Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin

Gleb Schmidt
Radboud University
Nijmegen, Netherlands
gleb.schmidt@ru.nl

Svetlana Gorovaia
LEYA Lab, HSE University
St. Petersburg, Russia
sgorovaya@hse.ru

Ivan P. Yamshchikov
CAIRO, THWS
Würzburg, Germany
ivan.yamshchikov@thws.de

Abstract

This paper evaluates the performance of Large Language Models (LLMs) in authorship attribution and authorship verification tasks for Latin texts of the Patristic Era. The study showcases that LLMs can be robust in zero-shot authorship verification even on short texts without sophisticated feature engineering. Yet, the models can also be easily "mislead" by semantics. The experiments also demonstrate that steering the model's authorship analysis and decision-making is challenging, unlike what is reported in the studies dealing with high-resource modern languages. Although LLMs prove to be able to beat, under certain circumstances, the traditional baselines, obtaining a nuanced and truly explainable decision requires at best a lot of experimentation.

1 Introduction

Unlike in computational linguistics, authorship analysis in the field of digital humanities still largely relies on the complicated process of domain-specific manual feature engineering (Corbara et al., 2020; Manousakis and Stamatatos, 2023; Corbara et al., 2023; Clérice and Glaise, 2023; Beullens et al., 2024). This is mostly due to the fact that the predictions made by machine learning models with regard to philological and historical authorship problems are expected to be contextualized within long-standing scholarly traditions with their established views on what kind of features matter in the studied material (Clérice and Glaise, 2023). For this reason, although deep-learning approaches, especially those based on pre-trained language models (Fabien et al., 2020; Rivera-Soto et al., 2021; Ai et al., 2022; Huertas-Tato et al., 2022; Yamshchikov et al., 2022; Wang et al., 2023; Huertas-Tato et al., 2024), have recently demonstrated their reliability and high performance, even in the most sophisticated settings of virtually all authorship-related

tasks[1] offered at PAN competitions[2] (Stamatatos et al., 2023; Petropoulos, 2023; Guo et al., 2023).

Pre-trained language models offer valuable insights even in challenging scenarios such as authorship analysis with limited training data, cross-discourse type verification or attribution, style change detection, and cases of stylistic mimicry where authors deliberately disguise their writing style.Additionally, there has been a recent surge in explainable Artificial Intelligence (XAI) techniques, including feature ranking, probing, factual and counterfactual selection, attribution maps, and concept relevance propagation (Achtibat et al., 2023). While these methods are neither flawless (Setzu et al., 2024) nor exhaustive, they represent a significant advancement in the field of explainability.

The linguistic "knowledge" of LLMs, acquired through training on extensive multilingual textual datasets, along with their advanced inference capabilities and their ability to provide human-like natural language explanations for their outputs inevitably raise the question of how these systems can be leveraged for philological and historical investigations.

To promote the wider adoption of large language models (LLMs) as research tools in the digital humanities, this study assesses the zero-shot performance of several publicly available, state-of-the-art LLMs — namely GPT-4o, Gemini, Mistral, and Claude — in authorship verification and attribution tasks. In contrast to previous studies, which have primarily worked with modern languages, our research focuses on a historical language: Latin. To the best of our knowledge, this is one of the first studies to evaluate GPT-4o's "proficiency" in Latin and the first to report test results for the three

---

[1]As of 2024, the following tasks have been offered at least once: Authorship Attribution, Authorship Clustering, Generative AI Detection, Authorship Verification, Authorship Obfuscation Author Profiling, and Author Diarization.
[2]https://pan.webis.de/

additional models.

Our experiments seek to evaluate the zero-shot effectiveness of LLMs in authorship verification and attribution. We compare their performance against traditional baselines, including classical machine learning classifiers and models based on the pre-trained Latin transformer, LaBerta. Furthermore, we investigate how variations in the quantity and nature of instructions prompted to the LLMs impact the accuracy of their predictions.

2   Related Work

Authorship attribution and authorship verification are two fundamental tasks in authorship analysis. They are the most popular applications of stylometry — the modelling of writing style using statistical methods. Attribution, in its simplest form, aims to identify the author of a previously unseen text sample from a list of candidate authors. Verification, on the other hand, involves determining whether two given texts were written by the same person. Both tasks can vary in complexity, especially when additional challenges arise, such as cross-domain or cross-discourse type problems.

The origins of stylometric analysis for authorship-related problems go back to the 1960s (Mosteller and Wallace, 1963). Early work in stylometry for authorship attribution relied on extraction of hand-crafted features believed to represent the writing style (word frequency, sentence length, and syntactic patterns, etc.) and Beyesian inference (Mosteller and Wallace, 1984; Holmes, 1994, 1998). The study by Joachims (1998) on text categorization provided a foundation for applying Support Vector Machines (SVMs) to stylometric data. In 2002, Burrows suggested a distance-based technique, which became formative for the present-day stylometry, the method now known as Burrows' Delta. Since then, various features and classification methods were used to quantify stylistic differences and estimating the likelihood of shared authorship between texts. The work of Stamatatos (2009) provides a comprehensive overview of the classical methods used in stylometric analysis.

Since 2010s, the evolution of methodologies for solving these tasks can be traced through the overviews of PAN competitions. Until 2016, with rare exceptions, texts in authorship analysis were treated as bags-of-words (Stamatatos, 2009; Koppel et al., 2009). Research in this field revolved

around various stylistic features such as word and character n-grams, sentence lengths, word and punctuation frequencies, part-of-speech (POS) tag frequencies, and POS n-grams (Stamatatos, 2013). These features were often combined with feature selection or weighting mechanisms and utilized alongside distance measures and standard classifiers like Support Vector Machines (SVM) or Naive Bayes.

The rise of the neural networks marked the shift towards closer attention to the sequential nature of the text. Convolutional and Recurrent Neural Networks (CNNs and RNNs) and later transformers have proven outperform the previous methods, particularly in cases where the writing style is more nuanced and complex (Shrestha et al., 2017; Kestemont et al., 2018, 2020). Yet, this improvement was achieved at expense of the models's explainability.

Transformer models, such as BERT, RoBERTa, and T5, made authorship attribution and verification systems particularly robust. Since the advent of Siamese network architectures (Reimers and Gurevych, 2019) and the work presented by Fabien et al. (2020) fine-tuning pre-trained models to solve authorship problems has de facto become a standard approach (Rivera-Soto et al., 2021; Stamatatos et al., 2022; Ai et al., 2022; Huertas-Tato et al., 2022; Yamshchikov et al., 2022; Stamatatos et al., 2023; Wang et al., 2023; Huertas-Tato et al., 2024), although ensemble models integrating additional stylometric features (Fabien et al., 2020; Ai et al., 2022) and even independent use of manually engineered features remain quite common (Manousakis and Stamatatos, 2023; Corbara et al., 2023; Clérice and Glaise, 2023; Beullens et al., 2024; Camps et al., 2024).

Since the release of GPT-3.5 in late 2022 (Brown et al., 2020) and the subsequent emergence of GPT-4 (Achiam et al., 2023) and a pleiade of LLMs (Naveed et al., 2023), there have been numerous reports of their groundbreaking performance on various research tasks relevant for the humanities (Karjus, 2023). These tasks range from relatively simple data processing, cleaning, and structuring tasks (such as post-OCR correction, NER, and mark-up) to data augmentation and labeling (Törnberg, 2024), from semantic search to confirmatory topic analysis (Oiva et al., 2023), and from text summarization and translation (Volk et al., 2024) to multimodal processing. The examples of successful applications continue to proliferate, paving the

way for what Karjus has described as "machine-assisted mixed methods" (2023), which facilitate interaction with data and promise unprecedented scaling of research efforts.

Ironically, although the availability of LLMs made the detection of machine-generated text one of the most relevant real-world tasks for linguistic forensics and consequently the prevalent topic at PAN competitions (Bevendorff et al., 2024), the number of studies which explore the LLMs's own abilities to solve authorship-related problems or serve for feature extraction is rather limited so far.

Hicke and Mimno (2023) leveraged a pre-trained T5 model further fine-tuning for authorship identification in Early Modern English drama. Patel et al. (2023) tried to bridge the gap between stylometry and language models annotating examples of writing style and creating interpretable machine-generated writing style embeddings. A somewhat comparable approach was also proposed in Ramnath et al. (2024). The model is trained using a distillation process from GPT-4-Turbo to Llama-3-8B model. First, GPT is used to produce and standardize a corpus of structured writing style descriptions. Llama is then fine-tuned to produce similar descriptions. This approach addresses the challenges of interpretability in authorship analysis by trying to establish a clear and consistent framework for it.

An immediate source of inspiration for this study, the work by Huang et al. (2024), focuses on a direct prompting of different models with authorship-related questions. The authors arrived at the conclusion that guiding the model by explicitly providing specific linguistic features to pay attention to can significantly improve the precision of the model's prediction and the quality of the analysis.

We find a compelling reason to explore the use of LLMs to be the challenge posed by sample size. Traditional machine learning methods, such as those described by Eder (2015) and Eder (2017), often require samples of approximately 1000 words to achieve reliable results. Deep learning approaches typically require substantial amounts of training data, which can be difficult to obtain. In contrast, LLMs can perform effectively without extensive additional training, making them advantageous when dealing with limited or costly data resources.

While the impressive results reported in some studies (Fabien et al., 2020; Kestemont et al., 2019)

are noteworthy, it is important to stress that they were conducted using English-language datasets. Given the widespread use of English in the training data of state-of-the-art LLMs, there is a possibility that some of the datasets may overlap with the training data, potentially influencing the outcomes (Brown et al., 2020).

The case of Latin is very different. First of all, the overall amount of available data is incomparably less. Second, it remains unknown how much of it is actually in the training data of the major LLMs. In 2023, Burns evaluated the amount of Latin in the training dataset of GPT-3.5 as 339 million tokens, assuming that this number could be higher for GPT-4o. Although Latin is arguably the highest-resourced of all the historical languages, the extent of the easily-available Latin dataset hardly exceeds 700 million words (Bamman and Burns, 2020) (including Neo-Latin, Latin Wikipedia, and Internet Archive), while state-of-the-art language models for this language are trained on even smaller data, a clean and high-quality subset of the extant corpus, mostly (Roelli, 2014; Riemenschneider and Frank, 2023a; Ströbel, 2022; Bamman and Burns, 2020). Nevertheless GPT-4 excels in various tasks involving Latin, such as morpho-syntactic annotation (tagging), translation from and into Latin, as well as in text summarization and paraphrasing (Volk et al., 2024).

To the best of our knowledge, no comprehensive study has yet been conducted on the performance of major LLMs, such as GPT-4, in the specific tasks of authorship verification and attribution in Latin. Similarly, the capabilities of other mature LLMs released after GPT-4, such as Gemini, Claude, and Mistral, have also not been thoroughly examined in this context.

For this reason, in our investigation, we tried not only to measure LLMs's performance and compare it to conventional baselines but also to study the discrepancy between different LLMs.

3 Methodology

We conducted a series of experiments on two tasks: authorship attribution and authorship verification. The experiments utilized direct prompting of the flagship versions[3] of four major LLMs: GPT-4o, Claude, Mistral-Large, and Gemini-1.5. However, only the authorship verification experiments involved all four models, as only GPT-4o demon-

---

[3]As of July-August 2024.

strated competitive results in the preliminary authorship attribution tests.

All prompting was implemented in a model-agnostic manner using LangChain library. The choice was mostly dictated by the fact that this library offers a unified API to interact with many different models and facilitates crucial operations such as rate limiting, error handling (request retries), fallbacks, and, most importantly, obtaining structured output from the models.

Each run assessed the performance of a specific model on a given task within a particular setting, defined by the prompt used.

We tested three settings differed by the level of guidance the models received in addition to the default task definition:

1. BASE: the models get only a general description of the task;

2. LIMITED: the models get a general description of the task and explicit instruction to pay attention to writing style;

3. HIP: historically informed prompting, when the models get a general description of the task and a concise list of features to pay attention to formulated by a domain expert and anchored into the scholarly tradition.

Each of the aforementioned settings was tested in two different variants: basic and topic-ignorant, in which the models were explicitly instructed to avoid taking the content and theme into account. For the exact formulation of the prompts, see Appendix B.

To gain further insight into the models's decision-making processes and compare their performance, we undertook two additional steps: (1) we investigated the influence of semantic similarity on the predictions, and (2) we measured the agreement between the models . For the former, the texts used in our Authorship Verification experiments were vectorized using OpenAI's text-embedding-3-large model, and a pairwise cosine similarity was calculated between them. We then computed the correlation between these similarity scores and the the models' predictions across various prompt settings. For the latter, we calculated pairwise joint probability of agreement between models, the pairwise agreement scores are presented in A.

Author	Word count
Augustine of Hippo	2,519,484
Gregory the Great	794,955
Origen (tr. by Rufinus)	385,346
Caesarius of Arles	311,965
Petrus Chrysologus	189,864
Jerome	178,704
Optatus of Milevis	169,280
Quoduultdeus	132,160
Chromatius of Aquileia	115,446
Venerable Bede	114,282
Leo the Great	113,373
Maximus of Turin (I)	73,836
Gregory of Elvira	70,712
Zeno of Verona	48,077
Gaudentius of Brescia	46,127
Ambrose of Milan	43,118
Valerian of Cimiez	31,352
Basil of Caesarea (tr. by Rufinus)	31240
Priscillian of Avila	23,165
Fulgentius of Ruspa	14,804
Leander of Seville	13,077
Potamius of Lisbon	9,190

Table 1: Dataset.

3.1 Metrics and Baselines

To evaluate the performance of the models, we relied on accuracy, precision, recall, and F1 score.

Furthermore, the performance of the LLMs on each of the two tasks was compared against two different baselines, (four baselines in total). For each task, one baseline features a classical machine learning approach, while the other builds upon a state-of-the-art pre-trained transformer model for Latin, LaBerta (Riemenschneider and Frank, 2023b). For details, see the Tables 3, 7, and 4.

4 Data

In this study, we focus on a subset of the Patristic Sermon Textual Archive (PaSTA), a corpus of Latin homiletic literature of the Patristic era. We prefer this corpus to a seemingly more conventional Classical Roman prose for a reason. Indeed, the very nature of the genre of sermon (or homily) — oral and written — provides a wide spectrum of styles depending on the occasion on which sermons were delivered, the intended audience, underlying material, etc. At the very same time, the act of preaching was always framed by the scriptural and liturgical context. As the goal of the preacher was to ex-

plain the message of the Scripture, demonstrate its relevance to the everyday lives of the flocks, and make clear the symbolic and moral meaning of the sacraments and feasts of the Church, the creativity of the preacher was constantly confronted with the canons of the established genre, which suggested themes as well as discursive and rhetorical devices (Boodts and Schmidt, 2022). Such relative thematic homogeneity of the homiletic corpus makes it a particularly interesting and complex benchmark.

4.1 Preparation and General Preprocessing

For the sake of quality, the data was extracted not only from various open (*Patrologia Latina* as available in the *Corpus Corporum*) resources but also proprietary ones (*Corpus Christianorum Series Latina*), which is why we cannot publish the full texts along with all the associated rich metadata. However, we provide all the data used in the described experiments — the randomly sampled textual fragments with the corresponding author labels. All the data is published on GitHub [4].

Out of the 62 distinct authors currently represented in PaSTA, we selected 22 authors featured in Dupont et al. (2018) , a standard reference work to survey Latin preaching from the 3rd to the 7th centuries, see Table 1. This selection covers all regions of the Late Antique Latin West and encompasses all homiletic subgenres.

Since most of the texts used in the study constitute composite entities (e.g., collections of sermons, epistles, gatherings of treatises, etc.), we first divided all the material into units (henceforth, work-units) representing self-contained acts of preaching (e.g., *sermo*, *homilia*, *tractatus*, epistle, *dictio*). Subsequently, for different experiments, the texts were split into chunks of *approximately* (1) 250 and (2) 500 words. We opted for an oscillating chunk length to respect sentence boundaries. Therefore, some chunks are slightly longer or shorter than the target length.

Sampling texts from the pools of chunked examples was done for each task independently.

4.2 Authorship Verification

Before conducting the first authorship verification experiment, we sampled 5 positive and 5 negative pairs for each of the 22 authors. This process was repeated three times, allowing us to perform each

experiment with three distinct sets of pairs. While each pair was unique, the same passage could appear in multiple pairs. This yielded a balanced corpus of 660 pairs, with 30 pairs per author — 15 positive and 15 negative. This same set of 660 pairs[5] was used across all subsequent authorship verification experiments, with 220 pairs evaluated in each iteration, though the content of each iteration could vary depending on the model employed.

4.3 Authorship Attribution

Authorship attribution experiments were conducted using varying numbers of candidate authors: 5, 10, 15, and 22. For each of these configurations, we randomly selected the required number of authors. To ensure diversity and enhance the reliability of the results, this selection process was repeated five times, generating distinct sets of candidate authors for each iteration.

The sampling of text examples proceeded as follows. For each randomly selected author, we randomly picked two text fragments. The first fragment was designated as the query text, while the second fragment, drawn from a different work by the same author, served as the target text (i.e., the text forming a positive pair with the query). This was further supplemented with texts by other authors, which created negative pairs with the query text. The task for the model was then to match each query text with the correct target text from the provided set.

5 Results

5.1 Author Verification

Table 3 presents the performance of each model in the tested settings, averaged over three iterations. Only two models — GPT-4o and Claude-3.5 — demonstrated accuracy comparable to the results reported by Huang et al. (2024) for English texts. Both models outperformed the LaBerta-based baseline[6] in terms of accuracy, with notably high positive predictive values. Although Claude-3.5 did not outperform the baselines in terms of recall and F1 scores, its numbers were higher than those of Mistral and Gemini.

Contrarily to what was expected based on the results yielded by the so-called linguistically-informed prompt reported by Huang et al. (2024),

---

[4]`https://github.com/glsch/sui_generis`.

[5]`https://github.com/glsch/sui_generis/blob/main/data/authorship_verification_dataset.pkl`.
[6]The model was used without fine-tuning.

Experiment	Parameter	Value
Authorship Verification	Total authors	22
	Pairs per author	30 (15 positive, 15 negative)
	Total pairs	660
	Repetitions	3
	Pairs per iteration	220
	Text length	app. 500 words
Authorship Attribution	Sizes of candidate author sets	5, 10, 15, 22
	Repetitions per configuration	5
	Texts per author	2 (1 query, 1 target)
	Pair types	1 positive, multiple negative
	Text length	app. 500 words and app. 250 words

Table 2: Sampling.

Model	Prompt/Parameters	Accuracy	Precision	Recall	F1
claude-3-5-sonnet-20240620	BASE	72 (±1%)	98(±2%)	45 (±4%)	62 (±4%)
	BASE_TOPIC_IGNORANT	68 (±1%)	99 (±1%)	37 (±4%)	54 (±4%)
	HIP	67 (±4%)	98 (±2%)	34 (±10%)	50 (±11%)
	HIP_TOPIC_IGNORANT	61 (±3%)	100 (±0%)	22 (±4%)	36 (±5%)
	LIMITED	70 (±5%)	99 (±1%)	40 (±6%)	57 (±6%)
	LIMITED_TOPIC_IGNORANT	67 (±3%)	99 (±1%)	34 (±1%)	51 (±1%)
gemini-1.5-pro	BASE	56 (±2%)	73 (±6%)	18 (±1%)	29 (±2%)
	BASE_TOPIC_IGNORANT	52 (±4%)	57 (±11%)	21 (±4%)	31 (±6%)
	HIP	57 (±2%)	77 (±4%)	21 (±5%)	33 (±6%)
	HIP_TOPIC_IGNORANT	54 (±1%)	84 (±7%)	11 (±2%)	19 (±4%)
	LIMITED	56 (±4%)	84 (±9%)	15 (±4%)	25 (±5%)
	LIMITED_TOPIC_IGNORANT	55 (±7%)	82 (±11%)	12 (±5%)	20 (±8%)
gpt-4o	**BASE**	78 (±2%)	90 (±6%)	63 (±1%)	74 (±2%)
	BASE_TOPIC_IGNORANT	70 (±3%)	95 (±1%)	43 (±3%)	59 (±3%)
	HIP	75 (±1%)	88 (±0%)	57 (±3%)	69 (±3%)
	HIP_TOPIC_IGNORANT	71 (±1%)	96 (±5%)	43 (±2%)	59 (±3%)
	LIMITED	**80 (±2%)**	**89 (±5%)**	**68 (±5%)**	**77 (±3%)**
	LIMITED_TOPIC_IGNORANT	70 (±3%)	95 (±2%)	43 (±1%)	59 (±1%)
mistral-large-latest	BASE	56 (±3%)	54 (±3%)	76 (±5%)	63 (±4%)
	BASE_TOPIC_IGNORANT	54 (±3%)	54 (±4%)	56 (±3%)	55 (±3%)
	HIP	54 (±6%)	53 (±7%)	75 (±3%)	62 (±6%)
	HIP_TOPIC_IGNORANT	54 (±5%)	53 (±7%)	63 (±5%)	58 (±6%)
	LIMITED	56 (±1%)	54 (±1%)	76 (±1%)	63 (±1%)
	LIMITED_TOPIC_IGNORANT	53 (±6%)	53 (±9%)	47 (±7%)	50 (±7%)
TF-IDF + Random Forest	char, ngram_range=2,9, max_features=5000	58	59	61	60
LaBerta + Mean pooling + Cosine similarity		69	54	93	68

Table 3: Results for Authorship Verification task on full dataset (22 authors, 5 positive and 5 negative pairs per author in each iteration).

explicit philological and historical features generally deteriorated results compared to the BASE setting for all models except Gemini.

5.2 Author Attribution

Tables 4 and 7 present the results of the authorship attribution task conducted on subsets of 5, 10, and 15 authors, as well as on the full dataset of 22 authors, using text fragments of 500 and 250 words, respectively. Only the GPT-4o model was tested for this task, as it had demonstrated the best performance in the simpler authorship verification setting.

Since multi-class classification is generally more challenging than binary classification, it is unsurprising that GPT-4o did not surpass the LaBerta baseline when the text length was sufficient (500 words). However, in one setting — fragments of 250 words with 5 authors (see Table 7 in Appendix C) — GPT-4o outperformed both baselines. In all other cases, as the number of candidate authors increased, GPT-4o's performance declined, and at a faster rate than that of the LaBerta baseline.

Consistent with the observations from the authorship verification experiments, explicit instructions regarding philological and histori-

Model	Prompt/Setting	5 Authors		10 Authors		15 Authors		22 Authors	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o	BASE	48	37	32	24	37	28	21	13
	BASE_TOPIC_IGNORANT	**68**	**62**	**36**	**27**	37	28	21	12
	HIP	56	48	32	23	35	28	**21**	**13**
	HIP_TOPIC_IGNORANT	44	36	32	22	**39**	**29**	17	8
	LIMITED	52	42	28	19	29	22	20	11
	LIMITED_TOPIC_IGNORANT	56	46	34	25	35	26	14	7
TF-IDF		44	37	26	19	12	7	6	4
LaBerta + Mean pooling + Cosine		**72**	**65**	42	34	**41**	**35**	**36**	**29**

Table 4: Results for Authorship Attribution task on subsets of 5, 10, 15 and 22 (full dataset) authors with fragments of 500 words in terms of Accuracy and Weighted F1. The results of GPT-4o model are compared with several baseline pre-trained models.

cal features had a negative impact on performance. Prompts with fewer constraints, such as BASE_TOPIC_IGNORANT or LIMITED, yielded better results. As expected, the length of text fragments had a predictable effect on prediction quality, with accuracy generally decreasing as the texts became shorter (except in the 5-author setting). This suggests that longer fragments provide more information beneficial for authorship attribution.

6 Discussion

The experiment have provided interesting insights into the capabilities of the LLMs and the way how they approach the tasks of authorship verification and attribution.

In Authorship Verification, the strong performance of GPT-4o in the basic setting was largely anticipated due to its advanced capabilities. However, the comparable results achieved by Claude-3.5 are noteworthy, indicating its potential effectiveness in authorship verification tasks.

We were initially concerned about the high performance of the GPT-4o model in the Authorship Verification task, assuming the possibility that parts of our dataset could be simply memorized during the training and merely recalled in our experiment. The decrease in the GPT-4o's performance observed in the Authorship Attribution task, especially with an increasing number of candidate authors, suggests that the model's decisions were guided by underlying processes other then reproducing memorized content.

In this respect, the observation that more detailed instructions, crafted by a domain expert based on scholarly tradition, actually deteriorated performance contrasts the performance of the linguistically-informed prompt used by Huang et al. (2024) and is perhaps particularly noteworthy.

While the models are capable of detecting and describing philological features within the texts, this ability does not necessarily translate into accurate predictions. The connection between the features mentioned in HIP to the prediction is much subtler and less straightforward than that of, for example, orthography or punctuation mistakes so successfully used by Huang et al. (2024). This possibly suggests that when a model can leverage its intrinsic knowledge, it achieves better results than when formal instructions are provided for tasks that are resistant to formalization (Ouyang et al., 2022; Liu et al., 2021).

A closer examination of the models' output[7] highlights this issue. When explicitly instructed, the models generally perform well in identifying the specified features. For instance, they demonstrated notable "attention" to syntactical patterns such as anaphora (repetition of a word or phrase at the beginning of successive clauses), asyndeton (omission of conjunctions), polysyndeton (repetition of conjunctions), and hyperbaton (disruption of normal word order through the insertion of other words). However, in many cases, the models tend to overinterpret these features, often assuming a deterministic relationship between the presence of such patterns and the final prediction.

For example, when comparing different passages from Leander of Seville, GPT-4o generated the following description of the rhetorical devices in the two texts: "The first text uses rhetorical questions and exclamations to emphasize its points (e.g., 'O infinita humilitatis documenta!'). The second text, however, relies more on a narrative and descriptive style, with extensive use of quotations from Solomon to build its argument. The rhetori-

---

[7]All responses are available on GitHub: https://github.com/glsch/sui_generis/blob/main/data/authorship_verification_responses.tsv

cal strategies differ significantly between the two texts." Although this succinct characterization is adequate, the conclusion reached by the model is incorrect.

Similarly, GPT-4o was perplexed by a discrepancy in two different sermons by Caesarius of Arles, stating: "Text 1 employs a more complex and formal structure, with longer sentences and a higher frequency of subordinate clauses. For example, phrases like *ut modestiae tuae non desit auctoritas, constantiam mansuetudo commendet, iustitiam lenitas temperet* show a sophisticated use of parallelism and balance. Text 2, while still formal, uses shorter sentences and simpler structures. It often employs direct questions and answers, such as *quis est hic, et laudabimus eum?* and *absit, ut desperem hic esse aliquem, immo non aliquem, sed aliquos.* This creates a more conversational tone." Similar example can be multiplied at random.
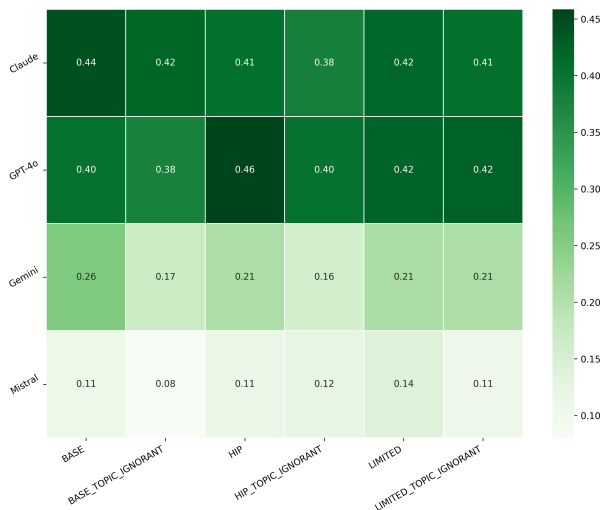
6.1 The Role of Semantic Similarity

Figure 1: Cosine Similarity Correlation Heatmap by Model and Prompt

Given the high number of such cases, we tried to analyze how semantic similarity influenced the models's decisions in the authorship verification setting. Figure 1 represents the Pearson correlation coefficient between the cosine similarity of the prompted texts and the correctness of the model's answers. The responses of the best-performing models, GPT-4o and Claude, seem to align well with semantic similarity across various prompt settings, with only marginal variation. Even when explicitly instructed not to take the content into consideration, the models largely relied on the meaning

of the texts. While writing style and semantics are inherently connected, in authorship analysis, the challenge lies precisely in discerning writing style independently of the subject matter. Our results suggest that LLMs struggle with this distinction, at least in a zero-shot setting.

LLMs are designed to follow human instructions closely, which probably explains why the settings with a lot of explicit guidance show a higher precision. However, the inherently intuitive nature of authorship analysis, especially for short texts, is not easily formalizable, which is in contradiction with what the models are trained for.

When models are given strict prompts, they tend to follow them closely but may overinterpret features, resulting in deteriorated prediction quality. Overprompting seems to limit the models' ability to leverage their intrinsic knowledge effectively.

It is particularly clear in the case of Gemini. The model seems to have responded positively to provided instructions. With more detailed prompts, the precision of the answers increased, whereas correlation with the semantic similarity diminished. Yet, the instructions — although formulated by a domain expert and synthesized the criteria commonly applied to authorship analysis in the field (Mutzenbecher, 1962; Dolbeau, 2017; Weidmann, 2018) — do not cover all possible stylistic subtleties, limiting its effectiveness. Larger models like GPT-4o and Claude benefit from less constrained prompts, allowing them to apply their extensive intrinsic knowledge more freely and leverage their capability to discern semantic similarities. We suggest that this is the reason, why the LIMITED setting, which gives provides the models with some hint to what to pay attention to and does not constrain them too much, performed that well on both tasks.

7 Conclusion

This study highlights the potential of large language models in performing authorship verification and attribution for Latin texts. The LLMs, particularly GPT-4o, exhibited robust performance, often surpassing traditional baselines. However, our results also highlight the challenges in steering these models' "decision-making" processes. While LLMs are capable of handling complex linguistic tasks in low-resource historical languages like Latin, there is still significant room for improvement in their interpretability and adaptability to domain-specific nuances. Enhancing their ability

to disentangle style from content without relying overly on thematic similarities is crucial.

By addressing these challenges, we can unlock the full potential of LLMs in philological and historical investigations, contributing valuable tools to the fields of computational linguistics, stylometry, and the digital humanities.

Limitations

This study and the very approach it explores have several limitations one has to keep in mind. First, in zero-shot setting we fully rely in how the models were trained by their creators, and none of the used state-of-the-art LLMs was specifically trained (or tuned) on extensive Latin datasets especially on a rather peculiar and niche task such as authorship analysis. Therefore, experimenting with it might not fully capture the potential of these models. Second, the dataset used in this study is relatively small and, as mentioned in Section 4, is very peculiar from a thematic point of view. While being an interesting benchmark, it might yield observations which are difficult to generalize for texts of other epochs or genres, e.g. Latin poetry, scientific or legal prose. Third, in study, only a very superficial qualitative analysis of the output was performed. Although we present working hypotheses on the models's decision-making based on quantitative observations, the real extent of the relevance of the analysis generated by the models is yet to be determined in a close reading. We intend to investigate this in our future research.

Ethics Statement

This research adheres to the ethical guidelines established by the Association for Computational Linguistics (ACL). We acknowledge the limitations inherent in the use of LLMs, particularly concerning their potential biases and the ethical implications of using proprietary datasets. Care was taken to ensure that the data used did not violate any privacy or copyright concerns. The broader impact of this research is considered in terms of its contribution to the digital humanities, particularly in enhancing the tools available for studying historical texts in low-resource languages. We encourage further research that critically examines the ethical dimensions of applying LLMs to historical and cultural datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? Learning to Contrast for Authorship Attribution. Publisher: arXiv Version Number: 2.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology.

Pieter Beullens, Wouter Haverals, and Ben Nagy. 2024. The Elementary Particles: A Computational Stylometric Inquiry into the Mediaeval Greek-Latin Aristotle. *Mediterranea. International Journal on the Transfer of Knowledge*, 9:385–408.

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 CLEF 2024 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS.org.

Shari Boodts and Gleb Schmidt. 2022. Sermon/Homiletics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Patrick Burns. 2023. How much Latin does ChatGPT "know"?

John F. Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Jean-Baptiste Camps, Benedetta Salvati, Gonzalo Freijedo, Donghan Bian, Gaëtan Drouet, Eglantine Gaglione, Émilie Guidi, Carolina Macedo, Yaelle Zribi, and Florian Cafiero. 2024. The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination. In *DH Benelux 2024*.

Thibault Clérice and Anthony Glaise. 2023. Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world. In *Computational Humanities Research Conference 2023*, Proceedings of the Computational Humanities Research Conference 2022, Paris, France.

Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Syllabic quantity patterns as rhythmic features for Latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141.

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, Mirko Tavoni, and others. 2020. L'Epistola a Cangrande al vaglio della computational authorship verification: Risultati preliminari (con una postilla sulla cosiddetta "XIV Epistola di Dante Alighieri"). In *Nuove inchieste sull'epistola a Cangrande: atti della giornata di studi*.

François Dolbeau. 2017. Sermons « africains » : critères de localisation et exemple des sermons pour l'Ascension. In Gert (Ed) Partoens, A. Dupont, and S. Boodts, editors, *Praedicatio patrum: Studies on preaching in late antique North Africa*, Instrumenta Patristica et Mediaevalia (IPM)_, pages 9–35. Brepols, Turnhout.

Anthony Dupont, Johan Leemans, Gert Partoens, and Shari Boodts, editors. 2018. *Preaching in the Patristic Era: sermons, preachers, and audiences in the Latin West*. Number 6 in A new history of the sermon. Brill, Leiden ; Boston.

Maciej Eder. 2015. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Maciej Eder. 2017. Short Samples in Authorship Attribution: A New Approach. In *International Conference on Digital Health*.

Maël Fabien, Esa ú 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*. ACL.

Mingcan Guo, Zhongyuan Han, Haoyang Chen, and Haoliang Qi. 2023. A contrastive learning of sample pairs for authorship verification. *Working Notes of CLEF*.

Rebecca M. M. Hicke and David Mimno. 2023. T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models.

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

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117. Publisher: Oxford University Press.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can Large Language Models Identify Authorship? ArXiv:2403.08213 [cs].

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. PART: Pre-trained Authorship Representation Transformer. ArXiv:2209.15373 [cs].

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*, page 111867. Publisher: Elsevier.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

Andres Karjus. 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. Publisher: arXiv Version Number: 1.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working notes of CLEF 2020-Conference and Labs of the Evaluation Forum, 22-25 September, Thessaloniki, Greece*, pages 1–14.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, and Walter Daelemans. 2019. Overview of the cross-domain authorship verification task at pan 2019. In *Working Notes Papers of the CLEF 2019 Evaluation Labs*.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. 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.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing.

407

Nikos Manousakis and Efstathios Stamatatos. 2023. Authorship Analysis and the Ending of Seven Against Thebes: Aeschylus' Antigone or Updating Adaptation? *Classical World*, 116(3):247–274.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309. Publisher: Taylor & Francis.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer-Verlag New York Inc.

A. Mutzenbecher, editor. 1962. *Sermonum collectio antiqua, nonnullis sermonibus extravagantibus adiectis (Maximus Taurinensis)*. Number 23 in Corpus Christianorum Series Latina. Brepols, Turnhout. Ed.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Mila Oiva, Ksenia Mukhina, Vejune Zemaityte, Tillmann Ohm, Mikhail Tamm, Andres Karjus, Mark Mets, Daniel Chavez Heras, Mar Canet Sola, Helena Hanna Juht, and Maximilian Schich. 2023. A Framework for the Analysis of Historical Newsreels.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning Interpretable Style Embeddings via Prompting LLMs. ArXiv:2305.12696 [cs].

Panagiotis Petropoulos. 2023. Contrastive learning for authorship verification using BERT and bi-LSTM in a Siamese architecture. *Working Notes of CLEF*.

Sahana Ramnath, Kartik Pandey, Elizabeth Boschee, and Xiang Ren. 2024. CAVE: Controllable Authorship Verification Explanations. *arXiv preprint arXiv:2406.16672*.

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*. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023a. Exploring Large Language Models for Classical Philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023b. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*. Association for Computational Linguistics, Toronto, Canada. To appear.

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.

Philipp Roelli. 2014. The corpus corporum, a new open latin text repository and tool. *Archivum latinitatis medii aevi: Bulletin du Cange*, 72:289–304.

Mattia Setzu, Silvia Corbara, Anna Monreale, Alejandro Moreo, and Fabrizio Sebastiani. 2024. Explainable authorship identification in cultural heritage applications. *ACM Journal on Computing and Cultural Heritage*.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar 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.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *JASIST*, 60:538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21:421.

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 *CEUR workshop proceedings*, volume 3180, pages 2301–2313. CEUR-WS. org.

Estathios Stamatatos, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2023. Overview of the Authorship Verification Task at PAN 2023. *Working Notes of CLEF*.

Phillip Benjamin Ströbel. 2022. Roberta base latin cased v2.

Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. Version Number: 1.

Martin Volk, Dominic P Fischer, Lukas Fischer, Patricia Scheurer, Phillip Ströbel, Rachele Sprugnoli, and Marco Passarotti. 2024. LLM-based Machine Translation and Summarization for Latin. Publisher: University of Zurich.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can Authorship Representation Learning Capture Stylistic Features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Clemens Weidmann. 2018. Maximus of Turin. Two Preachers of the Fifth Century. In Anthony Dupont, Shari Boodts, Gert Partoens, and Johan Leemans, editors, *Preaching in the Patristic Era. Sermons, Preachers, and Audiences in the Latin West*, number 6 in A New History of the Sermon, pages 347–372.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. Bert in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080.

A Intra-model agreement

The intra-model agreement scores reflect the reproducibility and reliability of results across models. High agreement scores, particularly with prompts incorporating topic-ignorance instruction, suggest that these prompts encourage models to make more predictions rather based on features unrelated to the subject matter of the texts.

Table 5 presents the intra-model agreement scores across different prompts for each model comparison. Generally, we observe that models demonstrate higher agreement scores when using prompts with TOPIC_IGNORANCE instruction compared to the generic prompts. The LIMITED_TOPIC_IGNORANT prompt consistently yields higher agreement, especially between Claude and GPT-4o, as well as between GPT-4o and Mistral, suggesting that topic ignorance instructions positively influence intra-model consistency in predictions. Conversely, lower agreement scores are observed between Claude and Mistral, indicating that certain model-prompt pairs may interpret and respond to stylistic cues differently, even when following similar instructions.

B Experiment Settings

Table 6 summarizes the prompts used in the study.
.

C Authorship Verification 250 words

Table 7 shows the Authorship Attribution results for fragments of 250 words.

Prompt	Claude vs Gemini	Claude vs GPT-4o	Claude vs Mistral	Gemini vs GPT-4o	Gemini vs Mistral	GPT-4o vs Mistral
BASE	79.09	81.06	44.85	68.48	38.03	50.00
HIP	79.09	81.06	44.85	68.48	38.03	50.00
LIMITED	80.91	80.61	43.79	66.67	36.52	51.97
BASE_TOPIC_IGNORANT	72.73	85.61	53.03	69.85	46.82	54.09
HIP_TOPIC_IGNORANT	88.03	84.85	44.55	78.64	42.73	49.09
LIMITED_TOPIC_IGNORANT	81.97	87.12	56.97	78.48	53.48	55.61

Table 5: Intra-model agreement scores across different prompts for model comparisons.

Prompt structure

System message

You are an experienced philologist who specializes in post-Classical Latin and has a deep knowledge of Latin patristic literature. Your task is to verify the authorship of texts.

Taske definition

Authorship Verification	**Authorship Attribution**
You will be given a pair of texts, and you will have to analyze them in order to decide whether they are written by the same author or not. Importantly, you do **not** have to guess who the author is, but only decide whether the provided texts are likely to be written by the same person or not.	Given a set of texts with known authors and a query text, determine the author of the query text.

Optional parameter

TOPIC IGNORANCE As the texts are thematically similar and all of them feature religious, theological, and philosophical content, you should disregard in your decision the topic and content (an additional instruction which can be prepended to any other).

Guidance levels

BASE *Task definition only, no further guidance provided except for optional TOPIC IGNORANCE.*

LIMITED Base your reasoning on the analysis of the writing style of the input texts.

HIP (historically-informed prompt) Carry out your analysis by examining the philological and historical elements of the writing style found in the input texts. Consider, but do not limit your analysis to, the following features:

- Morphology: affixes, declination, and verbal endings
- Syntax: sentence structure, use of tenses and moods
- Rhetorical figures: tropes and figures of speech which alter the ordinary meaning or order of words to produce rhetorical effects or rhythmical patterns.
- The use of the Bible: how biblical quotations are introduced, framed, and/or connected to each other
- Vocabulary of the text: compound and modal verbs; the words authors use to make evident the structure of the argument as well as various function words (conjunctions, pronouns, interjections, and particles) and the so-called hapax legomena (rare word and expressions)
- The tone of the text (moralizing, philosophical, exegetical, high-flown, affectionate, chunky, simplistic, etc.)

Human message

Authorship Verification	**Authorship Attribution**
- Text 1 - Text 2	- Query text - Texts of candidate authors

Table 6: Prompt structure and experiment settings.

Model	Prompt/Setting	5 Authors		10 Authors		15 Authors		22 Authors	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o	BASE	64	55	30	21	31	23	18	11
	BASE_TOPIC_IGNORANT	56	47	30	24	25	17	18	12
	HIP	60	50	28	19	21	14	22	15
	HIP_TOPIC_IGNORANT	52	40	28	19	21	15	17	11
	LIMITED	**68**	**61**	30	27	24	16	23	17
	LIMITED_TOPIC_IGNORANT	68	59	32	22	28	19	17	11
TF-IDF		20	15	14	11	13	10	8	5
LaBerta + Mean pooling + Cosine		48	39	**66**	**57**	**40**	**31**	**39**	**33**

Table 7: Results for Authorship Attribution task on subsets of 5, 10, 15 and 22 (full dataset) authors with fragments of 250 words in terms of Accuracy and Weighted F1. The results of GPT-4o model are compared with several baseline pre-trained models.