.17	0.76	1.86	**0.15**	0.60
	deepseek	**1.76**	**0.06**	**0.52**	**1.85**	**0.16**	**0.62**	**2.21**	**0.15**	**0.53**

Table 3: Evaluation on obfuscation. KL (↑), SIM (↓), and Verification Accuracy (ACC) (↓) between the mimicked and original text.

5.2 Pairwise Interdependency (RQ2)

From the isolation evaluation (Sec. 5.1), we select *o3-mini* as both the obfuscation and verification judge, and *4o-mini* as the mimicking judge to assess the interplays among the authorship tasks. Fig. 2 presents a comprehensive comparison of the four models' influence capabilities across AO, AM, and AV. Overall, *gemini*, *deepseek*, and *o3-mini* are the most influential or effective with mimicking, obfuscation, and verification, respectively. We analyze each authorship task in detail as follows.

Influence of Obfuscation. To quantify AO, we employ a mimicking judge (*4o-mini*) and a verification judge (*o3-mini*). Table 3 reports the KL and SIM between mimicked and original texts and the verification accuracy on original texts when using obfuscated texts as the ground truth. Overall, among all models, *deepseek* consistently demonstrates the strongest obfuscation influence across all datasets, achieving the highest KL and lowest SIM scores. This indicates that its obfuscated outputs *deviate the most from the original writing style*.

	Models	Speech			Quora			Essay		
		KL	SIM	ACC	KL	SIM	ACC	KL	SIM	ACC
w. meta	*4o-mini*	3.25	**0.05**	0.73	2.51	0.17	0.78	2.32	0.20	0.68
	o3-mini	2.95	0.06	0.70	2.30	0.19	0.73	2.14	0.19	0.65
	gemini	**3.29**	**0.05**	**0.87**	**3.20**	**0.15**	**0.89**	**2.98**	**0.18**	**0.71**
	deepseek	2.95	0.07	0.65	2.18	0.18	0.82	1.97	0.21	0.67
w.o. meta	*4o-mini*	**3.32**	0.06	0.70	2.13	0.16	0.79	2.16	0.20	0.63
	o3-mini	3.26	0.06	0.62	2.24	0.19	0.64	1.98	0.22	0.60
	gemini	3.28	**0.05**	**0.82**	**2.48**	**0.15**	**0.87**	**2.79**	**0.19**	**0.69**
	deepseek	2.58	0.07	0.59	2.37	0.17	0.81	2.03	0.22	0.62

Table 4: Evaluation on mimicking. KL (↑), SIM (↓), and Verification Accuracy (ACC) (↑) between the obfuscation and original text.

In addition, the results also show that *obfuscation without user metadata generally outperforms the versions that incorporate metadata*. This suggests that metadata may inadvertently constrain the models, making it more difficult to mask the original writing style. In other words, the mimicking judge can utilize the same user metadata to reconstruct the original author's writing style, making the obfuscation less impactful. Furthermore, the performance gap between the with-metadata and without-metadata settings is most pronounced in the Speech dataset, which features more well-known authors. This gap progressively narrows in the Quora and Essay datasets, reflecting a trend: *it is easier to conceal the identity of less well-known authors, regardless of metadata inclusion*.

Influence of Mimicking. To quantify AM, we evaluate the mimicked texts using two distinct judges: an obfuscation judge (*o3-mini*) and a verification judge (*o3-mini*). Table 4 reveals several consistent trends across datasets. Gemini achieves the strongest overall performance in text obfuscation and verification, followed by 4o-mini, with Gemini leading in most KL (↑), SIM (↓), and ACC (↑) metrics. *Contrast with previous AO evaluation, incorporating user metadata to AM significantly enhances verification quality specially on Speech data*. Notably, the performance gap between settings with and without metadata narrows from well-known to lesser-known authors, suggesting that metadata plays a more critical role in capturing and disguising distinctive writing styles. Specifically, in the Speech dataset, the gap in KL divergence

	Models	$\overrightarrow{VO}$				$\overrightarrow{VM}$			
		Speech		Quora		Speech		Quora	
		KL	SIM	KL	SIM	KL	SIM	KL	SIM
w. meta	*4o-mini*	1.47	0.24	1.89	0.19	0.21	0.33	0.39	0.26
	o3-mini	**1.08**	**0.27**	**1.57**	**0.24**	**0.19**	**0.34**	**0.30**	**0.28**
	gemini	1.65	0.22	1.80	0.18	0.22	0.30	0.40	0.25
	deepseek	1.21	0.24	1.74	0.21	0.20	0.33	0.35	0.26
w.o. meta	*4o-mini*	1.72	0.22	1.91	0.17	0.34	0.29	0.41	0.25
	o3-mini	**1.24**	**0.24**	**1.60**	**0.23**	**0.24**	**0.31**	**0.36**	**0.27**
	gemini	1.71	0.18	1.83	0.17	0.33	0.28	0.43	0.25
	deepseek	1.45	0.21	1.72	0.20	0.29	**0.31**	0.38	0.26

Table 5: Evaluation on verification. KL (↓) and SIM (↑) measure similarity between two obfuscated texts. Full results are shown in Table A7.

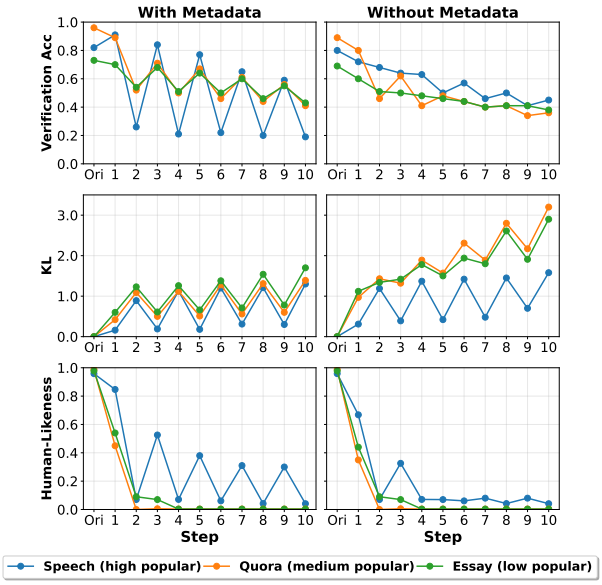

Figure 3: Verification accuracy (↑), KL (↓), and Human-likeness scores of mimicked and obfuscated texts compared to original texts across datasets, both with and without metadata. The x-axis represents the step order, ranging from 1 to 10 for 5 iterations *alternating between AM→AO→AM→...→AO*. AV is used as an intermediate step after AO and does not generate any texts, so we hide it for clarity. We refer to Table A8 for the detailed results.

and SIM metrics between the metadata and without-metadata settings is substantially larger for AO than for AM. This implies that metadata is more influential in AO or that AO is generally more effective than AM. One possible explanation is that the input text contains many identifiable linguistic patterns, making it easier to alter (for obfuscation) than to replicate (for mimicking).

Influence of Verification. We construct noisy samples $\overline{C}_a$ by doing AV across the 4 models, which then serve as inputs for obfuscation and mimicking judge. Overall, *o3-mini* achieves the highest precision and recall, with *deepseek* showing strong recall, while *4o-mini* and *gemini* perform less effectively in AV. We refer to Appendix. A.6 for detailed setup and results.

Table 5 reports how AV influences AO and AM when feeding AV with perfect (C_a) and noisy samples ($\overline{C}_a$). Overall, models with higher precision, indicating fewer *false positives* in $\overline{C}_a$ (Eq. 8, Eq. 9) and reduced noise in the few-shot ground truth, exhibiting smaller divergence between obfuscation texts generated with perfect and imperfect samples. This suggests that cleaner sample ground truth examples make the obfuscation texts more indistinguishable. Moreover, removing metadata during obfuscation amplifies the divergence between obfuscated texts, potentially because the obfuscation judge can utilize the metadata to force the obfuscated texts to be similar. Lastly, across datasets, the gap in KL and SIM becomes narrower as the author becomes less well-known, reflecting the *diminishing influence of author-specific features in obfuscation*.

In terms of $\overrightarrow{VM}$, overall, mimicked texts derived from ground-truth examples of LLMs with higher precision exhibit lower divergence, reflected by smaller KL and higher SIM, because higher precision reduces false positives and thus introduces less noise during the mimicking process. Additionally, *AV's access to metadata consistently improves the AM judge's ability to perform accurate text mimicking compared to settings without metadata, although this benefit diminishes as the authors become less well-known*. The reason might be LLMs' familiarity with famous people, and hence able to effectively utilize metadata.

5.3 Triplet-wise Interdependency (RQ2, RQ3)

This section analyzes five *iterative cycles* of AO, AM, and AV to evaluate how LLMs progressively shape stylometric patterns over time. Without loss of generality, we begin with mimicking followed by obfuscation, as their outputs are iteratively used as inputs for the subsequent task throughout the evaluation process. An interesting observation is the emergence of zig-zag patterns in all plots in Fig. 3, suggesting an ongoing "tug-of-war" between mimicking and obfuscation. Obfuscation appears to be more dominant, though the nature of this interplay varies depending on (1) the dataset and (2) the

presence or absence of metadata.

Authorship Verification. Overall, mimicking demonstrates the ability to recover the original text to some extent (first plot in Fig. 3). However, its effectiveness diminishes over successive iterations, due to the cumulative noise introduced by repeated obfuscation steps, which makes it increasingly difficult for the mimicker to reconstruct the original content. This degradation is particularly evident in the Quora and Essay datasets, where mimicking accuracy drops sharply after the first iteration. In terms of obfuscation, we observe a substantial reduction in verification accuracy for the Speech dataset compared to Quora and Essay. This suggests that *obfuscation is more effective when author identity is strongly encoded in the text*, as is the case for public figures whose speech styles are easily recognizable. Notably, *removing metadata from AO/AM consistently decreases verification accuracy across all datasets and iterations*, further demonstrating the value of auxiliary information in authorship verification.

Language Naturalness. Overall, KL divergence increases over iterations, mirroring verification trends and signaling growing linguistic drift from the original text (second plot in Fig. 3). Mimicking degrades over time, especially without metadata, while obfuscation consistently drives text away from its original form. Mimicking works best on shorter, structured texts like Speech, whereas obfuscation excels on longer, more variable texts like Quora and Essay due to richer linguistic features for distortion.

Anthropomorphism Analysis. We investigate whether generated text becomes more human-written or machine-generated through successive iterations of AM→AO (Cheng et al., 2025). To quantify this, we employ GPTZero[2], one of the most popular commercial deepfake text detectors, to assess the degree to which a given text *resembles human writing*. Fig. 3 reports the human-likeness score- GPTZero's estimated probability that a given text is written by a human. The first mimicked texts often appear most human-like, especially on the Speech dataset, while obfuscated texts consistently score low. Mimicking after obfuscation can partially restore human-like style, but this effect *fades over time as the text becomes increasingly machine-generated*. For Quora and Essay, texts generated after the second iteration are generally classified as machine-generated. This may be attributed to the lower popularity and variability in writing styles within these datasets, making it harder for mimicking models to recover stylistic patterns. Without metadata, this effect intensifies across all datasets, texts quickly adopt machine-like traits after two iterations, with minimal recovery by AM even in the Speech dataset.

Topic Distribution. We analyze how mimicking and obfuscation alter topic distributions using LDA (Blei et al., 2003) and find that iterative authorship tasks gradually shift texts away from their original themes. For instance, in the Speech dataset, the initial texts cover topics such as *politics*, *elections*, *health/life*, *war/terror*, and *economy/jobs* are replaced by more generic, repetitive content over time. This degradation may result from the *compounding effects of generation*, as LLMs tend to produce less specific and more repetitive content (Holtzman et al., 2020). Detailed topic trends are in Appendix A.9.

6 Discussions

Relationship between Authors' Popularity and Metadata's Effectiveness. Including metadata significantly boosts AV effectiveness, especially for well-known individuals, heightening privacy risks through easier re-identification or impersonation. Otherwise, lesser-known authors are less affected, indicating that popularity increases identifiability. While obfuscation helps, it does not reliably ensure anonymity. These results carry important implications for LLM providers like OpenAI, Google: (1) LLMs may unintentionally erode user privacy by leveraging publicly available or leaked metadata; second, (2) incorporating privacy-preserving mechanisms into authoring and editing tools; (3) providing transparency and safeguards around how metadata is used or inferred in LLM-driven authorship tasks.

The Double-edged Sword of LLMs: Empowering Privacy or Enabling Threats? LLMs are double-edged tools. On one hand, users can utilize LLMs for privacy-preserving purposes. For instance, whistleblowers or vulnerable individuals may rely on LLM-powered obfuscation tools to share sensitive content anonymously. On the other hand, the same technology can be misused for impersonation or misinformation. Our results show that LLMs can convincingly mimic writing styles, especially when metadata such as demographics

[2] https://gptzero.me/

is available, opening the door for social engineering attacks or deepfake text generation. Therefore, individuals must be aware that their public user-generated content, even absent explicit identifiers, can leave behind implicit rich digital traces. This raises an urgent need for tools that proactively evaluate and adjust online writings to minimize their digital traces.

Impersonation and Misuse at Scale. The interplay between AO and AM reveals that obfuscated text can still be reverse-engineered by powerful LLMs, especially with demographic cues. This poses real risks: malicious actors could impersonate public figures or institutions at scale to spread misinformation. As a result, stronger authorship detection tools are essential to identify AI-generated impersonations and trace their origins.

7 Conclusion

In this work, we introduce a unified framework for studying how LLMs engage with three interrelated dimensions of authorship: obfuscation (hiding identity), mimicking (imitating style), and verification (detecting authenticity). By examining these tasks together, we highlight how demographic metadata and model capabilities shape outcomes. Our findings show that obfuscation is generally more effective than mimicking at disrupting recognizable writing patterns, although mimicking can gradually restore elements of an author's style. Importantly, models with stronger reasoning skills are more successful at detecting and concealing authorship but less reliable at faithfully reproducing an individual's unique voice. These results have implications not only for computational linguistics and AI safety, but also for digital privacy, authorship studies, forensic linguistics, and the social sciences, where questions of identity, authenticity, and trust in machine-generated text are of central importance.

Limitation

Despite presenting a comprehensive evaluation framework for the three core authorship privacy tasks, namely authorship verification, obfuscation, and mimicking, using diverse linguistic metrics across a range of real-world datasets, our study is limited by the absence of human-centered evaluation. While automated metrics offer scalability and consistency, incorporating human judgment would provide valuable insights into the perceived naturalness, fluency, and effectiveness of obfuscated or mimicked text. This is especially important in assessing whether generated text truly conceals authorship or convincingly imitates another writing style from a human perspective. Future work could benefit from human-in-the-loop studies to better align evaluation with real-world perceptions and practical usability.

Broader Impacts and Ethics Statement

This work raises important ethical considerations regarding authorship privacy. While our framework supports the evaluation and improvement of privacy-preserving techniques, it also demonstrates how large language models (LLMs) can be leveraged to deanonymize or impersonate writers. Such capabilities pose risks to vulnerable individuals and create opportunities for misuse, including the spread of misinformation. We emphasize the need for safeguards, including tools that alert users to identifiability risks and more robust systems for detecting AI-generated content. All data used in this study are publicly available and were handled in accordance with established ethical research standards.

Acknowledgments

The authors thank the reviewers for their detailed feedback on this work. This work used Jetstream2 at Indiana University through allocations $\#CIS250090$, $\#CIS240570$ from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The authors also acknowledge the use of ChatGPT and Grammarly for minor editorial assistance.

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*.

Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2024. Measuring gender and racial biases in large language models. *arXiv*.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *arXiv*.

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR*.

Tom B Brown. 2020. Language models are few-shot learners. *NeurIPS*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *ICLR*.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *WWW*.

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. *arXiv*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. *ACL*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. Compost: Characterizing and evaluating caricature in llm simulations. *EMNLP*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *EMNLP*.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *NAACL*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell Gordon, Zaid Harchaoui, and Yejin Choi. 2024. Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements. *EMNLP*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv*.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. 2025. Cos: Enhancing personalization and mitigating bias with context steering. *ICLR*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ICLR*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *EMNLP*.

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

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Ka-Wei Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. *EMNLP*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024a. Evaluating and inducing personality in pre-trained language models. *NeurIPS*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. Personallm: Investigating the ability of large language models to express personality traits. *NAACL*.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv*.

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. *Conference and Labs of the Evaluation Forum*.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *NeurIPS*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arXiv*.

Jiatao Li and Xiaojun Wan. 2025. Who writes what: Unveiling the impact of author roles on ai-generated text detection. *arXiv*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv*.

Tuc Nguyen and Thai Le. 2024. Adapters mixup: Mixing parameter-efficient adapters to enhance the adversarial robustness of fine-tuned pre-trained text classifiers. *EMNLP*.

David Pape, Sina Mavali, Thorsten Eisenhofer, and Lea Schönherr. 2024. Prompt obfuscation for large language models. *arXiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *ACL*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. *ACL*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *NeurIPS*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv*.

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

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *SIGKDD*.

Adaku Uchendu, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Catch me if you gpt: Tutorial on deepfake texts. In *NAACL: Human Language Technologies (Volume 5: Tutorial Abstracts)*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv*.

Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Alison: Fast and effective stylometric authorship obfuscation. In *AAAI*.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *NeurIPS*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*.

A Appendix

A.1 Practical applications on authorship privacy

In this section, we present some applications of our research questions related to real-world authorship privacy.

RQ1: A practical application of this research question in authorship privacy is enabling users, such as whistleblowers, activists, or social media participants, to select the most suitable LLM for their goals. For instance, if a user seeks to mask their identity when writing sensitive content, the analysis can guide them toward models with strong authorship obfuscation (AO) performance. Conversely, a journalist or researcher aiming to emulate a public figure's writing style might benefit from models that excel in authorship mimicking (AM). Similarly, platforms concerned with detecting AI-generated or impersonated text can rely on models with high authorship verification (AV) accuracy. Thus, understanding isolated LLM performance informs the deployment of tailored models in real-world authorship privacy scenarios.

RQ2: A practical application of this research question in authorship privacy lies in improving the design and security of multi-step text processing pipelines used in sensitive communications. Specifically, in scenarios like anonymous online forums, whistleblower disclosures, or secure messaging, texts often undergo multiple transformations—generation, obfuscation, and verification—each performed by different LLMs. Understanding how these models influence one another and the interdependencies that arise helps identify potential privacy risks, such as:

1. whether obfuscation techniques are truly effective in concealing an author's style. For instance, whistleblowers and journalists who rely on textual obfuscation to anonymize their writing may still be at risk if LLMs can reverse-engineer their original style, allowing adversaries to trace the obfuscated text back to them.
2. anonymizing sensitive documents, e.g, legal testimonies or medical records, where ensuring that downstream mimicking models cannot recover the original author's style is critical for privacy protection.
3. evaluating the potential misuse of LLMs in impersonation attacks, such as forging stylistically similar content for deception or misinformation.
4. forensic investigations, where reliable verification must distinguish genuine statements from adversarially altered or mimicked texts. Additionally, content moderation systems can leverage these insights to detect and flag deceptive or impersonated content, enhancing online platform safety and trust.

RQ3: A practical application of this question is in developing robust authorship privacy tools that account for real-world scenarios where text undergoes multiple rounds of transformation. For instance, in environments like anonymous publishing platforms or secure communication channels, text might be repeatedly mimicked, obfuscated, and verified using different LLMs. Understanding how these iterative cycles influence each other helps identify how privacy can degrade or be preserved over successive edits. This knowledge allows designers to build more effective multi-stage pipelines that maintain author anonymity, prevent unintended leakage of writing style, and improve the reliability of verification methods, ultimately enhancing the security and trustworthiness of authorship privacy systems.

A.2 Additional statistics on evaluation dataset

We present the statistics on the evaluation dataset in Table A1.

Attribute	Value	Count
CEFR	B1_1	914
	B1_2	881
	A2_0	470
	B2_0	231
	XX_0	73
Acad. Genre	Sciences & Tech.	1,034
	Social Sciences	762
	Humanities	674
	Life Sciences	99
Lang. Env.	EFL	1,886
	ESL	610
	NS	73
Sex	F	1,430
	M	1,139

Table A1: Distribution of author attributes across 2,569 learners.

A.3 Prompt Construction

Prompt Construction. Author identification can be generated based on the attributes of each learner, including sex, academic background, level of English proficiency, and country of origin, to build a

more targeted background persona. For example: *The author is female. Her academic background is in the Humanities. Her English proficiency level is CEFR B1 (lower). She is from Singapore, an ESL environment (English as a Second Language).* The prompt construction for mimicking, attribution, and obfuscation are available in our GitHub repository.

Prompt Ablation Study. As part of our prompt ablation study, we designed a simplified prompt for the authorship verification (AV) task to ensure clarity and accessibility, even for non-expert users. The prompt explicitly instructs the model to decide whether a given text matches the style of a specific author based on a short description and balanced examples (10 from the target author and 10 from other authors). The task is constrained to a binary "yes" or "no" response, minimizing ambiguity and enforcing consistency. The full AV prompt is shown below:

> You are helping decide whether a piece of writing was written by a specific person. You will be shown 10 examples of writing by that person and 10 examples by other people. Based on the writing style, like word choice and structure, decide if the new input text matches the person's style. You will also be given a short description of the person. Your job is to answer only "yes" or "no", with no explanation. Here is some information about the person: *author identification*. Here are 10 examples of their writing: text from author. Here are 10 examples of writing by others: *text from others*. Now, here is the text you need to evaluate: *input text*.

The results of different LLMs on the authorship verification (AV) task using our simplified prompt are as follows: 4o-mini achieved an accuracy of 0.42, o3-mini achieved 0.84, Gemini achieved 0.37, and Deepseek achieved 0.72. While we observed slight variations in absolute accuracy compared to the original prompt, the relative ranking of the models remained consistent—o3-mini and Deepseek continued to outperform 4o-mini and Gemini. This reinforces our finding that the comparative effectiveness of different LLMs in AV is robust to prompt variations, underscoring the importance of focusing on relative rankings rather than exact performance figures.

Models	Speech		Quora		Essay	
	Precision	Recall	Precision	Recall	Precision	Recall
4o-mini	0.36	0.50	0.36	0.50	0.33	0.50
o3-mini	**0.67**	**0.80**	**0.54**	**0.70**	**0.50**	**0.70**
gemini	0.36	0.50	0.33	0.50	0.27	0.40
deepseek	0.62	**0.80**	**0.54**	**0.70**	0.43	0.60

Table A2: Authorship verification precision and recall of the four LLMs on the Speech, Quora, and Essay datasets.

A.4 Additional evaluation metrics

Word Distribution. We employ TF-IDF to quantify each word's significance within a document relative to the entire corpus. TF counts word occurrences, while IDF down-weights common terms. We extract TF-IDF vectors from our text sources and compute cosine similarity to assess stylistic and thematic alignment.

Language Naturality. Perplexity (PPL) evaluates how well a language model predicts a given text, with lower PPL reflecting greater confidence and closer adherence to learned linguistic patterns. Since LMs capture typical language structures from large corpora, PPL is a proxy for naturalness. Here, we fine-tune GPT-2 (Radford et al., 2019) on the original corpus and compute text-level PPL for both human-written and generated texts.

Attack Success Rate (ASR). ASR is used to evaluate the effectiveness of attacks in both tasks: for authorship verification, it measures how successfully an adversarial text deceives the verifier into making an incorrect authorship decision, while for authorship obfuscation, it captures how effectively the obfuscation hides the true author by misleading the verifier into misattributing the text.

BLEU. BLEU (Papineni et al., 2002) is used as a measure of linguistic similarity, capturing surface-level overlap between original and obfuscated text. By computing n-gram matches, it reflects how much of the lexical structure is preserved after transformation, offering insights into the degree of textual alteration introduced by an authorship privacy technique.

BERTScore. BERTScore (Zhang et al., 2020) is employed to measure semantic similarity, leveraging contextual embeddings from pretrained language models to compare the meaning of original and obfuscated text. Unlike BLEU, which focuses on exact lexical overlap, BERTScore evaluates

	Models	$\overrightarrow{AO}$			$\overrightarrow{AM}$		
		Speech	Quora	Essay	Speech	Quora	Essay
w. meta	*4o-mini*	0.29	0.33	0.41	0.27	0.22	0.32
	o3-mini	0.42	0.41	0.49	0.30	0.27	0.35
	gemini	0.34	0.31	0.39	0.13	0.11	0.29
	deepseek	0.47	0.43	0.49	0.35	0.18	0.33
w.o. meta	*4o-mini*	0.37	0.38	0.38	0.30	0.21	0.37
	o3-mini	0.41	0.37	0.47	0.38	0.36	0.40
	gemini	0.30	0.24	0.40	0.18	0.13	0.31
	deepseek	0.48	0.38	0.47	0.41	0.19	0.38

Table A3: ASR on AO, AM.

		1		2		3		4		5	
		AM	AO	AM	AO	AM	AO	AM	AO	AM	AO
w meta	*Speech*	0.09	0.74	0.16	0.79	0.23	0.78	0.35	0.80	0.40	0.81
	Quora	0.11	0.48	0.29	0.50	0.33	0.54	0.39	0.56	0.44	0.59
	Essay	0.30	0.46	0.32	0.49	0.36	0.50	0.40	0.54	0.45	0.57
wo meta	*Speech*	0.27	0.32	0.36	0.37	0.50	0.43	0.54	0.50	0.59	0.55
	Quora	0.20	0.54	0.38	0.59	0.52	0.56	0.60	0.59	0.66	0.64
	Essay	0.40	0.49	0.50	0.52	0.54	0.56	0.60	0.59	0.59	0.62

Table A4: ASR on the Triplet-wise Interdependency.

Models	$\overrightarrow{AO}$		$\overrightarrow{AM}$	
	BLEU ↓	BERTScore ↓	BLEU ↑	BERTScore ↑
4o-mini	0.22	0.85	0.22	0.88
o3-mini	0.20	0.83	0.21	0.87
gemini	0.24	0.86	0.23	0.89
deepseek	0.21	0.84	0.20	0.87

Table A5: Average BLEU, BERTScore.

whether the underlying semantics are preserved, providing a deeper assessment of how obfuscation techniques maintain or distort the intended meaning.

A.5 Additional experimental results

ASR on AO, AM. We provide ASR-based evaluations for AO and AM across all datasets and conditions (with/without metadata), as well as within the full iterative authorship pipeline (triplet-wise interdependency) in Table A3, Table A4, respectively. Overall, these ASR results support our original finding that obfuscation is generally more effective than mimicking in reducing verification accuracy, especially with models like Deepseek and o3-mini. Removing metadata modestly improves ASR (i.e., makes obfuscation more effective), particularly on the Speech dataset, reinforcing our claim that metadata aids verification and limits obfuscation success. On the task of AM, mimicking becomes less effective without metadata, as the model lacks key stylistic and demographic signals to guide imitation. Notably, Gemini and 4o-mini consistently achieve the lowest ASR (i.e., best mimicking), confirming our earlier conclusion that it is better suited for style replication despite weaker performance in reasoning-heavy tasks.

For the triple-wise interdependency evaluation, the ASR surges after each obfuscation step (even steps), and drops after mimicking (odd steps), but the trend shows that obfuscation gradually dominates, with ASR increasing overall. This supports our claim that obfuscation more effectively perturbs authorial signals than mimicking can recover them, especially when metadata is included.

BLEU, BERTScore on AO, AM. We report the average BLEU and BERTScore of AO and AM across the three datasets in Table A5. Overall, AO with Deepseek and o3-mini proves more effective for obfuscation, producing outputs with reduced lexical overlap and lower semantic similarity to the original texts. In contrast, Gemini performs best in mimicking, achieving the highest BLEU and BERTScore, which reflects stronger surface-level and semantic alignment with the target author. Notably, despite their strong reasoning capabilities, Deepseek and o3-mini yield slightly lower scores in this setting, highlighting a trade-off in which they are more adept at concealing than imitating writing style.

Additional baselines for AV, AO. We report additional baselines for AV and AO, including the non-LLMs methods in Table A6. For AV, we additionally use the PAN author verification model and n-grams, BERT. For AO, we report additional results for different non-LLMs authorship obfuscation techniques, including back-translation and synonym swapping. Specifically, for the back-translation method, we use round-trip machine translation by translating a text from English to German, then to French, and then back to English (Keswani et al., 2016). We use M2M translation models from (Fan et al., 2021). For synonym swapping, we utilize the PWWS word synonyms substitution strategy (Ren et al., 2019) for obfuscation.

A.6 Detailed results on Precision and Recall

We construct the imperfect ground truth examples $\overline{x}_p$ by sampling 20 examples from the original texts, including 10 from the author and 10 from others. The target model will be used to verify authorship.

Models	$\overrightarrow{AV}$	$\overrightarrow{AO}$			
	ACC ↑	PPL	SIM	BLEU	BERT
4o-mini	0.41	0.66	0.16	0.25	0.88
o3-mini	0.85	2.53	0.13	0.22	0.84
gemini	0.35	0.25	0.17	0.26	0.89
deepseek	0.72	0.98	0.14	0.23	0.85
PAN(non-LLMs	0.78	-	-	-	-
n-grams (non-LLMs)	0.68	-	-	-	-
BERT-based	0.65	-	-	-	-
Synonym Swapping	-	0.91	0.21	0.31	0.92
Back-Translation	-	1.03	0.19	0.34	0.90

Table A6: Additional baselines comparison of AV, AO.

All the examples classified as correct verification will be used as the ground truth for the obfuscation and mimicking processes. Table A2 shows detailed results on Precision and Recall.

A.7 Additional results for VO and VM

We present detailed evaluation results of VO and VM in Table A7.

A.8 Additional evaluation results on triplet-wise interdependency

We present detailed evaluation results on triplet-wise interdependency in Table A8.

A.9 Detailed results on topic distribution analysis

Table A9 shows detailed results on topic distribution analysis on the mimicking and obfuscation process.

	Models	VO						VM					
		Speech		Quora		Essay		Speech		Quora		Essay	
		KL	SIM	KL	SIM	KL	SIM	KL	SIM	KL	SIM	KL	SIM
w. meta	*4o-mini*	1.47	0.24	1.89	0.19	1.34	0.28	0.21	0.33	0.39	0.26	0.69	0.18
	o3-mini	**1.08**	**0.27**	**1.57**	**0.24**	**1.26**	**0.31**	**0.19**	**0.34**	**0.30**	**0.28**	**0.52**	**0.19**
	gemini	1.65	0.22	1.80	0.18	1.51	0.28	0.22	0.30	0.40	0.25	0.63	0.17
	deepseek	1.21	0.24	1.74	0.21	1.32	0.29	0.20	0.33	0.35	0.26	0.64	0.17
wo. meta	*4o-mini*	1.72	0.22	1.91	0.17	1.35	0.27	0.34	0.29	0.41	0.25	0.66	0.17
	o3-mini	**1.24**	**0.24**	**1.60**	**0.23**	**1.32**	**0.29**	**0.24**	**0.31**	**0.36**	**0.27**	**0.54**	**0.18**
	gemini	1.71	0.18	1.83	0.17	1.49	0.28	0.33	0.28	0.43	0.25	0.63	**0.18**
	deepseek	1.45	0.21	1.72	0.20	1.34	0.28	0.29	**0.31**	0.38	0.26	0.65	0.17

Table A7: Merged results from both evaluations: **Verification Obfuscation** and **Verification Mimicking**. KL (↓) and SIM (↑) measure similarity between two obfuscated texts. **Bold** and underline indicate best and second-best performance per category.

			Verification										KL									
		Original	1		2		3		4		5		1		2		3		4		5	
			AM	AO	AM	AO	AM	AO	AM	AO	AM	AO	AM	AO	AM	AO	AM	AO	AM	AO	AM	AO
w meta	*Speech*	0.82	0.91	0.26	0.84	0.21	0.77	0.22	0.65	0.20	0.59	0.19	0.16	0.89	0.19	1.13	0.18	1.20	0.31	1.21	0.30	1.30
	Quora	0.96	0.89	0.52	0.71	0.50	0.67	0.46	0.61	0.44	0.56	0.41	0.42	1.09	0.50	1.12	0.51	1.28	0.56	1.31	0.60	1.39
	Essay	0.73	0.60	0.54	0.58	0.51	0.50	0.49	0.46	0.45	0.46	0.43	0.60	1.23	0.61	1.26	0.66	1.38	0.71	1.54	0.78	1.70
wo meta	*Speech*	0.80	0.72	0.68	0.64	0.63	0.50	0.57	0.46	0.50	0.41	0.45	0.31	1.19	0.39	1.37	0.42	1.42	0.48	1.45	0.70	1.58
	Quora	0.89	0.80	0.46	0.62	0.41	0.48	0.44	0.40	0.41	0.34	0.36	0.97	1.43	1.32	1.89	1.57	2.31	1.89	2.80	2.17	3.20
	Essay	0.69	0.60	0.51	0.50	0.48	0.46	0.44	0.40	0.41	0.41	0.38	1.12	1.34	1.42	1.78	1.50	1.98	1.80	2.61	1.91	2.90

Table A8: Performance analysis across 5 iterations (AM: mimicking, AO: obfuscation) for Verification and KL Divergence metrics.

Topic	Top Words
0	day, election, going, people, help, votes, working, could, got, better
1	weapons, tax, best, let, people, made, could, plan, give, think
2	people, country, time, right, look, together, one, border, want, believe
3	iraq, health, costs, people, team, looking, war, year, care, working
4	people, jobs, american, time, america, states, think, right, work, put
5	new, nation, america, american, years, right, peace, workers, great, drug
6	want, people, terrorists, important, college, enforcement, asking, terror
7	one, security, people, country, war, life, let, never, america, american
8	going, government, economy, world, america, afghanistan, iraq, getting, history, go
9	want, going, people, americans, think, true, test, save, health, support

Table A9: Top 10 words for each LDA topic on the original Speech dataset