Benchmarking Arabic Authorship Attribution and Style Transfer with Large Language Models

Injy Hamed[1] **Bashar Alhafni**[1] **Nizar Habash**[1,2] **Thamar Solorio**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]New York University Abu Dhabi
{injy.hamed,bashar.alhafni,thamar.solorio}@mbzuai.ac.ae
{nizar.habash}@nyu.edu

Abstract

Writing style is a fundamental component of natural language. However, significant research gaps remain in two key style-centric tasks: authorship attribution (AA) and authorship style transfer, particularly for Arabic. In this work, we revisit both tasks in that context. We introduce a new AA dataset comprising texts in Modern Standard and Dialectal Arabic. We train transformer-based AA models using dual cross-entropy and contrastive learning loss objectives, and validate model performance through human evaluation. We then utilize the trained AA model to benchmark a range of large language models (LLMs) on style recognition and generation tasks, providing new insights into their capabilities in modeling Arabic writing styles. Our work reveals limitations of current models and provides resources to advance research in this direction.

Keywords: Arabic, Authorship Attribution, Authorship Style Transfer, Benchmark, LLMs

1. Introduction

Authorship Attribution (AA) is the task of identifying the author of a given text based on the writing style. AA plays a vital role in preserving the credibility of textual content and enforcing authorial accountability (Huang et al., 2025). Moreover, researchers also use AA models to evaluate text style transfer (TST) systems, assessing the extent to which their generated output is stylistically aligned with a target author (Horvitz et al., 2024; Mukherjee et al., 2025).

Despite the importance of AA, studies focusing on Arabic remain limited across several dimensions. The majority of previous research has focused on Classical Arabic or specific formal domains, relying on small datasets and outdated machine learning approaches (Alqahtani and Dohler, 2023). Dialectal variation – a distinctive stylistic feature in Arabic – is also frequently overlooked, and human evaluations of AA models are often lacking.

Another notable research gap lies in the area of TST, particularly in the context of LLMs. While numerous benchmarks evaluate LLMs across various NLP tasks in Arabic and other languages (Singh et al., 2024; Mousi et al., 2025; Baucells et al., 2025), their capabilities in modeling writing styles remains underexplored. A few studies have evaluated LLMs across multiple TST tasks (Reif et al., 2022; Patel et al., 2022; Mukherjee et al., 2024; Liu et al., 2024b; Yang and Carpuat, 2025), however with a primary focus on English. Thus, our understanding of how well these models perform across languages is still limited, particularly for authorship style transfer, which has received little attention.

In this work, we revisit Arabic authorship attribution and style transfer, aiming to fill these research gaps. Specifically, we evaluate the extent to which LLMs understand and generate stylistic features in Arabic across two tasks: ranking texts with regards to style and authorship style transfer. To this end, we present the following contributions:

- A new Arabic AA dataset spanning Modern Standard Arabic (MSA) and Dialectal Gulf Arabic, sourced from datasets containing modern Arabic literature and forum novels, respectively.
- A set of fine-tuned transformer-based AA models trained using cross-entropy and contrastive learning dual loss objective.
- A human evaluation study on style similarity ranking, supported by detailed annotation guidelines and publicly released annotations.
- An evaluation of LLMs on both style ranking and style transfer tasks, offering new insights into their stylistic capabilities in Arabic.

Our findings shed light on the limitations of current models in handling Arabic writing styles. To support further research in Arabic authorship attribution and style transfer, we make the AA dataset, human annotations, LLMs' generations, AA models, and scripts publicly available.[1]

[1]https://github.com/mbzuai-nlp/arabic-authorship-attribution

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7262–7278
11–16 May 2026.

2. Related Work

In this section, we review prior work on authorship attribution, with a focus on Arabic. We cover existing datasets, modeling approaches, stylometric features, and human evaluation studies. We also highlight current research gap in text style transfer.

2.1. Authorship Attribution Datasets

Several AA datasets have been developed for English covering a range of domains, including Reddit posts (Baumgartner et al., 2020), Amazon reviews (Ni et al., 2019), blogs (Koppel et al., 2006), emails (Klimt and Yang, 2004), and IMDb (Seroussi et al., 2014). In the scope of literature, Project Gutenberg and fanfiction have been used as popular sources for AA datasets utilized by multiple researchers (Bischoff et al., 2020; Terreau et al., 2021; Tyo et al., 2023; Silva et al., 2023).

In the context of Arabic, the coverage of language variants (MSA versus dialectal) varies across domains. Dialectal Arabic is primarily represented in tweets (Albadarneh et al., 2015; Altakrori et al., 2018; Alsager, 2020), whereas MSA is more prevalent in forum messages (Benjamin et al., 2013), articles (Khalil et al., 2020), and Islamic contexts (Al-Sarem et al., 2020). In the literary domain, Arabic corpora remain limited, being relatively small and focused on Classical Arabic. Early efforts include the Ancient Arabic Texts (AAAT) corpus created by Ouamour and Sayoud (2012). The dataset comprises 10 authors each having one document, written between 921 and 1852. Altheneyan and Menai (2014) also introduced a dataset with 30 books written by 10 authors and Shaker and Corne (2010) collected a dataset of 14 books written by six authors, all in Classical Arabic.

2.2. Authorship Attribution Methodologies

Early work on AA relied on analyzing stylometric features that capture distinctive aspects of writing style (Holmes, 1994). With advances in machine learning (ML), these features were paired with classifiers, such as support vector machines, logistic regression, and naive bayes (Diederich et al., 2003; Aborisade and Anwar, 2018). This was followed by the use of deep learning architectures such as RNNs, LSTMs, and CNNs (Jafariakinabad et al., 2020). More recently, researchers utilized pretrained models – particularly BERT-based (Devlin et al., 2019) models, bringing improvements while removing the reliance on hand-crafted features (Rivera-Soto et al., 2021). A few recent studies have explored LLMs' capabilities in AA (Adewumi et al., 2025; Huang et al., 2024). For an overview on AA studies, we refer the readers to several survey papers (He et al., 2024; Huang et al., 2025; Medh and Sarma, 2025; Habib et al., 2025).

In the context of Arabic, Alqahtani and Dohler (2023) provide an Arabic-focused survey. The majority of reported efforts utilized stylometric features with ML approaches (Howedi and Mohd, 2014; Alanazi, 2015; Altakrori et al., 2018), with few studies investigating deep learning techniques (Ouamour and Sayoud, 2013; Al-Sarem and Emara, 2019). Transformer models have been explored in a few studies, covering limited domains in Classical Arabic and MSA, including poetry (El-Halees, 2022; Alqurashi et al., 2025) and Islamic law (AlZahrani and Al-Yahya, 2023). Our work represents a valuable step towards addressing a research gap in Arabic AA, introducing a new dataset and models spanning MSA and Dialectal Arabic.

2.3. Stylometric Features

Researchers have explored a wide range of features that characterize writing styles (Altheneyan and Menai, 2014; Alhafni et al., 2024; Nitu and Dascalu, 2024). Among stylometric categories that have been widely studied are surface, lexical, syntactic, as well as stylistic and aesthetic features. Surface features capture low-level attributes including sentence length and punctuation usage. Arabic-specific elements include elongation, hamzas, and the usage of diacritics. Lexical features include vocabulary usage, word frequency distributions, as well as vocabulary richness and complexity. Syntactic features assess the grammatical structure and morphosyntactic properties of the text, such as part-of-speech (POS) distribution, verb tenses, and narrative voice. Stylistic and aesthetic features encompass literary aspects of writing such as the use of vivid imagery, metaphor, humor, and poetic tone. In Arabic, the dialectness level also plays a crucial role in shaping stylistic expression, yet this has received limited attention in literature, which has primarily focused on Classical Arabic and MSA.

2.4. Authorship Attribution Human Studies

Previous studies have investigated human performance on AA tasks in comparison to trained classifiers. In Patel et al. (2022), participants were provided with example texts from two authors and asked to identify which author wrote a randomly sampled text across both authors. Humans were outperformed by a trained AA model (78% versus 84% accuracy). In another human evaluation study, Patel et al. (2024) presented annotators with a reference text from a target author along with two candidates; another text belonging to the same author and a style transfer output. The task was to

	Original Snippet (portion)	Translation
(a)	فقالت أمينة معربة عن نفس الإحساس: طبعًا، طبعًا يا سيدي، إنها صداقة العمر، وليست لهوًا ولعبًا.	"Amina said, expressing the same feeling, 'Of course, of course, sir. It is a friendship of a lifetime, not mere amusement or play.'
	عاوده التثاؤب مرة أخرى، فتمتم قائلًا: خذي المصباح خارجًا.	He yawned again and murmured, 'Take the lamp outside.'
	قامت أمينة لتنفيذ أمره فأغمض عينَيه قليلًا، ثم نهض دفعة واحدة كأنما ليُقاوم الكسل، واتجه نحو الفراش فاستلقى عليه …	Amina rose to carry out his order. He closed his eyes for a moment, then suddenly got up as if to resist his drowsiness, walked toward the bed, and lay down on it…"
(b)	قالت هنادي تستهبل :دانوه بعترفلك بشي....	Hanadi said jokingly, "Dano, I want to confess something to you…"
	قالت دانه بسرعه لأنها عارفه بلاوي بنت عمها :أقول أنطمي حنا عند الباب بالله اطلعي...	Dana replied quickly, since she knew the trouble her cousin could cause, "Hey, be quiet - we're at the door. Come on, get out."
	هنادي:هههههههههههههه جايه 5 mints	Hanadi: "Hahahahaha, coming in 5 minutes."
	قال خالد يستخف دمه :طيب ورى ماخليتيها تعترف	Khalid joked, "So why didn't you let her confess?"

Table 1: Portions of snippets from the A3D corpus, demonstrating differences in writing styles. These are drawn from (a) *Palace of Desire* by Naguib Mahfouz from the Hindawi dataset and (b) *Gather Me with Longing and Hold Me . . . Your Distance Has Scattered Me* by Amirat al-Ward from the Gumar dataset.

discriminate which of the two texts is written by the author. Humans were also outperformed by AA models on this task (37% versus 38–46% accuracy). Hallinan et al. (2023) investigated a different setup where they sampled 10 texts from each of 11 distinct writing styles. Annotators were trained on the AA task by providing them with descriptors of the 11 styles, in addition to 20 examples from each. The annotators were then asked to assign a style label to each of the 110 provided texts. Again, humans were significantly outperformed by a trained classifier (40% versus 85% accuracy). Collectively, these studies highlight how challenging style-based annotation can be for humans, with trained models consistently outperforming human judgments.

2.5. Text Style Transfer

TST is the task of modifying stylistic properties of text while preserving its meaning. Style transformations can be performed on various dimensions, including sentiment, formality, toxicity, and authorship (Jin et al., 2022). A key challenge in training TST models is the scarcity of parallel data. Recently, the high performance of LLMs on diverse tasks has inspired researchers to investigate prompting strategies for TST, utilizing LLMs' in-context learning capabilities (Patel et al., 2022; Mukherjee et al., 2024; Horvitz et al., 2024; Yang and Carpuat, 2025). However, this line of research remains limited, and the full extent of LLMs' capabilities in TST across different languages and TST tasks is not yet well understood. In this work, we focus on authorship style transfer, exploring LLMs' capabilities in Arabic, a setting that remains largely underexplored.

3. The Arabic Authorship Attribution Dataset (A3D)

Given that the majority of AA corpora in the literary domain consist of Classical Arabic literature, we construct the Arabic Authorship Attribution Dataset (A3D), a more contemporary dataset spanning MSA and Dialectal Arabic. For MSA, we utilize books from the Hindawi Foundation website,[2] collected by Elmadani et al. (2025a,b). From this collection, we use books from the novel and literature genres. For Dialectal Arabic, we use the Gumar corpus (Khalifa et al., 2016) containing Gulf Arabic forum novels collected from an online platform[3]. We provide the list of documents included in A3D and their respective authors in Appendix B. In Table 1, we present portions of two snippets from the A3D corpus, demonstrating stylistic differences.

In developing this corpus, we make several decisions to ensure effective training and robust evaluation, guided by preliminary experiments. For training, we sample text snippets from multiple novels per author to avoid training models that classify documents rather than authorial styles. For evaluation, we sample snippets from documents observed during training in addition to author snippets from unseen documents to assess model generalization.

Accordingly, we limit the dataset to the 40 authors having at least four documents, where 27 are sampled from Hindawi and 13 from Gumar. For each author, we utilize the top four documents with the highest word counts. Three documents are used for training, development and testing. From each document, we sample 60 train, 10 tuning, 10 development, and 10 test snippets. We refer to these as in-document evaluation. From the fourth document, we sample 10 tuning, 10 development, and 10 test snippets, referred to as cross-document evaluation. The tuning set is used to choose the best model checkpoints, the development set is used for hyperparameter tuning, and the test set is used as a blind evaluation set. The snippets are formed by adding document lines until a minimum threshold of 200 words is reached. The resulting snippets have an average length of 234 words (σ=56.3). Corpus statistics are provided in Table 2.

[2] https://www.hindawi.org/books/

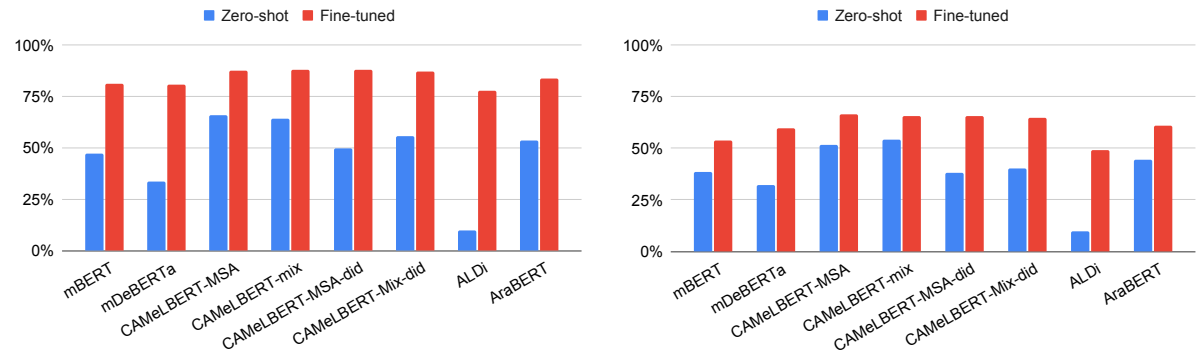
[3] https://www.graaam.com/

	Overall			Hindawi			Gumar		
	#snippets	#tokens	#words	#snippets	#tokens	#words	#snippets	#tokens	#words
Train	7,200	2,102,516	1,675,149	4,860	1,429,349	1,187,133	2,340	673,167	488,016
Tune	1,600	472,088	372,192	1,080	319,069	262,725	520	153,019	109,467
Dev	1,600	462,228	367,061	1,080	310,508	257,291	520	151,720	109,770
Test	1,600	464,834	367,286	1,080	313,821	258,299	520	151,013	108,987
Test in-doc	1,200	349,703	276,742	810	236,926	194,857	390	112,777	81,885
Test cross-doc	400	115,131	90,544	270	76,895	63,442	130	38,236	27,102

Table 2: A3D Corpus Statistics. We report the number of snippets, tokens, and words sampled from the Hindawi and Gumar resources for the train, development, and test sets. For the test set, we provide a breakdown of the statistics for each of the in-document and cross-document subsets.

	Macro Average			In-document			Cross-document		
	Acc	R@5	Rank	Acc	R@5	Rank	Acc	R@5	Rank
Zero-Shot Setting									
mBERT	42.8	75.1	4.59	47.0	77.8	4.10	38.5	72.5	5.09
mDeBERTa	32.8	63.9	5.76	33.6	67.3	5.52	32.0	60.5	6.01
CAMeLBERT-MSA	58.7	**85.6**	**3.06**	**65.7**	**89.8**	**2.44**	51.8	**81.5**	**3.68**
CAMeLBERT-Mix	**59.2**	84.3	3.12	64.3	88.7	2.53	**54.0**	80.0	3.71
CAMeLBERT-MSA-DID	43.9	76.8	4.25	49.8	82.8	3.57	38.0	70.8	4.93
CAMeLBERT-Mix-DID	48.0	80.1	4.02	55.8	85.3	3.15	40.3	75.0	4.89
ALDi	9.7	36.3	10.3	9.7	36.7	9.54	9.8	36.0	11.1
AraBERT	49.0	78.5	4.05	53.5	82.3	3.48	44.5	74.8	4.63
Fine-tuned Models									
mBERT	67.4	87.6	3.06	81.1	94.7	1.90	53.8	80.5	4.22
mDeBERTa	70.0	89.9	2.99	80.6	94.3	2.01	59.5	85.5	3.97
CAMeLBERT-MSA	**76.9**	91.0	2.80	87.6	97.3	1.66	**66.3**	84.8	3.95
CAMeLBERT-Mix	76.7	**91.5**	**2.39**	**87.9**	97.3	**1.57**	65.5	**85.8**	**3.22**
CAMeLBERT-MSA-DID	76.7	90.8	2.75	87.8	**97.5**	1.60	65.5	84.0	3.90
CAMeLBERT-Mix-DID	75.8	88.7	2.96	86.9	96.9	1.61	64.8	80.5	4.32
ALDi	63.4	84.8	3.52	77.6	93.8	2.30	49.3	75.8	4.75
AraBERT	72.4	87.9	2.79	83.8	95.8	1.77	61.0	80.0	3.80

Table 3: Results of zero-shot and fine-tuning approaches for AA across the different pretrained models on the test set, reporting accuracy (Acc), recall-at-5 (R@5), and mean rank (Rank). For each of the zero-shot and fine-tuning settings, we bold the best results. The overall best results are underlined.



(a) In-document Evaluation

(b) Cross-document Evaluation

Figure 1: Accuracy results achieved using zero-shot and fine-tuning approaches for AA across the different pretrained models on in-document and cross-document test sets.

4. Authorship Attribution Models

4.1. Modeling Approaches

We explore two modeling approaches: (1) zero-shot, where we rely on the representations learnt in pretrained models, and (2) fine-tuning models on the AA classification task while optimizing a dual cross-entropy and contrastive learning loss function. We elaborate on the approaches below.

We utilize multilingual and Arabic-focused pretrained transformer encoders. The models are chosen to include state-of-the-art models as well as models fine-tuned on tasks that are relevant to the AA task, including dialect identification and Arabic dialectness level estimation. We utilize mBERT [4] (Devlin et al., 2019), mDeBERTa[5] (He et al., 2023), CAMeLBERT-MSA,[6] CAMeLBERT-Mix,[7] CAMeLBERT-MSA DID NADI,[8] CAMeLBERT-Mix DID NADI[9] (Inoue et al., 2021a), ALDi (Keleg et al., 2023), and AraBERT[10] (Antoun et al., 2020).

Zero-shot Utilizing each pretrained model, we obtain the embeddings of the training text snippets belonging to each author and compute their mean to construct an author embedding. During evaluation, we obtain the embedding of each test snippet and compute its cosine similarity with all author embeddings. We assign the author label based on the closest author in the authorial embedding space.

Fine-tuning Pretrained encoder-only models, especially BERT-based architectures, have proven effective for AA (Alqurashi et al., 2025; Huang et al., 2025), leveraging their strong discriminative capabilities for this classification task. While fine-tuned AA models are typically optimized using cross-entropy loss, contrastive learning has also been explored to help models in learning more discriminative representations that capture authors' unique stylistic features (Rivera-Soto et al., 2021; Ai et al., 2022). This approach aims to create a latent embedding space where texts from the same author are closely clustered, while those from different authors remain well separated.

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

[5] https://huggingface.co/microsoft/mdeberta-v3-base

[6] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa

[7] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix

[8] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa-did-nadi

[9] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix-did-nadi

[10] https://huggingface.co/aubmindlab/bert-base-arabert

We utilize the work of Ai et al. (2022) where contrastive learning is integrated with cross-entropy fine-tuning. This combined objective has shown to be effective on several English AA datasets. The model is jointly optimized using both cross-entropy (L_{CE}) and contrastive learning (L_{CL}) loss functions as follows:

$$L = L_{CE} + \lambda L_{CL},$$

where λ (balancing coefficient) is set to 1. When training the AA models, we use AdamW optimizer (Loshchilov and Hutter, 2019) and a dropout of 0.35, following the implementation in Ai et al. (2022). We set the learning rate to $1e-5$ and train for 10 epochs. We provide further implementation details in Appendix A.

4.2. Results

We evaluate the models using accuracy, recall-at-5 (R@5), and mean rank. Recall-at-5 represents the proportion of cases where the correct author label is found among the top five predictions. We present the accuracy results in Figure 1 and the full results in Table 3. Comparing zero-shot to fine-tuned approaches, we report overall absolute improvements for fine-tuning across both in-document and cross-document evaluations of 22% and 12% on accuracy. Overall, we observe the superiority of CAMeLBERT models, where our best AA model is achieved by fine-tuning CAMeLBERT-Mix model, achieving accuracies of 87.9% and 65.5%, R@5 of 97.3% and 85.8%, and mean rank of 1.57 and 3.22 on in-document and cross-document evaluations, respectively.

5. Style Ranking Human Study

Setup We conduct a human evaluation study to further evaluate the AA model through a style ranking task. Given a reference text written by a specific author, the task is to rank seven other candidate text snippets based on their stylistic similarity to the reference. For the reference texts, we sample a random snippet from each author, while the seven candidates are selected to include a range of author, document, and corpus settings. This included variations in authorship (three same vs. four different authors from the reference text), document familiarity (sampled from in-document vs. cross-document subsets, covering seen and unseen documents during training), and corpus source (sampling from Hindawi vs. Gumar to match or differ from the reference). This sampling strategy ensures a balanced mix of stylistic similarity levels. All snippets are sampled from the test set and are truncated to the first 40 words to keep the task manageable for humans. We demonstrate an example of the task in Appendix C.

Model	Hindawi+Gumar		Hindawi		Gumar	
	QWK	%Same Author in Top-3	QWK	%Same Author in Top-3	QWK	%Same Author in Top-3
AA Model	**0.75**	**88.3**	**0.73**	**86.4**	**0.79**	**92.3**
GPT	0.42	57.1	0.41	55.6	0.43	60.3
DeepSeek	0.29	47.9	0.35	50.0	0.18	43.6
Gemini	0.24	52.1	0.26	51.9	0.20	52.6
Llama	0.13	45.4	0.12	43.8	0.15	48.7
Fanar	0.12	45.4	0.08	43.8	0.20	48.7
Jais	0.11	46.2	0.09	44.4	0.17	50.0

Table 4: Results of the style ranking task. We report the average Quadratic Weighted Kappa (QWK) between the rankings obtained from both annotators and each model, covering our best AA model and the LLMs. We also report the percentage of candidate snippets written by the same author as the reference snippet that are assigned a ranking of 1-3 (% Same Author in Top-3).

We provide the annotators with annotation guidelines, explaining key stylistic attributes that should be considered when ranking, including surface features, lexical features, syntactic features, level of dialectness, and stylistic and aesthetic features. The annotators were instructed to focus on style, not on topic or content. They were also asked, if unsure, to consider which snippet could have been written by the same author as the reference text. The annotation guidelines are provided in Appendix D.

We recruited two native Arabic speakers with prior writing experience.[11] The annotators ranked the same set of 80 samples, each consisting of one reference and seven candidate snippets, resulting in a total of 1,120 ranking annotations (2 annotators × 80 samples × 7 snippets). We also asked the annotators to provide explanatory justifications for their rankings, which we include in the release.[1]

Inter-annotator Agreement (IAA) We report a high agreement between annotators, where the Quadratic Weighted Cohen's Kappa (Cohen, 1968) score is 0.80 (substantial agreement).

Performance Analysis We assess the ability of annotators and our best AA model at ranking the three snippets belonging to the same reference author as the top three similar snippets. We find that the AA model outperforms humans, with successfully identifying 88% of these snippets compared to 75% and 77% achieved by the two annotators. The superiority of the AA model on this task is in-line with previous studies outlined in Section 2.4.

[11]The annotation task was conducted on Upwork, a crowdsourcing platform. The task took an average of 20 hours from each annotator, with a compensation of $240, which is in accordance with their hourly rates.

6. Benchmarking LLMs

We benchmark multilingual and Arabic-focused LLMs on style ranking and authorship style transfer tasks. We evaluate the following models: GPT-4.1 (Achiam et al., 2023), DeepSeek-Chat (Liu et al., 2024a), Gemini 2.5 Flash (Comanici et al., 2025), Llama 3 (Grattafiori et al., 2024), Fanar (Team et al., 2025), and Jais (Sengupta et al., 2023).[12]

6.1. Style Ranking Task

We evaluate LLMs on the same style ranking setup as the human annotation study. The prompt is constructed following the same instructions provided in the annotation guidelines (prompt provided in Appendix E). To assess alignment with human judgments, we report Quadratic Weighted Kappa between human rankings and those generated by our best AA model and the LLMs. Kappa scores are computed separately for each annotator and the average is reported. For the AA model, rankings are based on cosine similarity between embeddings of reference and candidate snippets obtained using the model.

We present the results in Table 4. Overall, LLMs perform poorly, with GPT scoring highest, followed by DeepSeek. In terms of ranking snippets written by the same reference author in the top three, GPT, DeepSeek, and Gemini perform best, with the other LLMs scoring close to random (42.9%). Comparing results across Hindawi and Gumar subsets, we report higher performance on Gumar for the majority of models. We believe the informal nature of Gumar may have provided easier surface-level stylistic cues, such as the overuse of punctuation and letter repetition.

[12]We utilize deepseek-chat, gemini-2.5-flash, Meta-Llama-3.1-8B-Instruct, QCRI/Fanar-1-9B-Instruct, and inceptionai/jais-family-13b-chat models.

		Macro Average			In-document			Cross-document		
Input Text	**BERTScore**	**Acc**	**R@5**	**Rank**	**Acc**	**R@5**	**Rank**	**Acc**	**R@5**	**Rank**
Original Text										
Original Text		76.7	91.5	2.39	87.9	97.3	1.57	65.5	85.8	3.22
Neutralizing Style: Original → Neutral										
Neutral_Rewrite	80.9	25.7	54.3	9.01	34.1	61.9	7.60	17.3	46.8	10.43
Neutral_MT	78.3	19.0	45.9	11.13	24.7	51.0	9.62	13.3	40.8	12.64
Authorship Style Transfer: Neutral_MT → Styled										
GPT	76.9	42.3	68.8	5.97	51.1	73.8	5.13	33.5	63.8	6.81
DeepSeek	**78.5**	46.6	75.0	5.11	54.3	80.3	4.32	**39.0**	69.8	5.90
Gemini	77.5	**48.1**	**76.7**	**4.91**	**58.3**	**82.1**	**3.93**	38.0	**71.3**	**5.88**
Llama	74.4	27.2	52.8	9.57	33.1	56.1	8.53	21.3	49.5	10.62
Fanar	70.2	33.8	57.3	9.00	38.6	61.4	7.95	29.0	53.3	10.05
Jais	68.3	13.5	34.1	14.30	16.7	37.4	13.36	10.3	30.8	15.24

Table 5: Results of authorship style transfer. We present the results of our best AA model on the original texts as reported in Table 3 followed by the results achieved by inferencing the model on the neutralized and styled texts. We evaluate style transfer accuracy using accuracy (Acc), recall-at-5 (R@5), and mean rank (Rank). We also report BERTScore (F1) for the neutralized and styled texts against the original texts.

6.2. Authorship Style Transfer Task

To evaluate LLMs on authorship style transfer, we first neutralize the writing style of the text snippets in the test set using GPT-4.1. Afterwards, we prompt the LLMs to rewrite each neutralized text to mimic the style of its original author.

Style Neutralizing Step We explore two approaches for neutralizing writing style, as outlined in the prompt templates in Figure 2. In the first approach, we prompt GPT-4.1 to rewrite the text in simple style. In the second approach, we apply a translation pipeline to deviate further from the original style. We first translate the original Arabic text to English, run a rewrite prompt on the translated text, then translate back to Arabic.

Style Transfer Step For style transfer, we use the prompt outlined in Figure 3, following Horvitz et al. (2024). As outlined in the prompt, for each snippet in the test set, we provide an author reference snippet. This snippet is sampled from the text snippets in the training set belonging to the same author. The reference snippets are used consistently when prompting different LLMs.

6.2.1. Authorship Style Transfer Results

To evaluate the style transfer outputs, following previous work (Mukherjee et al., 2025), we assess *style transfer accuracy* and *content preservation*. For style transfer accuracy, a common approach is to use a trained classifier to evaluate whether the style transfer output reflects the intended style (Krishna et al., 2020; He et al., 2020). Accordingly, we utilize our best AA model, reporting accuracy, recall-at-5, and mean rank of the neutralized and styled texts. For content preservation, we report BERTScore (Zhang et al., 2020). Results are shown in Table 5.

Prompt: Neutralizing Style - Rewrite

```
Text: {input_text}
Paraphrase the following text in a
simple neutral style in Arabic.
No information should be lost in
paraphrasing.
Rewrite:
```

Prompt: Neutralizing Style - MT

```
Prompt #1:
Translate the following text to
English:
{input_text}

Prompt #2:
Text: {input_text}
Paraphrase the following text in a
simple neutral style in English. No
information should be lost in
paraphrasing.
Rewrite:

Prompt #3:
Translate the following text to
Arabic:
{input_text}
```

Figure 2: Style neutralization prompts.

Prompt: Authorship Style Transfer

```
The following text is written by a
single author:
{author_reference_text_snippet}

Rewrite the following text to make it
look like the above author's style.
Only provide the rewritten text
without explanation or extra text.
Rewrite the following:
{neutralized_text_snippet}
```

Figure 3: Prompt used for authorship style transfer.

For **style neutralization**, we opt for the translation pipeline as it results in a higher accuracy reduction (from 76.7% to 19.0%), with slight effect on BERTScore compared to the rewrite prompt. For **authorship style transfer**, Gemini achieves the highest accuracy with a 29% absolute improvement over the neutralized text, followed by DeepSeek (28%), GPT (23%), Fanar (15%) and Llama (8%). For Jais, we report deterioration in accuracy. In terms of content preservation, we observe strong performance by Deepseek and Gemini, followed by GPT then Llama. Fanar and Jais show lower performance, with absolute BertScore reductions of 8% and 10%, respectively.

6.2.2. Stylometric Analysis of TST Output

To gain insights into the style transfer outputs, we analyze their similarity to the author reference texts provided in the prompts in terms of stylometric features. We represent each text snippet as a vector of linguistic attributes spanning the following features.

Surface Features We cover these five attributes: the *number of tokens per line*, the *percentage of words, punctuation marks, and digits*, as well as the *ratio of diacritics count over Arabic words*.

Syntactic Features Given that Arabic is a morphologically rich language (Habash, 2010), we capture morphological features in Arabic words covering: *inflectional and cliticization morphology*, *POS tags*, and *morphological richness*. All syntactic features are obtained using CAMeL Tools BERT-based model (Obeid et al., 2020; Inoue et al., 2021b). For inflectional morphology, we cover the eight inflectional features in Arabic: aspect, mood, person, voice, case, state, gender, and number. For cliticization morphology, we include the different types of proclitics and enclitics. We also include morphological richness, calculated as the average number of morphomes per word. In total, syntactic features are represented with 160 attributes.

Lexical Features We include two main features: *token type ratio* and *readability level*. The token type ratio is defined as the number of unique words divided by the total number of words. The readability level of texts is assessed using two approaches: (1) the readability level estimation model[13] provided by Elmadani et al. (2025a), where the authors define 19 readability levels, ranging from kindergarten to postgraduate comprehension and (2) SAMER lexicon (Al-Khalil et al., 2020), where the authors developed a lexicon of Arabic lemmas with an assigned readability level on a five-level scale. The SAMER readability level of a snippet is calculated by averaging the levels of mapped Arabic words, where mapping is performed using the word's undiacritized lemma and associated POS tag. Lexical features are represented using three attributes.

Language Choice We cover three features with regards to language choice: *Arabic variant*, *dialectness level*, and *percentage of foreign words*. For the Arabic variant, we utilize the Dialect Identification system provided in CAMeL Tools (Salameh et al., 2018). The model provides probabilities of the text belonging to MSA and 25 Arabic city dialects, which we utilize as separate attributes. We also analyze the texts in terms of dialectness level, defined as the extent by which a sentence diverges from MSA. For this, we utilize the ALDi model developed by Keleg et al. (2023) that quantifies the level of dialectness of text on a continuum from 0 (MSA) to 1 (high colloquialism). Finally, given the prevalence of code-switching in the Arab world (Hamed et al., 2025), we also include the percentage of foreign words, as few code-switching cases are present across snippets. Language choice is thus represented with a total of 28 attributes.

Analysis Results We report cosine similarities between the author reference texts used in the style transfer prompts and the neutralized and styled outputs in Table 6, using vectors constructed for each of the four stylometric features. **Surface** and **lexical** features show minimal change with style neutralization, with cosine similarities exceeding 0.999 between both the original and neutralized texts, and the reference and neutralized texts. For **surface** features, GPT, Gemini, and Llama preserve this similarity, while the other models show slight declines. For **lexical** features, GPT, Gemini, and Deepseek preserve the high similarity, while Llama and Fanar show slight declines, and greater declines are shown by Jais. For **syntactic** features, relative improvements are only achieved by Gemini (25.2%), GPT (23.0%), and Deepseek

[13] https://huggingface.co/CAMeL-Lab/readability-arabertv02-word-CE

	Overall				Hindawi				Gumar			
	Surface	**Syntactic**	**Lexical**	**Lang.**	**Surface**	**Syntactic**	**Lexical**	**Lang.**	**Surface**	**Syntactic**	**Lexical**	**Lang.**
Neutral	0.999	0.992	0.999	0.770	1.000	0.994	0.999	0.958	0.999	0.989	0.999	0.381
GPT	1.000	0.994	0.999	0.951	1.000	0.995	0.999	0.970	0.999	0.992	0.999	0.911
DeepSeek	0.998	0.994	0.999	0.939	1.000	0.995	0.999	0.965	0.996	0.991	0.999	0.885
Gemini	1.000	0.994	0.999	0.950	1.000	0.995	0.999	0.974	0.999	0.992	0.999	0.901
Llama	1.000	0.991	0.998	0.795	1.000	0.993	0.999	0.963	0.999	0.985	0.995	0.445
Fanar	0.999	0.991	0.998	0.823	0.999	0.994	0.998	0.956	0.998	0.987	0.997	0.546
Jais	0.998	0.941	0.955	0.693	0.998	0.938	0.950	0.913	0.998	0.948	0.966	0.236

Table 6: Cosine similarities between the author reference texts provided in authorship style transfer prompts and the neutralized and styled texts by LLMs across stylometric features.

(15.2%). The largest gains appear in the **language choice** feature, where the largest transformations are achieved through style neutralization. On Hindawi, relative improvements are achieved by Gemini (37.7%), GPT (29.8%), Deepseek (16.9%), and Llama (13.2%), while slight and greater declines are reported for Fanar and Jais, respectively. On Gumar, the effect of style transfer is even more pronounced. All LLMs yield improvements except Jais. GPT achieves the highest relative improvement (85.7%), followed by Gemini (84.0%), DeepSeek (81.5%), Fanar (26.7%), and Llama (10.4%). Overall, only GPT and Gemini perform consistently well across all features. While this analysis provides insights into the style transfer outputs, the analyzed features are limited by the available models and tools (especially for Dialectal Arabic), and does not fully capture all stylistic dimensions.

7. Conclusion and Future Work

In this work, we focused on authorship attribution and authorship style transfer in Arabic, a language that remains underexplored in stylistic NLP research. For AA, we introduced a new benchmark spanning Modern Standard and Dialectal Arabic and demonstrated the effectiveness of transformer-based models fine-tuned with a dual cross-entropy and contrastive learning objective. To assess the ability of LLMs in understanding and generating Arabic writing styles, we used the train authorship attribution model to benchmark a range of LLMs on style ranking and authorship style transfer tasks. While GPT, Gemini, and DeepSeek showed relative strength, overall results indicate that current LLMs' performance remains limited in capturing Arabic stylistic nuances. We release all resources to support further research in this direction.

Limitations

While this work provides new resources and insights into Arabic authorship attribution and style transfer, the study has some limitations. First, the stylistic analysis of LLM-generated style transfer outputs relies on Arabic NLP tools. This constrains the range of covered stylometric features and also introduces potential inaccuracies tied to the performance of the tools. Second, given that LLMs are sensitive to prompt design, the prompts used may not be fully representative of the capabilities of all the investigated LLMs. Third, although our dataset extends existing authorship attribution resources with more contemporary texts spanning Modern Standard Arabic and Dialectal Gulf Arabic, further research is needed to explore a wider range of authors, dialects, and genres.

Ethics Statement

This work involves authorship attribution and style transfer, which may raise concerns around privacy, misuse, or impersonation. To mitigate such risks, we focus exclusively on publicly available literary texts and avoid personal or sensitive data. Our dataset is constructed from two sources: books published by the Hindawi Foundation and the Gumar corpus of Arabic forum novels. Within the Hindawi collection, 59% of the books are released under open licenses, while the rest remain under copyright held by their respective authors. In our dataset, we only utilize a subset of the original documents, with sampled snippets comprising 234 words on average (σ=56.3). The snippets amount to 36% (σ=19.3) of the documents on average (Hindawi) and 16% (σ=9.2) of the documents on average (Gumar). In order to avoid potential copyright concerns, we distribute the dataset in encrypted form generated using the Muddler tool developed by CAMeL Lab.[14] Access to the dataset thus relies on the availability of the original source documents, for which we provide web archive links in our release.[1] The human study for the style ranking task was conducted through Upwork, where annotators were compensated in accordance with their hourly rates.

[14] https://github.com/CAMeL-Lab/muddler

8. Acknowledgements

We thank Khalid N. Elmadani and Salam Khalifa for their efforts in collecting the datasets used in this work and Nadine El-Naggar for her helpful discussions. We also thank the reviewers for their valuable comments and constructive feedback.

9. Bibliographical References

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *Proceedings of IRI*, pages 269–276.

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

Tosin Adewumi, Nudrat Habib, Lama Alkhaled, and Elisa Barney. 2025. On the limitations of large language models (LLMs): False attribution. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 11–21.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? learning to contrast for authorship attribution. In *Proceedings of AACL-IJCNLP*, pages 1142–1157.

Mohammed Al-Sarem and Abdel-Hamid Emara. 2019. The effect of training set size in authorship attribution: application on short Arabic texts. *International Journal of Electrical and Computer Engineering*, 9(1):652.

Mohammed Al-Sarem, Faisal Saeed, Abdullah Alsaeedi, Wadii Boulila, and Tawfik Al-Hadhrami. 2020. Ensemble methods for instance-based Arabic language authorship attribution. *IEEE Access*, 8:17331–17345.

Saad Alanazi. 2015. Classical Arabic authorship attribution using simple features. *Natural Language Processing and Cognitive Science*, 4551.

Jafar Albadarneh, Bashar Talafha, Mahmoud Al-Ayyoub, Belal Zaqaibeh, Mohammad Al-Smadi, Yaser Jararweh, and Elhadj Benkhelifa. 2015. Using big data analytics for authorship authentication of Arabic tweets. In *Proccedings of the IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, pages 448–452.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101.

Fatimah Alqahtani and Mischa Dohler. 2023. Survey of authorship identification tasks on Arabic texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–24.

Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. 2025. BERT-based classical Arabic poetry authorship attribution. In *Proceedings of COLING*, pages 6105–6119.

Haroon Nasser Alsager. 2020. Towards a stylometric authorship recognition model for the social media texts in Arabic. *Arab World English Journal*, 11(4):490–507.

Malik H Altakrori, Farkhund Iqbal, Benjamin CM Fung, Steven HH Ding, and Abdallah Tubaishat. 2018. Arabic authorship attribution: An extensive study on twitter posts. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(1):1–51.

Alaa Saleh Altheneyan and Mohamed El Bachir Menai. 2014. Naïve bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 26(4):473–484.

Fetoun Mansour AlZahrani and Maha Al-Yahya. 2023. A transformer-based approach to authorship attribution in classical Arabic texts. *Applied Sciences*, 13(12):7255.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, José Javier Saiz, Robiert Sepúlveda-Torres, et al. 2025. Iberobench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of COLING*, pages 10491–10519.

Victor Benjamin, Wingyan Chung, Ahmed Abbasi, Joshua Chuang, Catherine A Larson, and Hsinchun Chen. 2013. Evaluating text visualization: An experiment in authorship analysis. In *Proceedings of the International Conference on Intelligence and Security Informatics*, pages 16–20.

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. *arXiv preprint arXiv:2005.14714*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence*, 19(1):109–123.

Alaa M El-Halees. 2022. Arabic poetry authorship attribution and verification using transfer learning. *Egyptian Computer Science Journal*, 46(1).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Nudrat Habib, Tosin Adewumi, Marcus Liwicki, and Elisa Barney. 2025. Trends and challenges in authorship analysis: A review of ML, DL, and LLM approaches. *arXiv preprint arXiv:2505.15422*.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In *Findings of EMNLP*.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of COLING*.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *Proceedings of ICLR*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of ICLR*.

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

David I Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen Mckeown, and Zhou Yu. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of EMNLP*, pages 13376–13390.

Fatma Howedi and Masnizah Mohd. 2014. Text classification for authorship attribution using naive bayes classifier with limited training data. *computer engineering and intelligent systems*, 5(4):48–56.

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

Weihang Huang, Akira Murakami, and Jack Grieve. 2024. ALMS: Authorial language models for authorship attribution. *arXiv preprint arXiv:2401.12005*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021a. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Arabic Natural Language Processing Workshop*, pages 92–104.

Go Inoue, Salam Khalifa, and Nizar Habash. 2021b. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of ACL*, pages 1708–1719.

Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A Hua. 2020. Syntactic neural model for authorship attribution. In *FLAIRS*, pages 234–239.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of EMNLP*, pages 10597–10611.

Heba M Khalil, Taha Ahmed, and Tarek El-Shistawy. 2020. Authorship authentication of political Arabic articles based on modified TF-IGF algorithm. *Journal of Theoretical and Applied Information Technology*, 98(17):3575–3583.

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

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of EMNLP*, pages 737–762.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024b. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of LREC-COLING*, pages 15285–15295.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Smriti Priya Medh and Shikhar Kumar Sarma. 2025. Automatic author attribution of different languages: A review. *Edelweiss Applied Science and Technology*, 9(2):1245–1259.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of COLING*, pages 4186–4218.

Sourabrata Mukherjee, Atul Kr Ojha, and Ondřej Dušek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the International Natural Language Generation Conference*, pages 523–539.

Sourabrata Mukherjee, Atul Kr Ojha, John Philip McCrae, and Ondřej Dušek. 2025. Evaluating text style transfer evaluation: Are there any reliable metrics? In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 418–434.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP-IJCNLP*, pages 188–197.

Melania Nitu and Mihai Dascalu. 2024. Authorship attribution in less-resourced languages: A hybrid transformer approach for romanian. *Applied Sciences*, 14(7):2700.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of LREC*, pages 7022–7032.

Siham Ouamour and Halim Sayoud. 2012. Authorship attribution of ancient texts written by ten Arabic travelers using a smo-svm classifier. In *Proceedings of the International Conference on Communications and Information Technology (ICCIT)*, pages 44–47.

Siham Ouamour and Halim Sayoud. 2013. Authorship attribution of ancient texts written by ten Arabic travelers using character n-grams. In *Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer: Can non-famous authors be imitated? *arXiv preprint arXiv:2212.08986*.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2024. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. *arXiv preprint arXiv:2410.12757*.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of ACL*, pages 837–848.

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 EMNLP*, pages 913–919.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of COLING*, pages 1332–1344.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

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

Kareem Shaker and David Corne. 2010. Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In *Proceedings of the UK Workshop on Computational Intelligence (UKCI)*, pages 1–6.

Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov. 2023. Authorship attribution of late 19th century novels using gan-bert. In *Proceedings of ACL (Volume 4: Student Research Workshop)*, pages 310–320.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of ACL*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2021. Writing style author embedding evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2023. Valla: Standardizing and benchmarking authorship attribution and verification through empirical evaluation and comparative analysis. In *Proceedings of IJCNLP-AACL*.

Xinchen Yang and Marine Carpuat. 2025. Steering large language models with register analysis for arbitrary style transfer. *arXiv preprint arXiv:2505.00679*.

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

10. Language Resource References

Muhamed Al-Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for standard Arabic. In *Proceedings of LREC*, pages 3053–3062.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, pages 830–839.

Khalid Elmadani, Nizar Habash, and Hanada Taha. 2025a. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of ACL*, pages 16376–16400.

Khalid N. Elmadani, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025b. BAREC shared task 2025 on Arabic readability assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 239–252.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of LREC*, pages 4282–4289.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*, pages 217–226.

A. Authorship Attribution Modeling Details

When training the authorship attribution models, each experimental setup is run once, where training takes approximately an hour on a RTX 2000 Ada GPU. The number of parameters of the trained models are provided in Table 7.

Model	#Parameters
mBERT	278,537,768
mDeBERTa	378,903,080
CAMeLBERT-MSA	209,765,672
CAMeLBERT-Mix	209,765,672
CAMeLBERT-MSA-DID	209,765,672
CAMeLBERT-Mix-DID	209,765,672
ALDi	263,525,672
AraBERT	235,877,672

Table 7: Number of parameters of trained AA models.

B. Authors and Documents

The following table presents the list of authors and documents included in A3D.

Author Name		Document Titles
		Hindawi
Naguib Mahfouz	نجيب محفوظ	أولاد حارتنا، الحرافيش، بين القصرين، قصر الشوق
Yusuf Idris	يوسف إدريس	أرخص ليالي، البيضاء، جمهورية فرحات، مدينة الملائكة
Tharwat Abaza	ثروت أباظة	ثم تشرق الشمس، خيوط السماء، قصر على النيل، هارب من الأيام
Marun Abboud	مارون عبود	أدب العرب، جدد وقدماء، رواد النهضة الحديثة، سبل ومناهج
Abdel-Ghaffar Makkawi	عبد الغفار مكاوي	بكائيات، ثورة الشعر الحديث من بودلير إلى العصر الحاضر (1)، ثورة الشعر الحديث من بودلير إلى العصر الحاضر (2)، شعر وفكر
Nawal El Saadawi	نوال السعداوي	إنه الدم، رحلاتي في العالم، زينة، كسر الحدود
Muhammad Farid Abu Hadid	محمد فريد أبو حديد	أزهار الشوك، أنا الشعب، الوعاء المرمري، مع الزمان
Zaki Mubarak	زكي مبارك	البدائع، المدائح النبوية في الأدب العربي، مدامع العشاق، وحي بغداد
Mustafa Sadiq al-Rafi'i	مصطفى صادق الرافعي	إعجاز القرآن والبلاغة النبوية، تحت راية القرآن، كتاب المساكين، وحي القلم
Muhammad Sa'id al-Aryan	محمد سعيد العريان	بنت قسطنطين، شجرة الدر، على باب زويلة، من حولنا
Abdelaziz Baraka Sakin	عبد العزيز بركة ساكن	الجنقو مسامير الأرض، الطواحين، رماد الماء، مسيح دارفور
Ali al-Jarim	علي الجارم	الشاعر الطموح، سيدة القصور، غادة رشيد، هاتف من الأندلس
Salama Moussa	سلامة موسى	الأدب للشعب، برنارد شو، عقلي وعقلك، كيف نربي أنفسنا
May Ziadeh	مي زيادة	المساواة، بين الجزر والمد، رجوع الموجة، سوانح فتاة
Maher al-Battouti	ماهر البطوطي	الرواية الأم، بين الفن والأدب، روايات وروائيون من الشرق والغرب، قاموس الأدب الأمريكي
Taha Hussein	طه حسين	ألوان، صحف مختارة من الشعر التمثيلي عند اليونان، صوت باريس، من بعيد
Abbas Mahmoud al-Aqqad	عباس محمود العقاد	دراسات في المذاهب الأدبية والاجتماعية، ساعات بين الكتب، يسألونك، يوميات
Ibrahim Abdel Qader al-Mazni	إبراهيم عبد القادر المازني	إبراهيم الثاني، أحاديث المازني، صندوق الدنيا، في الطريق
Jurji Zaydan	جرجي زيدان	أبو مسلم الخراساني، أسير المتمهدي، فتاة غسان، فتح الأندلس
Ibrahim Zaydan	إبراهيم زيدان	النوادر المطربة، نوادر الأدباء، نوادر العشاق، نوادر الكرام
Shawqi Abdel Hakim	شوقي عبد الحكيم	الحكايات الشعبية العربية، السير والملاحم الشعبية العربية، سيرة الملوك التَبَاعِنَة، مدخل لدراسة الفولكلور والأساطير العربية
Muhammad Husayn Haykal	محمد حسين هيكل	زينب، في أوقات الفراغ، في منزل الوحي، هكذا خلقت
Zaki Naguib Mahmoud	زكي نجيب محمود	جنة العبيط، شروق من الغرب، قصة نفس، وجهة نظر
Ahmad Amin	أحمد أمين	فيض الخاطر (1)، فيض الخاطر (6)، فيض الخاطر (7)، قاموس العادات والتقاليد والتعابير المصرية
Nicola Haddad	نقولا حداد	آدم الجديد، الصديق المجهول، حواء الجديدة، فاتنة الإمبراطور
Ahmad Taymur Pasha	أحمد تيمور باشا	الأمثال العامية، الحب والجمال عند العرب، الكنايات العامية، رسالة لغوية عن الرتب والألقاب المصرية
Sonallah Ibrahim	صنع الله إبراهيم	برلين ٦٩، بيروت بيروت، ذات، يوميات الواحات
		Gumar
Baqaya Shatat	بقايا شتات	احلى صدفة بحياتي (2)، احلى صدفة بحياتي (3)، الا ليت القدر (1)، الا ليت القدر (2)
Amirat al-Ward	اميرة الورد	جروح قلبي، لمني بشوق و أحضني....بعادك عني بعثرني (1)، لمني بشوق و أحضني....بعادك عني بعثرني (2)، و طال انتظاري
Anjal	أنجال	أعد لي هويتي، في صمتي كلام، وش الله مكلفني بحبن تاليه فراق، وش مكلفني بحبن تاليه فراق
Al-Jouri	الجوري	بقراري ضيعت نفسي وأحبابي، ذنبي إني لقيط، غفلتي دفعتني حياتي، والله لا اجيب راسك (1)
The Salt of Life and Its Sugar	ملح الحياة وسكرها	شرعك اللهم ولااعتراض (1)، شرعك اللهم ولااعتراض (3)، كل يوم لك حال جديد .. مره قريب و مره بعيد (1)، من أعماق أحضان مجنونة (1)
My Heart Is Enchanted	مفتون قلبي	أنجبرت فيك وماتوقعت أحبك و أموت فيك، رماني حظي العاثر عليه.. وقلت ما أبيه وأتاريني ميته فيه، و اخيرا حبينا بعض (1)، واخيرا حبينا بعض (2)
Julie	جولي	أضمك وين من حسادي و أحميك حبيبي بقلبي لو بعيوني أخليك (1)، أضمك وين من حسادي و أحميك حبيبي بقلبي لو بعيوني أخليك (2)، مجبور احبك دام عمري ومصيري انكتب معاك، باحظ عينك تنام اليل مرتاحة
The Flower Seller	بائعة الورد	خنفت الورد يآ يمه وبيدي انكسر ذبلآن (1)، خنفت الورد يا يمه وبيدي إنكسر ذبلان (2)، ما رجيت من الفرح إلا رضاك و ما بكيت إلا علشانك قهر (1)، ما رجيت من الفرح إلا رضاك و ما بكيت إلا علشانك قهر (2)
Zuhur Hussein	زهور حسين	يا خاطفي وين ألقى عزتي في زمان المذله (1)، يا خاطفي وين القى عزتي في زمان المذلة (2)، يا خاطفي وين القى عزتي في زمان المذلة (3)، يا خاطفي وين القى عزتي في زمان المذلة (4)
My Mother's Heart and Soul	قلب امي وروحها	أنين , كره , إنتقام , حب ,نهاية مجهولة لعبة الشطرنج، توأمي لكن انا غير ومالي مثيل وراسي عنيده (1)، توأمي لكن انا غير ومالي مثيل وراسي عنيده (2)، شامخة و بنت رجال و جاء من يغلبني
Wedeema al-Ata	وديمة العطا	فرسان على جمر الغضى (1)، فرسان على جمر الغضى (2)، فرسان على جمر الغضى (3)، نور
Hold Me Between the Eyelashes	ضمني بين الاهداب	بعض العيون حقدها في نظرها (1)، بعض العيون حقدها في نظرها (2)، ثرثرة أرواح متوجعة (1)، ثرثرة أرواح متوجعة (2)
Omani and Proud	عمانية وأفتخر	حبه بصدري سالفه سر في بير، طفلة أحلامي (1)، طفلة أحلامي (2)، طفلة أحلامي (3)

C. Style Ranking Task Example

In Table 8, we provide an example of the style ranking task.

	Texts
Ref	وأجاب صاحبي: «أنت يا سيدتي التي أوحيت إلى القمر كل هذا الشعر الذي يوقع لنا الليلة أنغامه، وسترينه على سفح الأهرام، وعلى وجه أبي الهول أروع شعرًا وأبدع إيقاعًا بفضل وحيك وإلهامك.» واتصل بيننا بعد ذلك حديث رقيق حرصت ما [And my friend responded: <<It is you, my lady, who inspired the moon with all this poetry that tonight sings its melodies for us. You shall see it etched upon the slope of the pyramids, and upon the face of the Sphinx—most splendid poetry and most exquisite rhythm, thanks to your inspiration.>>]
C1	لسنا نَدَّعي إقامة المقارنة بين الشاعرين الكبيرين في هذه الكلمة. ومن شاء الوقوف عليها فليرجع إلى الصحف المختارة ويرى تطور العصر والموازنة بين مختلف ما كتب كل منهما. وإنما نريد أن نشير إلى أن شعرهما جميعًا بلغ من السمو والعظمة [We do not presume, in this brief word, to draw a comparison between the two great poets. Whoever wishes to delve into such a comparison may turn to the selected journals and trace the evolution of the era, weighing the diverse writings of each. What we seek only to affirm is that the poetry of both has reached heights of nobility and grandeur]
C2	فقال أحمد: أنا يا حضرة العمدة لا أصلح للزواج. فيقول العمدة ساخطًا: لعن الله الزواج وسني الزواج، اسمع يا ولد، أقسم بالله العلي العظيم، إن سمعت أنك ذهبت إلى الحارة التي فيها سعدية لأقطعن أسبابك بالقرية جميعًا. أتسمع؟ ويرتجف أحمد [So Ahmed said: Sir, I am not fit for marriage. The mayor replied angrily: Curse marriage and the age of marriage! Listen, boy—I swear, if I hear that you've gone to the alley where Saadia lives, I'll sever all your ties to this village. Do you hear me? And Ahmed trembles]
C3	قلت ملقية الكلام إلى الحاضرين من غير أن أوجهه إلى أحد بذاته: «والغيرة، ألها صلة بالحب؟ أم أنها مستقلة عنه قائمة بذاتها؟» قالت الأمريكية وكأنما حرك هذا السؤال عندها شجنًا دفينًا: «غيرة المرأة عاطفة طبيعية باعثها الدفاع عن النفس، وعن [I said, casting the words to those present without addressing anyone in particular: <<And jealousy—does it have a connection to love? Or is it an independent feeling, standing on its own?>> The American woman said, as though the question had stirred a buried sorrow within her: <<A woman's jealousy is a natural emotion, born out of the instinct to defend herself, and to...]
C4	لكن الأيام المملوءة بالعمل الجد، وأحلامه الطويلة للمستقبل، جعلت تقضي على هذه الفكرة رويدًا رويدًا، وأصبح الوجود الذي كان يتخيله من قبل معطرًا بالزهور وبسكرات الحب وجودًا هادئًا ساكنًا ألذ ما فيه العمل والفكر، وانهمك بكله في مطالعات مختلفة بلغت [But the days, filled with earnest work and his long dreams of the future, gradually began to wear down that idea, little by little, and the life he once imagined as scented with flowers and love turned into a calm, quiet existence, its sweetest pleasures being work and thought, and he completely got immersed in varied readings that reached...]
C5	: لا حبيبي بس في وحده بتخطب زوجته ومو مصدقه انها متزوجه قلت أعطيها رقمه تتأكد ماجد وما مشت عليه الكذبه وإنتي وش دخلك فيهم بقلعتها لا صدقت همست بزعل لا تدرني حبيبي قبل لا أنام بقولك كل شي ماجد [No, my love, it's just that there's a woman talking to his wife and doesn't believe she's married, I thought I'd give her his number so she checks Majid and he didn't fall for the lie, and you—why are you getting involved? Let her be." I whispered sadly, don't be upset with me, my love, before I sleep, I'll tell you everything. Majid…]
C6	تساءل ملك بعلبك: من يدري؟ فقد نخسر كل شيء في حربنا مع اليمنيين، وما عرف عنهم من بأس وجلد على المنازلة والقتال. وأعاد الجميع إلى الأذهان سيرة تبعهم حسان اليماني وتجبره في حروبه مع «آل مرة» والتغلبيين، وما أقدمت عليه [The King of Baalbek wondered aloud: Who knows? We may lose everything in our war with the Yemenis, known as they are for their valor and endurance in battle and combat. And all were reminded of the legacy of their ancestor, Hassan al-Yamani, and his tyranny in his wars against the Al Murrah and the Taghlibis, and the measures taken...]
C7	عتيبه وهو مايدري وش به:تعالي واركبي ورااي ركب فرسه وركبت قوت وراه وهي مبتسمه بانتصارها للي تريده الا ان عتيبه شاف انها ما مسكت فيه ومسكت بحواف السرج الي على الفرس حرك حواجه مستغرب منها لفوق وانطلق لاحق بغيث وهي [Utaybah, not quite sure what had come over him: Come on, ride behind me. He mounted his horse, and Qoot climbed up behind him, smiling for her victory in getting what she wanted, but Utaybah noticed she hadn't held on to him; but clutched the edges of the saddle, he lifted his brows in surprise, and took off riding after Ghaith, and she...]

Table 8: An example of the style ranking task. The example shows a reference text (Ref) and seven candidate snippets (C1-C7) which annotators rank based on their stylistic similarity to the reference text. Translations are provided in square brackets.

D. Style Ranking Annotation Guidelines

We provide the annotation guidelines for the style ranking task below.

Project overview: This project aims to advance Natural Language Processing (NLP) in tasks involving writing styles. Your annotations will help us evaluate how well trained authorship attribution models as well as large language models (LLMs) understand writing styles and assess their ability in ranking texts based on stylistic similarity in alignment with human judgments.

Annotation Task: You are given a reference text written by a specific author. Your task is to rank 7 other text snippets based on their stylistic similarity to the reference text. Each piece of text consists of 40 words. The most stylistically similar snippet should be placed at the top of the list, and the least similar should be placed last.

General Instructions:

- Focus on **style**, not on topic or content.
- Use the provided **stylometric features** to guide your judgment.
- Be **consistent and thorough** in your comparison.
- If unsure, consider which snippet could have been written by the same author as the reference.

Stylometric Features to Consider:

1. **Surface Features:** examine low-level characteristics of the text such as the average sentence length, punctuation usage, repetition of letters, as well usage of diacritics, elongations, hamzas. For diacritics, assess their usage, distinguishing between syntactic and lexical diacritics.
2. **Lexical Features:** look at word choices and how varied they are, checking word frequencies distribution, vocabulary richness and complexity, and orthographic spelling of words.
3. **Syntactic Features:** assess the grammatical structure and complexity of text, including morphosyntactic features, part-of-speech distribution, sentence structure complexity, tense and narrative voice (first, second, or third person).
4. **Level of Dialectness:** evaluate the level of dialectness, ranging from formal Modern Standard Arabic to high degree of informal colloquialism.
5. **Stylistic and Aesthetic Features:** notice expressive elements and literary style such as presence of poetic tone as well usage of vivid imagery, metaphors, sarcasm and humor.

For each pair of texts, compare the texts with regards to all these features to judge the extent of style similarity and be able to compare this pair against other pairs for ranking.

E. Prompt for Style Ranking Task

In Figure 4, we provide the prompt used in the style ranking task.

Prompt for Style Ranking Task

```
You are an Arabic professional writer. You are given a reference text written
by a specific author. Your task is to rank 7 other text snippets based on their
stylistic similarity to the reference text. The most stylistically similar
snippet should receive a ranking of 1, and the least similar should receive a
ranking of 7.
General Instructions:
Focus on style, not on topic or content.
Use the provided stylometric features to guide your judgment.
Assign each rank only once (i.e., no two snippets should have the same rank).
Be consistent and thorough in your comparison.
If unsure, consider which snippet could have been written by the same author as
the reference.
Stylometric Features to Consider:
Surface Features: examine low-level characteristics of the text such as the
average sentence length, punctuation usage, repetition of letters, as well usage
of diacritics, elongations, hamzas. For diacritics, assess their usage,
distinguishing between syntactic and lexical diacritics.
Lexical Features: look at word choices and how varied they are, checking word
frequencies distribution, vocabulary richness and complexity, and orthographic
spelling of words.
Syntactic Features: assess the grammatical structure and complexity of text,
including morphosyntactic features, part-of-speech distribution, sentence
structure complexity, tense and narrative voice (first, second, or third person).
Level of Dialectness: evaluate the level of dialectness, ranging from formal
Modern Standard Arabic to high degree of informal colloquialism.
Stylistic and Aesthetic Features: notice expressive elements and literary style
such as presence of poetic tone as well usage of vivid imagery, metaphors,
sarcasm and humor.
Provided the following reference text and 7 other text snippets, return the
ranking of the 7 snippets based on their stylistic similarity to the reference
text. The output should only contain comma-separated rankings in the same order
of the text snippets.
Only provide the rankings without explanation or extra text.
Reference text:
{input_reference_text}
Text Snippet 1
{input_text_snippet_1}
Text Snippet 2
{input_text_snippet_2}
Text Snippet 3
{input_text_snippet_3}
Text Snippet 4
{input_text_snippet_4}
Text Snippet 5
{input_text_snippet_5}
Text Snippet 6
{input_text_snippet_6}
Text Snippet 7
{input_text_snippet_7}
Now provide the comma-separated rankings without explanation or extra text.
```

Figure 4: Prompt used for style ranking task.