The Impostors Method as a Statistical Test for Authorship Verification with Large Language Models as Sampling Source for Hard Negatives

Anonymous ACL submission

Abstract

Authorship verification asks whether two given texts were written by the same author. Since supervised verification methods often generalize poorly across domains, the unsupervised Impostors Method, inspired by police lineups, reduces verification to in-domain authorship attribution. Using one text as a query, it hides the other among a sample of hard negatives and measures how reliably a style similarity function retrieves it at the top rank. Although widely applied in practice, the Impostors Method still suffers when no in-domain impostor documents are available. Resorting instead to sampling from open-domain collections provides much less control over confounding stylistic variables such as genre or register. Moreover, its statistical foundations remain largely unexplored compared with those of police lineups. In this paper, we recast the Impostors Method as a statistical hypothesis test, enabling a principled assessment of evidential strength without training a decision threshold. We further investigate large language models (LLMs) as a controllable source for generating in-domain impostor documents.

1 Introduction

The Impostors Method is among the most influential unsupervised approaches to authorship verification (Koppel and Winter, 2014). It continues to be widely used in digital humanities and digital text forensics, where in-domain training data is often scarce and therefore limits the applicability of deep learning. Inspired by police lineups, the Impostors Method casts authorship verification as an unsupervised attribution problem. Specifically, it likens a style similarity measure to a witness asked to identify a suspect from memory in a lineup of fillers (Figure 1): To test the authorship of a document, Koppel and Winter assess how reliably a document by the known suspect author is ranked above documents by impostor authors in terms of stylistic similarity. The confidence an investigator may derive from this test depends on the consistency of the outcome across repeated tests, that is, on whether the suspect document sufficiently often succeeds (or fails) to outrank the impostors.

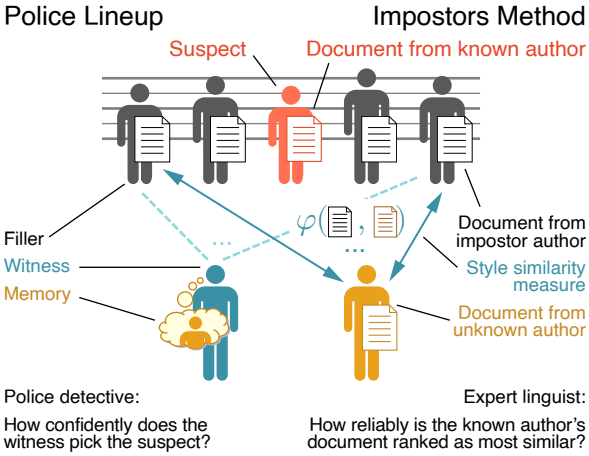


Figure 1: A police lineup tests a witness's memory by asking them to identify the suspect among similar-looking fillers. Inspired by this approach, the Impostors Method for authorship verification tests a style similarity function's ("witness's") reliability to rank the document from a known author ("suspect") as most similar to the document of unknown authorship ("memory") among similar documents from impostor authors ("fillers").

Despite its practical relevance, the Impostors Method has two key limitations: its decision rule relies on empirically chosen thresholds that do not quantify uncertainty within a formal statistical framework, and sampling impostors still works best when sufficient in-domain data is available.

We address both limitations. First, we reinterpret the Impostors Method as a statistical hypothesis test (Section 3), enabling a principled assessment of evidential strength instead of empirically tuning thresholds. Second, we investigate the usefulness of LLMs for generating impostor documents when in-domain data are unavailable, controlling for relevant contextual variables while varying stylistic realization (Section 4). In our evaluation (Section 5), we assess the strengths and weaknesses of both contributions against the state of the art.[1]

[1]Code: https://anonymous.4open.science/r/arr26DDD3

2 Related Work

Authorship analysis emerged in the 19th century as a way to resolve disputes over literary authorship (Stamatatos, 2009; Neal et al., 2018). In closed-set authorship attribution, given a disputed document and texts by several candidate authors, the task is to decide which candidate wrote the disputed text. Authorship verification asks whether two documents were written by the same author (Koppel and Schler, 2004). Unlike attribution, verification is not restricted to a fixed candidate set. Instead, it must distinguish pairs of same-author documents from pairs of different-author documents from a virtually unbounded population of authors. For that reason, verification is commonly viewed as a one-class classification problem (Schölkopf et al., 2001) and considered a core challenge to generalizing authorship analysis (Koppel and Schler, 2004).

Supervised authorship verification is dominated by deep learning (Bevendorff et al., 2025a, 2024; Stamatatos et al., 2023, 2022). However, despite significant efforts to scale training corpora, LLM-based verification approaches still exhibit overfitting and poor cross-domain generalization (Bevendorff et al., 2025b). This limits their applicability in the humanities and in forensics, where in-domain training data is typically scarce. Consequently, real cases are often tackled with unsupervised methods.

Early Unsupervised Approaches Unmasking Koppel and Schler (2004); Bevendorff et al. (2019b) measures the accuracy degradation of a style classifier, which is iteratively re-trained to distinguish passages drawn from two book-length documents as the previous iteration's most discriminative feature is removed. Compression-based methods (e.g., based on Shkarin's 2002 string compression algorithm Prediction by Partial Matching, PPM) measure the joint compression rate of two documents (Veenman and Li, 2013; Halvani et al., 2017). Since documents written by the same author are expected to exhibit fewer stylistic differences than documents written by different authors, stronger evidence for same authorship is indicated by more efficient joint compression.

Both approaches competed on pre-LLM benchmarks (Juola and Stamatatos, 2013; Stamatatos et al., 2014, 2015), yet no generalizable decision threshold has been found. Forensic casework therefore rarely relies on fully automated use of these methods; instead, experts inspect their results and interpret them alongside other evidence.

Impostors Method		**Effectiveness Scores**						
Authors	Year	Acc.	F1	Prec.	Rec.	ROC	c@1	ROC · c@1
Student Essays Dataset								
Koppel, Winter	2014	**0.721**	**0.713**	**0.734**	0.695	**0.772**	**0.737**	**0.570**
Khonji, Iraqi	2014	0.502	0.666	0.501	**0.994**	0.674	0.676	0.457
Hernández et al.	2015	0.642	0.702	0.602	0.845	0.701	0.672	0.473
Kestemont et al.	2016	0.673	0.711	0.638	0.805	0.759	0.709	0.539
Potha, Stamatat.	2017	0.678	0.706	0.651	0.774	0.760	0.703	0.535
Nagy	2024	0.706	0.712	0.699	0.727	0.767	0.725	0.557
LLM impostors	2026	0.538	0.659	0.523	0.896	0.684	0.676	0.464
Min-max (B)	Baselines	0.502	0.660	0.501	0.968	0.659	0.662	0.438
Cosine (B)	Baselines	0.498	0.662	0.499	0.983	0.625	0.624	0.392
Unmasking (B)	Baselines	0.518	0.671	0.509	0.982	0.645	0.620	0.402
PPMd (B)	Baselines	0.516	0.668	0.509	0.973	0.664	0.677	0.451
Blog Posts Dataset								
Koppel, Winter	2014	**0.722**	0.700	**0.761**	0.648	0.762	0.740	0.564
Khonji, Iraqi	2014	0.683	0.687	0.679	0.695	0.735	0.713	0.524
Hernández et al.	2015	0.614	0.677	0.584	0.809	0.675	0.668	0.451
Kestemont et al.	2016	0.717	0.720	0.712	0.728	0.780	0.743	0.580
Potha, Stamatat.	2017	0.671	0.708	0.636	0.800	0.756	0.714	0.540
Nagy	2024	0.705	0.699	0.714	0.685	0.760	0.732	0.557
LLM impostors	2026	0.671	0.700	0.647	0.764	0.758	0.712	0.541
Min-max (B)	Baselines	0.662	0.699	0.631	0.785	0.759	0.716	0.544
Cosine (B)	Baselines	0.589	0.670	0.560	0.837	0.718	0.682	0.490
Unmasking (B)	Baselines	0.609	0.683	0.575	**0.843**	0.712	0.670	0.478
PPMd (B)	Baselines	**0.722**	**0.729**	0.710	0.750	**0.796**	**0.749**	**0.597**

Table 1: Reproduction of the Impostors Method variants and baselines. The scores are computed as mean over ten 10-fold cross-validations.

The Impostors Method Starting from a high-dimensional style representation, the Impostors Method (Koppel and Winter, 2014) repeatedly applies random projections into lower-dimensional spaces and measures the proportion of projections in which the known author's text is more similar to the disputed text than texts written by distractor authors ("impostors"). Stronger evidence for same authorship is indicated by a higher proportion of comparisons favoring the known author over the impostors. Due to its success (Seidman, 2013), subsequent work further analyzed and extended the Impostors Method, showing that more rounds and swapped document roles can improve effectiveness (Juola, 2015). To determine which variant to use for our experiments, we reimplemented existing extensions in a unified framework and carried out a comparative evaluation the results of which are shown in Table 1 (details are found in Appendix C and D). Surprisingly, the original Impostors Method outperforms all its extensions and the baselines on the Student Essays Dataset originally used by Koppel and Winter for in-domain impostor sampling.

However, the Impostors Method and its extensions still rely on empirically chosen thresholds, which lack a formal statistical interpretation. That limitation is especially important in forensic settings, where experts must substantiate and defend their decision and the uncertainty attached to it.

Police Lineups In a police lineup, a witness is asked to identify a suspect from a group that includes several distractors, called foils or fillers. Lineups constitute controlled recognition tests intended to reduce the likelihood of biased accusations (Tunnicliff and Clark, 2000; Colloff et al., 2021). As witnesses typically select the person with the highest ecphoric similarity to their memory of the perpetrator (Levi and Jungman, 1995), a successful identification constitutes probabilistic evidence linking the suspect to the crime (Yilmaz et al., 2022). Crucially, the evidential value of a lineup depends on the likelihood of selecting the suspect among the fillers, motivating a statistical interpretation of lineup outcomes. This highlights the importance of selecting fillers: a poor selection biases decisions, whereas appropriate fillers are similar to the suspect (Valentine and Heaton, 1999; Peska and Trojanova, 2018; Dokoupil and Peska, 2021). Common filler selection strategies are based on the witness's description, similarity to the suspect, or a hybrid of both (Levi and Jungman, 1995; Tunnicliff and Clark, 2000; Colloff et al., 2021). The probabilistic foundation of police lineups has not been investigated for the Impostors Method.

Forensic Admissibility Forensic authorship investigations subject to judicial scrutiny require methods that are both reliable and transparent. While approaches that depend on human perception and subjective judgment are difficult to audit and challenge, data-driven approaches grounded in quantitative measurements and statistical modeling provide greater transparency and reproducibility (Morrison, 2022). Forensic methods are only suitable for legal use if they meet demanding standards of rigor and replicability (Voice et al., 2025). Adversarial legal systems also give the accused the right to confront and challenge the evidence presented against them (Grant, 2025). In many jurisdictions, admissibility is assessed using the Daubert criteria, i.e., testability, peer review, known error rates, and general acceptance in the scientific community (Berger, 1993). Approaches that produce conclusions without transparent reasoning are considered ipse dixit assertions in violation of these criteria (Juola, 2015) and are therefore inadmissible in court (Jurs and DeVito, 2024). For authorship verification approaches, these principles imply two requirements: rigorous empirical assessment of evidential strength and transparent quantification of the uncertainty associated with that evidence.

3 The Impostors Method as a Statistical Test

This section recaps the Impostors Method and reinterprets it as an approach to statistical inference.

3.1 The Traditional Impostors Method

The Impostors Method reduces open-set verification to closed-set attribution (Figure 2). In verification, the question is whether a document a of unknown or disputed authorship has been written by the candidate author B, who is known to have written document b. In the first step of the Impostors Method, Koppel and Winter (2014) introduce a sample of documents $S \subset D$ from a corpus D, each written by an author different from the candidate author B and from all other authors in S. These documents are referred to as impostors.

In the second step, the 100 000 most frequent space-free character 4-grams in the reference corpus D are determined as writing style features and all documents are embedded into this space as $tf \cdot idf$-vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{S}$. The third step consists of a random projection of these representations using half the feature set k and ranking all documents $\mathbf{s} \in \mathbf{S} \cup \{\mathbf{b}\}$ with respect to their similarity to $\mathbf{a}$ using the min-max similarity measure:

$$\varphi(\mathbf{a}, \mathbf{s}) = \frac{\sum_{i=1}^{k} \min(a_i, s_i)}{\sum_{i=1}^{k} \max(a_i, s_i)} .$$

Fourth, the third step is repeated m times, counting how often b is ranked as most similar to a. The entire procedure is repeated with the roles of a and b reversed. If the average proportion of both rounds where b is ranked first exceeds a threshold τ, a is considered to have been written by author B.

Convincing Impostors The crucial step of the Impostors Method is sampling *convincing* impostors S, i.e., hard negatives. Like in police lineups, where convincing fillers serve to distract the witness, the impostor documents must have the capacity to "distract" the style similarity measure.

If an in-domain corpus D for document b is available (e.g., a collection of student essays), sampling impostors from it is a natural choice. However, such data is typically not available. Koppel and Winter therefore propose to retrieve impostors from the web instead. Using queries of 3-5 medium-frequency words from b, the top 25 retrieved documents of each query are pooled. In either case, sampling proceeds in two stages: The n documents most similar to b are identified, and then $n' < n$ impostors are sampled at random.

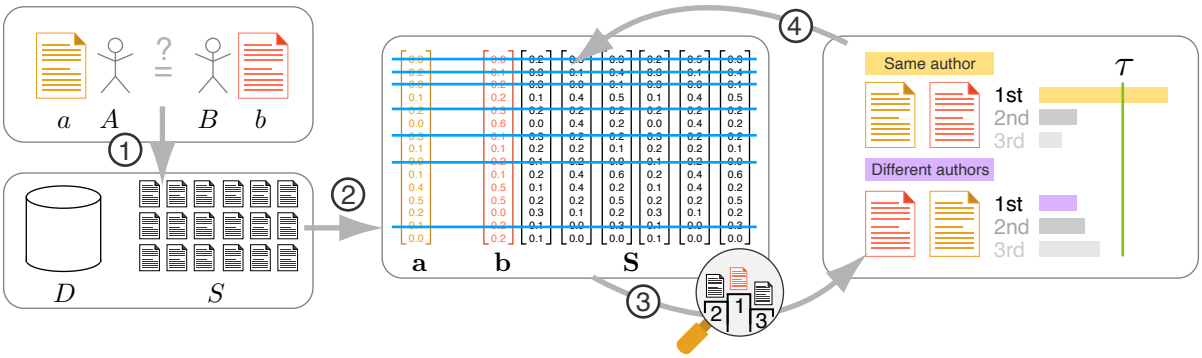


Figure 2: Schematic of the Impostors Method. Steps 1-4 are described in Section 3.1. If documents a and b have the same author, a is expected to be ranked as more similar to b than the impostor documents S across different style representations. A same-author classification is made if the proportion of positive rankings exceeds τ.

3.2 Limitations to Evidential Strength

The related work on the Impostors Method missed two key limitations: the need for determining a confidence threshold τ, and the lack of sufficient control over impostor selection and fairness.

Koppel and Winter already remark "a single uniform threshold for all pairs is not likely to work especially well." Since the Impostors Method tests if the average stylistic similarity between a to b exceeds that between a and the impostors in S, only experienced analysts can assign the threshold on a case-by-case basis. We resolve this limitation by drawing a closer analogy to the statistical inference underlying police lineups and transferring it to authorship analysis (Section 3.3).

The second limitation stems from the lack of style models that disentangle different components of writing style, which particularly affects retrieval-based impostor sampling. Though indexing style at web scale has been proposed (Potthast et al., 2016), it has not implemented beyond genres such as news or shopping sites. Koppel and Winter's use of Google controls for topic, but controls for style only insofar it correlates with topic. Section 4 introduces a new LLM-based approach that offers more fine-grained control over impostor style.

3.3 The Impostors Method: A Hypothesis Test

The strength of the Impostors Method does not lie in its potential for full automation. Its strength lies in the probabilistic evidence it supplies if it succeeds: In that case, with a high likelihood, the two documents were not written by different authors. However, unlike intuition might suggest, this does not guarantee same authorship beyond doubt. The reason for this is that the Impostors Method is a hypothesis test, suggesting an alternative framing.

Although police lineups have been in use since the mid-19th century, their strong likeness to hypothesis tests has been outlined only comparably recently (Wells and Luus, 1990), taking inspiration from experiments in social psychology. In Table 2, we show that this analogy naturally extends to the Impostors Method. According to Valentine and Heaton (1999), the fairness of a police lineup and thus its validity as evidence depends on the witness's description of the suspect and the composition of the lineup. Likewise, the validity of the Impostors Method depends on the style model and the sample of impostor documents. If both are well-chosen, the statistical inference is canonical.

Under H_0 that document a is not authored by B, and assuming a *fair* impostor sample S, document b is exchangeable with the impostors in *one* iteration of the Impostors Method. Hence, φ ranks b highest among $S \cup \{b\}$ with probability $p_0 = 1/(|S| + 1)$. Let X be the number of iterations, out of m, where this happens. Under H_0, $X \sim \text{Binomial}(m, p_0)$, and the directional upper-tail p-value for an observed count x is $P(X \geq x)$.

The same applies after exchanging the roles of a and b, yielding two directional tests based on m runs each. Because the two directions are computed over partially shared feature spaces, for example overlapping vocabularies, the resulting statistics need not be independent. Bonferroni control is therefore used, noting that independence is not required. With directional p-values $p^{(1)}$ and $p^{(2)}$, direction $d \in \{1, 2\}$ is declared significant at family-wise level α when $2p^{(d)} < \alpha$, equivalently $p^{(d)} < \alpha/2$. We reject H_0 only when both directional tests are significant, providing bidirectional evidence against different authorship of the disputed and suspected author's documents.

Experiment Design	Police Lineup (Wells and Luus, 1990)	Impostors Method
Hypothesis	Suspect is the perpetrator	Author B has written document a
Null hypothesis H_0	Suspect is not the perpetrator	Author B has not written document a
Subject(s)	Eyewitness(es)	Style similarity measure(s) φ
Stimulus	Suspect	Document b, written by author B
Confederates	Fillers in the lineups	Impostor documents S
Parameters	Number of fillers, appearance of fillers, presence/absence of suspect	Number of impostors $\|S\|$ and their style similarity $\varphi(\mathbf{b}, \mathbf{d}), d \in S$, to document b
Procedure	Instructions to eyewitness(es)	Feature subsampling and similarity measure
Outcome	Identification of the suspect/ a filler/ no one	Identification of document b/ an impostor

Table 2: Mapping between hypothesis test components, police lineups and *one* iteration of the Impostors Method.

4 LLM-based Impostor Synthesis

This Section introduces a new approach to impostor synthesis based on LLMs.

Requirements for Fair Impostors If impostors differ from the document of known authorship, e.g., in topic, genre, or other contextual variables, the resulting comparison may reflect those differences rather than authorial style, potentially compromising lineup fairness. A reliable analysis requires a linguistically *relevant* comparison corpus that accounts for relevant sources of linguistic variation (Grant, 2012). Retrieval-based impostor sampling from the web as only alternative impostor sampling approach other than using an in-domain corpus offers limited control over multiple style variables simultaneously. It often fails to align genre with the candidate text (Koppel and Winter, 2014). Below, we introduce instruction-tuned LLMs as a novel approach to impostor sampling.

Impostor Synthesis via Two-Step Paraphrasing Given document b as input, the task of impostor synthesis is to generate a document d which closely resembles b, yet does not copy it directly. The approach we settled on for our experiments operates in two steps: (1) extraction of a bullet-point list summary of main ideas as well as style-affecting variables from document b, and (2) generation of document d based on this information. This way, relevant information about document b is communicated to the generator, yet the generator is not directly influenced by b. The generated document is then subjected to a post-hoc sanity check. The entire pipeline is implemented using DSPy (Khattab et al., 2024), with field descriptions specified in Table 4, Appendix A. Experimenting with a number of models, we opted for OpenAI's GPT-5 Nano.

Extraction. In the first step, the LLM extracts stylistic and contextual attributes from the input text, including heading, topic, genre, tone, authorial task, register, target audience, author description, and a bullet-point summary.

Generation. In the second step, paraphrases are generated conditioned on these attributes using the following DSPy output description:

> A single, complete paraphrased text written in the style defined by the provided parameters. The text must integrate **all** bullet points into one coherent composition, follow the specified **heading**, **topic**, **genre**, **tone**, **authorial task** **register**, **century**, **audience**, and **author profile**, and approximately match the requested word length.

Sanity Check. The definition of paraphrases varies across tasks and applications (Gohsen et al., 2024). In our case, a generated document d is a suitable impostor if it preserves the extracted variables of document b, while introducing sufficient lexical and syntactic variation to avoid phrasal duplication. We assess semantic similarity and lexical diversity jointly using the $\Delta_{\text{sem,syn}}$ measure proposed by Gohsen et al. (2023).

Example The following shows an excerpt of "The Declaration of Independence" by T. Jefferson:

> "[...] When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation. [...]"

Our pipeline generated a corresponding paraphrase:

> "In the course of human events it hath become necessary for these United Colonies to declare the causes which impel them to sever the political ties by which they have long been bound to the Crown of Great Britain, and to assume among the powers of the earth a station of independence and equal rank, to which the Laws of Nature and the Creator of mankind entitle them. [...]"

More details about this, including a failure case, are provided in Appendix A.

5 Evaluation

We introduces two changes to the Impostors Method: reframing it as a hypothesis test and a new impostor synthesis approach. We reimplement the original Impostors Method, its variants, and evaluate how statistical inference and LLM-generated impostors affect verification effectiveness.

5.1 Data

We obtained the two original corpora in (Koppel and Winter, 2014). The Student Essays corpus (Ireland and Pennebaker, 2010) is not publicly available due to the presence of sensitive student information, but was provided under a non-disclosure agreement by its creator, J. W. Pennebaker. Its restricted availability makes the data very useful, as it is unlikely that the essays appeared in LLM training data. The dataset consists of 7052 English essays written for five assignments by 950 university students in 2006. The assignments are (1) a stream-of-consciousness task, (2) a reflections on childhood, (3) a self-assessment of personality, (4) a thematic apperception test, and (5) four examples of four different theories. Following Koppel and Winter, we include only the first four assignments. Metadata includes author ID, author name, gender, political orientation, ethnicity, task, teacher, and year.

The Blog Posts corpus (Schler et al., 2006) consists of blog posts collected from *blogger.com* before September 2004, with each blog authored by a single user. The corpus comprises 681 288 English posts written by 19 320 bloggers aged 13-47 years, averaging approximately 35 posts and 7250 words per author. Metadata includes author ID, gender, age, topic, zodiac sign, and date.

Most publicly available large authorship verification corpora are rife with string artifacts and unsuitable text fragments, strongly affecting experimental outcomes (Bevendorff et al., 2019a, 2025b). We therefore adopt a rigorous preprocessing pipeline for cleansing, including removal of HTML, text register artifacts (e.g., character names in play scripts), text structure elements, and whitespace formatting, replacing Unicode with ASCII equivalents, and lower casing. The Blog Posts corpus was reduced by 23.33%, the Student Essays dataset by $<1\%$.

From the cleaned corpora, we constructed two evaluation datasets: Student Essays, comprising 1284 document pairs from 568 authors authors with an average document length of 708 words, and Blog Posts, comprising 7423 document pairs from 5378 authors and an average document length of 1154 words. For both, 50 % of the document pairs are from the same author, the remainder from different authors. Author pairs are selected to ensure matched metadata. Detailed dataset and preprocessing information is provided in Appendix B.

5.2 Reproducing All Impostors Methods

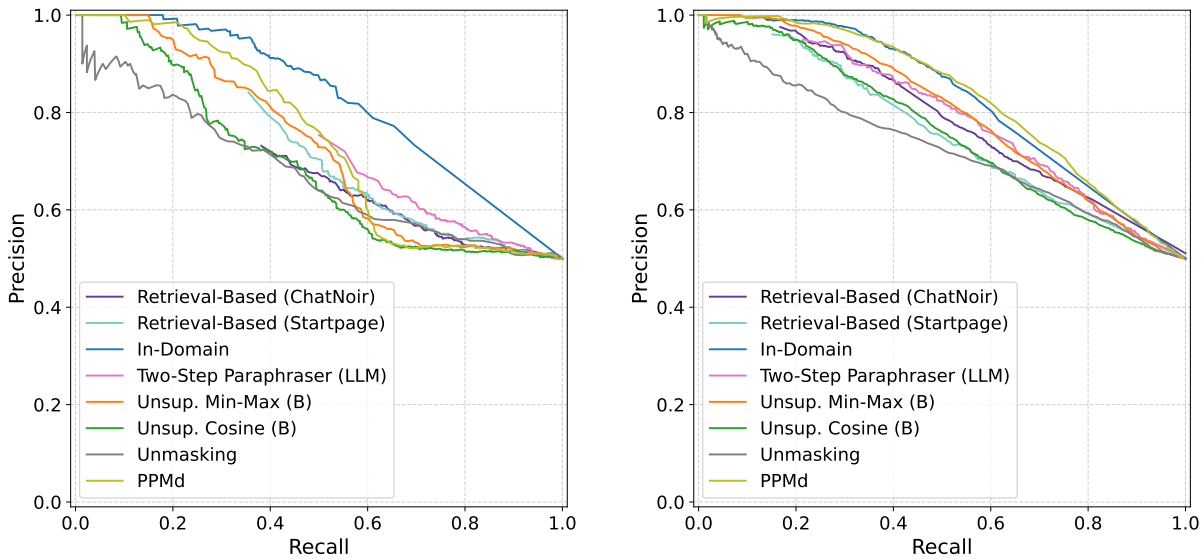
Since neither implementations of the original Impostors Method nor its variants, except for that of Nagy (2024), are available, we opted for reproducing all of them in a unified way. A key success indicator for our reproduction of the original method is its ability to achieve comparable effectiveness on the two corpora, barring a presumably more rigorous preprocessing than applied in the original implementation by Koppel and Winter.

For in-domain impostor generation, impostors are sampled from the same topic or task within the Student Essays and Blog Posts datasets, respectively, while ensuring that the reference document itself is excluded. For retrieval-based impostors, we replace Google with StartPage (which uses on Google's search results) and ChatNoir (Bevendorff et al., 2018) to enable large-scale automated experiments. Apart from these substitutions, our implementation follows the original approach.

Figure 3a and 3b show precision–recall curves for the same-author class on the Student Essays and the Blog Posts datasets. Compared with the original plots of Koppel and Winter, we observe a similar effectiveness profile: in-domain impostor sampling consistently outperforms retrieval-based sampling. We further observe that the retrieval index used for retrieval-based impostor sampling (Startpage vs. ChatNoir) has a substantial impact on effectiveness. Appendix C reports more details (e.g., hyperparameters) and other reproduced plots.

We further find that in-domain scores differ significantly from those of other impostor generation approaches. Significance is assessed using a Bonferroni corrected paired t-test (Bouckaert and Frank, 2004) on aligned metric scores from ten 10-fold cross-validation with $\alpha = 0.005$. Refer to Appendix E for details on the statistical tests.

Based on the successful reproduction, we implemented literature variants of the Impostor Method that replace individual original components (e.g., using rank fusion instead of expecting document b to be ranked first). In a comparative evaluation of all variants, the original Impostors Method still performed best (see Table 1 and Appendix D).



(a) Student Essays dataset.

(b) Blog Posts dataset.

Figure 3: Precision–recall curves of the Impostors Method for different impostor sampling approaches.

5.3 Statistical Inference with LLM Impostors

Table 3 reports the PAN metrics obtained when using the original Impostors Method as a hypothesis test at $\alpha = 0.05$. In the experiment, we compare different impostor sampling approaches, including our LLM-based impostors.

As can be seen, both the in-domain impostors and the LLM-based impostors compete with each other. For $c@1$ and ROC $\cdot\, c@1$, which incorporate the option to abstain from low-confidence decisions, the in-domain impostors work significantly better on the Student Essays dataset. On the Blog Posts dataset, however, the LLM-generated impostors achieve comparable effectiveness. We hypothesize that the difference is due to the Student Essays dataset being more homogeneous in terms of writing style while also including learner English. The Blogs Posts resembles the style of modern web pages (despite the dataset's age) on which LLMs have likely been trained. Interestingly, the two retrieval-based approaches exhibit comparable performance, which supports this hypothesis.

Another observation is that the distribution of effectiveness between precision and recall is reversed between in-domain and LLM-based impostors. Namely, the in-domain variant achieves the highest precision at low recall, whereas the LLM-based variant attains the highest recall at lower (mediocre) precision, suggesting that LLM impostors increase sensitivity to same-author pairs at the expense of precision.

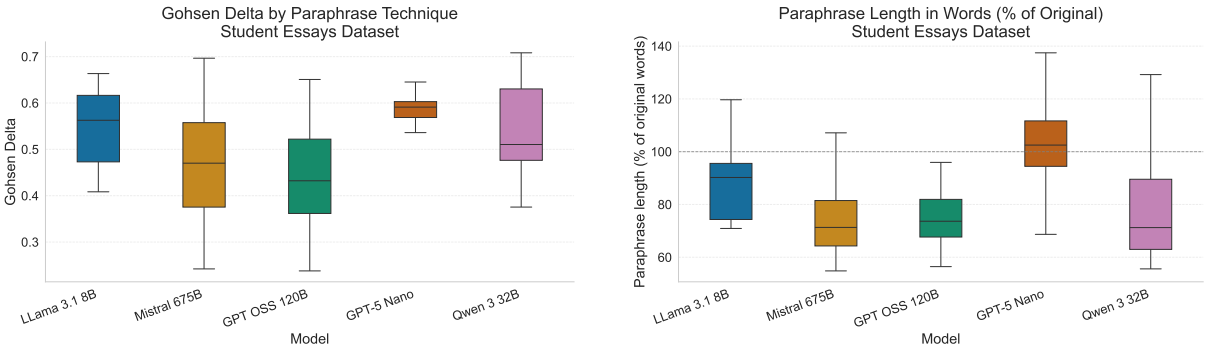
Impostor Sampling	**Effectiveness Scores**						
	Acc.	F1	Prec.	Rec.	ROC	c@1	ROC · c@1
Student Essays Dataset							
ChatNoir	0.581	0.626	0.566	0.701	0.581	0.581	0.339
Startpage	0.585	0.617	0.573	0.671	0.585	0.585	0.344
In-Domain	**0.648**	0.465	**0.962**	0.308	**0.648**	**0.648**	**0.421**
LLM impostors	0.548	**0.668**	0.528	**0.908**	0.548	0.548	0.301
Blog Posts Dataset							
ChatNoir	0.667	0.576	0.821	0.444	0.672	0.667	0.448
Startpage	0.645	0.513	0.813	0.376	0.645	0.645	0.416
In-Domain	0.671	0.520	**0.962**	0.356	0.671	0.671	0.451
LLM impostors	**0.687**	**0.677**	0.703	**0.655**	**0.687**	**0.687**	**0.473**

Table 3: PAN metrics for statistical-inference formulation of the Impostors Method ($\alpha = 0.05$). ChatNoir and Startpage are retrieval-based approaches.

5.4 Comparison with Baselines

In this experiment, we compare our approach with well-established baselines for unsupervised verification. We employ Unmasking (Koppel and Schler, 2004; Bevendorff et al., 2019b) and PPMd (Veenman and Li, 2013), where Unmasking's feature subsampling can be viewed as a conceptual precursor to the Impostors Method.

Figures 3a and 3b show that LLM-generated impostors consistently outperform Unmasking on both datasets, suggesting that generating hard negatives is more effective than feature subsampling. In contrast, PPMd clearly surpasses the LLM-based approach on the Blog Posts dataset. However, on the Student Essays dataset, our approach achieves higher precision in high-recall regions.



(a) $\Delta_{sem,syn}$ proposed by (Gohsen et al., 2023).

(b) Percentual paraphrase length differences.

Figure 4: Boxplots comparing different paraphrasing approaches for the Student Essays dataset.

5.5 Open LLMs for Impostor Synthesis

We use GPT-5 Nano for impostor synthesis. Here, we analyze alternative open source LLMs using different prompting strategies for Llama 3.1 8B, Mistral 675B, GPT OSS 120B, and Qwen 3 32B.[2]

For our use case, high-quality paraphrases should preserve topic but alter style, which requires a high semantic similarity at a low syntactic similarity with respect to the original text. To capture this trade-off, we report Gohsen et al.'s (2023) $\Delta_{\text{sem,syn}}$, which measures the difference between semantic and syntactic similarity. It combines BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), the cosine similarity between SBERT embeddings (Reimers and Gurevych, 2019), and the Word Mover's Distance (Kusner et al., 2015).

Figure 4a shows that prompt-engineered models (e.g., Qwen 3 and Llama 3.1) appear to be competitive, with high $\Delta_{\text{sem,syn}}$ suggesting strong paraphrasing. However, further analysis reveals substantial length reduction (Figure 4b): Qwen and Llama return on average only $\sim$70% and $\sim$90% of the original length, respectively. The resulting drop in k-gram overlap lowers syntactic similarity, inflating $\Delta_{\text{sem,syn}}$ and yielding a misleading impression of paraphrase quality. Length deviations are undesirable, since they introduce stylistic confounders unrelated to authorship.

GPT-5 Nano as cheap closed-source alternative achieves consistently high $\Delta_{\text{sem,syn}}$, while maintaining paraphrase length. In favor of experiment scale and for lack of budget, we leave experimenting with frontier models to future work. Suffice it to say that using a model close to the state of the art in open source models (more powerful open source models have been released since our experiments), already demonstrate the potential of our approach.

6 Conclusion

This paper revisits the Impostors Method. First, we reinterpret the method as a statistical hypothesis test that evaluates whether a similarity function identifies a candidate author more often than expected by chance. This replaces heuristic thresholds with a principled inferential framework and improves quantification of evidential strength.

Second, we explore LLMs as a controllable source for generating impostor documents. By producing paraphrases that preserve contextual properties while varying stylistic realization, LLMs can generate effective hard negatives. Our experiments show that LLM-generated impostors can achieve competitive verification performance, though their effectiveness depends on prompt design and domain. While traditional in-domain impostors remain the strongest baseline across both threshold-based and inferential settings, LLM-based impostors may provide a viable alternative to retrieval-based impostor generation approaches when suitable in-domain impostor corpora are unavailable.

Framing the Impostors Method as a statistical test clarifies its evidential interpretation, while LLM-based impostor generation expands the method's applicability when natural impostor pools are limited. In future work, we expect that the gap between in-domain and LLM-based impostors can be further reduced, e.g., by fine-tuning LLMs to the task of style paraphrasing.

[2]Refer to Appendix A.1 for prompting and model details.

Limitations

A number of limitations should be noted. We compared variants of the Impostors Method on the original datasets from 2004 and 2006, which predate LLM-generated text. Consequently, we did not evaluate the approach on more diverse domains such as PAN datasets or LLM-generated input pairs, even though LLM-generated text is becoming increasingly common in real-world settings. Its behavior under such conditions therefore remains to be investigated. Moreover, we cannot share the Student Essays dataset due to the presence of sensitive student information. It was provided under a non-disclosure agreement by J. W. Pennebaker.

It also remains unclear whether certain documents or authorial styles provide systematically better references for LLM-generated paraphrases, potentially reducing lineup fairness and introducing bias into authorship analysis. Moreover, LLM-generated paraphrases may introduce model-specific stylistic regularities, normalized spelling or grammar patterns. Such effects may violate the assumption that generated impostors preserve only author-independent variation, potentially affecting both fairness and interpretability. While LLM-based impostor generation provides greater control than retrieval-based approaches, it does not in itself resolve broader fairness issues in authorship analysis.

Moreover, we have not analyzed whether negative results should be interpreted as absence of evidence or evidence of absence, i.e., whether failure to identify stylistic similarity reflects insufficient evidential strength or indicates different authorship.

Ethical Considerations

Our contribution to the statistical inference of the Impostors Method is not intended to assess lineup fairness itself. We explicitly assume lineup fairness, as our impostor construction is inspired by description-based police lineups. Instead, our focus is on determining whether a witness identifies the candidate significantly more often than expected by chance.

Automated authorship verification systems should only be used as decision-support tools rather than as standalone evidence for authorship attribution. In particular, authorship verification should complement, rather than replace, careful human evaluation in contexts such as forensic analysis, academic integrity investigations, or historical authorship studies. Overreliance on automated systems without expert review may lead to erroneous conclusions or unfair accusations.

Risks We acknowledge that authorship analysis technologies could potentially be misused for intrusive surveillance or attempts to deanonymize individuals. Responsible deployment therefore requires clear purpose limitation, transparency about methodological limitations, and adherence to applicable legal and ethical standards. Our work aims to improve methodological rigor in authorship verification research and should not be interpreted as enabling automated attribution decisions.

Third Party Artifacts The Blog Posts corpus may be used for non-commercial research purposes provided that (Schler et al., 2006) is cited. The Student Essays corpus (Ireland and Pennebaker, 2010) is not publicly available due to the presence of sensitive student information, but was provided under a non-disclosure agreement by J. W. Pennebaker, the official custodian. We do not share any sensible identifiable author data with third-party APIs and the texts is stored separated from that identifiable data. Refer to the code repository's[1] readme file for details on data access. Our paper and code do not expose any personal data beyond the metadata and dataset description already disclosed in the original publication by Koppel and Winter (2014). Dataset statistics are reported in Appendix B.2.

Details on licensing and terms of use for the retrieval-based impostor generation and the LLM-based paraphrase generation are as follows. For the ChatNoir (Bevendorff et al., 2018) retrieval-based impostor generation, we used the `msmarco-v2.1` index without modification. This index is released under custom terms that restrict its use to non-commercial research. According to its privacy policy, Startpage[3] does not collect personal data as defined by the General Data Protection Regulation (GDPR).[4] Consequently, the queries issued in our experiments for Startpage retrieval-based impostor generation do not involve the processing of personal data.

For the LLM-based paraphrase generation, OpenAI's terms of use state that rights to model-generated outputs are assigned to the user.[5] This permits the use of generated paraphrases for research purposes. All open-source models used for

[3] https://www.startpage.com/ (06.01.2026)
[4] https://www.startpage.com/en/privacy-policy (12.03.2026)
[5] https://openai.com/policies/terms-of-use/ (12.03.2026)

paraphrasing are released under the Apache License 2.0, except for Llama, which is distributed under the Llama 3.1 Community License Agreement. These licenses allow the use and evaluation of the models and their outputs in non-commercial research settings.

We cited the third party artifacts we used where appropriate in the paper. With our repository[1] we also provide a pyproject.toml file with versions used in our work. Beyond those, we also made use of the following frameworks: We used the evaluation frameworks by PyTorch (Ansel et al., 2024) under Modified BSD license, Gensim (Řehůřek and Sojka, 2010) version 4.3.3 under GNU LGPLv2.1 license, Huggingface Evaluate (Face, 2022) version 0.4.3 (for several metrics including ROUGE and BERTScore) under Apache-2.0 licence, Scikit-learn (Pedregosa et al., 2011) version 1.6.1 under 3-Clause BSD license, correctipy[6] under MIT licence with modifications resolving some variance computation error and adding a two-sided test, statsmodels (Skipper and Josef, 2010) under 3-Clause BSD license, Transformers (Wolf et al., 2020) version 4.53.3 under Apache-2.0 licence, text processing by SpaCy (Honnibal et al., 2020) version 3.8.7 under the MIT licence and NLTK (Bird et al., 2009) version 3.9.1 (among others, for several metrics including BLEU and METEOR) under Apache-2.0 licence, embeddings by Sentence-BERT (Reimers and Gurevych, 2019), NumPy (Harris et al., 2020) version 1.26.4 under BSD-3-Clause, Pandas (The pandas development team) version 2.1.3 under BSD-3-Clause, and Matplotlib (Hunter, 2007) version 3.10.3.

References

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Margaret A. Berger. 1993. Procedural Paradigms for Applying the Daubert Test. *Minn. L. Rev.*, 78:1345.

Janek Bevendorff, Matthias Hagen, Benno Stein, and Martin Potthast. 2019a. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6301–6306. Association for Computational Linguistics.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 820–824. Springer.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019b. Generalizing Unmasking for Short Texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 654–659. Association for Computational Linguistics.

Janek Bevendorff, Yuxia Wang, Jussi Karlgren, Matti Wiegmann, Maik Fröbe, Akim Tsivgun, Jinyan Su, Zhuohan Xie, Mervat Abassy, Jonibek Mansurov, Rui Xing, Minh Ngoc Ta, Kareem Ashraf Elozeiri, Tianle Gu, Raj Vardhan Tomar, Jiahui Geng, Ekaterina Artemova, Artem Shelmanov, Nizar Habash, Efstathios Stamatatos, Iryna Gurevych, Preslav Nakov, Martin Potthast, and Benno Stein. 2025a. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 3504–3534. CEUR-WS.org.

Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2024. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF*

[6] https://github.com/hendersontrent/correctipy (08.04.2026)

2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, pages 2486–2506. CEUR-WS.org.

Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. 2025b. The Two Paradigms of LLM Detection: Authorship Attribution vs Authorship Verification. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3762–3787. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The Importance of Suppressing Domain Style in Authorship Analysis. *CoRR*, abs/2005.14714.

Remco R. Bouckaert and Eibe Frank. 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings*, Lecture Notes in Computer Science, pages 3–12. Springer.

Jean-Baptiste Camps and Benedetta Salvati. 2023. On Burgundian (di)vine orators and other impostors: Stylometry of Late Medieval Rhetoricians. In *Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2022, Graz, Austria, July 10-14, 2023, Conference Abstracts*.

Isobelle Clarke and Tim Grant. 2024. Impostors tending to the wild: Purposes of authorship analysis and a specific impostors method in the Karvounakis terrorism case. *Language and Law / Linguagem e Direito*, 11(2).

Melissa F. Colloff, Brent M. Wilson, Travis M. Seale-Carlisle, and John T. Wixted. 2021. Optimizing the selection of fillers in police lineups.

Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.*, 10(7):1895–1923.

Patrik Dokoupil and Ladislav Peska. 2021. LiGAN: Recommending Artificial Fillers for Police Photo Lineups (Short paper). In *Joint Workshop Proceedings of the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), Virtual Event, Amsterdam, The Netherlands, September 25, 2021*, volume 2960 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Hugging Face. 2022. Evaluate: A library for easily evaluating machine learning models. https://github.com/huggingface/evaluate.

Marcel Gohsen, Matthias Hagen, Martin Potthast, and Benno Stein. 2023. Paraphrase Acquisition from Image Captions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3340–3350. Association for Computational Linguistics.

Marcel Gohsen, Matthias Hagen, Martin Potthast, and Benno Stein. 2024. Task-Oriented Paraphrase Analytics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15640–15654. ELRA and ICCL.

Tim Grant. 2012. TXT 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages. *Journal of Law and Policy*, 21:467–494.

Tim Grant. 2025. The future of AI in court. https://aifl-blog.com/the-future-of-ai-in-court-f39ff71954da. [Accessed 14-12-2025].

Vladimir Gurvich and Mariya Naumova. 2021. Logical Contradictions in the One-Way ANOVA and Tukey-Kramer Multiple Comparisons Tests with More Than Two Groups of Observations. *Symmetry*, 13(8):1387.

Oren Halvani, Christian Winter, and Lukas Graner. 2017. On the Usefulness of Compression Models for Authorship Verification. In *Proceedings of the 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, August 29 - September 01, 2017*, pages 54:1–54:10. ACM.

Charles R. Harris, K. Jarrod Millman, Stéfan van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nat.*, 585:357–362.

Josué Gerardo Gutiérrez Hernández, José Casillas, Paola Ledesma, Gibran Fuentes-Pineda, and Iván Vladimir Meza Ruíz. 2015. Homotopy Based Classification for Author Verification Task: Notebook for PAN at CLEF 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

John D. Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, 9(3):90–95.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Patrick Juola. 2015. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digit. Scholarsh. Humanit.*, 30(Suppl-1):i100–i113.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Andrew W Jurs and Scott DeVito. 2024. Machines Like Me: A Proposal on the Admissibility of Artificially Intelligent Expert Testimony. *PEPPERDINE LAW REVIEW*, 51.

Mike Kestemont, Justin Anthony Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Syst. Appl.*, 63:86–96.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *The Twelfth International Conference on Learning Representations*.

Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*, pages 977–983. CEUR-WS.org.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a One-Class Classification Problem. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM.

Moshe Koppel and Yaron Winter. 2014. Determining If Two Documents Are Written by the Same Author. *J. Assoc. Inf. Sci. Technol.*, 65(1):178–187.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, 4:62627.

A. M. Levi and Noam Jungman. 1995. The Police Lineup: Basic Weaknesses, Radical Solutions. *Criminal Justice And Behavior*, pages 347–369.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL Workshop*.

Geoffrey Stewart Morrison. 2022. Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy*, 5:100270.

Ben Nagy. 2024. Bootstrap Distance Imposters: High precision authorship verification with improved interpretability. In *Proceedings of the Computational Humanities Research Conference 2024, Aarhus, Denmark, December 4-6, 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 482–493. CEUR-WS.org.

Tempestt J. Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon L. Woodard. 2018. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.*, 50(6):86:1–86:36.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *CoRR*, arXiv:2508.10925.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ladislav Peska and Hana Trojanova. 2018. Personalized Recommendations in Police Photo Lineup Assembling Task. In *Proceedings of the 18th Conference Information Technologies - Applications and Theory (ITAT 2018), Hotel Plejsy, Slovakia, September 21-25, 2018*, volume 2203 of *CEUR Workshop Proceedings*, pages 157–160. CEUR-WS.org.

Nektaria Potha and Efstathios Stamatatos. 2017. An Improved Impostors Method for Authorship Verification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*, volume 10456 of *Lecture Notes in Computer Science*, pages 138–144. Springer.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike

Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. 2016. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 393–407, Berlin Heidelberg New York. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, 13(7):1443–1471.

Shachar Seidman. 2013. Authorship Verification Using the Impostors Method Notebook for PAN at CLEF 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Dmitry A. Shkarin. 2002. PPM: One Step to Practicality. In *2002 Data Compression Conference (DCC 2002), 2-4 April, 2002, Snowbird, UT, USA*, pages 202–211. IEEE Computer Society.

Seabold Skipper and Perktold Josef. 2010. Statsmodels: Econometric and Statistical Modeling with Python. *9th Python in Science Conference*.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the Author Identification Task at PAN 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*, pages 877–897. CEUR-WS.org.

Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2022. Overview of the Authorship Verification Task at PAN 2022. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2301–2313. CEUR-WS.org.

Efstathios Stamatatos, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2023. Overview of the Authorship Verification Task at PAN 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2476–2491. CEUR-WS.org.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *J. Assoc. Inf. Sci. Technol.*, 67(1):239–242.

The pandas development team. pandas-dev/pandas: Pandas.

Jennifer L. Tunnicliff and Steven E. Clark. 2000. Selecting Foils for Identification Lineups: Matching Suspects or Descriptions? *Law and Human Behavior*, 24(2):231–258.

Tim Valentine and Pamela Heaton. 1999. An evaluation of the fairness of police line-ups and video identifications. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 13(S1):S59–S72.

Cor J. Veenman and Zhenshi Li. 2013. Authorship Verification with Compression Features. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Matthew Voice, Chloe Harrison, Tim Grant, and Marcello Giovanelli. 2025. Towards a cognitive forensic stylistics: An intercoder reliability test for replicable feature finding in the Operation Heron corpus. *Language and Literature: International Journal of Stylistics*, 34(4):327–348.

Zeev Volkovich and Renata Avros. 2025. Comprehension of the Shakespeare authorship question through deep impostors approach. *Digit. Scholarsh. Humanit.*, 40(1):308–328.

Gary L. Wells and C.A. Elizabeth Luus. 1990. Police Lineups as Experiments: Social Methodology as a Framework for Properly Conducted Lineups. *Personality and Social Psychology Bulletin*, 16(1):106–117.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics.

Anne S. Yilmaz, Taylor C. Lebensfeld, and Brent M. Wilson. 2022. The reveal procedure: A way to enhance evidence of innocence from police lineups. *Law and Human Behavior*, 46(2):164–173.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, pages 45–50, Valetta, MT. University of Malta.

Field	Description	Extractor	Generator
Heading	A short, reader-facing title or header that names or frames the text for presentation. It reflects how the content is labeled or introduced, and may use stylistic, rhetorical, or attention-directing language. Multiple different headings could plausibly describe the same topic.	Output	Input
Topic	The core subject matter or thematic focus of the text, expressed abstractly and independently of wording or presentation. This should describe what the text is about at a conceptual level, typically as a concise noun phrase, and should remain stable even if the heading or phrasing changes.	Output	Input
Authorial Task	The primary communicative or rhetorical purpose the author is pursuing in writing the text, i.e., what the author is trying to accomplish (e.g., to inform, persuade, narrate, explain, criticize, reflect, instruct, or document). This reflects intention rather than emotional style.	Output	Input
Tone	The emotional, attitudinal, or stylistic coloring of the language, i.e., how the author sounds while carrying out the authorial task (e.g., neutral, urgent, ironic, solemn, sarcastic, optimistic). Tone describes manner or affect, not purpose.	Output	Input
Genre	The literary or content genre of the text (e.g., essay, poem, news article, academic paper, novel, speech, letter, religious text, etc.).	Output	Input
Register/Style	The social level of language used in the text (e.g., casual, colloquial, academic, bureaucratic, poetic, technical).	Output	Input
Century	The century in which the text was likely written (e.g., '18th century', '20th century', '21st century'). Use language, style, and context clues to infer.	Output	Input
Audience	The intended audience or readership (e.g., general public, scholars, children, political leaders, students, religious followers).	Output	Input
Author	A concise author profile including likely traits such as gender (if implied), profession, education level, nationality, or perspective (e.g., 'a 19th-century British poet', 'a modern journalist', 'an academic researcher'), using idiosyncrasies in the text (e.g., spelling or grammar errors) to guide the inference.	Output	Input
Bulletpoints	A list of concise bullet points summarizing the main ideas or arguments of the text, written in plain language, focusing on meaning rather than wording.	Output	Input
Length	The approximate target length of the generated text, in number of words.	-	Input
Text	see above	Input	Output

Table 4: DSPy descriptions of all input and output fields for the extractor and generator signatures, including the role of each field.

A Impostor Synthesis Approaches

This section describes the implementation details of the evaluated impostor synthesis approaches, including LLM-based paraphrasing (Section A.1), which generates stylistically controlled impostors directly from a reference text.

A.1 LLM-Based Paraphrasing

This section presents the proposed two-step DSPy paraphrasing pipeline (Section A.1.1) and an ablation based on direct prompting open-source LLMs (Section A.1.2).

A.1.1 Two-Step Paraphrasing

Implementation. We use DSPy (Khattab et al., 2024) with the field descriptions from Table 4 with OpenAI's `openai/gpt-5-nano-2025-08-07`. This model provides a 400 000-token context window, supports up to 128 000 output tokens, and has a knowledge cutoff of May 31, 2024.[7]

[7] https://developers.openai.com/api/docs/models/gpt-5-nano (14.03.2026)

Language	Relative percentage (%)	Count
English	99.6537	210 099
Tagalog	0.1579	333
Indonesian	0.0783	165
Spanish	0.0403	85
German	0.0294	62
French	0.0194	41
Norwegian	0.0109	23
Hindi	0.0052	11
Greek	0.0014	3
Hungarian	0.0014	3
Dutch	0.0009	2
Romanian	0.0005	1
Swedish	0.0005	1

Table 5: Language distribution of paraphrases.

Paraphrase Language. We used the Python library `langdetect`, a port of Google's language-detection library[8] to identify the languages of the generated paraphrases (Table 5). Manual inspection revealed that non-English paraphrases occurred when authors were identified as non-English. We therefore explicitly specified English as the target paraphrase language in subsequent runs.

[8] https://github.com/shuyo/language-detection (24.05.2026)

Example Paraphrasing Failure. Table 6 shows a failure case of the proposed two-step prompting strategy using `openai/gpt-5-nano-2025-08-07`, where the generated paraphrase achieved low semantic similarity to the source text despite moderate thematic overlap.

The bullet-point extraction stage correctly captures several high-level themes from the original text, including details on friends and relationships. However, the extraction inferred a highly standardized and reflective blogging style from a heavily abbreviated and non-standard original text, effectively normalizing the author's idiosyncratic writing style instead of preserving it (Table 7).

The following excerpt shows the original source text:

> "Well ds s e 1st tme dat mo is wrtng a blog.. lolx.. werd hor y start arh.. mo also dont noe.. juz haf e feel of wrtng somethng here to express moself ba... Lfe s full 0f ups n downs.. haha.. frends,f@mly,relatonshps... [...] he wont juz throw hS temper n volence at mo too.. how to take it wor.. took t for another 3yrs when moi fnally made e decson break up was on mo nnd alotz of tmes.. [...] heartbreak coz we used to meet almost everyday den now he go le cant meet muche le... cant get used to t at 1st but slowly got used to t le... e seperate was hard for mo but soon mo got used to t le.. frends was around to acc mo... but thngs start to change.. rte after hs army.. he arh used to b a carng , understandng n dotes on mo alotz.. but change.. hs famly to hm was muche more mpt den mo.. he stll dote on mo but turn to no lfe.. always ask mo stay at hs house dong nothng.. after 10pm mo not allow to haf sms or fone calls.. not allow to go out wf frends when not wkng or studyng at all.. e changes was hard.. [...]"

The following excerpt shows the corresponding generated paraphrase:

> "Hey there, welcome to my little corner of the internet. I'm trying this blog thing out so I can spill the lifey bits-the wins, the mess, the mundane, and the moments that somehow end up shaping who I am. This time, it's all about the people who orbit me: friends, family, and two very different relationships that have colored the last few years in bright and muddy shades alike. [...] We eventually grew apart, and the breakup was less a dramatic twist and more a slow, necessary quieting of a relationship that couldn't survive the weight of constant pressure. [...] a six-year chapter that taught me a lot with very loud, very messy lessons. The two of us rode the rollercoaster up and down for a long time: arguments that felt like storms, periods when we talked in harsh tones like it would fix things, and days when it felt easier to retreat than to resolve. There were threats—moments when I worried about the safety of crossing wires with family and expectations [...] "

Metric	Score
Semantic Similarity Average	0.532
Syntactic Similarity Average	0.102
$\Delta_{\text{sem,syn}}$	0.429

Table 6: Similarity scores for low semantic-similarity paraphrase in the Blog Posts dataset.

DSPy field	Content
audience	general blog readers
author	a young, informal blogger (likely female) writing in casual online English with slang, from Southeast Asia
authorial_task	to share personal experiences and feelings about friendships, family, and romantic relationships
bulletpoints	• The writer starts a blog to express herself and share life 2019s ups and downs, especially about friends and relationships. • Family life at home is rough and feels distant despite living together. • Family life at home is rough and feels distant; despite being together, communication often feels like strangers.. • Two close friends stand out: Charbo (a long-time, supportive friend) and Jeme (a caring and understanding friend). • Old Guy: a six-year relationship that involved arguments and periods of tough behavior, leading to a breakup; there were threats and family concerns, but they eventually part ways • New Guy: a newer, long-distance romance with a fun, caring guy who may return soon; feelings are growing and hopeful. • Overall, the post mixes humor with honesty, showing hope for the future while acknowledging challenges, and it signals a plan to keep writing.
century	21st century
genre	personal blog post
heading	Blogging My Life: Friends, Family, and Two Relationships
length	1138
register_style	colloquial, informal, conversational
tone	confessional, humorous, reflective
topic	Personal relationships and social life

Table 7: Extracted metadata from Blog Posts text.

Example Paraphrasing success. Table 8 shows a successful case of the proposed two-step prompting strategy using `openai/gpt-5-nano-2025-08-07`, where the generated paraphrase achieved high semantic similarity to the source text.

The following excerpt shows the original source text:

> "[...] When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation. [...] He is at this time transporting large Armies of foreign Mercenaries to compleat the works of death, desolation and tyranny, already begun with circumstances of Cruelty & perfidy scarcely paralleled in the most barbarous ages, and totally unworthy the Head of a civilized nation. He has constrained our fellow Citizens taken Captive on the high Seas to bear Arms against their Country, to become the executioners of their friends and Brethren, or to fall themselves by their Hands. He has excited domestic insurrections amongst us, and has endeavoured to bring on the inhabitants of our frontiers, the merciless Indian Savages, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions. [...] "

The following excerpt shows the corresponding generated paraphrase:

> "In the course of human events it hath become necessary for these United Colonies to declare the causes which impel them to sever the political ties by which they have long been bound to the Crown of Great Britain, and to assume among the powers of the earth a station of independence and equal rank, to which the Laws of Nature and the Creator of mankind entitle them. [...] To prove that we do not act hastily, but with earnest purposes and sober resolve, we here declare that our isolation from a government so violent and perverse is necessary for our protection, and that we are now to be known as free and independent States. We claim the full power to wage war, to conclude peace, to contract alliances, to establish commerce, and to do all other acts and things which Independent States may of right do. And for the support of this Declaration, with a firm reliance on the protection of Divine Providence, we mutually pledge to each other our Lives, our Fortunes, and our sacred Honor. [...] "

Metric	Score
Semantic Similarity Average	0.887
Syntactic Similarity Average	0.594
$\Delta_{\text{sem,syn}}$	0.292

Table 8: Similarity scores for high semantic-similarity paraphrase in the Blog Posts dataset.

DSPy field	Content
audience	the American colonists and a global audience (the world)
author	Thomas Jefferson, 18th-century American statesman and principal drafter for the Continental Congress
authorial_task	to declare independence from Britain and justify the secession to an international audience
bulletpoints	• Declares the thirteen American colonies free and independent from Britain. • Affirms self-evident truths: all men are created equal and endowed with unalienable rights, including life, liberty, and the pursuit of happiness. • Affirms that governments derive their just powers from the consent of the governed, and that people may alter or abolish them when they become destructive. • Lists grievances and usurpations by the British Crown to justify separation. • Declares the colonies' separation is necessary and that they are free and independent states with power to wage war, conclude peace, and establish commerce. • Appeals to divine Providence for support and pledges the lives, fortunes, and sacred honor of the signers.
century	18th century
genre	historical political document
heading	The Declaration of Independence (1776)
length	1339
register_style	archaic, formal, legalistic 18th-century prose
tone	formal, solemn, and resolute
topic	Political independence declaration and justification based on unalienable rights

Table 9: Extracted metadata from Blog Posts text.

The bullet-point extraction stage correctly captures the topic and writing style of the original text (Table 9).

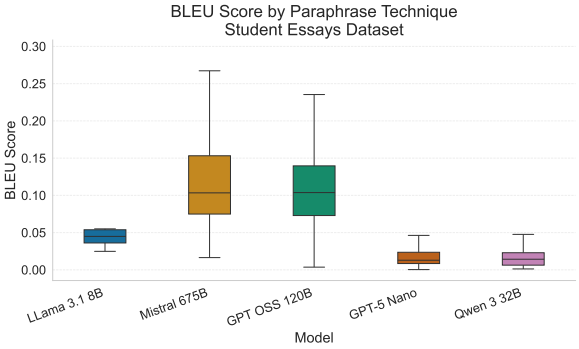


Figure 5: Boxplots comparing different paraphrasing approaches for BLEU for the Student Essays dataset.

A.1.2 Prompted paraphrasing

Open Source Models We accessed and evaluated several open-source models through the GDWG AI services platform.[9] The models evaluated in our experiments are Meta Llama 3.1 8B Instruct[10] Mistral Large Instruct 3 675B Instruct 2512[11] OpenAI GPT OSS 120B (OpenAI, 2025)[12] Qwen 3 32B[13]

Prompts We used the following prompts to instruct open-source LLMs. The prompts were iteratively refined during experimentation; however, the resulting paraphrases remained unsatisfactory overall, primarily because the models tended to generate overly short outputs and struggled to consistently follow the required output format.

- **Prompt 1.** Paraphrase the text above without changing its meaning. Use different words and vary the sentence structure while maintaining a consistent tone. Your paraphrase should be three times as long than the original. Output only the paraphrased sentence, with NO explanations or extra text.
- **Prompt 2.** Rewrite the text to preserve meaning but restructure sentences. Avoid copying original phrasing. Output only the paraphrase.
- **Prompt 3.** Paraphrase the text with different syntax and sentence order. Keep all facts and named entities. No explanations.
- **Prompt 4.** Express the same ideas with different wording and clause structure. Keep tone and meaning. Output only the paraphrase.

[9] https://docs.hpc.gwdg.de/services/ai-services/chat-ai/models/index.html (14.03.2026)
[10] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (15.03.2026)
[11] https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512 (15.03.2026)
[12] https://huggingface.co/openai/gpt-oss-120b (15.03.2026)
[13] https://huggingface.co/Qwen/Qwen3-32B (15.03.2026)

B Data

In the following, we outline our preprocessing pipeline (Section B.1) and describe the text-pair selection procedure (Section B.2).

B.1 Dataset Preprocessing

Minimum Length. All input texts shorter than a pre-defined minimal length (i.e., 700 whitespace-separated tokens), which are not supposed to be upsampled, are skipped. The choice of minimum text length was informed by literature research. While Bevendorff et al. (2019b) used text chunks of at least 700 words for their unmasking approach, Koppel and Schler (2004) set the minimum to 500 words. Recent work (Bevendorff et al., 2025b) identifies 2500–4000 characters to be sufficient for effective LLM detection framed as Authorship Verification.

Preprocessing Pipeline. The preprocessing pipeline removes formatting and layout information to obtain plain text. To determine appropriate preprocessing steps, we examined multiple datasets from different domains and analyzed their characteristic artifacts (Figure 6). In addition to the Student Essays and Blog Posts datasets, we included the PAN'20 dataset (Bischoff et al., 2020) and a Gutenberg dataset in this analysis. The Gutenberg dataset consists of a curated selection of 19 works by 7 authors from the 16th to 19th centuries, manually collected from the Project Gutenberg website[14] and Wikipedia. These additional datasets were used solely for this analysis to capture cross-domain effects of different preprocessing steps. Based on this analysis, we identified eight potentially useful preprocessing operations.

These include removing HTML artifacts, play artifacts, and newlines, converting UTF-8 characters to ASCII, and stripping leading and trailing whitespace. Domain-specific artifacts are removed using regular expressions (Table 10). We deliberately avoid lowercasing the texts, as our preliminary analysis showed no measurable benefit across datasets (Figure 6) while potentially discarding meaningful authorial capitalization choices.

B.2 Selection of Text Pairs

For the Blog Posts and Student Essays datasets, we selected pairs of texts according to specific criteria to control potential confounding variables.

[14] https://www.gutenberg.org/ (12.03.2026)

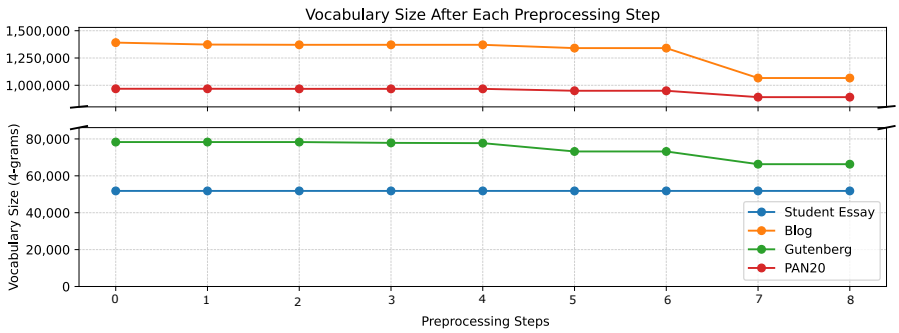


Figure 6: Effect of preprocessing steps on vocabulary size. The vocabulary (step 0) contains unique space-free character 4-grams. Step 1 decodes any HTML entities, 2 strips out HTML tags, 3 removes play artifacts (e.g., KING), 4 removes chapter artficats, 5 removes other play artifcats (e.g., ACT, SCENE), 6 collapses all whitespaces, 7 lowercases text. For this experiment, there was no minimum text length. For PAN, train and test split were combined.

Description	Regular Expression
HTML tags	`<[^>]+>`
Play artifacts	`^[A-Z\s]+[.:]?$`
Chapter artifacts	`^\s*Chapter\s+\w+.*\n\s*\n`
Play artifacts	`\s*ACT\s+\w+\b\.?`
	`^\s*SCENE\s+\w+\b\.?`
	`\[(.*?)\]` → `\1`
	`_+"`
Whitespace	`\s+`

Table 10: Overview of the regular expressions and substitution rules applied during text preprocessing (see Figure 6). Each rule targets specific artifacts.

Dataset	# pairs	# authors	# p_s	# $p_{\neg s}$	$\varnothing l_c$	$\varnothing l_w$	max l_w	σ_{l_w}	median l_w
Blog Posts	7423	5378	3944	3479	6245.69	1154.80	70374	1419.49	905
Student Essays	1284	568	642	642	3667.17	708.06	1634	174.48	660

Table 11: Statistics of the preprocessed Blog Posts and Student Essays datasets. p_s and $p_{\neg s}$ denote same-author and different-author pairs, respectively. l_w and l_c denote text length in words and characters. Reported statistics include mean ($\varnothing$), maximum (max), standard deviation (σ), and median text length. # denotes number of.

For the Blog Posts dataset, text pairs were matched by topic, year, gender, and author age. Within each pair, one text was sampled from the earlier writings of the author and the other from the later writings. For the Student Essays dataset, following Koppel and Winter (2014), we drew all text pairs from different tasks, with authors matched by sex, ethnicity, and political orientation. Descriptive statistics for all preprocessed datasets are provided in Table 11.

Effective Dataset Sizes. Although the curated datasets contain the pair configurations reported in Table 11, the effective number of evaluated pairs (Table 12) is smaller in some cases. This reduction results primarily from balancing the number of same-author and different-author pairs (e.g., $6958 = 2 \times \min(3944, 3479)$ instead of 7423 Blog Posts pairs).

Impostor Generation Approach	Blog Posts	Student Essays
Retrieval-Based (ChatNoir)	6032	1284
Retrieval-Based (StartPage)	2078	1284
In-Domain	6958	1284
Two-Step Paraphrasing (LLM)	1705	1284
Baselines	6958	1284

Table 12: Number of pairs per dataset (i.e., Blog Posts and Student Essays) used in experiments.

C Experiments

The following sections provide additional experimental details and results. Section C.1 compiles the used hyperparameters and the used computing resources. Section C.2 summarizes PAN metrics across cross-validation splits, Section C.3 reports effectiveness measures for the different-author setting, Section C.4 compares effectiveness of reimplementations of existing Impostors Method extensions, and Section C.5 compares directional consistency of hypothesis test for different impostor generation variants.

C.1 Hyperparameters

Model Configurations. The Unmasking detector was configured with 10-fold cross-validation, 60 chunks, 30 iterative feature-removal rounds, the top 250 features removed per round and upsampling as proposed by Bevendorff et al. (2019b). The Impostors Method was configured with 50 impos-

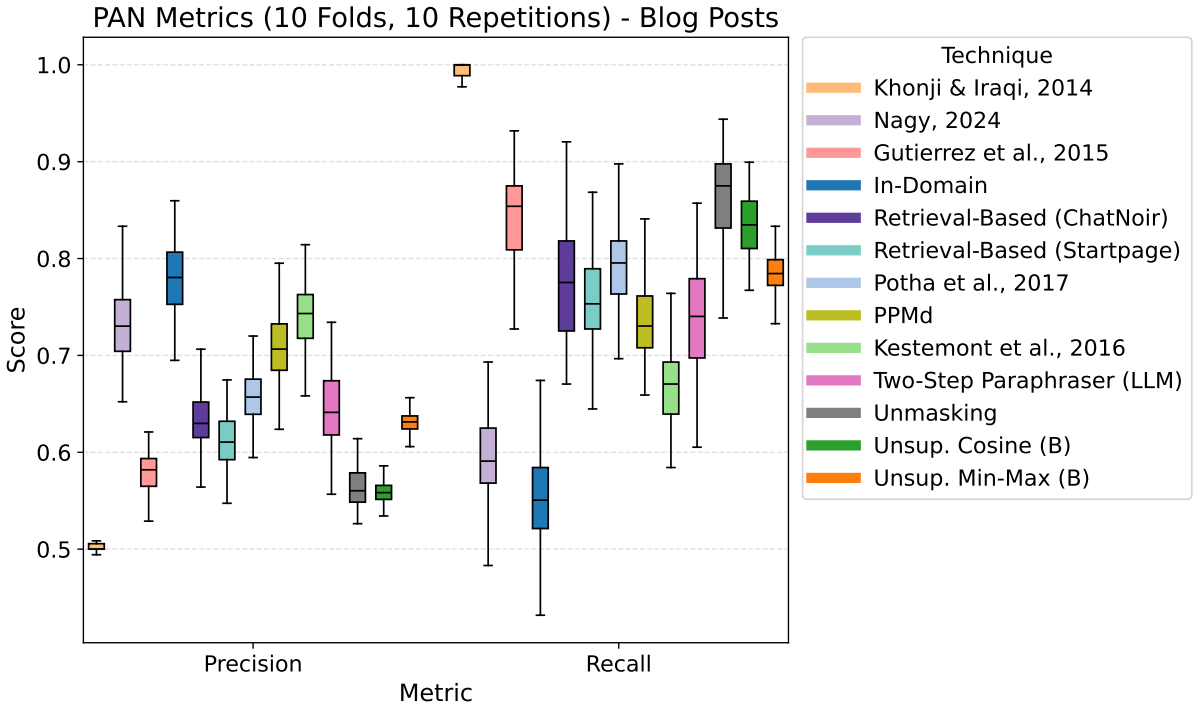


Figure 7: Boxplots over different metrics across 10 times 10-fold CV on the Blog Posts dataset in the same-author setting.

tors sampled from a pool of 75 potential impostors across 100 rounds. Documents were represented using the top 100 000 features, with upsampling enabled, and 50% of the feature space was randomly removed per iteration.

Computing Infrastructure. All experiments were conducted on a MacBook Pro equipped with an Apple M5 chip and 32 GB RAM. Generated artifacts, including scores, paraphrases, and dataset pairs, were stored in a 32 GB MongoDB instance deployed on a Kubernetes cluster using a Helm chart. Additional details regarding the infrastructure and deployment configuration are provided in the project repository.[1]

C.2 PAN Metrics

Inspection of repeated CV scores (Figure 7) further reveals distinct trade-offs: LLM-based impostors yield consistently high recall but lower precision with relatively low variance compared to in-domain impostors. Retrieval-based approaches exhibit greater variability in recall and perform slightly worse than the LLM-based method in both precision and recall.

C.3 Different-Author Setting

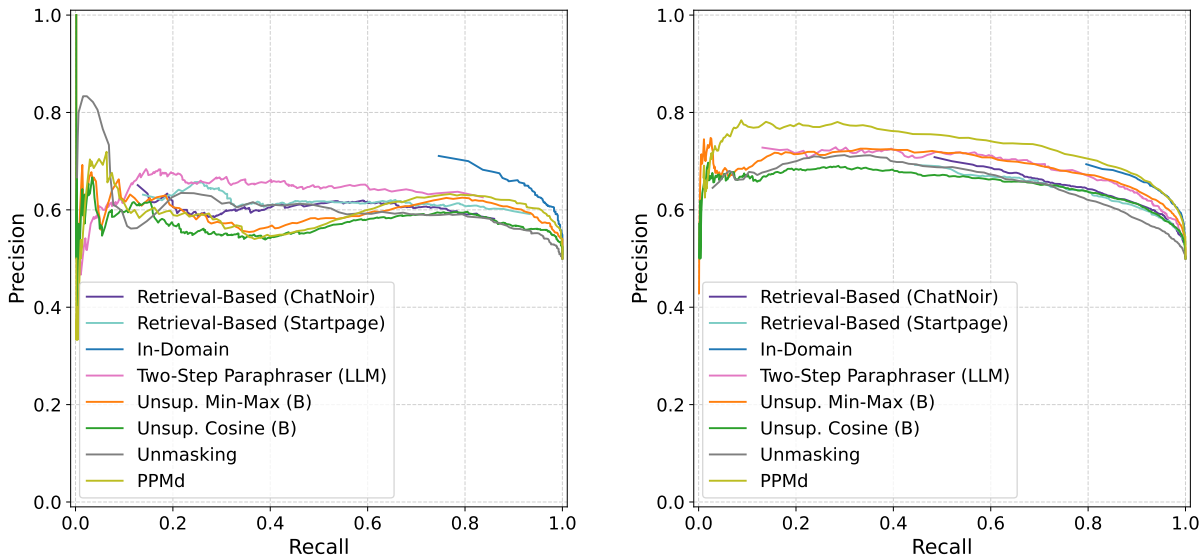
Koppel and Winter (2014) report precision and recall for different impostor generation approaches under two settings that differ in the class considered positive (i.e., same-author vs. different-author). Their results show that the in-domain approach achieves superior performance in both settings.

Figures 8b and 8a present precision-recall curves for the different-author setting on the Blog Posts and Student Essays datasets, respectively. Consistent with the original findings, in-domain impostor generation yields the highest overall effectiveness among the evaluated traditional impostor generation approaches. We further observe that effectiveness in the different-author setting is lower than in the same-author setting.

C.4 Impostors Method Extensions

We reimplemented five existing extensions of the Impostors Method and evaluated them on the Blog Posts and Student Essays datasets in terms of precision and recall (Figure 9b and Figure 9a). The hyperparameters for these extensions are summarized in Table 13. We do not optimize hyperparameters, but use the values reported in the respective papers and for (Nagy, 2024) existing implementations.

The same-author scores reported in Table 1 are



(a) Student Essays dataset.

(b) Blog Posts dataset.

Figure 8: Precision–recall curves for the different-author setting of the Impostors Method for different impostor sampling approaches.

Author	# Rounds	Proportion to delete	# Impostors	# Potential Impostors
(Khonji and Iraqi, 2014)	50	0.60	20	30
(Hernández et al., 2015)	50	–	2	3
(Kestemont et al., 2016)	50	0.60	20	30
(Potha and Stamatatos, 2017)	10	0.50	5	50
(Nagy, 2024)	100	0.67	30	45

Table 13: Hyperparameter for extensions of original Impostors Method.

Author	Dataset	τ_{low}	τ_{high}	τ_{F1}
(Koppel and Winter, 2014)	Student Essays	0.01 ± 0.00	0.04 ± 0.01	0.01 ± 0.00
	Blog Posts	0.01 ± 0.00	0.09 ± 0.01	0.01 ± 0.00
(Khonji and Iraqi, 2014)	Student Essays	0.85 ± 0.01	0.88 ± 0.00	0.43 ± 0.01
	Blog Posts	0.88 ± 0.00	0.94 ± 0.00	0.83 ± 0.00
(Hernández et al., 2015)	Student Essays	0.49 ± 0.03	0.66 ± 0.02	0.36 ± 0.04
	Blog Posts	0.62 ± 0.01	0.68 ± 0.02	0.44 ± 0.09
(Kestemont et al., 2016)	Student Essays	0.03 ± 0.00	0.09 ± 0.01	0.01 ± 0.00
	Blog Posts	0.02 ± 0.00	0.13 ± 0.03	0.01 ± 0.00
(Potha and Stamatatos, 2017)	Student Essays	0.04 ± 0.00	0.06 ± 0.01	0.04 ± 0.00
	Blog Posts	0.06 ± 0.00	0.09 ± 0.00	0.04 ± 0.00
(Nagy, 2024)	Student Essays	0.01 ± 0.01	0.03 ± 0.02	0.01 ± 0.00
	Blog Posts	0.02 ± 0.00	0.16 ± 0.00	0.01 ± 0.00

Table 14: Average thresholds optimized on ten 10 fold CV for different extensions of the original Impostors Method. The lower and upper thresholds define the c@1 rejection band, whereas τ_{F1} is the threshold optimized for F1, precision, recall and accuracy. Values are reported as mean $\pm$ standard deviation over the per-fold thresholds from cross-validation. These dataset-specific thresholds were used to obtain the scores reported in Table 1.

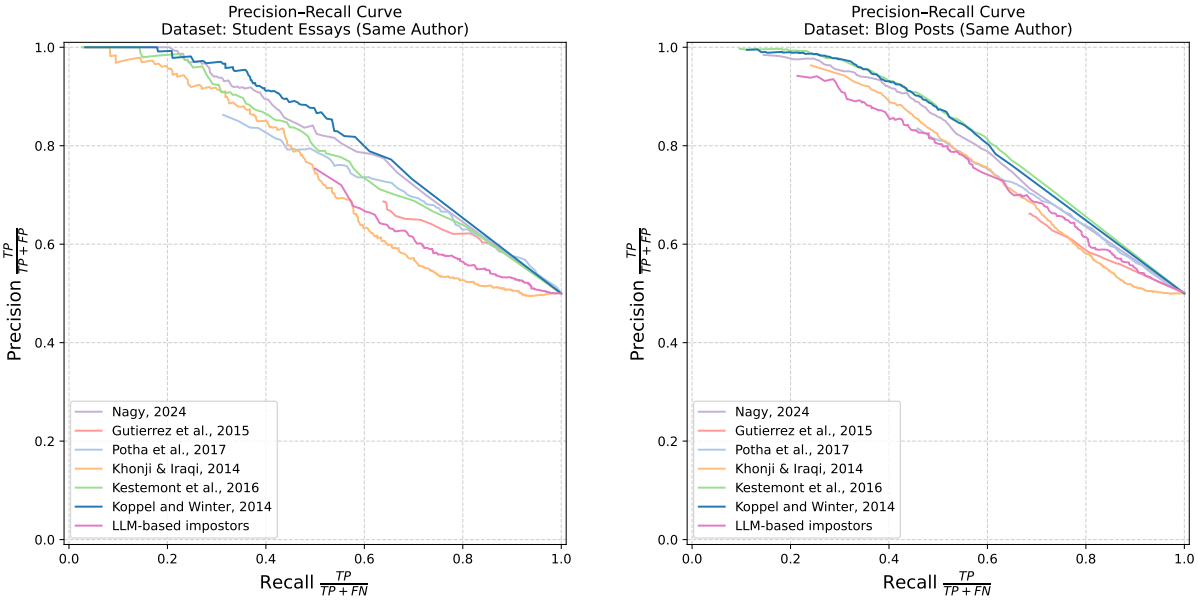
averaged over ten 10 fold cross-validation runs (i.e., 10% test data per split). For each split, decision thresholds were optimized separately for c@1 and F1, with the F1-optimized thresholds also used to compute precision and recall. The corresponding dataset-specific threshold means and standard deviations are reported in Table 14.

Dataset	Method	n	Cohen's κ
Blog Posts	Retrieval-Based (ChatNoir)	6032	0.374
	Retrieval-Based (Startpage)	2158	0.300
	In-Domain	6958	0.587
	Two-Step Paraphraser (LLM)	1761	0.184
Student Essays	Retrieval-Based (ChatNoir)	1284	0.389
	Retrieval-Based (Startpage)	1284	0.400
	In-Domain	1284	0.524
	Two-Step Paraphraser (LLM)	1284	0.226

Table 15: Cohen's κ for agreement between the two directional statistical-inference decisions; n denotes the number of aligned pairs.

C.5 Directional Consistency

Directional consistency between the two hypothesis tests is measured using Cohen's κ (Table 15). The in-domain variant achieves the strongest directional agreement, indicating comparatively stable impostor selection across both test directions. In contrast, the LLM-based variant yields substantially lower agreement but the highest recall on both datasets (Table 3), suggesting that its generated impostors increase sensitivity to same-author pairs at the cost of directional stability and precision. The two retrieval-based approaches exhibit similar agreement patterns across both datasets.



(a) Student Essays dataset.

(b) Blog Posts dataset.

Figure 9: Precision–recall curves for the same-author setting of the Impostors Method for reimplemented extensions of the Impostor Method.

D Impostors Method Variants

In the following, we outline existing extensions of and case studies using the Impostors Method.

Extensions The General Impostors Method (GenIM) (Seidman, 2013) is an extension to the Impostors Method and ranked first in PAN'13 authorship identification competition. This work optimized different hyperparameters, e.g., features, similarity function, number of iterations, proportion of features per iteration, thresholds, on different language corpora, and extended the authorship verification method to work with AA by running the method on each pair and averaging the results.

Another extension of the Impostors Method ranked first in PAN'14 (Potha and Stamatatos, 2017; Clarke and Grant, 2024): ASGALF by Khonji and Iraqi (2014) modifies the score computation during random feature subsampling to measure the degree of similarity between input vectors rather than whether they are sufficiently similar. The method also incorporates additional features for document representation.

Juola (2015) introduced a modification of the Impostors Method in which random feature subsampling is replaced by the explicit selection of predefined feature sets. They also reformulated the decision criterion: the procedure was considered to select the candidate author if the author appeared within a specified range of top similarity-based ranks, rather than only at rank one. The authors further examined how the likelihood of a false rejection error varies as a function of two parameters: (1) the similarity-based rank up to which the candidate author is considered selected, and (2) the number of *assumedly independent* elimination steps performed by the method.

Hernández et al. (2015) frame the Impostors Method's random feature subsampling procedure, subsequently aggregated across iterations, as Homotopy-Based Classification (HBC). Texts are represented using lexical and character-level features and combined into a feature matrix. Using an L_1-homotopy sparse reconstruction algorithm, the disputed text is approximated as a linear combination of candidate and impostor texts. Reconstruction residuals are computed for candidate and impostor class, and authorship is attributed to the class yielding the lowest residual. The procedure is repeated with different randomly sampled feature subsets and impostors.

Kestemont et al. (2016) used the original proposed Impostors Method to detect the authorship of Caesar's texts. They compare different vector space models (i.e., TF, TF-IDF, TF-STD) and distance metrics (i.e., Manhattan, cosine, minmax distance) on PAN'14 dataset. Overall TF-IDF and TF-STD similarity performed best.

An extension of the original Impostors Method

Author	Year	Dataset(s)
Juola	2015	Poe short stories Newspaper articles *The Cuckoo's Calling* & J. K. Rowling novels Bitcoin documents
Stover et al.	2016	Text from the 2nd century
Camps and Salvati	2023	French poetry from the 16th century
Clarke and Grant	2024	Paroxysm of Chaos I Paroxysm of Chaos II (anarchist texts)
Volkovich and Avros	2025	Shakespeare's work from Project Gutenberg

Table 16: Summary of prior applications of the Impostors Method.

by Potha and Stamatatos (2017) selects impostors based on maximal min-max similarity to the suspect rather than randomly. In addition, score computation considers only the similarity between the disputed document and the fillers relative to the suspect, without swapping the roles of disputed document and suspect. Rather than simply checking whether the suspect is the most similar document, similarities are treated as a ranking, and the position of the candidate author's text within this ranking is taken into account.

Clarke and Grant (2024) build on the work of Potha and Stamatatos (2017) by extending the original Impostors Method with impostors that are similar to the suspect and by adopting the idea of incorporating ranking positions into the scoring procedure. As their study is a case study, impostors did not need to be artificially generated but were selected from texts within the same social community. Instead of previously used statistical distance metrics, they apply a ranking method based on the relative frequencies of shared n-grams.

Volkovich and Avros (2025) retain the idea of using impostors but propose a new framework. Texts are segmented and embedded, and a deep neural network is fine-tuned to distinguish segments from different impostors. Test segments are assigned to impostor classes, their similarity is measured using dynamic time warping, and finally an Isolation Forest is used to cluster and score the texts.

Case Studies The datasets used in the following case studies are summarized in Table 16.

The traditional Impostors Method was used by Stover et al. (2016) to attribute a second-century text to Apuleius, a North African writer in the Roman Empire (Table 16).

Camps and Salvati (2023) applied the traditional method, implemented in an R package, to a corpus of 16th-century French poetry.

E Statistical Tests

We assess whether differences in PAN metrics between impostor generation approaches are statistically significant (Table 17) irrespective of data split.

Significance Test. The choice of statistical tests was guided by prior literature examining the strengths and limitations of different approaches in comparable settings.

Both the McNemar test and the difference of proportions test rely on a single fixed train-test split and therefore fail to account for variability introduced by data partitioning. In addition, the difference of proportions test violates certain independence assumptions (Dietterich, 1998).

The resampled paired t-test addresses this limitation by performing multiple trials with randomly sampled train and test sets. However, this procedure itself violates independence assumptions (Dietterich, 1998) and tends to underestimate the variance. While extensions of this test have been proposed to mitigate the variance underestimation, the issue of dependence remains (Bouckaert and Frank, 2004).

The k-fold cross-validation paired t-test avoids overlap in test sets, but for $k > 2$ the training sets overlap substantially, again introducing dependencies. Dietterich (Dietterich, 1998) therefore recommends the 5×2-fold cross-validation paired t-test. In this setting, $k = 2$ minimizes overlap between training and test sets, while $r = 5$ repetitions provide a balance between reducing noise ($r < 5$) and limiting violations of independence assumptions ($r > 5$).

Bouckaert and Frank (Bouckaert and Frank, 2004) further examine the issue of replicability and report that the 5×2-fold test exhibits high sensitivity to specific data partitions, leading to low replicability. Instead, they find that the ten 10-fold cross-validation paired t-test offers a better trade-off, reducing both Type I and Type II errors while improving replicability (Dietterich, 1998).

ANOVA and MANOVA are statistical tests used to assess whether there are significant differences between the means of random samples, for one and multiple response variables (here: metric scores), respectively. However, if these tests indicate a significant effect, they do not identify which specific pairs of samples differ; therefore, additional post-hoc tests are required for pairwise comparisons. Both ANOVA and MANOVA assume that the sam-

Method	Retrieval-Based (ChatNoir)			Retrieval-Based (Startpage)			In-Domain			Two-Step Paraphraser (LLM)			Nagy, 2024			Gutierrez et al., 2015			Potha et al., 2017			Khonji & Iraqi, 2014			Kestemont et al., 2016		
	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1
Blog Posts Dataset																											
Retrieval-Based (ChatNoir)	–	–	–				***	***	***				***	***	***	***	***	**		***		***	***		***	***	***
Retrieval-Based (Startpage)				–	–	–	***	***					***	***		***				***		***	***		***	***	***
In-Domain							–	–	–	***	***		***	***		***	***	***	***	***	***	***	***	***	***	***	
Two-Step Paraphraser (LLM)										–	–	–	***	***			***						***		***	***	
Nagy, 2024													–	–	–	***	***	***	***	***	***		***	*	***		
Gutierrez et al., 2015																–	–	–		***	***	***	***	***	***	***	***
Potha et al., 2017																			–	–	–	***	***		***	***	***
Khonji & Iraqi, 2014																						–	–	–	***	***	***
Kestemont et al., 2016																									–	–	–
Student Essays Dataset																											
Retrieval-Based (ChatNoir)	–	–	–				***	***	***				***	***	***	***	***			***					***	***	***
Retrieval-Based (Startpage)				–	–	–	***	***					***	***		***				***					***	***	***
In-Domain							–	–	–	***	***					***	***	***	*	***		***	***	***	***	***	
Two-Step Paraphraser (LLM)										–	–	–	***	***			***						***		***	***	
Nagy, 2024													–	–	–	***	***	*		*			***	*	***		
Gutierrez et al., 2015																–	–	–		*		***	***				
Potha et al., 2017																			–	–	–	***	***				
Khonji & Iraqi, 2014																						–	–	–	***	***	
Kestemont et al., 2016																									–	–	–

Table 17: Corrected paired t-test based on ten 10-fold CV for the Blog Posts and Student Essays datasets. Asterisks indicate statistically significant differences ($*$, $**$, $***$ for $p_{\text{corr}} < 0.05, 0.01, 0.005$, respectively).

Method	Retrieval-Based (ChatNoir)			Retrieval-Based (Startpage)			In-Domain			Two-Step Paraphraser (LLM)			Nagy, 2024			Gutierrez et al., 2015			Potha et al., 2017			Khonji & Iraqi, 2014			Kestemont et al., 2016		
	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1	Recall	Precision	c@1
Blog Posts Dataset																											
Retrieval-Based (ChatNoir)	–	–	–	0.28	0.32	-0.49	-7.51	7.19	1.88	0.12	-0.42	-0.69	-5.42	5.08	1.49	1.25	-2.53	-1.43	0.26	-1.15	-0.72	-4.32	3.01	0.34	3.70	-5.54	-2.34
Retrieval-Based (Startpage)				–	–	–	-5.01	4.15	0.47	0.28	-0.83	-0.64	-4.37	2.94	0.27	1.66	-0.64	-0.48	-0.50	-0.80	-0.92	5.31	-3.81	0.17	2.87	-3.24	-1.34
In-Domain							–	–	–	-3.62	2.76	-0.23	1.66	-2.14	-0.57	2.92	-5.59	-4.19	-4.99	5.86	1.67	2.25	-3.49	-1.57	-3.41	2.23	-0.22
Two-Step Paraphraser (LLM)										–	–	–	-2.77	2.01	-0.30	1.18	-1.73	-1.02	0.47	0.25	0.10	3.08	-4.06	-0.68	-1.55	1.95	0.57
Nagy, 2024													–	–	–	-2.15	4.16	3.63	-4.92	3.77	1.46	0.42	-1.40	-1.15	-2.15	0.09	-0.74
Gutierrez et al., 2015																–	–	–	0.14	-1.87	-2.53	-2.07	3.19	2.23	1.46	-4.64	-4.18
Potha et al., 2017																			–	–	–	-3.51	1.84	-0.09	2.63	-3.64	-2.06
Khonji & Iraqi, 2014																						–	–	–	-1.51	-1.69	-1.87
Kestemont et al., 2016																									–	–	–
Student Essays Dataset																											
Retrieval-Based (ChatNoir)	–	–	–	1.69	-0.85	-0.60	-4.33	4.95	2.74	0.66	-0.40	-1.44	-3.61	4.63	2.42	-1.65	2.55	0.80	2.34	-3.30	-1.74	1.96	-1.16	1.23	2.66	-3.85	-1.82
Retrieval-Based (Startpage)				–	–	–	-3.49	4.68	2.39	-0.56	0.57	-1.03	-2.66	4.62	2.12	-0.39	2.08	0.50	1.47	-2.83	-1.47	3.62	-1.83	0.76	1.25	-3.00	-1.49
In-Domain							–	–	–	-3.20	4.79	1.50	0.64	-0.76	-0.39	2.63	-2.90	-1.51	-1.19	1.67	0.92	5.55	-5.31	-1.72	-1.89	2.03	0.76
Two-Step Paraphraser (LLM)										–	–	–	-2.54	4.89	1.26	-0.79	2.52	-0.07	-1.54	3.27	0.66	1.75	-1.51	-0.00	-1.51	3.50	0.74
Nagy, 2024													–	–	–	-1.93	2.30	1.15	-0.86	1.19	0.72	5.08	-5.27	-1.33	-1.39	1.43	0.41
Gutierrez et al., 2015																–	–	–	1.10	-1.20	-0.66	3.16	-3.47	0.08	0.61	-0.83	-0.69
Potha et al., 2017																			–	–	–	3.79	-3.70	-0.59	-0.45	0.28	-0.13
Khonji & Iraqi, 2014																						–	–	–	4.14	-4.41	-0.71
Kestemont et al., 2016																									–	–	–

Table 18: Pairwise effect sizes (Cohens d_z) for each metric. Effect sizes are reported for the same corrected paired tests used in the significance table.

ples are independent and normally distributed (Gurvich and Naumova, 2021). Since our data is generated using repeated k-fold cross-validation, the independence assumption is violated, making these tests unsuitable in this context.

Consequently, we adopt the ten 10-fold cross-validation paired t-test as the most suitable compromise for repeated cross-validation experiments, while acknowledging that all such resampling-based tests remain approximate rather than strictly valid statistical procedures (Dietterich, 1998).

Effect Size. We use Cohen's d_z to quantify effect size (Table 18), defined as the mean difference between paired scores normalized by their standard deviation (Lakens, 2013). This choice is motivated by the within-subject (paired) nature of our data, where observations are not independent but instead linked across conditions (e.g., identical data splits or instances evaluated under different methods). In such settings, the variability of interest is the variability of the within-pair differences, rather than the variability between independent groups.