

One-shot Style Transfer LLM log-probabilities for Authorship Attribution and Verification

Pablo Miralles-González¹, Javier Huertas-Tato¹, Alejandro Martín¹, David Camacho²

¹Department of Computer Systems, Technical University of Madrid, Madrid, Spain

²United Arab Emirates University, Al Ain, United Arab Emirates

Correspondence: pablo.miralles@upm.es

Abstract

Computational stylometry studies writing style through quantitative textual patterns, enabling applications such as authorship attribution, identity linking, and plagiarism detection. Despite the relevance of language modeling to these tasks, the pre-training of modern large language models (LLMs) has been underutilized in authorship attribution and verification. We introduce an unsupervised framework that uses the log-probabilities of an LLM to measure style transferability between two texts. This framework takes advantage of the extensive Causal Language Modeling (CLM) pre-training, one-shot capabilities and scale of LLMs, avoiding explicit supervision. Our methods substantially outperform prompting-based unsupervised baselines in authorship verification at similar model sizes, and is competitive with or improves contrastive baselines in most settings with sufficient model scale. We further observe strong performance across non-English languages. The effectiveness of the proposed framework improves consistently with increasing model scale. In the case of authorship verification, we propose an additional mechanism that increases test-time computation to improve accuracy; enabling flexible trade-offs between computational cost and task performance. Code is available at pablmiralles22/OSST.

1 Introduction

Computational stylometry is the quantitative and computational analysis of writing style, using patterns in vocabulary, syntax, punctuation, and character sequences to identify, verify, or cluster authors based on textual data (Daelemans, 2013). Authorship analysis has been widely applied across diverse domains, with applications spanning both forensic and humanities contexts. In forensics, stylistic and linguistic patterns have been used to analyze terrorist discourse (Abbasi and Hsinchun Chen, 2005), link user identities across online plat-

forms (Shu et al., 2017), detect plagiarism (Stamatatos and Koppel, 2011), and identify coordinated disinformation campaigns (Stiff and Johanson, 2022). Authorship methods also play a key role in the humanities, supporting tasks such as the attribution of pseudographic texts (Koppel and Seidman, 2013) and the analysis of anonymous literary works (Stover et al., 2016).

Authorship analysis is a long-standing challenge in NLP. Early approaches used hand-crafted stylistic features (Stamatatos, 2016; Lagutina et al., 2019). But the advent of pre-trained transformers (Vaswani et al., 2017; Devlin et al., 2019) led to the use of supervised encoder models (Fabien et al., 2020; Zangerle et al., 2023a). We also find work fine-tuning individual language models for candidate authors (Huang et al., 2025; Barlas and Stamatatos, 2020), measuring the probability assigned to texts by the language model. While this approach tries to take advantage of the pre-training of LLMs, it is expensive to apply, and not useful in low resource contexts like authorship verification tasks.

Recent work has increasingly focused on learning author embeddings (see, e.g., (Rivera-Soto et al., 2021; Sawatphol et al., 2022; Patel et al., 2025)), mapping texts to dense vector representations that encode stylistic features. These approaches also train on labeled data, but optimize contrastive objectives. These techniques yield representations that are interpretable and transferable to many authorship tasks, including author profiling.

In this work we focus on non-supervised approaches. The emergence of large language models (LLMs) has led to prompt-based unsupervised stylistic analysis (Hung et al., 2023; Huang et al., 2024), offering an alternative without supervision but facing poor performance (at least at moderate model sizes; see Table 3) and challenges in attribution due to context length constraints.

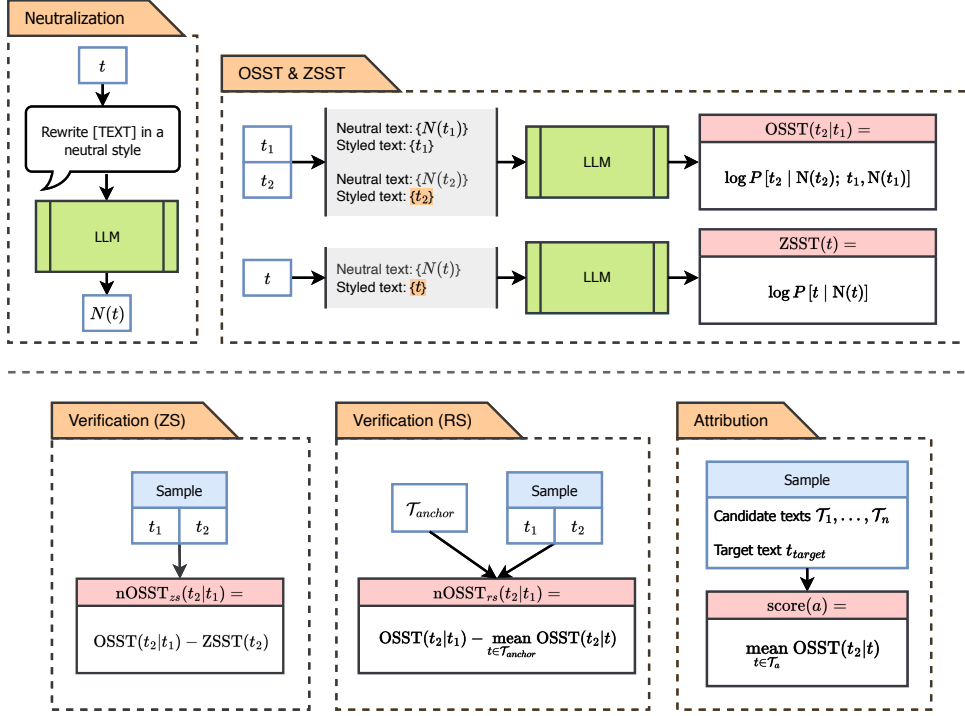


Figure 1: High-level overview of OSST-based methods for authorship attribution and verification. **Text highlighted in orange** means that we obtain the log probabilities of the corresponding tokens, and {bracketed text} means variable interpolation. The top row defines common operations, while the bottom row defines the application of those operations to authorship verification and attribution. Random sampling (RS) and Zero-shot (ZS) are two different variants of our method for authorship verification.

Proposal. We propose a novel unsupervised approach to authorship attribution and verification that leverage the extensive Causal Language Modeling (CLM) pre-training, one-shot learning capabilities and scale of LLMs. For a given text t , we obtain a paraphrased version $N(t)$ in neutral style via LLM prompting. Next, we consider the task of predicting the styled text t from its neutral version $N(t)$. LLMs can perform this task, especially with in-context examples. Given a single in-context example, the average of the LLM log-probabilities of the target text t (or *OSST score*) reflects how helpful the style of the in-context example was. Therefore, we use this metric to measure style similarity between the texts. The *OSST score* can be used in both authorship attribution and verification. We provide a more in-depth explanation in Figure 1 and Section 3.

Our results show OSST-based approaches consistently outperform comparably sized LLM prompting methods on authorship verification. With sufficiently large backbone models, they are also competitive with or surpass contrastive baselines across most of our experimental settings, and present good

results in authorship attribution in non-English languages. We further observe a scaling behavior: performance increases with model size and, for the authorship verification, with an additional mechanism that increases test-time compute.

Contributions. The main contributions of this work are the following.

1. We proposed a novel unsupervised approach to authorship attribution and verification taking advantage of the extensive CLM pre-training, one-shot learning capabilities and scale of LLMs.
2. We empirically validated its performance for authorship attribution and verification across several datasets and compared with contrastive and prompting-based baselines.
3. We studied the performance of our methods across multiple base LLM sizes, and performed ablations on another verification-specific hyperparameter that scales test-time compute.

4. We validated the multilingual performance of OSST-based methods in attribution.

2 Background

2.1 Authorship tasks

In this work, we focus on two core authorship tasks: verification and attribution. Other authorship tasks, such as style change detection, can be reduced to them, and our method is not suited for authorship profiling (identifying features of the person who wrote a text) or clustering.

Authorship verification (AV). Given a set of texts T_a known to be written by the author a , the task is to make a binary decision on whether a target text t was written by a .

Authorship attribution (AA). Given a target text t and a set of candidate authors $A = \{a_1, a_2, \dots, a_n\}$, the task is to determine the most likely author of t . To do this, a set of training texts $T = \{T_{a_1}, T_{a_2}, \dots, T_{a_n}\}$ is provided, where T_{a_i} is a set of texts known to be written by the author a_i . In *closed-set* AA the author of the text is always among the candidates. In *open-set* AA this might not be the case, and methods must detect it.

2.2 Available data

Authorship datasets are foundational for advancing research in computational authorship analysis. Diverse corpora exist across multiple domains, including news articles (Lewis et al., 2004), opinion pieces (Stamatatos, 2013), product reviews (Seroussi et al., 2014), blogs (Schler et al., 2006), literary fiction (Gerlach and Font-Clos, 2020), and social media platforms such as Reddit (Völske et al., 2017) and Twitter (Cheng et al., 2010). For evaluation, this work focuses on datasets that have been cleaned and standardized for AA and AV tasks.

A major driver of dataset quality and diversity in authorship analysis is the CLEF Initiative’s annual PAN competitions, which provide standardized benchmarks for digital text forensics. Early AA datasets (2011–2012) drew on corporate emails and fiction (Argamon and Juola, 2011; Juola, 2012). The 2018 AA task introduced a cross-fandom Fan-Fiction setting (Kestemont et al., 2018b,a), and the 2019 edition extended this to an open-set scenario (Kestemont et al., 2019b,a).

For AV, the 2020–2021 editions focused on Fan-Fiction, with 2021 adding out-of-distribution au-

thors and topics (Bischoff et al., 2020; Bevendorff et al., 2020). In 2022, the task expanded to multiple discourse types (e.g., essays, emails, text messages, business memos) (Stamatatos et al., 2022).

We also consider style-change detection tasks (2022–2024), which require determining whether adjacent paragraphs share authorship. The 2022 dataset is based on StackExchange posts (Zangerle et al., 2022), while the 2023 and 2024 editions use topic-controlled Reddit data to reduce semantic cues and emphasize stylistic signals (Zangerle et al., 2023b, 2024).

2.3 Content and style

An open question in authorship is whether the style and the content of a text can be disentangled. Topic information is known to be predictive in some authorship datasets (see e.g. (Sari et al., 2018)). Zangerle et al. (2023a) showed worse results in sub-tasks where topical correlation was accounted for. Some work in the literature (e.g. (Wegmann et al., 2022; Sawatphol et al., 2022; Patel et al., 2025)) attempts to learn representations that are more independent of content. However, it is not clear that this can be done. Mikros and Argiri (2007) studied several stylometric features that were topic-neutral in theory, and found many of them to be correlated with text topic. Jafaritazehjani et al. (2020) found transferring stylistic features to a text while preserving content difficult, concluding content and style might be entangled.

2.4 Authorship methods

Traditional authorship analysis methods relied on handcrafted lexical (e.g., character and word n-grams, word frequency, word-length distribution), syntactic (e.g., POS distributions, syntactic n-grams), and semantic features (Stamatatos, 2016; Lagutina et al., 2019). These features were typically used in supervised settings with classifiers such as SVMs (Stamatatos, 2016), or with early neural architectures including RNNs (Bagnall, 2016) and CNNs (Shrestha et al., 2017).

With the introduction of transformers (Vaswani et al., 2017), pre-trained encoder-only models such as BERT (Devlin et al., 2019) have been widely adopted for authorship-related tasks (Fabien et al., 2020; Zangerle et al., 2023a).

An alternative line of work fine-tunes an individual language model for each candidate author using a corpus of known texts, and attributes authorship by evaluating perplexity under these author-specific

models (Huang et al., 2025; Barlas and Stamatatos, 2020). However, the need to fine-tune a language model for each author makes these methods expensive and not applicable in low-resource environments like authorship verification tasks.

There is a growing interest in learning author embeddings, mapping each text to a vector representation capturing stylistic properties. These methods still learn from supervised data, but rely on contrastive losses. They have the advantage of being more interpretable and simple to apply in all attribution tasks, including author profiling. Rivera-Soto et al. (2021) trained a convolutional model with contrastive learning. Wegmann et al. (2022) studied the use of the triplet contrastive loss with a hard negative from the same conversation to obtain embeddings that are more independent from content. Sawatphol et al. (2022) used topic similarity (via a semantic similarity model) to modify the loss function and regularize training, reducing reliance on topical cues. Huertas-Tato et al. (2024) scaled training by efficiently implementing the use of large batches. Kim et al. (2025) try to improve multilingual performance by using Language-Aware Batching (LAB), where examples in a batch are from the same language. They also used Probabilistic Content Masking (PCM) to randomly mask content words (words carrying meaning). Patel et al. (2025) tried to reduce content-dependence by creating a synthetic dataset with LLMs. They create near-exact paraphrases with controlled style variations for 40 known style features. Agarwal et al. (2025) focused on cross-genre authorship attribution, trying to avoid topical cues. They implement hard negative mining in their contrastive pipeline, and curate a dataset to select topically dissimilar positive pairs.

A recent trend explores the use of large language models (LLMs) for human-like stylistic analysis based on prompts (Hung et al., 2023; Huang et al., 2024). They are promising unsupervised methods, but they require very large models in AA settings due to context limitations, and empirical evaluation remains scarce. Our results indicate that they yield limited performance in AV.

2.5 Causal Language Modeling and in-context learning

Causal Language Modeling (CLM) refers to predicting the next token based on prior tokens in a unidirectional, autoregressive manner, ie, modeling $P(x_t \mid x_{<t})$ (Bengio et al., 2003). It is the

pre-training objective used by GPT-style decoder-only models (Radford et al., 2019). A landmark discovery in large language models (LLMs) is their in-context learning ability (Wei et al., 2022). GPT 3 (175B parameters) was shown to perform new tasks with just a few textual examples provided in the prompt, without any explicit gradient updates or fine-tuning, demonstrating competitive performance with fully fine-tuned systems (Brown et al., 2020).

3 One-shot Style Transfer (OSST) LLM log-probabilities for authorship attribution and verification

In this section we provide an in-depth explanation of the main components and intuitions behind our OSST-based methods for AA and AV. A high-level overview of all the components is given in Figure 1.

3.1 General components

Paraphrasing texts. Our methods begins by paraphrasing a given text in a neutral style. To do this, we prompt an instruction-tuned LLM using the template shown in Appendix B. More formally, given a text t , we consider the random variable $N_{M,\theta}(t)$ of generating a neutral version with LLM M and sampling parameters θ . For the rest of the work, we omit the subscript and abuse the notation to signify a sample of the random variable.

The neutralization process merits discussion, since the existence of such an style is debatable and there is not a well defined process to obtain it. *We note that we are not arguing for the existence of a neutral style or a neutralization process, only for the empirical usefulness of this line of prompting.* The intuition behind the design is to transform the text into a consistent and different anchor style while keeping as much semantic information as possible. This way, scores are comparable between different texts. We ablate other lines of prompting in Appendix C, comparing with simply asking the model to paraphrase or to rewrite in the style of Shakespeare. We find that the difference in performance between neutralizing and paraphrasing is not significant, while the Shakespeare style lags behind. It seems that the prompt choice is not irrelevant, but other sensitive choices may also be acceptable.

One-shot style transfer LLM log-probabilities. Our method relies on the one-shot capabilities of LLMs to transfer the style of one text to another.

In particular, for a pair of texts t_1 and t_2 , consider the following input:

Neutral text: $\{N(t_1)\}$
 Styled text: $\{t_1\}$

Neutral text: $\{N(t_2)\}$
 Styled text: $\{t_2\}$

We frame the problem as a style transfer task from the neutral version to the styled version of the second text, based on the style of the one-shot example. In this way, the LLM can predict semantic information from the neutral version while obtaining stylistic information from the one-shot example. To measure the degree of transferability from the style of the first text to the second text, we retrieve the log-probabilities from a given *foundational LLM*:

$$\text{OSST}(t_2|t_1) = \log P[t_2 | N(t_2); t_1, N(t_1)]. \quad (1)$$

In other words, the model is measuring if the second text could come from transferring the style of the first text to the neutral version of the second text.

3.2 Variability and normalization of the OSST score

We observe that the OSST score defined above may have a varying distribution depending on the target text t_2 (its length, topic, genre...). For example, initial tokens are harder to predict because there is limited preceding context within the styled target text (only the neutral version and one-shot example are available). Thus, shorter target texts will naturally have a higher OSST score. Therefore, we want to *normalize* these scores if we are to compare scores obtained for different target texts.

3.3 Authorship attribution

Closed-set AA. Given a target text t_{target} , a set of candidate authors $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$, and a set of training texts $\mathcal{T} = \{T_{a_1}, T_{a_2}, \dots, T_{a_n}\}$, our method assigns a score to each candidate author as

$$\text{score}(a) = \text{mean}_{t \in T_a} \text{OSST}(t_{\text{target}}|t), \quad (2)$$

and the prediction is the candidate with the highest score. Since we only compare scores for the given target text, we do not need any form of normalization in this setting as per Section 3.2.

Open-set AA. The open-set attribution setting includes the possibility that the target text is not written by any of the candidate authors. Unlike in the closed-set setting, we must select a decision boundary such that *if the highest score of the candidates is below this threshold, then we predict that no candidate is the author*. This boundary must be selected globally for all inputs, and this requires comparing OSST scores for different target texts. Therefore, we want to apply some form of normalization as per Section 3.2.

The OSST score is computed for the same target text across several candidate texts, providing a simple way to *standardize the scores* without additional computations:

$$\begin{cases} \mu(t_{\text{target}}) = \text{mean}_{a \in \mathcal{A}, t \in T_a} \text{OSST}(t_{\text{target}}|t), \\ \sigma(t_{\text{target}}) = \text{std}_{a \in \mathcal{A}, t \in T_a} \text{OSST}(t_{\text{target}}|t). \end{cases} \quad (3)$$

This alleviates the natural discrepancies between target texts and makes the scores more comparable.

3.4 Authorship verification

We consider the task deciding whether two texts t_1 and t_2 were authored by the same person. Since the task is symmetric across texts, we can consider the score

$$\text{score}(t_1, t_2) = \frac{1}{2} (\text{OSST}(t_1|t_2) + \text{OSST}(t_2|t_1)). \quad (4)$$

AV presents the same difficulty as open-set AA. We have to define a global decision boundary such that a pair is predicted to be by the same author if their OSST score exceeds this threshold. *This requires some form of normalization* as per Section 3.2. But unlike open-set AA, we only have one OSST score for each target text. Thus, we have to produce other OSST (or OSST-like) scores for the same target text to normalize. We present two options to solve this problem.

Zero-shot baseline (nOSST-ZS). We *normalize* the OSST score by comparing it with the log-probabilities of the zero-shot problem, without any in-context examples. This is shown graphically in Figure 1, and mathematically as follows:

$$\text{nOSST}_{zs}(t_2|t_1) = \text{OSST}(t_2|t_1) - \log P[t_2 | N(t_2)] \quad (5)$$

This normalization technique has some problems. Figure 2 shows the median nOSST_{zs} score (without averaging across tokens of the target text) by token position in the PAN23 style change subsets. As we can see, the first few tokens have a much higher median value, and it progressively stabilizes. This still presents a challenge in comparing texts of different lengths, because shorter texts will have a higher overall score.

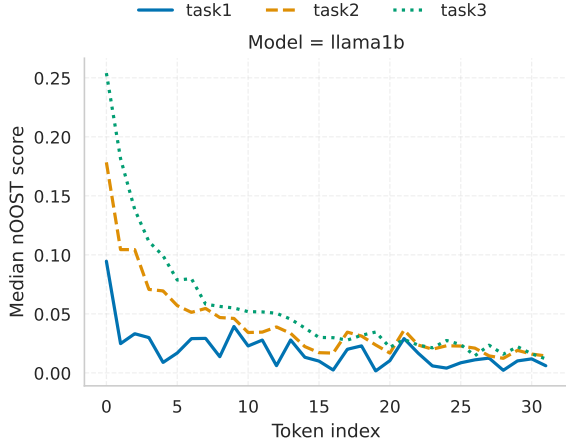


Figure 2: Median nOSST_{zs} score (without averaging across tokens of the target text) by token position in the PAN23 test sets.

Random sampling baseline (nOSST-RS). To mitigate the previous problems, we can select a fixed set of *anchor texts* $\mathcal{T}_{anchor}$, and compute the normalized score as

$$\text{nOSST}_{rs}(t_2|t_1) = \text{OSST}(t_2|t_1) - \frac{\text{mean}_{t \in \mathcal{T}_{anchor}} \text{OSST}(t_2|t)}{|\mathcal{T}_{anchor}|}. \quad (6)$$

In this case, the difficulty lies in sampling appropriate anchor texts. In this work, we sample in-distribution from the corresponding training set. This is a limitation, as it is not feasible in real-world applications. We may be able to use contrastive models to sample in-distribution anchors, but this is not studied here.

4 Experimental results

We conducted several experiments with the purpose of evaluating and studying OSST-based methods in attribution and verification. Since OSST-based methods are unsupervised, we focus on baselines that can be applied without supervised training, such as contrastive and prompt-based methods.

4.1 Authorship attribution

We evaluated our method in the fiction domain with the PAN12 (Juola, 2012), PAN18 (Kestemont et al., 2018b), and PAN19 (Kestemont et al., 2019b) attribution datasets, and in the corporate email domain with PAN11 (Argamon and Juola, 2011). As contrastive learning baselines we use STAR (Huertas-Tato et al., 2024), LUAR (Rivera-Soto et al., 2021), StyleEmbedding (Wegmann et al., 2022), StyleDistance (Patel et al., 2025) and the multilingual models from (Kim et al., 2025) (with XLM-RoBERTa and Llama3.2-1B backbones). We do not compare against LLM prompting approaches such as LIP (Huang et al., 2024), as these methods require large context windows in AA, resulting in markedly lower performance at moderate model sizes. Finally, OSST experiments are run three times.

Table 1 presents the results in closed-set attribution. In the fiction domain, our method with 4B and 8B parameters outperforms the baselines across all datasets, with the smaller Llama1B backbone performing competitively with all baselines except STAR. There is an increase in performance with larger model sizes, suggesting that further gains may be achievable with continued scaling. In the PAN11 dataset, all OSST models achieve higher accuracy but lower rank than some of the baselines. Interestingly, model performance does not improve with model size in PAN11. A likely explanation is its noisier nature compared to other datasets, as it contains erroneous messages, irregular candidate set sizes, and other inconsistencies.

Table 2 summarizes results in the open-set setting on the PAN19 dataset. A small held-out subset is used to select the decision threshold for whether the true author is among the candidates. Scores are standardized across candidate texts (for all methods) to improve comparability of scores for different target texts. Our method exhibits better performance in the open-set scenario than contrastive baselines with all parameter counts.

4.2 Authorship verification

For authorship verification (AV), we evaluate on the 2020 and 2021 editions of the PAN AV task (Bischoff et al., 2020; Bevendorff et al., 2020), which focus on FanFiction, and in the PAN style-change detection tasks of 2022–2024 (Zangerle et al., 2022, 2023b, 2024). The 2022 edition contains StackExchange posts, while the 2023 and 2024 editions comprise Reddit posts and explic-

Table 1: Closed-set authorship attribution results. Accuracy and average rank of correct candidate are reported.

	PAN11		PAN12		PAN18-en		PAN19-en	
	Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank
OSST-Gemma1B	23.71% $\pm 0.23\%$	0.82 ± 0.00	33.33% $\pm 1.03\%$	0.70 ± 0.03	46.41% $\pm 1.66\%$	0.84 ± 0.01	69.96% $\pm 1.24\%$	0.90 ± 0.00
OSST-Llama1B	22.64% $\pm 0.20\%$	0.83 ± 0.00	44.05% $\pm 8.25\%$	0.76 ± 0.04	52.15% $\pm 2.09\%$	0.86 ± 0.02	72.40% $\pm 0.88\%$	0.91 ± 0.01
OSST-Gemma4B	23.71% $\pm 0.49\%$	0.83 ± 0.00	60.71% $\pm 3.09\%$	0.87 ± 0.02	63.32% $\pm 1.54\%$	0.93 ± 0.01	84.49% $\pm 0.24\%$	0.96 ± 0.00
OSST-Llama8B	21.45% $\pm 0.29\%$	0.83 ± 0.00	71.43% $\pm 3.09\%$	0.94 ± 0.01	70.49% $\pm 1.10\%$	0.95 ± 0.00	89.00% $\pm 0.71\%$	0.97 ± 0.00
STAR	17.83%	0.85	57.14%	0.87	62.68%	0.93	76.07%	0.94
LUAR _{CRUD}	16.21%	0.84	39.29%	0.76	47.85%	0.88	69.55%	0.91
LUAR _{MUD}	17.99%	0.85	42.86%	0.78	54.55%	0.88	62.02%	0.91
StyEmb	6.69%	0.74	26.79%	0.72	54.55%	0.88	58.04%	0.86
StyDist	10.47%	0.74	37.50%	0.74	48.33%	0.87	60.08%	0.87
MLangRob	14.32%	0.82	46.43%	0.79	49.76%	0.88	45.01%	0.81
MLangLlama	19.83%	0.81	39.29%	0.79	47.37%	0.84	41.65%	0.82

Table 2: Open-set authorship attribution evaluation in PAN19. BinF1 refers to the binary classification of true author presence among candidates, while the rank of the correct candidate is calculated only for instances where the true author is included. All similarity scores are standardized across candidate texts.

	Accuracy	BinF1Macro	Mean rank
OSST-Gemma1B	63.10% $\pm 0.58\%$	64.59% $\pm 1.60\%$	0.90 ± 0.00
OSST-Llama1B	64.82% $\pm 0.08\%$	67.05% $\pm 0.14\%$	0.91 ± 0.01
OSST-Gemma4B	74.28% $\pm 0.67\%$	74.91% $\pm 0.72\%$	0.96 ± 0.00
OSST-Llama8B	76.13% $\pm 1.96\%$	75.46% $\pm 2.40\%$	0.97 ± 0.00
STAR	59.39%	61.36%	0.93
LUAR _{CRUD}	60.63%	65.52%	0.92
LUAR _{MUD}	53.13%	62.07%	0.91
StyEmb	53.06%	58.39%	0.86
StyDist	51.75%	58.98%	0.87
MLangRob	31.95%	45.03%	0.81
MLangLlama	42.21%	56.76%	0.82

itly control for topic in the second and third tasks, thereby preventing detectors from exploiting semantic information.

We use the same constrastive baselines as in AA. We include LIP (Huang et al., 2024) and PromptAV (Hung et al., 2023) with a 24B Mistral LLM as unsupervised baselines. For contrastive and OSST-based methods, we select a decision boundary with 2000 examples from the training set. For nOSST methods, we work with the LLaMA base models and a 24B Mistral model to test larger-capacity base models. *In the nOSST-RS setting, we use 10 anchors sampled in-distribution from the training*

set.

The results on authorship verification datasets are summarized in Table 3. We make several observations: (i) In nOSST methods, larger backbone models tend to achieve higher accuracy. (ii) nOSST-ZS with the larger backbone models performs competitively with contrastive baselines on average and in the harder PAN23SC-T3 and PAN24SC-T3 tasks where there is less topical correlation. (iii) nOSST-RS generally improves performance with respect to nOSST-ZS across multiple datasets, though a few exceptions remain. (iv) nOSST-RS with a Mistral24B backbone is the best overall, and it pushes performance on the more difficult sets. (v) At comparable model scales (up to 24B parameters), nOSST methods outperform LLM prompting approaches.

4.3 Effect of the number of anchor samples in nOSST-RS

Figure 3 presents the mean F1 classification score on all verification datasets as a function of the number of anchor examples provided to nOSST-RS. The point at zero corresponds to nOSST-ZS, and, for computational efficiency, we re-sample from a fixed pool of 10 examples. We observe that introducing a single anchor example yields a big improvement over nOSST-ZS with minimal additional cost. While F1 gains exhibit a pattern of diminishing returns up to a plateau, we observe that the Mistral24B base model shows an upward final trajectory. This may provide a guideline to choose the number of anchor examples for nOSST-RS in practice.

Table 3: Results on authorship verification tasks. F1 score is reported. Best result per dataset is marked in bold.

	22ST			23ST			24ST			20AV	21AV	Avg
	T1	T2	T3	T1	T2	T3	T1	T2	T3			
nOSST ^{1B} _{ZS}	62.26%	62.57%	58.30%	69.01%	55.82%	58.97%	73.21%	67.13%	61.27%	67.65%	67.34%	63.96%
nOSST ^{8B} _{ZS}	64.57%	66.87%	59.99%	68.63%	61.54%	62.93%	75.42%	69.28%	63.36%	76.22%	76.50%	67.75%
nOSST ^{24B} _{ZS}	66.38%	69.55%	62.02%	71.49%	63.06%	64.54%	76.56%	75.84%	65.73%	79.18%	79.30%	70.33%
nOSST ^{1B} _{RS}	63.55%	68.08%	61.32%	73.79%	61.38%	61.93%	81.03%	61.99%	63.02%	73.22%	72.37%	67.43%
nOSST ^{8B} _{RS}	65.16%	71.46%	62.23%	76.23%	70.25%	66.67%	83.07%	68.58%	66.17%	80.84%	80.01%	71.88%
nOSST ^{24B} _{RS}	67.73%	74.99%	64.10%	77.76%	69.83%	68.30%	84.42%	80.00%	68.66%	84.56%	83.50%	74.89%
STAR	64.62%	66.85%	58.06%	79.46%	72.86%	59.77%	79.79%	76.87%	57.97%	76.57%	79.22%	70.19%
LUAR _{CRUD}	66.78%	70.67%	62.41%	80.76%	73.66%	62.25%	78.32%	76.36%	60.10%	68.74%	69.98%	70.00%
LUAR _{MUD}	66.66%	69.85%	62.21%	79.94%	72.13%	61.14%	77.66%	67.23%	60.03%	72.82%	74.86%	69.50%
StyEmb	63.42%	66.12%	58.78%	69.28%	69.91%	63.90%	70.17%	71.03%	64.95%	66.09%	67.98%	66.51%
StyDist	59.91%	63.77%	55.95%	64.85%	65.33%	61.96%	69.00%	70.81%	63.44%	64.02%	67.20%	64.20%
MLangRob	61.08%	65.21%	58.30%	80.50%	69.18%	59.21%	83.52%	71.37%	57.37%	72.65%	74.34%	68.43%
MLangLlama	55.75%	59.49%	54.42%	81.38%	64.66%	52.04%	81.66%	69.68%	49.76%	65.31%	68.55%	63.88%
PromptAV	56.33%	50.35%	54.16%	67.19%	62.43%	54.23%	61.56%	62.02%	53.61%	54.29%	54.58%	57.34%
LIP	46.23%	27.30%	31.24%	11.76%	36.30%	34.43%	9.25%	30.22%	35.22%	35.24%	33.33%	30.05%

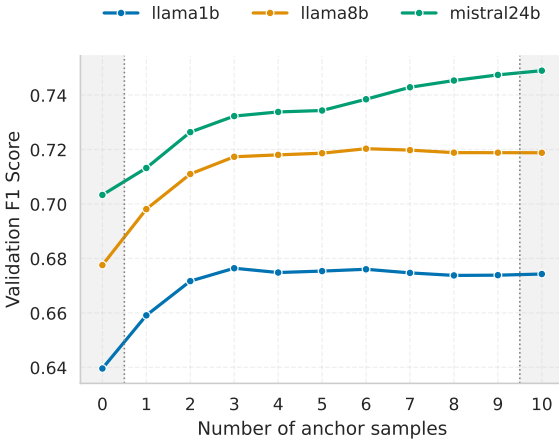


Figure 3: Average classification F1 score across all verification datasets by number of anchor examples in nOSST-RS. At 0 we show the performance of nOSST-ZS, and at 10 a single set of anchor samples was used.

4.4 AA multi-language performance

In Table 4 we present results across various languages for AA, using the PAN18 and PAN19 multi-language AA datasets. The same contrastive models as in section 4.1 are included as baselines, and the results of OSST methods are averaged across three runs. As anticipated, performance declines for all methods outside of English. OSST-based methods with the larger backbones consistently outperform contrastive baselines in non-English settings, even for baselines trained on multiple

languages. These results show OSST’s multilingual capabilities in AA, inherited from multilingual CLM pre-training, without requiring labeled authorship data in other languages.

5 Advantages and disadvantages of OSST-based and contrastive methods

OSST-based methods present several advantages: (i) They are fully unsupervised with respect to authorship data, and can directly leverage any large open-source language model without fine-tuning. (ii) In our limited experiments (verification datasets, up to 24B backbone models) they outperform prompting-based unsupervised baselines at similar model scales. (iii) With larger backbones they are able to compete with or outperform contrastive models in most of our experimental settings. (iv) They generalize across languages without language-specific authorship labeled data.

However, contrastive methods also have advantages. They are much faster in inference. Since they produce embeddings, they allow for vector retrieval, are easily applied to other authorship tasks (e.g. profiling and clustering) and are more interpretable. Finally, they do not have the limitation of requiring anchor samples as nOSST-RS. In fact, they might provide a solution to retrieve appropriate anchor samples, but we do not study this.

Table 4: Accuracy and rank of correct candidate by language and method in the PAN18 and PAN19 AA datasets.

	English		French		Italian		Polish		Spanish	
	Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank
OSST-Gemma1B	58.19%	0.87	36.26%	0.76	30.69%	0.69	25.00%	0.66	42.74%	0.80
OSST-Llama1B	62.28%	0.88	37.43%	0.76	41.43%	0.81	30.22%	0.73	53.31%	0.85
OSST-Gemma4B	73.90%	0.95	54.46%	0.87	58.26%	0.87	40.56%	0.80	71.27%	0.93
OSST-Llama8B	79.75%	0.96	55.23%	0.86	61.66%	0.89	51.89%	0.88	74.39%	0.94
STAR	69.37%	0.93	38.59%	0.77	24.59%	0.71	42.33%	0.81	35.19%	0.80
LUAR _{CRUD}	58.70%	0.90	26.39%	0.71	23.56%	0.68	29.67%	0.73	21.88%	0.65
LUAR _{MUD}	58.28%	0.89	32.32%	0.76	27.04%	0.69	26.67%	0.74	26.99%	0.75
StyEmb	56.30%	0.87	29.98%	0.76	28.24%	0.70	26.67%	0.77	25.12%	0.73
StyDist	54.20%	0.87	41.42%	0.79	34.73%	0.75	38.33%	0.77	35.27%	0.79
MLangRob	47.39%	0.84	37.57%	0.82	30.33%	0.75	23.00%	0.78	28.03%	0.78
MLangLlama	44.51%	0.83	28.46%	0.73	19.10%	0.66	19.00%	0.79	23.95%	0.71

6 Conclusions

We introduced OSST-based methods, a family of unsupervised approaches to authorship attribution and verification that take advantage of the scale and extensive CLM pre-training of LLMs.

In our experiments, OSST-based methods outperform LLM prompting alternatives of comparable scale in authorship verification. With sufficiently large backbone models, they are competitive with or surpass contrastive baselines in most of our experiments, and also present good results in authorship attribution in non-English languages. Our findings further suggest a scaling behavior: performance improves with model size and, for the verification random sampling variant (nOSST-RS), with the number of anchor examples, up to a saturation point. This behavior allows a practical trade-off between computational resources and stylistic accuracy.

Limitations

We find the following limitations in our work. (i) Model sizes were limited by hardware requirements. While small models are interesting in resource-constrained scenarios, further work needs to be conducted with larger models to properly observe scaling laws and push performance. We also restricted ourselves to non-commercial models due to economical constraints. Experiments with commercial models could be of interest due to the current gap with open-source ones. (ii) Empirical comparisons are limited by lack of standardization and maturity of datasets and pre-processing in the

field. (iii) Contrastive baselines are trained for authorship with labeled data, and there may be intersection with the evaluation domains/distributions (perhaps even data), which may affect evaluation results. (iv) We do not have larger contrastive models to compare performance at similar scales, although this is part of the advantage of unsupervised methods like OSST-based ones. (v) The random sampling (RS) variant for verification is studied with in-distribution anchors from the training set, which limits real-world applicability. Future research could study performance using a surrogate model to select anchors, such as already existing contrastive models.

Acknowledgments

This work has been supported by the following projects: H2020 TMA-MSCA-DN TUAI project "Towards an Understanding of Artificial Intelligence via a transparent, open and explainable perspective" (HORIZON-MSCA-2023-DN-01-01, Grant agreement n°: 101168344); by project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program; by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI; and by Comunidad Autonoma de Madrid, CIRMA-CAM Project (TEC-2024/COM-404).

References

- A. Abbasi and Hsinchun Chen. 2005. [Applying Authorship Analysis to Extremist-Group Web Forum Messages](#). *IEEE Intelligent Systems*, 20(5):67–75.
- Shantanu Agarwal, Joel Barry, Steven Fincke, and

- Scott Miller. 2025. [Cross-Genre Authorship Attribution via LLM-Based Retrieve-and-Rerank](#). *Preprint*, arXiv:2510.16819.
- Shlomo Argamon and Patrick Juola. 2011. [PAN11 author identification: Attribution](#).
- Douglas Bagnall. 2016. [Author Identification using Multi-headed Recurrent Neural Networks](#). *Preprint*, arXiv:1506.04891.
- Georgios Barlas and Efstathios Stamatatos. 2020. [Cross-Domain Authorship Attribution Using Pre-trained Language Models](#). In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, volume 583, pages 255–266. Springer International Publishing, Cham.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Janek Bevendorff, Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Martin Potthast, and Benno Stein. 2020. [PAN20 Authorship Analysis: Authorship Verification](#).
- 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](#). *Preprint*, arXiv:2005.14714.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. [You are where you tweet: A content-based approach to geo-locating twitter users](#). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768, Toronto ON Canada. ACM.
- Walter Daelemans. 2013. [Explanation in Computational Stylometry](#). In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, and Alexander Gelbukh, editors, *Computational Linguistics and Intelligent Text Processing*, volume 7817, pages 451–462. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for Authorship Attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLP AI).
- Martin Gerlach and Francesc Font-Clos. 2020. [A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics](#). *Entropy*, 22(1):126.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Baixiang Huang, Canyu Chen, and Kai Shu. 2024. [Can Large Language Models Identify Authorship?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.
- Weihang Huang, Akira Murakami, and Jack Grieve. 2025. [Attributing authorship via the perplexity of authorial language models](#). *PLOS One*, 20(7):e0327081.
- Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2024. [Understanding writing style in social media with a supervised contrastively pre-trained transformer](#). *Knowledge-Based Systems*, 296:111867.
- Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. [Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084, Singapore. Association for Computational Linguistics.
- Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John Kelleher. 2020. [Style versus Content: A distinction without a \(learnable\) difference?](#) In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2169–2180, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Patrick Juola. 2012. [PAN12 author identification: Attribution](#).

- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019a. Overview of the cross-domain authorship attribution task at {PAN} 2019. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, pages 1–15.
- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019b. [PAN19 authorship analysis: Cross-domain authorship attribution](#).
- Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018a. Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [Edit.]; et Al.*, pages 1–25.
- Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018b. [PAN18 author identification: Attribution](#).
- Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. [Leveraging Multilingual Training for Authorship Representation: Enhancing Generalization across Languages and Domains](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34855–34880, Suzhou, China. Association for Computational Linguistics.
- Moshe Koppel and Shachar Seidman. 2013. Automatically identifying pseudonymous texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1449–1454, Seattle, Washington, USA. Association for Computational Linguistics.
- Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P.G. Demidov. 2019. [A Survey on Stylometric Text Features](#). In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195, Helsinki, Finland. IEEE.
- David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.
- George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*, volume 276, pages 29–35. CEUR Workshop Proceedings.
- Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. [StyleDistance: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. [Learning Universal Authorship Representations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? Exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jitkapat Sawatphol, Nonthakit Chaiwong, Can Udomcharoenchaikit, and Sarana Nutanong. 2022. [Topic-Regularized Authorship Representation Learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1076–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.
- Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. [Authorship Attribution with Topic Models](#). *Computational Linguistics*, 40(2):269–310.
- Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Tamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.
- Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. [User Identity Linkage across Online Social Networks: A Review](#). *ACM SIGKDD Explorations Newsletter*, 18(2):5–17.
- Efstathios Stamatatos. 2013. On the Robustness of Authorship Attribution Based on Character N-gram Features. *Journal of Law and Policy*, 21(2).
- Efstathios Stamatatos. 2016. [Authorship Verification: A Review of Recent Advances](#). *Research in Computing Science*, 123(1):9–25.

- Efstathios Stamatatos and Moshe Koppel. 2011. [Plagiarism and authorship analysis: Introduction to the special issue](#). *Language Resources and Evaluation*, 45(1):1–4.
- Efstathios Stamatatos, Krzysztof Kredens, Piotr Pezik, Annina Heini, Mike Kestemont, Janek Bevendorff, Martin Potthast, and Benno Stein. 2022. [PAN22 Authorship Analysis: Authorship Verification](#).
- Harald Stiff and Fredrik Johansson. 2022. [Detecting computer-generated disinformation](#). *International Journal of Data Science and Analytics*, 13(4):363–383.
- 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](#). *Journal of the Association for Information Science and Technology*, 67(1):239–242.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#). Preprint, arXiv:2503.19786.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. [TL;DR: Mining Reddit to Learn Automatic Summarization](#). In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same Author or Just Same Topic? Towards Content-Independent Style Representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2023a. Overview of the multi-author writing style analysis task at PAN 2023. In *CLEF 2023 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR Workshop Proceedings.
- Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2023b. [PAN23 Multi-Author Writing Style Analysis](#).
- Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2024. [PAN24 Multi-Author Writing Style Analysis](#).
- Eva Zangerle, Maximilian Mayerl, Michael Tschuggnall, Martin Potthast, and Benno Stein. 2022. [PAN22 Authorship Analysis: Style Change Detection](#).

A Decision boundary under distribution shifts in nOSST

Figure 4 reports cross-dataset performance of nOSST scores for authorship verification, where the decision threshold is tuned on one dataset and evaluated on another. We consider three domains: FanFiction, StackExchange, and Reddit. As expected, thresholds selected in-distribution yield the best results. We observe the performance degradation is not too large in most pairs of datasets. However, this is not the case with the PAN20 dataset. This great difference could be partially explained by the great difference in text lengths (see appendix Figure 5), but not fully, as training in PAN22’s first task does generalize to PAN20 in the RS variant.

B Model and prompt to obtain a neutrally styled version of a text

We use a google/gemma-3-12b-it (Team et al., 2025) model to obtain neutral versions of a given text. This decision was made for its multilingual capabilities and the inclusion of multilingual texts in some of the datasets used.

The prompt we used is the following:

```
Text: [TEXT]
---
Paraphrase the previous text using the driest, simplest,
and most neutral style possible.
Remove all stylistic features, emotion, or rhetorical
devices. Replace as many words as possible with
simpler or more common alternatives.
Preserve as much of the original meaning as possible. Write
in [LANGUAGE]. Output only the rewritten version.
```

Listing 1: Prompt to obtain a neutrally styled version of a text.

C Ablation of the paraphrasing prompt

In this section we investigate the impact of our particular line of paraphrasing prompting in the performance of OSST-based methods. We compare against two baselines:

```
Text: [TEXT]
---
```

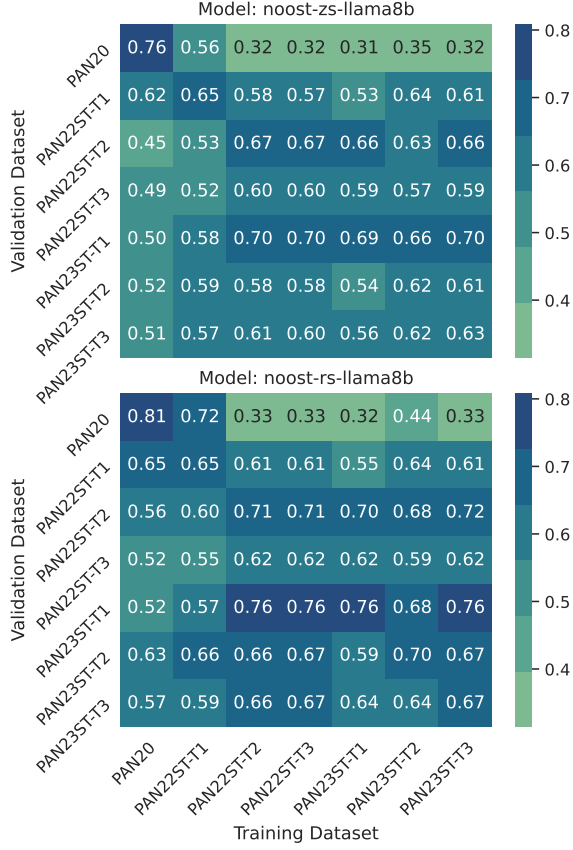


Figure 4: Cross-dataset macro-F1 of nOSST scores for authorship verification with a LLaMA-8B base model. The training set is used to select the decision boundary.

```
Paraphrase the previous text in a natural way, preserving
the original meaning but changing the wording and
structure. [LANGUAGE]. Output only the rewritten
version.
```

Listing 2: Version of the paraphrasing prompt asking to paraphrase the text.

```
Text: [TEXT]
---
Paraphrase the previous text in the style of Shakespeare,
using archaic language and poetic devices. Preserve
the original meaning but rewrite it in a way that
evokes the tone and style of Shakespeare's works.
Write in [LANGUAGE]. Output only the rewritten version.
```

Listing 3: Version of the paraphrasing prompt asking to rephrase the text in the in the style of Shakespeare.

In Table 5 we show the results of our ablation for some AA datasets. Results are reported across 3 seeds. The neutralizing prompt seems to be the best performing one generally.

In Table 6 we show some results in AV datasets. In this case the neutral and paraphrasing prompts are very competitive on average, with superiority varying between datasets and models. The Shakespeare strategy seems to lag slightly behind.

D Measuring semantic similarity with different paraphrasing prompts

In Tables 7 and 8 we show semantic similarity statistics in AA and AV sets for paraphrases with three different prompt versions: neutralizing, paraphrasing and rephrasing in the style of Shakespeare. This is measured with a sentence-transformers/all-mpnet-base-v2 model for semantic similarity.

E Datasets

In this section we list and briefly describe the datasets that we used in this paper for evaluation.

Authorship attribution datasets. The PAN-AA-2011 (Argamon and Juola, 2011) and PAN-AA-2012 (Juola, 2012) draw from corporate emails and fiction domain, respectively. We also use the PAN-AA-2018 dataset (Kestemont et al., 2018b,a), which established a challenging cross-fandom scenario within FanFiction, where all unknown documents originate from a single fandom, while candidate authors’ documents span non-overlapping fandoms. This design amplifies complexity by introducing domain shift and reducing stylistic overlap, as authors often emulate the source material. The PAN-AA-2019 task (Kestemont et al., 2019b,a) further advanced these challenges through an open-set evaluation.

Authorship verification datasets. The PAN-AV-2020 (Bischoff et al., 2020) and PAN-AV-2021 (Bevendorff et al., 2020) authorship verification tasks centered on FanFiction, with the latter introducing an out-of-distribution test set featuring unseen authors and novel topics.

We also consider style-change detection tasks. These involve determining whether consecutive paragraphs in a text are authored by the same person. We transform these datasets to standard authorship verification format by constructing independent examples for each pair of consecutive paragraphs.

We consider data from the PAN-SC-2022 (Zangerle et al., 2022), PAN-SC-2023 (Zangerle et al., 2023b), and PAN-SC-2024 (Zangerle et al., 2024) tasks. The 2022 dataset features StackExchange posts, while the 2023 and 2024 datasets focus on Reddit. The 2023 and 2024 datasets provide three subtasks, curating them to control for topical cues. The first tasks don’t account for this bias, while the second and

Table 5: Authorship attribution results of OSST-based models with three different paraphrasing prompts: neutralizing, paraphrasing and rephrasing in the style of Shakespeare. We highlight in bold the best result across paraphrasing prompts for each dataset and base model.

Model	Paraph. prompt	PAN12		PAN18-en		PAN19-en	
		Accuracy	Mean rank	Accuracy	Mean rank	Accuracy	Mean rank
OSST-Gemma1B	neutral	33.93% $\pm 1.03\%$	0.69 ± 0.03	47.37% $\pm 1.66\%$	0.84 ± 0.01	69.35% $\pm 1.24\%$	0.90 ± 0.00
	paraphrase	21.43% $\pm 1.03\%$	0.67 ± 0.00	34.93% $\pm 0.73\%$	0.76 ± 0.01	63.24% $\pm 0.21\%$	0.87 ± 0.00
	shakespeare	30.36% $\pm 1.79\%$	0.73 ± 0.00	46.89% $\pm 0.48\%$	0.81 ± 0.00	67.01% $\pm 0.26\%$	0.89 ± 0.00
OSST-Llama1B	neutral	39.29% $\pm 8.25\%$	0.74 ± 0.04	51.20% $\pm 2.09\%$	0.85 ± 0.02	71.89% $\pm 0.88\%$	0.91 ± 0.01
	paraphrase	30.36% $\pm 0.00\%$	0.72 ± 0.00	40.67% $\pm 1.44\%$	0.81 ± 0.01	69.96% $\pm 0.12\%$	0.90 ± 0.00
	shakespeare	39.29% $\pm 0.00\%$	0.80 ± 0.00	42.11% $\pm 0.28\%$	0.82 ± 0.00	64.66% $\pm 0.06\%$	0.88 ± 0.00
OSST-Gemma4B	neutral	62.50% $\pm 3.09\%$	0.86 ± 0.02	62.68% $\pm 1.54\%$	0.93 ± 0.01	84.62% $\pm 0.24\%$	0.96 ± 0.00
	paraphrase	55.36% $\pm 0.00\%$	0.87 ± 0.00	62.20% $\pm 3.19\%$	0.90 ± 0.00	85.64% $\pm 0.16\%$	0.96 ± 0.00
	shakespeare	50.00% $\pm 0.00\%$	0.86 ± 0.00	62.68% $\pm 0.73\%$	0.91 ± 0.00	75.05% $\pm 0.46\%$	0.92 ± 0.00
OSST-Llama8B	neutral	69.64% $\pm 3.09\%$	0.93 ± 0.01	69.86% $\pm 1.10\%$	0.95 ± 0.00	89.41% $\pm 0.71\%$	0.98 ± 0.00
	paraphrase	58.93% $\pm 0.00\%$	0.91 ± 0.00	74.16% $\pm 1.10\%$	0.95 ± 0.00	87.47% $\pm 0.71\%$	0.97 ± 0.00
	shakespeare	71.43% $\pm 0.00\%$	0.92 ± 0.00	78.47% $\pm 0.00\%$	0.96 ± 0.00	86.25% $\pm 0.00\%$	0.97 ± 0.00

Table 6: Authorship verification results (F1 score) of OSST-based models with three different prompt versions: neutralizing (N), paraphrasing (P) and rephrasing in the style of Shakespeare (S). Base models are Llama1B and Llama8B. We highlight in bold the best result across paraphrasing prompts for each dataset and base model.

Model	Prompt	22ST-T1	22ST-T2	22ST-T3	23ST-T1	23ST-T2	23ST-T3	24ST-T1	24ST-T2	24ST-T3	20AV	21AV	Avg
ZS-1B	N	62.26%	62.57%	58.30%	69.01%	55.82%	58.97%	73.21%	67.13%	61.27%	67.65%	67.34%	63.96%
	P	60.89%	61.13%	58.31%	70.78%	60.82%	57.38%	74.24%	66.36%	60.00%	65.96%	65.47%	63.76%
	S	61.75%	59.69%	56.24%	71.61%	54.31%	53.57%	75.13%	73.75%	56.05%	63.62%	64.59%	62.76%
ZS-8B	N	64.57%	66.87%	59.99%	68.63%	61.54%	62.93%	75.42%	69.28%	63.36%	76.22%	76.50%	67.75%
	P	62.81%	67.32%	59.92%	71.52%	65.49%	63.53%	74.94%	73.33%	64.08%	77.20%	78.19%	68.94%
	S	63.57%	63.88%	58.29%	68.57%	57.59%	57.92%	71.23%	75.87%	60.51%	75.75%	76.61%	66.34%

third tasks progressively increase in difficulty in this regard.

F Data preprocessing

Processing FanFiction data. To preprocess the FanFiction data used in the PAN18 and PAN19 authorship attribution tasks, we first remove all lines that do not contain alphabetic characters, thereby eliminating chapter separators that should not inform authorship prediction. Next, we remove line breaks to normalize paragraph formatting across authors, as the number of line breaks can vary. Finally, we truncate text at the paragraph level to a maximum of 384 tokens.

Enron emails: To preprocess the Enron emails data used in the PAN11 authorship attribution tasks, we remove surrounding whitespace, truncate to 384

tokens, and remove author texts under 20 characters, and select a maximum of 50 texts per author to avoid massive computational costs.

Cleaning style change data. As some paragraphs in the style change data where purely random symbols, we remove any sample where the paragraph does not contain at least one alphabetic character.

G Summary of candidates and number of texts in attribution datasets

In Table 9 we show the summary of the datasets used for authorship attribution, including the number of targets to predict for each amount of candidates and texts per candidate. *For problems with more candidates, the accuracy value will be naturally lower, but the mean rank value should be*

Table 7: Semantic similarity statistics in authorship attributions sets for paraphrases with three different prompt versions: neutralizing, paraphrasing and rephrasing in the style of Shakespeare. Measured with a sentence-transformers/all-mpnet-base-v2 model for semantic similarity.

dataset	variant	mean $\pm$ std	min	q25	median	q75	max
11AA	neutral	0.700 $\pm$ 0.158	0.002	0.604	0.724	0.818	1.000
	paraphrase	0.683 $\pm$ 0.158	-0.017	0.585	0.700	0.805	1.000
	shakespeare	0.451 $\pm$ 0.151	-0.023	0.346	0.450	0.558	0.904
12AA	neutral	0.790 $\pm$ 0.076	0.470	0.743	0.813	0.840	0.938
	paraphrase	0.877 $\pm$ 0.065	0.639	0.848	0.896	0.923	0.960
	shakespeare	0.753 $\pm$ 0.103	0.446	0.692	0.768	0.825	0.924
18AA	neutral	0.820 $\pm$ 0.097	0.309	0.776	0.839	0.890	0.970
	paraphrase	0.783 $\pm$ 0.165	0.235	0.694	0.849	0.906	0.973
	shakespeare	0.721 $\pm$ 0.166	0.232	0.603	0.766	0.855	0.971
19AA	neutral	0.794 $\pm$ 0.092	0.309	0.747	0.812	0.859	0.980
	paraphrase	0.748 $\pm$ 0.182	0.169	0.598	0.823	0.890	0.987
	shakespeare	0.661 $\pm$ 0.173	0.130	0.524	0.716	0.802	1.000

more stable across settings.

H Verification dataset lengths

Figure 5 presents the length distributions (in characters) of texts in different verification datasets after truncation.

I Base models used in OSST

The following models were used as base for the OSST-based methods in this work:

- google/gemma-3-1b-pt (Team et al., 2025)
- google/gemma-3-4b-pt (Team et al., 2025)
- meta-llama/Llama-3.2-1B (Grattafiori et al., 2024)
- meta-llama/Llama-3.2-8B (Grattafiori et al., 2024)
- mistralai/Mistral-Small-24B-Base-2501

With this choice we cover several model sizes and more than one model family. Llama, Gemma and Mistral models are some of the most well known open-source families. We also wanted to focus on models trained primarily on English data, since they will likely perform better on the languages we measure here.

J Hardware and software

All experiments were run on 1-2 NVIDIA GeForce RTX 3090 GPUs with 24GB of memory, using a small amount of compute and memory from an additional GPU for OSST with Mistral 24B. We used PyTorch 2.7.1 with CUDA 12.2. Code is available at [pablomiralles22/OSST](https://github.com/pablomiralles22/OSST), as well as the versions of the libraries as an uv environment.

K Do we need to frame as a rewriting problem instead of a continuation one?

A natural question is whether framing the problem as rewriting is actually required. To assess this, we compare our approach against a baseline (*NN*) that directly continues the input text without a neutral version to rewrite. As shown in Figure 6, omitting the neutral version has little impact (and can even be advantageous) on datasets where topical cues are strongly aligned. However, in settings with minimal topical correlation, performance drops sharply. This is unsurprising: generating a continuation of one text conditioned on another is substantially more difficult when their topics diverge, especially in the initial tokens, which makes distinguishing related from unrelated pairs significantly easier.

Table 8: Semantic similarity statistics in authorship verification test sets for paraphrases with three different prompt versions: neutralizing, paraphrasing and rephrasing in the style of Shakespeare. Measured with a sentence-transformers/all-mpnet-base-v2 model for semantic similarity.

dataset	variant	mean±std	min	q25	median	q75	max
20AV	neutral	0.786 ± 0.093	0.070	0.739	0.803	0.852	0.966
	paraphrase	0.883 ± 0.071	0.178	0.852	0.902	0.932	0.984
	shakespeare	0.754 ± 0.084	0.221	0.706	0.767	0.815	0.941
21AV	neutral	0.781 ± 0.096	0.021	0.733	0.799	0.849	0.969
	paraphrase	0.880 ± 0.072	0.004	0.850	0.900	0.931	0.987
	shakespeare	0.742 ± 0.090	0.052	0.692	0.755	0.807	0.948
22SC-T1	neutral	0.764 ± 0.139	0.079	0.697	0.792	0.865	0.994
	paraphrase	0.770 ± 0.131	0.084	0.708	0.796	0.867	0.981
	shakespeare	0.486 ± 0.170	-0.059	0.371	0.495	0.609	0.929
22SC-T2	neutral	0.780 ± 0.118	0.042	0.715	0.801	0.867	0.994
	paraphrase	0.795 ± 0.113	0.052	0.739	0.817	0.876	0.993
	shakespeare	0.499 ± 0.162	-0.010	0.388	0.511	0.618	0.930
22SC-T3	neutral	0.751 ± 0.142	-0.018	0.672	0.776	0.857	1.000
	paraphrase	0.750 ± 0.142	0.015	0.675	0.777	0.856	0.996
	shakespeare	0.452 ± 0.175	-0.072	0.322	0.454	0.582	0.949
23SC-T1	neutral	0.755 ± 0.124	0.133	0.691	0.781	0.835	0.984
	paraphrase	0.771 ± 0.118	0.117	0.699	0.796	0.854	0.976
	shakespeare	0.521 ± 0.155	0.045	0.398	0.523	0.637	0.898
23SC-T2	neutral	0.762 ± 0.120	0.031	0.694	0.784	0.850	0.989
	paraphrase	0.774 ± 0.120	0.209	0.704	0.796	0.867	0.988
	shakespeare	0.510 ± 0.168	0.012	0.384	0.515	0.641	0.907
23SC-T3	neutral	0.768 ± 0.107	0.171	0.706	0.785	0.848	0.977
	paraphrase	0.775 ± 0.110	0.185	0.708	0.792	0.858	0.977
	shakespeare	0.513 ± 0.150	-0.009	0.406	0.523	0.626	0.918
24SC-T1	neutral	0.735 ± 0.123	0.165	0.656	0.770	0.832	0.969
	paraphrase	0.749 ± 0.116	0.186	0.685	0.773	0.826	0.974
	shakespeare	0.496 ± 0.141	0.107	0.386	0.493	0.605	0.870
24SC-T2	neutral	0.714 ± 0.120	0.167	0.628	0.736	0.801	0.974
	paraphrase	0.739 ± 0.107	0.164	0.685	0.745	0.816	0.980
	shakespeare	0.492 ± 0.137	0.053	0.401	0.507	0.581	0.871
24SC-T3	neutral	0.755 ± 0.111	0.214	0.691	0.773	0.838	0.980
	paraphrase	0.754 ± 0.112	0.108	0.681	0.770	0.839	0.980
	shakespeare	0.527 ± 0.138	0.078	0.433	0.536	0.633	0.883

L To what extent does the neutralizer model influence performance?

In this work we have used a model of 12B parameters to create neutral versions of the given texts. This choice greatly influences the amount of resources needed to run our model overall. Is this an important choice? In Figure 7 we show a com-

parison between using a 4B and a 12B model in PAN22-24ST tasks. As we can see, this downgrade does affect performance slightly, but not to a great extent.

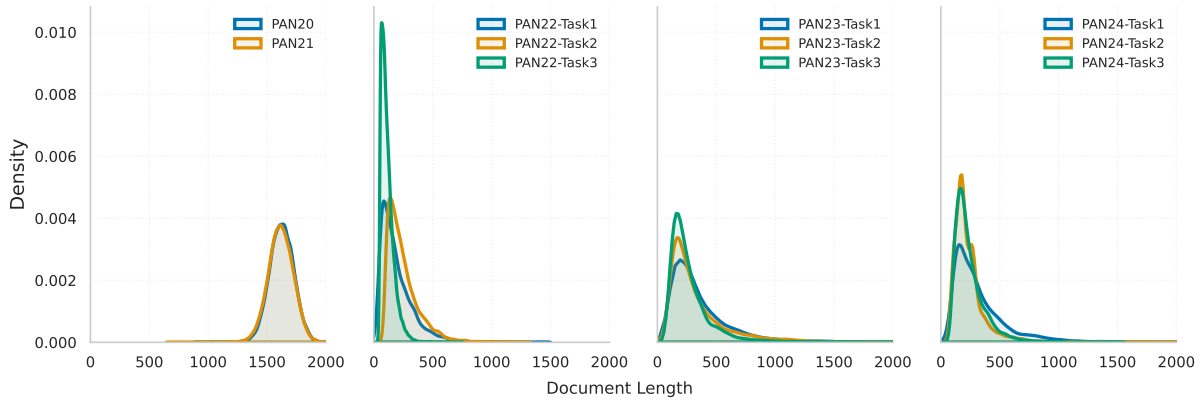


Figure 5: Length distributions (in characters) of texts in different verification datasets after truncation.

M Metrics

We use the following metrics to evaluate models in this work. For closed-set authorship attribution we report the *accuracy* (the ratio of times the correct candidate is picked) and the *rank of the correct candidate* (with 1 being the highest possible rank and 0 the lowest). For open-set attribution we also include the *macro F1 score* of the binary classification problem of deciding whether the true author is among the candidates. Finally, for verification we simply report the F1 score of the binary classification problem of deciding whether the two texts are written by the same author.

N Use of generative Artificial Intelligence

Generative artificial intelligence tools were used in a limited and supportive capacity during the preparation of this work. GitHub Copilot was employed to assist with software development tasks, including code completion and refactoring. Additionally, ChatGPT was used to help improve clarity and fluency of the English writing by suggesting rephrasings and edits. All scientific content, methodological decisions, analyses, and interpretations were conceived, verified, and approved by the author, who takes full responsibility for the final manuscript.

O Artifact licenses

In this work we made use of two kinds of artifacts from prior research: datasets (from PAN) and open-source LLMs (Gemma, Llama and Mistral). All of these were cited appropriately to the best of our knowledge. We discuss in this section their license and the appropriateness of our use of these artifacts.

All datasets were taken from the Zenodo platform. For all datasets, we either did not find a license records, or we found the default license assigned by Zenodo: Creative Commons Attribution 4.0 International (CC-BY). This license allows us to utilize their data provided that we credit them (e.g. citing). Thus, we assume all datasets under this license and consider our use appropriate.

The evaluation in this study was conducted using Llama 3.1/3.2 under the Llama 3.1/3.2 Community License Agreement. In accordance with Section 1(a), the model was used for research and performance testing. To the best of our knowledge, we have complied with the redistribution and attribution requirements set forth in Section 1(b) and adhered to the Acceptable Use Policy as required by the license.

Regarding the Gemma models, we used them in accordance with the Gemma Terms of Use. Our use case complies with the usage rights granted in Section 2.2 and adheres to the Use Restrictions outlined in Section 3.2.

In the case of Mistral models, they are licensed under the Apache License, Version 2.0. In accordance with Section 2 of the license, the software was reproduced and used for performance benchmarking.

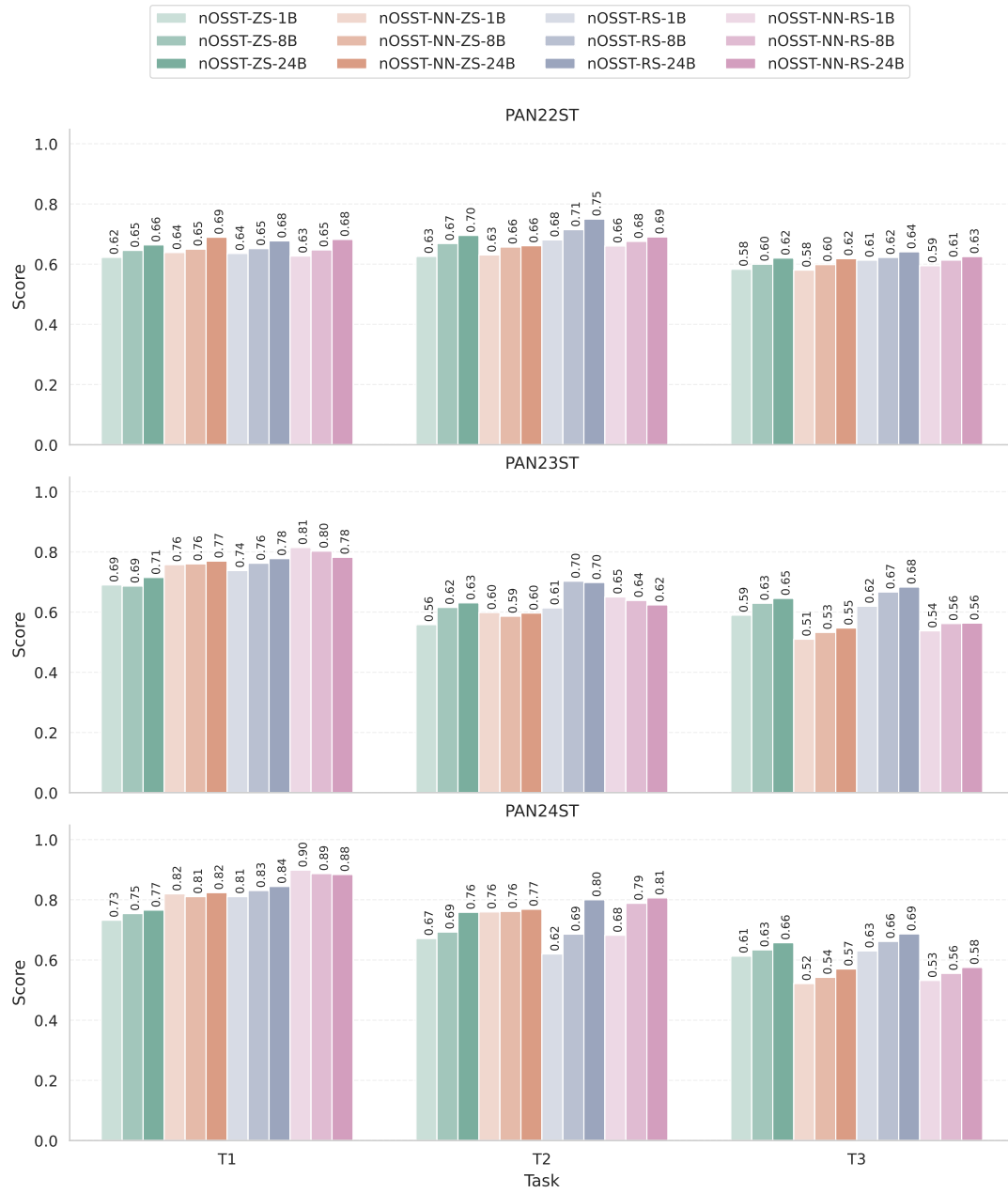


Figure 6: Comparison between using (nOSST) and not using (nOSST-NN) a neutral version to compute log-probabilities.

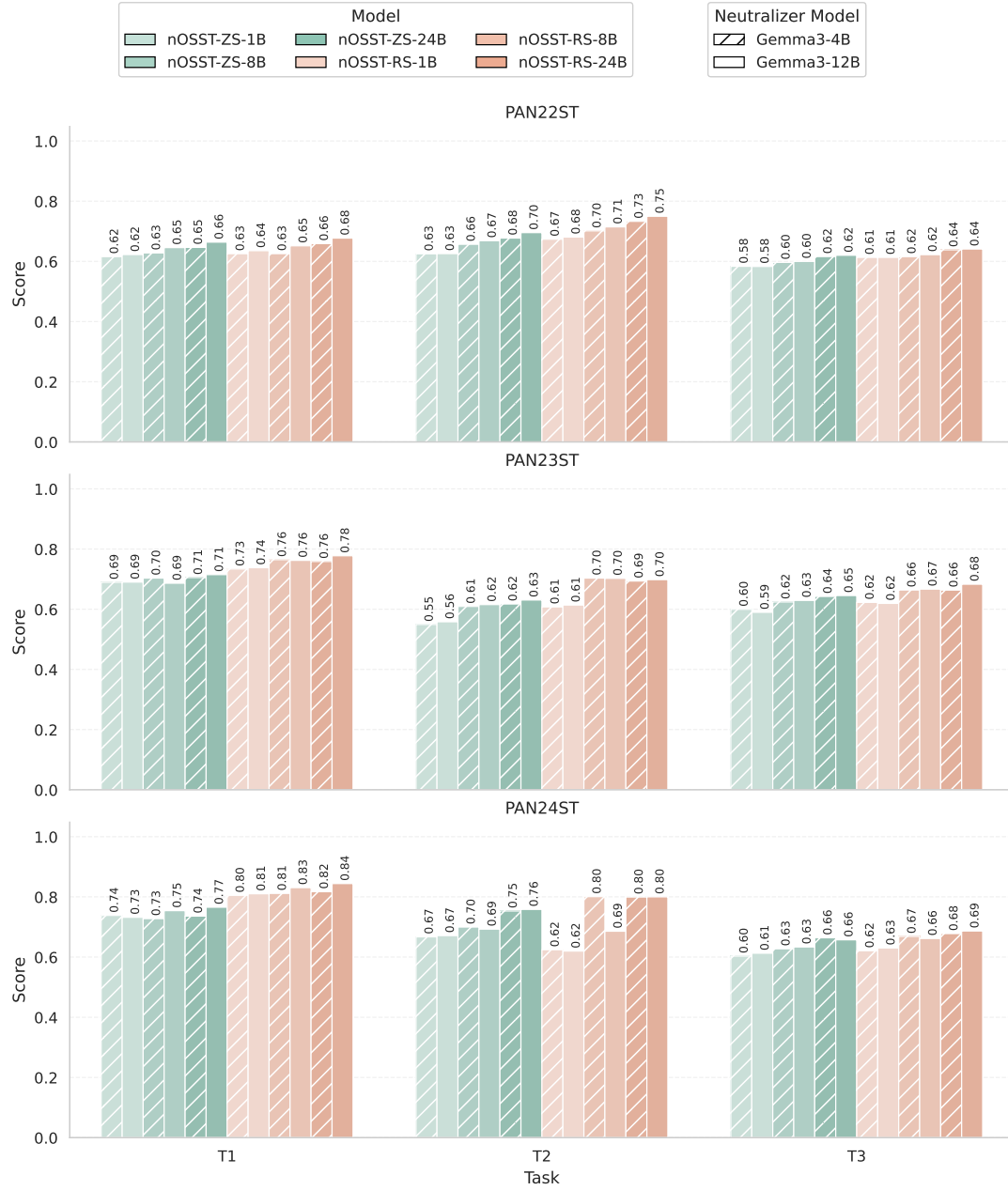


Figure 7: Comparison between using a google/gemma-3-4b-it google/gemma-3-12b-it (Team et al., 2025) model to create neutral versions of texts in PAN22ST, PAN23ST and PAN24ST. RS anchors are kept equal from the 12b neutralizer.

Table 9: Summary of the datasets used for authorship attribution, including the number of targets to predict for each amount of candidates and texts per candidate.

Dataset	Number of texts	# candidates	# targets
PAN11-en	Varies 1-50	26	495
		72	1300
PAN12-en	2	3	16
		8	25
		14	30
PAN18-en	7	5	16
		10	40
		15	74
		20	79
PAN18-fr	7	5	19
		10	47
		15	77
		20	95
PAN18-it	7	5	40
		10	30
		15	72
		20	95
PAN18-pl	7	5	33
		10	42
		15	100
		20	125
PAN18-sp	7	5	25
		10	70
		15	103
		20	133
PAN19-en	7	9	1446
PAN19-fr	7	9	920
PAN19-it	7	9	551
PAN19-sp	7	9	1134