De-Anonymization at Scale via Tournament-Style Attribution

Lirui Zhang
Beihang University
fysszlr@buaa.edu.cn

Huishuai Zhang*
Peking University
zhanghuishuai@pku.edu.cn

Abstract

As LLMs rapidly advance and enter real-world use, their privacy implications are increasingly important. We study an authorship de-anonymization threat: using LLMs to link anonymous documents to their authors, potentially compromising settings such as double-blind peer review. We propose De-Anonymization at Scale (DAS), a large-language-model–based method for attributing authorship among tens of thousands of candidate texts. DAS uses a sequential progression strategy: it randomly partitions the candidate corpus into fixed-size groups, prompts an LLM to select the text most likely written by the same author as a query text, and iteratively re-queries the surviving candidates to produce a ranked top-k list. To make this practical at scale, DAS adds a dense-retrieval prefilter to shrink the search space and a majority-voting–style aggregation over multiple independent runs to improve robustness and ranking precision. Experiments on anonymized review data show DAS can recover same-author texts from pools of tens of thousands with accuracy well above chance, demonstrating a realistic privacy risk for anonymous platforms. On standard authorship benchmarks (Enron emails and blog posts), DAS also improves both accuracy and scalability over prior approaches, highlighting a new LLM-enabled de-anonymization vulnerability.

1 Introduction

Large language models (LLMs) have seen rapid and widespread adoption due to their remarkable ability to generate human-like text and follow complex instructions across domains (Zhao et al., 2023; Chang et al., 2024; Aw et al., 2023). Such wide deployment in real-world systems, fundamentally reshapes how human interact with information.

Alongside their impressive capabilities, however, LLMs also introduce new *privacy concerns*. One emerging risk is the potential for LLMs to undermine anonymity in settings where privacy is paramount – for example, double-blind academic peer review[1], whistleblower forums, or anonymous communication platforms. These systems depend on keeping authors' identities hidden to protect fairness and safety, yet LLMs' strong text analysis abilities may enable them to infer identifying signals, such as distinctive writing patterns or domain expertise, and thereby de-anonymize content intended to remain anonymous (Staab et al., 2023; Nyffenegger et al., 2023)

Existing work on authorship attribution offers limited guidance for this privacy risk. Traditional authorship attribution (AA) in NLP and forensic linguistics (Stamatatos, 2009; Neal et al., 2017; He et al., 2024) is usually studied in a **closed-set** setting: a small, fixed list of candidate authors is given, and each author has labeled writing samples. This assumption underlies many stylometry benchmarks and PAN competitions (Bevendorff et al., 2022; Stamatatos et al., 2018; Bevendorff et al., 2023), where the candidate pool is often only tens of authors and building per-author profiles or classifiers is feasible.

Real-world anonymous systems look very different: there may be tens of thousands of possible authors and no pre-labeled profile texts for any of them. In double-blind peer review, for example, the reviewer pool for major AI venues can be extremely large, yet we typically lack labeled writing from each reviewer. Authorship attribution at this scale, under such minimal assumptions, remains underexplored. Recent work Huang et al. (2024a) prompts GPT-3/4 for author attribution on blogs and emails, it still considers relatively small candidate sets. In practice, de-anonymization may require searching tens of thousands of candidate texts with minimal supervision, a scenario where

*Corresponding author.

[1] https://openreview.net

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32270–32283

previous methods break down or become computationally infeasible, motivating new techniques for LLM-based de-anonymization at scale.

In this paper, we introduce **De-Anonymization at Scale (DAS)**, an LLM-based method for authorship matching in anonymous systems with tens of thousands of candidate texts. DAS uses a *progressive elimination strategy*: we randomly partition candidate texts into small groups, prompt the LLM to select the most likely match to a query within each group, i.e., a one-to-many identification process, and iteratively re-group and re-compare the surviving candidates until we obtain a ranked top-k list.

To scale under a restricted LLM token budget, DAS employs a *coarse-grained retrieval module* (Lewis et al., 2020) to shrink the search space. Given a query text, embedding-based retrieval narrows a corpus of up to 10^5 candidate texts to, say, the top 10^3. This retrieval step acts as a coarse but efficient sieve, ensuring that only the most likely candidates proceed to the expensive LLM comparisons. This design makes large-scale de-anonymization computationally and economically feasible.

To further improve the accuracy, we employ a *majority-voting-style scoring system* to enhance robustness with multiple independent runs. DAS repeats the progressive selection with different random partitions, and each time assigns scores to candidates that win comparisons. By aggregating these scores of multiple runs, DAS produces the final ranking, favoring texts that consistently match the query across runs.

We evaluate DAS on both real-world and benchmark datasets. On anonymized double-blind peer-review data, DAS identifies same-author reviews from pools of thousands at rates well above chance, providing evidence that modern LLMs can threaten reviewer anonymity. On standard benchmarks, DAS also improves accuracy and scalability over direct LLM prompting, including the Enron email corpus (Klimt and Yang, 2004) and a large blogger dataset (Schler et al., 2006). Overall, our results show that LLM-enabled de-anonymization is a practical risk, motivating the development of stronger mitigation and privacy safeguards in the era of omnipresent large language models.

The contributions are summarized as follows.

- **New Privacy Risk.** We identify a realistic and underappreciated privacy threat: by leveraging the power of modern LLMs, an adversary can de-anonymize texts in systems like double-blind peer review at rates significantly above chance. This finding urges a fundamental rethinking of how anonymous platforms are designed and secured against stylometric attacks.
- **Technical Contribution.** We propose **De-Anonymization at Scale (DAS)**, a two-stage pipeline combining (1) a coarse dense-retrieval filter to narrow down tens of thousands of candidates to a short list and (2) an LLM-based sequential progression with a majority-voting scoring mechanism for fine-grained authorship attribution. This design enables both efficiency and accuracy in massive open-set authorship attribution scenarios.
- **Empirical Contribution.** Through extensive experiments on anonymized peer-review data and public benchmarks (Enron emails, large blogging corpora), we show that DAS substantially outperforms random guessing and existing LLM-based stylometric methods, achieving high accuracy and scalability in challenging, real-world deanonymization tasks.

Finally, despite the difficulty of attributing authorship in massive anonymized collections, our method is remarkably simple, further underscoring the severity of this risk.

2 Related Work

Classical Authorship Attribution. Authorship attribution (AA) has a long history in statistical stylometry. Traditionally, AA methods relied on human-defined linguistic features that capture an author's writing style (Holmes, 1994; Stamatatos, 2009). Researchers engineered features such as character/word n-gram frequencies, vocabulary richness, function word usage, syntactic patterns, and other stylometric markers (Seroussi et al., 2014). These features, combined with machine learning classifiers (e.g., Bayesian or SVM-based techniques), proved effective on closed-world problems with small to medium author sets (Madigan et al., 2005; Koppel and Winter, 2014; Koppel et al., 2007; Bevendorff et al., 2022). Comprehensive overviews of authorship attribution, including taxonomies of tasks (closed-set vs. open-set, verification vs. profiling) and feature categories, are provided by Stamatatos (2009); He et al. (2024).

Authorship Attribution with Large Language Models. The recent generation of large language models (LLMs) has opened new avenues for authorship analysis. Unlike fixed feature extractors, LLMs can be prompted to perform complex NLP tasks zero- or few-shot, without fine-tuning on task-specific data. Initial studies have started to evaluate LLMs on authorship attribution and verification. Hung et al. (2023) prompt LLMs to produce explanations for authorship verification. Huang et al. (2024a) systematically tested GPT-3.5 and GPT-4 on both verification and closed-set attribution with up to tens of candidates, where zero-shot GPT-based models can match or even surpass fine-tuned BERT classifiers on certain datasets. By employing specially crafted prompts that encourage the LLM to explain its decisions, they also extract human-interpretable justifications for the model's predictions. Gorovaia et al. (2024) similarly showed that an LLM (GPT-3) could robustly verify authorship of Latin texts in a zero-shot manner. Overall, these works suggest LLMs hold promise for authorship tasks, especially when fine-tuning data is scarce.

De-anonymization, Privacy, and Adversarial Stylometry. Authorship attribution techniques pose dual-use concerns: the same tools that identify authors can undermine anonymity and privacy. The security community has long studied "stylometric attacks", where an adversary de-anonymizes an author by matching their writing style across anonymous texts. A landmark study (Koppel et al., 2006) illustrated that distinguishing tens of thousands of authors is theoretically feasible given sufficient text. Subsequent work demonstrated the practicality of such attacks and also how authors might evade them (Brennan et al., 2012; Emmery et al., 2021). At the same time, researchers have developed methods to detect when text has been manipulated or when a writing style is inconsistent (Bevendorff et al., 2024).

The broader implications of such de-anonymization capabilities have been highlighted in survey (Huang et al., 2025), which call for more research on privacy-preserving text rewriting and adversarial robustness in authorship analysis.

3 Method

This section presents the DAS framework for authorship attribution in large corpora. Section 3.1 formalizes the de-anonymization problem. Section 3.2 introduces DAS's three-phase processing pipeline. Section 3.3 elaborates on the key algorithmic modules, including Coarse Filtering, LLM-based Matching, and Progressive Elimination.

3.1 The De-Anonymization Problem

The de-anonymization task in this paper aims to identify documents sharing the same author within a large corpus of anonymized texts. Formally, let $\mathcal{C} = \{d_i\}_{i=1}^{N}$ denote a collection of documents where each document $d_i = (t_i, a_i)$ consists of text content t_i and hidden author identity $a_i \in \mathcal{A}$. Given a target document $d_t = (t_t, a_t)$, the task's objective is to obtain a subset of documents:

$$\mathcal{R}^* = \{d_j \in \mathcal{C} \setminus \{d_t\} \mid a_j = a_t\}\,. \quad (1)$$

3.2 De-Anonymization at Scale Framework

Our De-Anonymization at Scale (DAS) framework addresses the "needle in a haystack" challenge of authorship attribution through a dual-stage architecture. To maintain both computational feasibility and high precision, DAS decomposes the problem into two sequential objectives: **Stage 1 (Coarse Filtering)** rapidly narrows the search space from tens of thousands to hundreds of candidates, which keeps high recall rate and reduces the set of candidates significantly, and when the search space is manageable, **Stage 2 (Tournament-Style Attribution)** utilizes the deep reasoning capabilities of LLMs to pinpoint candidate texts of the same author. Notably, Stage 1 is optional and can be omitted when the candidate set is not large.

3.2.1 Coarse Filtering

In DAS, we use vectorized retrieval to accomplish the Coarse Filtering mission, in order to address the efficiency bottleneck. It transforms the entire corpus into a stylistic vector embedding. As detailed in Algorithm 1, we use a CoSENT-based model (Huang et al., 2024b; Xu, 2023)—a contrastive learning framework optimized for sentence embedding that normalizes cosine similarity scores to enhance semantic alignment—to generate fixed-dimension stylistic embeddings for every document. By computing the cosine similarity between the target document and the corpus, we solve the first subproblem: identifying a reduced candidate set with top-K maximized similarities.

Importantly, this phase has to retain the ground-truth matches, i.e., high recall rates, when reducing the size of candidate set.

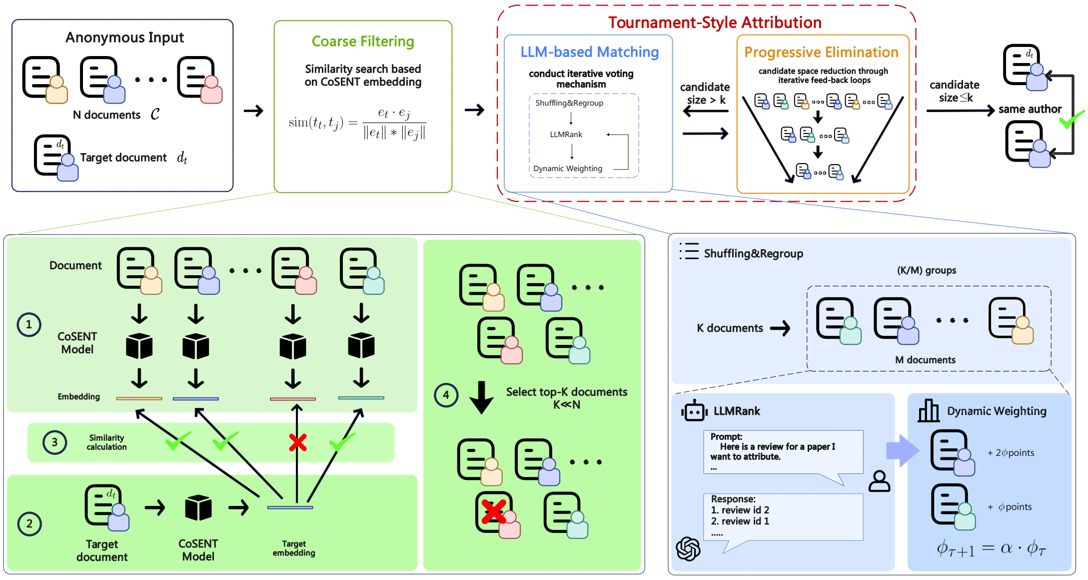

Figure 1: The DAS framework addresses the challenge of authorship attribution in large document corpora through a sequential progression strategy. The system operates in two core phases: (1) **Coarse Filtering** to narrow down candidate authors from thousands to a tractable subset, (2) **Tournament-Style Attribution (TSA)** that integrates LLMRank with dynamic weighting, multiple independent trials, and progressive elimination to iteratively reduce the candidate space and pinpoint the target author. This framework enables efficient analysis of long-form texts while maintaining computational feasibility at scale.

Algorithm 1 Coarse Filtering

Input: Target document d_t, Corpus $\mathcal{C}$, Size K
Output: Filtered pool $\mathcal{U}$

$e_t \leftarrow \mathrm{CoSENT}(t_t)$
for each document $d_j \in \mathcal{C}$ **do**
  $e_j \leftarrow \mathrm{CoSENT}(t_j), \quad s_j \leftarrow \frac{e_t \cdot e_j}{\|e_t\| \cdot \|e_j\|}$
  Store (d_j, s_j) in list L
end for
Sort L descending by s_j;
$\mathcal{U} \leftarrow L[1:K]$
return $\mathcal{U}$

3.2.2 Tournament-Style Attribution (TSA)

Once the search space becomes manageable, DAS initiates Tournament-Style Attribution (Algorithm 2). This phase addresses the second subproblem: determining the final identity via a "survival of the fittest" style comparison. Next, following Algorithm 2, we detail its critical design components.

Progressive Elimination. Rather than processing candidates simultaneously, the pool is partitioned into small groups. An LLM acts as a judge, performing deep linguistic analysis to rank candidates within each group. By immediately removing lower-ranked candidates after each LLM query, it prevents the system from wasting resources on unlikely matches. This recursive reduction allows the LLM to focus on the most subtle stylistic differences among the "finalists", significantly increasing attribution accuracy while enjoying the tournament's efficiency.

Specifically, in Algorithm 2, each group comparison retains 2 survivors and eliminates the remaining $l-2$ candidates. For ϕ iterates as the size of the candidate set decreases to give higher scores to trusted documents. For each group, top-ranked documents receive 2ϕ points, secondary choice ϕ points.

LLMRank. We randomly shuffle candidates and partition them into groups of size (l) to reduce ordering effects and fit the LLM context window. For each group $G_j = \{d_{j1}, \ldots, d_{jl}\}$, we prompt the LLM with (i) the target text (d_t), (ii) the texts in (G_j), and (iii) instructions to compare stylistic cues, and ask it to rank candidates by likelihood of sharing the same author. The prompt template for paper review de-anonymization is provided in Figure 2.

Dynamic Weighting. To reinforce confidence in survivors, we use an exponential weighting mechanism . As the rounds progress and the candidate pool shrinks, the scores awarded to survivors increase:

$$\phi_{\tau+1} = \alpha \cdot \phi_\tau, \tag{2}$$

where ϕ_τ is the current round and α is the accel-

Prompt

Here is a review for a paper I want to attribute.
{Target Review}
Below is a list containing the reviews of multiple papers.
{Reviews List}
Please analyze the writing style and structure of the review provided above and compare it with the reviews in the list. Specifically, focus on factors such as sentence length, complexity, tone, word choice, and any recurring stylistic features or habits in the writing.
Please try to identify and choose reviews in the list that most likely belong to the same author as the review I want to attribute.
Do not include as one of the options the self of the review I want to attribute!
Just output exactly five ID of the reviews you picked from the list directly, such as '1. review ID: number'!
Additionally, please sort the output based on confidence level, with the most likely reviews appearing first.

Figure 2: LLMRank prompt for paper review scenario.

Algorithm 2 Tournament-Style Attribution (TSA)

Input: target text t, Filtered pool $\mathcal{U}$, Trials m, Scaling factor α, group size l, final k;
Output: Ranked top-k IDs $\hat{\mathcal{R}}$;

Initialize global scores $\mathcal{S} \leftarrow \{d : 0 \mid d \in \mathcal{U}\}$
for Trial $t = 1$ **to** m **do**
  $\mathcal{O} \leftarrow$ shuffle($\mathcal{U}$) {A new random trail}
  $\phi \leftarrow 1$ {A new random trail}
  while $|\mathcal{O}| > k$ **do**
    Partition $\mathcal{O}$ into groups $\{G_1, \ldots, G_p\}$ of size l (last may be smaller),
    Let $\tilde{\mathcal{O}} \leftarrow \emptyset$,
    for each group G_j **do**
      $\{ID_1, ID_2\} \leftarrow$ LLMRank(d_t, G_j) {Top 2 IDs in G_j that matches d_t most}
      $\mathcal{S}[ID_1] \leftarrow \mathcal{S}[ID_1] + 2\phi$ {Top rank reward}
      $\mathcal{S}[ID_2] \leftarrow \mathcal{S}[ID_2] + \phi$ {Runner-up reward}
      $\tilde{\mathcal{O}} \leftarrow \tilde{\mathcal{O}} \cup \{ID_1, ID_2\}$
    end for
    $\mathcal{O} \leftarrow \tilde{\mathcal{O}}$, {Progressive Elimination}
    $\phi \leftarrow \phi \cdot \alpha$ {Dynamic Weighting}
  end while
end for
$\hat{\mathcal{R}} \leftarrow$ top-k IDs in $\mathcal{U}$ sorted by $\mathcal{S}[\cdot]$
return $\hat{\mathcal{R}}$

eration rate. This ensures that authors who consistently *win* their groups across multiple trials accumulate the highest global scores .

Multiple Trials. To be precise, Algorithm 2 forms a *multi-round survival tournament*. We run the tournament for (m) independent trials, each with a fresh random shuffle and regrouping of the candidate pool, which reduces sensitivity to any single partition and mitigates stochastic LLM behavior. In each trial, candidates that survive more rounds (and are selected more frequently within groups) accumulate higher scores. We then aggregate scores across trials (majority-voting style) and rank candidates by their total score, so texts that consistently match the query under different random groupings rise to the top while spurious matches are suppressed.

The score map persists across all trials. Even if a document is accidentally eliminated in one trial due to LLM noise, its consistent performance in other trials will keep its global ranking high.

So far, we have established the DAS framework and its underlying design philosophy. Next, we conduct experiments to validate its effectiveness.

4 Experiments

In this section, we evaluate the effectiveness and practical implications of DAS. Section 4.1 studies real-world risk on anonymized peer-review data. Section 4.2 evaluates DAS on cross-domain benchmarks, including blogs and emails. Section 4.3 analyzes sensitivity across genres by tracking rank progression over rounds. Finally, Section 4.4 reports controlled ablations to quantify the contribution of each component.

4.1 Evaluation on Anonymous Peer Reviews

To assess DAS in a realistic anonymous setting, we first evaluate it on anonymized conference peer-review data. This setup models an adversary attempting to compromise double-blind review by linking reviews of the same author via automated authorship analysis.

4.1.1 Experiment Setup

Specifically, we target on the following setup: Given a target annoymous review, where the underlying reviewer identity is hidden, we apply DAS to retrieve other reviews likely written by the same reviewer.

Dataset Construction. We build a benchmark from ICLR 2023–2025 OpenReview records, covering 23,889 submissions and 147,367 anonymized reviews collected via the OpenReview API. Each instance includes submission metadata (e.g., title and abstract) and an anonymized review with a randomized ID. Table 1 summarizes key statistics. Given a query review

Human Evaluation. Since reviewer identities are not observable, we validate DAS with a human study. Nine participants provided 25 test cases and supplied one-shot judgments (see Appendix A) indicating which retrieved candidates were written by themselves. Because the study involves sensitive anonymous reviews from real-world reviewers, we minimized data collection: we stored no personal information and retained only aggregated voting outcomes over candidate ranges for analysis. This protocol prioritized and relied on mutual trust—participants provide honest labels, and we did not collect or retain identifiable data.

Year	2023	2024	2025
Paper number	4955	7262	11672
Review number	27095	44771	75501
Mean Length	3175	2562	2524

Table 1: Statistical characteristics of the review dataset.

Implementation Details of DAS. Given a target review r_t, we first apply coarse filtering to retrieve the *top*-2000 candidates. We then run the fine-grained stage for $m = 5$ independent trials: in each trial, candidates are randomly grouped into clusters of $l = 10$ reviews and compared by an LLM (Gemini-2.0-flash unless noted), using exponential round weighting with $\alpha = 5$. The tournament stops once $k \leq 20$ candidates remain, and we aggregate scores across trials to produce the final ranking. All experiments run on a single server with 4 vCPUs (Xeon Platinum 8269CY) and 8GB RAM. Each trial uses approximately 6,000 tokens (about ~$1 USD) and, with parallelization, finishes in 2~3 minutes.

Evaluation Metric and Baseline. We use **Rank@k** to evaluate the effectiveness of the de-anonymization task, defined as the frequency that the top-k list contains at least one same-author document.

As baselines, repeated human trials are infeasible due to reviewer availability and our privacy protocol, where we retain only voting outcomes, not full interaction logs. We therefore use **random guessing** as the primary baseline. For a candidate pool of size N, with m same-author documents for a target review and cutoff k, the chance of retrieving at least one same-author document by random selection is:

$$\text{Rank@k(random)} = 1 - \frac{\binom{N-m}{k}}{\binom{N}{k}}. \quad (3)$$

We report Rank@k relative to this baseline and show that DAS performs substantially better than random guessing. We include additional algorithmic baselines in the open-benchmark experiments that follow.

4.1.2 Experiment Results

Table 2 shows that DAS is effective on large-scale anonymized reviews, achieving 28% Rank@5 about a 1,000-fold improvement over random guessing. In practical terms, an attacker could reach roughly a 44% success rate if inspecting only the top-20 candidates, indicating a substantive anonymity risk. Although the ICLR setting is challenging due to sparse stylistic cues and few same-author samples per reviewer, DAS remains robust, underscoring an emerging LLM-enabled privacy threat.

4.2 Evaluation on Existing Benchmarks

Besides evaluating the emerging privacy risk on anonymous peer review systems, we also conduct experiments on existing benchmarks that are widely used for distinguishing writing styles:

	Rank@5	Rank@10	Rank@15	Rank@20	Miss
DAS	28%	40%	44%	44%	56%
Random	0.03%	0.07%	0.10%	0.13%	99.88%

Table 2: Real scenario experiments on ICLR anonymous reviewing systems.

blogs and emails. This test examines the capability of DAS on non-academic settings and supports broader claims about LLM-enabled deanonymization.

4.2.1 Experiment Setup

Datasets & Baseline. We use two public corpora. (1) *Blog Authorship Corpus* (Schler et al., 2006), containing posts from 19,320 bloggers; we sample 1,500 authors, each with 10 posts. (2) *Enron Emails* (Klimt and Yang, 2004), a large email collection; following Huang et al. (2024a), we keep 174 authors with 50 emails each. As an algorithmic baseline, we report results using AIDBench (Wen et al., 2024), a recent framework for evaluating LLM-based authorship attribution under diverse conditions.

Experimental Design. We run two complementary settings: (1) a *one-to-many* test, where each query is mixed with a controlled number of distractors, and (2) a *in-the-wild* test, where an adversary attempts to identify an author from an entire corpus rather than a pre-filtered candidate pool. Both follow the DAS pipeline with domain-specific prompts: for blogs, we emphasize narrative voice and colloquial phrasing; for emails, we focus on greetings, sign-offs, and topic transitions. Each configuration uses five independent trials with randomized candidate ordering. For consistency with the baseline, we use Qwen1.5-72B-Chat for the one-to-many tests, while using Gemini-2.0-flash for he in-the-wild tests.

Evaluation Metrics. We report two metrics. **Rank@k** measures retrieval success, i.e., the fraction of queries for which the top-k list contains at least one same-author document. **Precision@k** measures ranking quality, i.e., the proportion of same-author documents among the top-k results.

4.2.2 Experiment Results

One-to-many Test. Tables 3 and 4 show that DAS consistently outperforms the AIDBench baseline across both benchmarks. On blogs, DAS improves Rank@1 by up to 28% in the two-author setting and maintains 75% Precision@k in five-author identification. On Enron emails, DAS also yields higher accuracy, reaching 80% precision in the five-author setting. Overall, DAS degrades more gracefully as the candidate pool grows, with substantially smaller performance drops than the baseline.

In-the-wild Test. Table 6 further confirms DAS in realistic search over the whole corpus. On blogs, Rank@k increases from 74% at k=5 to 94% at k=20, with only 6% misses, suggesting the iterative tournament effectively captures consistent narrative cues. On Enron emails, where texts are shorter and stylistically diverse, DAS achieves 74% Rank@5 and reaches 88% by k=20 (12% misses). Across both datasets, multi-trial aggregation improves stability and mitigates occasional LLM errors.

4.3 Ablation Studies on Progressive Elimination

To isolate the contribution of progressive elimination to DAS's performance, we conducted controlled ablation experiments on the Blog Authorship Corpus. We define a variant DAS-PE that disables the progressive elimination mechanism: instead of iteratively retaining top candidates and shrinking the pool, DAS-PE processes all candidates in a single round of grouping and scoring without iterative reduction. All other components remain identical to the full DAS framework. We evaluate both the full DAS and DAS-PE using the same metrics to ensure direct comparability.

Experiment Results Table 5 demonstrates the critical role of progressive elimination through controlled experiments. Disabling progressive elimination (see variant DAS-PE) caused 34% relative Rank@5 performance drop and doubled miss rates on blog data, with accuracy stagnating at 68% (vs. 94% full model). The complete system achieved noise suppression through iterative filtering, whereas DAS-PE demonstrated higher susceptibility to bias accumulation due to limited candidate screening. The results show the dual role of progressive elimination as a computational accelerator and a signal purifier.

4.4 Ablation Studies on Different Models

To further validate the robustness of DAS framework across different LLMs, we conducted additional experiments using two other popular language models DeepSeek-R1 and Claude-3.5-Sonnet. This extension aimed to assess the framework's adaptability to different models and reason-

	2 Authors			5 Authors		
	Rank@1	Rank@3	Rank@5	Rank@1	Rank@3	Rank@5
DAS	95.0	100.0	100.0	75.0	85.0	85.0
AIDBench	66.7	93.3	93.3	66.7	86.7	86.7
	Prec@1	Prec@3	Prec@5	Prec@1	Prec@3	Prec@5
DAS	95.0	85.0	76.0	75.0	66.7	60.0
AIDBench	66.7	73.3	70.0	66.7	60.0	56.7

Table 3: Evaluation of the *one-to-many* author identification on the Blog Authorship Corpus (with Qwen1.5-72B-Chat model).

	2 Authors			5 Authors		
	Rank@1	Rank@3	Rank@5	Rank@1	Rank@3	Rank@5
DAS	85.0	100.0	100.0	80.0	95.0	100.0
AIDBench	66.7	93.3	100.0	66.7	73.3	80.0
	Prec@1	Prec@3	Prec@5	Prec@1	Prec@3	Prec@5
DAS	85.0	83.3	77.0	80.0	71.7	72.0
AIDBench	66.7	68.9	66.0	66.7	53.3	49.3

Table 4: Evaluation of the *one-to-many* author identification on the Enron Email dataset (with Qwen1.5-72B-Chat model).

ing approaches. The evaluation employed identical experimental parameters as previous tests (see Section 4.2).

Experiment Results We evaluate the one-to-many author identification task on the Research Paper dataset. As shown in Table 11, Claude-3.5-Sonnet attained perfect 100% Rank@1 accuracy in two-author scenarios, while DeepSeek-R1 achieved 96.3%, both achieving excellent results. These results confirm DAS's ability to leverage different LLMs' strengths while ensuring robust author identification. This model-agnostic performance highlights the framework's practical viability in diverse deployment environments.

	Rank@5	Rank@10	Rank@15	Rank@20	Miss
DAS	74%	88%	92%	94%	6%
DAS-PE	40%	60%	66%	68%	32%

Table 5: Ablation test on blog dataset

	Rank@5	Rank@10	Rank@15	Rank@20	Miss
blog	74%	88%	92%	94%	6%
email	74%	86%	88%	88%	12%

Table 6: Evaluation of the *in-the-wild* tests on Blog Authorship Corpus and Enron Emails.

5 Conclusion

In this paper, we propose De-Anonymization at Scale (DAS), a progressive framework that enables scalable authorship matching with LLMs. Real-scenario evaluations show that DAS can recover same-author texts in anonymized peer reviews (44% Rank@20), blogs (94% Rank@20), and emails (88% Rank@20), exposing practical vulnerabilities in anonymous systems, where an attacker may succeed by inspecting only a small shortlist of candidates.

Looking ahead, methodologically, hybrid systems that combine LLM-based comparisons with classical stylometric signals may further strengthen attribution. More broadly, our results motivate standardized anonymity benchmarks and updated policies for deploying LLMs in privacy-sensitive settings, as well as closer collaboration between machine learning and privacy research to better protect anonymous communication in the LLM era.

6 Limitations

Our study demonstrates promising results, but has several limitations. First, our experiments focus mainly on Gemini-2.0-flash and Qwen1.5-72B-Chat, and may not fully reflect the behavior of other strong LLM architectures. Second, our prompt design is relatively limited, which may under-exploit

LLMs' capacity for fine-grained stylometric analysis.

7 Ethical Considerations

Purpose and dual-use risk. This work is a *risk demonstration*: it shows that large language models (LLMs) can de-anonymize texts in anonymity-critical systems (e.g., double-blind peer review) by exploiting stylometric signals at scale. Our intent is to show actionable privacy risks so platforms and venues can assess and strengthen defenses, not to target individuals. Accordingly, our evaluation mirrors realistic conditions in anonymous ecosystems (open-set attribution, large candidate pools, and no pre-built author profiles).

Data sources and scope. All peer-review texts analyzed in this paper come from publicly accessible, *anonymized* OpenReview records (ICLR 2023–2025). We use these texts only for evaluation and do not train or fine-tune any models on review content. DAS relies on retrieval and LLM prompting rather than supervised training on the review corpus, reducing risks of memorization or data leakage from these sources.

Human subjects and privacy. We complement offline evaluation with a small validation study (9 participants; 25 test cases) designed to protect participant privacy. We collect only the information needed to compute metrics and store no personally identifying information. Any potentially identifying inputs (e.g., paper identifiers) are hashed or deleted, and we retain only anonymized, aggregated voting outcomes (see Appendix A). The study design follows principles of notice, purpose limitation, data minimization, and storage limitation.

Safeguards and non-targeting. Because the study involves sensitive real-world material (anonymous reviews), we report only aggregated results and do not release per-document or per-user attributions. We also refrain from attempting to de-anonymize ongoing review threads outside controlled evaluation.

Responsible communication. To reduce misuse, we provide sufficient methodological detail for reproducibility while omitting operational instructions that would materially lower the barrier to abuse. We encourage venues to consider countermeasures such as stylometric audits, reviewer-style obfuscation tools, and clearer policies governing LLM use in privacy-sensitive workflows.

Use of AI assistants. We used LLMs only for language polishing (grammar and clarity). They were not used for code generation, data analysis, experimental design, or substantive content creation. All methods, results, and conclusions are the product of human work.

Acknowledgement

The authors sincerely thank the volunteers, who are kept anonymous for safety reasons, for evaluating the system and offering invaluable real-world feedback. This work would have been substantially weaker without their contributions. The authors would also like to thank Zichen Wen and Dadi Guo for their contributions to the preparation of AID-Bench and for their assistance in conducting the benchmark evaluation. The work was supported by National Natural Science Foundation of China (Grant No. 62576015), Beijing Major Science and Technology Project (Z251100008425004) and Beijing Natural Science Foundation (L253001).

References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*.

Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pezik, Martin Potthast, and 1 others. 2023. Overview of PAN 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 459–481. Springer.

Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Animesh Mukherjee, and 1 others. 2024. Overview of PAN 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification. In *European Conference on Information Retrieval*, pages 3–10. Springer.

Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reynier Ortega-Bueno, Piotr Pęzik, Martin Potthast, and 1 others. 2022. Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 382–394. Springer.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. *arXiv preprint arXiv:2101.11310*.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412.

Xie He, Arash Habibi Lashkari, Nikhill Vombatkere, and Dilli Prasad Sharma. 2024. Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information*, 15(3):131.

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

Baixiang Huang, Canyu Chen, and Kai Shu. 2024a. Can large language models identify authorship? *ArXiv*, abs/2403.08213.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S Yu. 2024b. Cosent: consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813.

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. In *Conference on Empirical Methods in Natural Language Processing*.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660.

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

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

David Madigan, Alexander Genkin, David D Lewis, and Dmitriy Fradkin. 2005. Bayesian multinomial logistic regression for author identification. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 803:509–516.

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.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at risk? assessing re-identification capabilities of large language models. *arXiv preprint arXiv:2308.11103*.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

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

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

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

Efstathios Stamatatos, Francisco Rangel, Michael Tschuggnall, Benno Stein, Mike Kestemont, Paolo Rosso, and Martin Potthast. 2018. Overview of pan 2018: Author identification, author profiling, and author obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 267–285. Springer.

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*.

Ming Xu. 2023. Text2vec: Text to vector toolkit. https://github.com/shibing624/text2vec.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendices

A Platform Functionalities

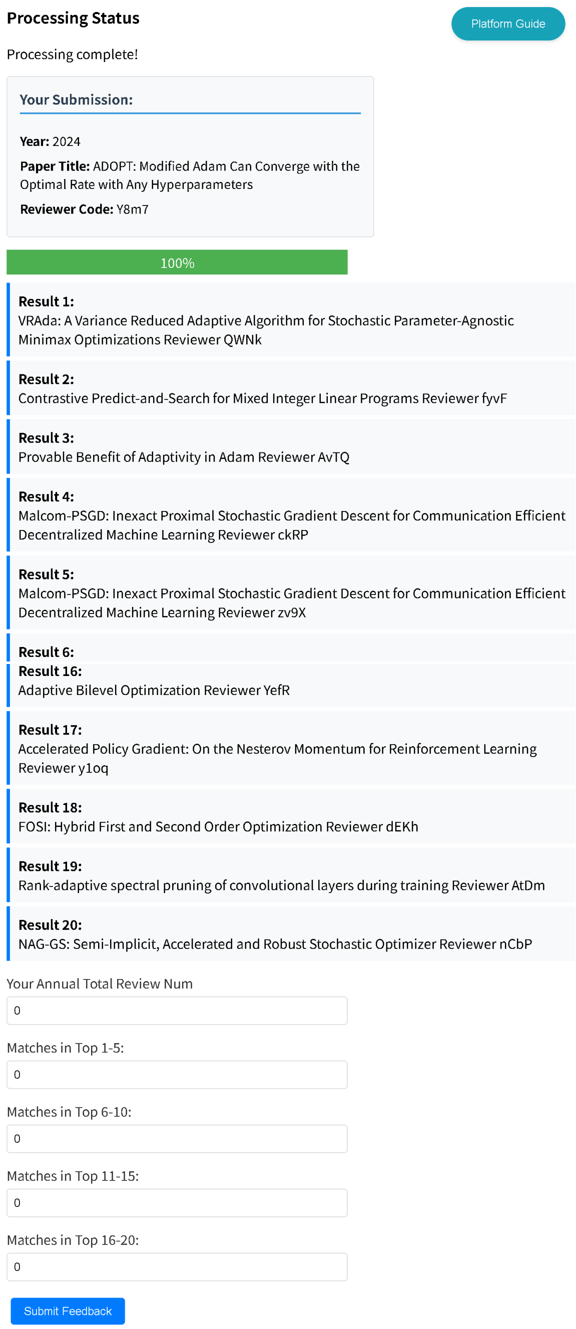

Figure 3: User interface of our platform

The web platform's workflow centers on processing anonymous review analysis tasks in real-world scenarios while providing a fluent user experience for our volunteer participants, who were not paid for their involvement. The process is presented in Figure 3 with a random query example.

First, participants submit a query specifying the year, paper title, and review code. That request enters a background task queue, which uses the DAS framework to retrieve the 20 most similar reviews and presents them as a ranked list from most to least likely.

Next, participants report both the total number of reviews they authored that year and how many of those appear within each ranking interval.

Finally, with these inputs, the system computes the necessary statistics. Under the hood, this workflow incorporates multiple technical components including asynchronous task processing, real-time status updates, dynamic result visualization, and privacy protection, ensuring efficient and safe evaluation.

We want to emphasize that this platform enabled researchers to evaluate the deanonymization risks inherent in anonymous review systems while strictly adhering to ethical guidelines and privacy protections.

Before participants use the platform, we explicitly inform users of the platform's functionalities through pop-up notifications, including the purposes, types of information collected, and privacy protection measures. Throughout the entire interaction process, we ensure that users remain completely anonymous, preventing potential breaches in the anonymity system's privacy protections. As shown in Table 7, the validation platform was designed in compliance with GDPR and COPPA regulations, incorporating multiple privacy safeguards.

Data Type	Description	Stored?	Protection Measures
User Feedback	Rage hit statistics	Yes	Anonymization
User Inputs	Year, paper title, review code	No	Hash, Automatic deletion
System Logs	Task status information	No	Automatic deletion

Table 7: Data handling protocol for human subject interactions

B Evaluation on Research Paper Dataset

To further investigate DAS's capability in identifying authorship of formal academic writing, we conducted comprehensive experiments on a large-scale research paper dataset. This evaluation extends our analysis to scholarly articles while addressing multi-author attribution challenges inherent in academic publications.

B.1 Experiment Setup

Benchmark Dataset & Baseline The research paper dataset (Wen et al., 2024) comprised titles, abstracts and introductions from 138,652 arXiv papers under the CS.LG tag (the field of machine learning in the computer science domain) published between 2019 and 2024. After removing duplicate entries and authors with fewer than ten papers, the dataset included 24,095 papers from 1,500 authors, ensuring that each author had at least ten papers. As detailed in Section 4.2, we implemented AID-Bench (Wen et al., 2024) as the baseline method.

Experiment Design Following the methodology established in Sections 4.1 and 4.2, we performed both one-to-many and real-scenario tests on the research paper dataset. The one-to-many test utilized Qwen1.5-72b-chat while the real-scenario test employed Gemini-2.0-flash. A successful identification was registered if the target author appeared in the author list of any retrieved result. This lenient evaluation criterion reflected real-world attack scenarios where attackers might only need to associate a document with one of its actual authors.

B.2 Experiment Results

One-to-many Test The experimental results demonstrated that the DAS framework exhibited significant advantages in research paper authorship attribution tasks. As shown in Table 8, in two-author scenarios, DAS achieved a Rank@1 accuracy of 93.3%, surpassing baseline methods by 23.3 percentage points. In the more challenging five-author scenarios, DAS attained a Rank@1 score of 60% - significantly higher than the baseline. Regarding precision metrics, DAS obtained 48.9% Precision@3 in five-author scenarios, representing a 38.9 percentage point improvement. This performance advantage proves particularly pronounced in long academic texts, likely stemming from DAS's proficiency in capturing discipline-specific terminological patterns, argumentative structures, and other deep stylistic features inherent to scholarly writing conventions.

Real-scenario Test This experiment demonstrated the remarkable efficacy of the DAS framework in academic author identification tasks. As shown in Table 9, DAS achieved 92% accuracy on the Rank@20 metric with a mere 8% missing rate, revealing that writing style-based author recognition could create substantial privacy vulnerabilities even in strictly anonymized academic scenarios. This finding carries significant implications for the academic publishing system, serving as a critical warning. Given the prevalent preprint culture in computer science, particularly in arXiv repositories, attackers could potentially compromise the

	2 Authors			5 Authors		
	Rank@1	Rank@3	Rank@5	Rank@1	Rank@3	Rank@5
DAS	93.3	93.3	100.0	60.0	80.0	86.7
AIDBench	70.0	76.7	86.7	13.3	23.3	33.3
	Prec@1	Prec@3	Prec@5	Prec@1	Prec@3	Prec@5
DAS	93.3	68.9	70.7	60.0	48.9	41.3
AIDBench	70.0	51.1	46.7	13.3	10.0	8.0

Table 8: One-to-many experiments on research paper dataset

double-blind review mechanism by analyzing authors' historical publications through DAS's sequential analysis.

	Rank@5	Rank@10	Rank@15	Rank@20	Miss
DAS	66%	80%	88%	92%	8%

Table 9: Real scenario experiments on research paper dataset

C Further explanation of random baseline

To rigorously establish the theoretical baseline for random guessing performance, we provide a detailed derivation of the combinatorial probability model. Consider a candidate pool containing N documents where each target document has m same-author documents (not including the target itself). The probability of selecting at least one same-author document when randomly choosing k candidates follows hypergeometric distribution principles.

Let X be the random variable representing the number of same-author documents in a random sample of size k. The probability of selecting exactly s same-author documents is:

$$P(X = s) = \frac{\binom{m}{s}\binom{N-m}{k-s}}{\binom{N}{k}}. \tag{4}$$

The probability of selecting at least one same-author document (Rank@k metric) is therefore:

$$\text{Rank@k} = 1 - P(X = 0) = 1 - \frac{\binom{N-m}{k}}{\binom{N}{k}} \tag{5}$$

Where $\binom{N}{k}$ represents total ways to choose k documents from N and $\binom{N-m}{k}$ counts the ways to choose k documents excluding all m same-author documents.

For concrete illustration, we used a set of typical data for calculation: N=44770 (total documents), m=4 (same author documents), k=20 (top candidates).

The random baseline probability calculates as:

$$\text{Rank@20} = 1 - \frac{\binom{44764}{20}}{\binom{44770}{20}} \approx 0.014\% \tag{6}$$

We ultimately aggregate the data from each year, where final result matches the 0.13% random baseline reported in Table 1, demonstrating that DAS's 44% Rank@20 accuracy represents a 338x improvement over random chance. The combinatorial approach provides exact theoretical expectations against which empirical results can be meaningfully compared.

D Dataset Details

We conducted comprehensive evaluations across four key datasets representing distinct textual domains and privacy challenges:

Blog Curated from the Blog Authorship Corpus (Schler et al., 2006), this dataset contains 15,000 posts by 1,500 active bloggers, capturing informal writing styles and personal expression patterns that enable authorship analysis of user-generated content.

ICLR Review The ICLR Review dataset aggregates 147,367 anonymized peer reviews from 2023-2025 conferences via OpenReview API, simulating real-world deanonymization attacks on academic review systems through structured metadata and textual analysis.

Enron Email Derived from the Enron corpus (Klimt and Yang, 2004), the benchmark includes 8,700 professionally written emails from 174 executives, preserving linguistic fingerprints while removing sensitive headers to study corporate communication privacy risks.

Research Paper The academic writing dataset comprises 24,095 single-author machine learning

	Domain	#Documents	#Authors	Avg.Length	Source
Review	Peer Review	147,367	N/A	2,655	OpenReview API
Blog	Personal Blogs	15,000	1,500	116	Schler et al. (2006)
Email	Corporate Communication	8,700	174	197	Klimt and Yang (2004)
Paper	Academic Writing	26,632	1,500	7,383	Wen et al. (2024)

Table 10: Detailed characteristics of evaluation datasets

	2 Authors			5 Authors		
	Rank@1	Rank@3	Rank@5	Rank@1	Rank@3	Rank@5
DeepSeek-R1	96.3	100.0	100.0	85.2	88.9	96.3
Claude-3.5-Sonnet	100.0	100.0	100.0	93.3	100.0	100.0
	Prec@1	Prec@3	Prec@5	Prec@1	Prec@3	Prec@5
DeepSeek-R1	96.3	97.5	90.4	85.2	76.5	75.6
Claude-3.5-Sonnet	100.0	95.6	93.3	93.3	97.8	90.7

Table 11: Evaluation of different models one-to-many identification capability on research paper dataset

papers from arXiv CS.LG (2019-2024) (Wen et al., 2024), filtered to include only authors with ≥ 10 publications, exposing stylistic consistency challenges in scholarly communication.

Table 10 provides comprehensive statistics for all datasets used in our experiments. The table includes key characteristics such as document domains, corpus sizes, author counts, average text lengths, and original data sources. All datasets were preprocessed to remove personally identifiable information while preserving linguistic patterns. Text lengths represent word counts after standard cleaning (stopword removal, punctuation stripping).

D.1 Sensitivity Analysis

Figure 4 reveals consistent performance improvement across text genres, with mean same-author rankings reduced by 47% after five iterations. Despite the initial variance caused by text length and stylistic differences, DAS maintaind downward trajectory in both free-form blogs and structured emails. This demonstrates the framework's dual-phase synergy: coarse filtering eliminates stylistic outliers while fine-grained analysis identifies stable authorial patterns, enabling effective cross-domain discrimination. The convergence trend validates DAS's capacity to extract personalized expression signatures regardless of text structure.

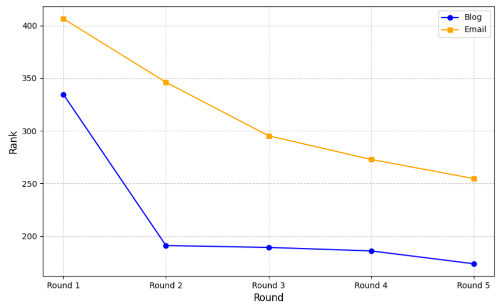

Figure 4: Mean rank of same author documents per round