MarsadLab at AraGenEval Shared Task: LLM-Based Approaches to Arabic Authorship Style Transfer and Identification

Md. Rafiul Biswas[1], Mabrouka Bessghaier[2], Firoj Alam[3], Wajdi Zaghouani[2]
[1]Hamad Bin Khalifa University, Qatar, [2]Northwestern University in Qatar, Qatar
[3]Qatar Computing Research Institute, Qatar
{mbiswas,fialam}@hbku.edu.qa
{mabrouka.bessghaier,wajdi.zaghouani}@northwestern.edu

Abstract

We present our system submitted to the AraGenEval Shared Task at ArabicNLP 2025, which addresses the tasks of *Authorship Style Transfer* and *Authorship Identification*. For Subtask 1 (Style Transfer), we fine-tuned instruction-following Arabic large language models using Low-Rank Adaptation (LoRA). Among the evaluated models, **Qwen2.5-7B-Instruct** achieved a BLEU score of **20.30** and a chrF score of **52.56**, ranking 3^{rd} on the official leaderboard. For Subtask 2 (Authorship Identification), **AraBERTv2** attained an accuracy of **86.49%** and a macro-F1 score of **82.82%**, demonstrating robust performance in multi-class author classification across 21 categories. Our approach integrates Arabic-specific preprocessing, task-oriented prompt design, and transformer-based architectures, which enables effective handling of both generative and discriminative aspects of authorship analysis. We have made experimental scripts publicly available for the community.[1]

1 Introduction

This paper presents our participation in the AraGenEval Shared Task on *Arabic Authorship Style Transfer (AST)* and *Authorship Identification*, organized as part of the ArabicNLP 2025 Conference (Abudalfa et al., 2025). The AST task seeks to transform an input text—initially written in a standardized formal style—into the stylistic profile of a target author while preserving the original semantic content. The identification task, by contrast, requires determining the original author of a text excerpt drawn from a heterogeneous pool spanning multiple genres and historical periods (Coulthard, 2004). These problems are especially challenging in Arabic due to linguistic diversity manifesting as diglossia, rich and productive morphology, and context-dependent variation (Alqahtani and Dohler, 2023; AlZahrani and Al-Yahya, 2023a).

Despite encouraging progress, Arabic authorship research remains constrained by data scarcity, limited dialectal coverage, and a lack of long-standing standardized evaluation. Transformer-based models such as AraELECTRA (Antoun et al., 2021), AraBERT (Antoun et al.), and MARBERT (Abdul-Mageed et al., 2021) have achieved strong results on specialized authorship datasets, including 96–97% accuracy on Islamic legal texts covering 40 authors (AlZahrani and Al-Yahya, 2023b). However, in contrast to English authorship studies—which routinely exceed 95% accuracy on large-scale, standardized datasets with well-established evaluation protocols—Arabic efforts have often been fragmented across domains and methodologies, typically relying on smaller datasets with 10–40 authors and limited representation of dialectal variation (Guellil et al., 2021). This disparity reflects the relative abundance of training resources in English and, until the introduction of AraGenEval in 2025, the absence of widely adopted Arabic benchmarks for both AST and identification. Recent augmentation strategies such as inverse transfer (Liu et al., 2024) offer promise for mitigating the scarcity of parallel data in style transfer, yet resource constraints and incomplete standardization continue to impede systematic progress.

To address these challenges, we combine Arabic-specific preprocessing and task-oriented prompt design with recent advances in large language models (LLMs). In particular, we leverage **Qwen2.5L** (Team, 2024; Yang et al., 2024), **Fanar** (Team et al.), **Jais** (Sengupta et al., 2023), and **AraBERTv2** (Antoun et al.), applying parameter-efficient fine-tuning (e.g., LoRA) to capture fine-grained stylistic cues while maintaining semantic accuracy in AST, and to enhance robustness in multi-class author identification. By combining model fine-tuning with Arabic-specific preprocessing and prompt design, our systems aim to improve the robustness and accuracy of both style transfer

[1]https://github.com/rafiulbiswas/AraGenEval

Proceedings of The Third Arabic Natural Language Processing Conference, pages 88–93

and author classification. The main contributions of this paper are:

- We formulate Arabic authorship style transfer as instruction-following generation and replace conventional encoder–decoder pipelines with parameter-efficiently fine-tuned LLMs (LoRA).
- We present a cost-effective recipe that leverages open-source LLMs and adapter-based tuning, enabling competitive performance under modest GPU budgets.
- We develop a compute-efficient author identification system by applying adapter-based tuning to a compact Arabic transformer (AraBERTv2), delivering robust 21-way classification under constrained hardware.

2 Background

Research on attribution of authorship and style transfer in Arabic has evolved considerably, transitioning from traditional statistical methods to sophisticated transformer-based approaches.

Authorship Style Transfer. This task has been explored extensively in English (e.g. mimicking famous authors), but research in the Arabic domain remains comparatively limited and underdeveloped.(Abudalfa et al., 2025). Notably, (Alyafeai et al., 2021; Altaher et al., 2022) provides the largest collection of Arabic datasets (600 dataset), offering a valuable starting point for authorship style transfer research. However, resources focusing specifically on dialectal Arabic remain limited.

Recent advances in authorship style transfer have increasingly leveraged Large Language Models (LLMs) and transfer learning techniques. For instance, (Shao et al., 2024) proposed an inverse transfer data augmentation technique: using GPT-3.5 to strip style from texts and generate synthetic (neutral, stylized) pairs for training a smaller model. Likewise, Horvitz et al. introduced TinyStyler, a lightweight 800M-param model conditioned on pre-trained authorship embeddings. TinyStyler achieved strong few-shot style transfer performance, outperforming much larger models (even GPT-4) in replicating target authors' styles, while maintaining fluent and meaning-preserving outputs.

Author Identification. Over the past five years, Arabic pretrained language models (PLMs)—including AraBERT, ARBERT, AraELECTRA, and MARBERT—have substantially advanced authorship identification via task-specific fine-tuning, as surveyed in (Alqahtani and Dohler, 2023). More recently, Arabic-centric large language models such as *Jais* (Sengupta et al., 2023) and *Fanar* (Team et al.), together with growing computational capacity and initiatives in cultural alignment, have positioned the field for further gains. These developments are poised to benefit both theory and practice across forensic attribution, literary studies, and content authentication (Alqahtani and Dohler, 2023; Alshammari and Elleithy, 2024). Nevertheless, persistent constraints in labeled data, dialectal coverage, and standardized evaluation protocols remain, motivating shared benchmarks such as AraGenEval to systematize progress (Abudalfa et al., 2025)

3 Dataset

Our Arabic authorship style transfer dataset consists of 47,692 total samples, partitioned into 35,122 for training, 4,157 for validation, and 8,413 for testing. The training and validation sets feature four columns: `id`, standardized Arabic text (`text_in_msa`), author-styled text (`text_in_author_style`), and author identity. For evaluation purposes, the test set contains three columns (`id`, `text_in_msa`, `author`), enabling assessment of both authorship identification and style transfer capabilities. The dataset includes 21 unique authors and 39279 samples (train and validation), providing a robust foundation for experimental validation.

Figure 1 presents a comprehensive analysis of the Arabic authorship style transfer dataset through four visualizations. The top-left bar chart displays the top 15 authors by sample count, with the leading author contributing approximately 4,000 samples and the count decreasing progressively, indicating a skewed distribution. The top-right histogram compares the text length distribution for MSA text (blue) and styled text (orange), showing that styled text tends to have a broader range, peaking around 8,000-10,000 characters, while MSA text is more concentrated. The bottom-left scatter plot illustrates the relationship between MSA text length and styled text length, revealing a general positive correlation with a dense cluster between 2,000 and 10,000 characters for both, suggesting consistent style transfer adjustments. Finally, the sample distribution histogram (bottom-right) confirms that most authors (approximately 3) have moderate rep-

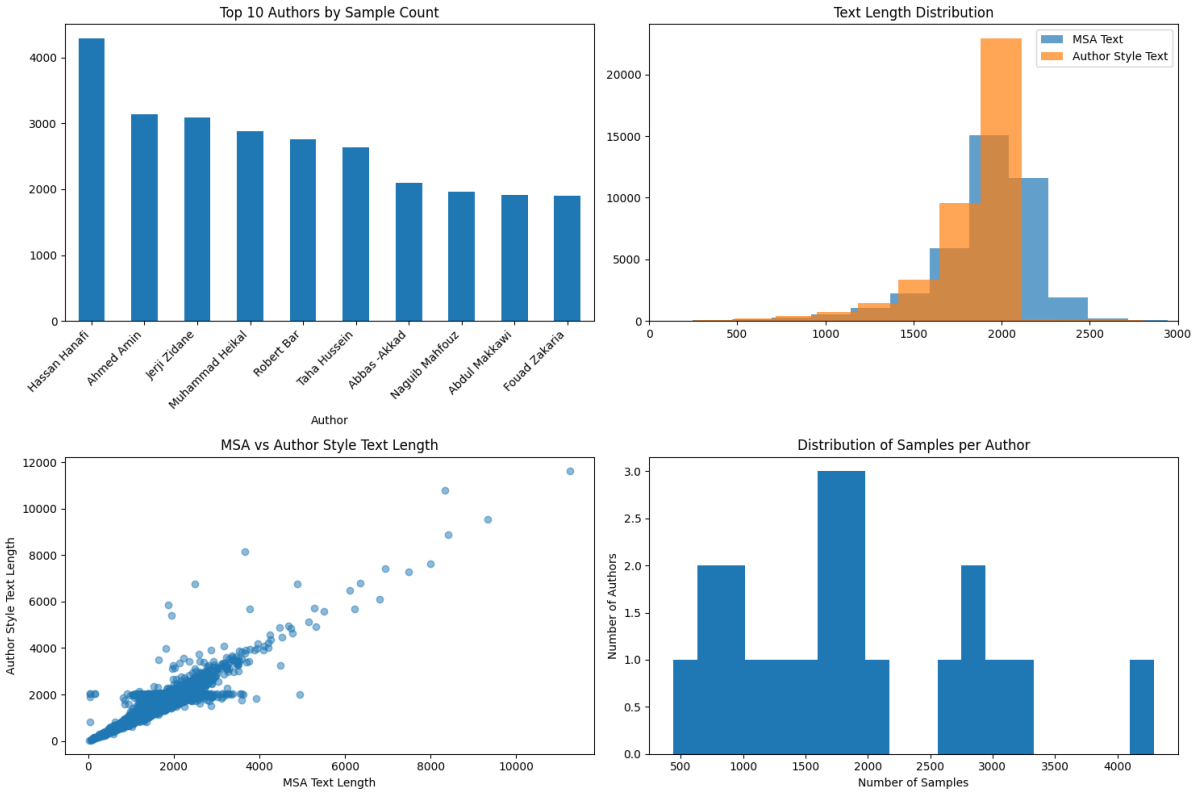

Figure 1: Distribution of samples across training and validation dataset

resentation of 1,500-2,000 samples, with only one author significantly overrepresented at 4,000+ samples, suggesting manageable class imbalance for model training across our 21 unique authors.

4 System Overview

4.1 Task 1: Authorship Style Transfer

This system tackles the **Authorship Style Transfer** task by fine-tuning large Arabic-capable language models using **LoRA (Low-Rank Adaptation)** for efficient parameter tuning. The model architecture centers around the `Qwen2.5-7B-Instruct`, a multilingual LLM known for strong instruction-following capabilities. Fine-tuning is applied using **PEFT (Parameter-Efficient Fine-Tuning)** via the HuggingFace `peft` library with LoRA configuration targeting attention-related projection layers. The model is optimized for `causal language modeling` (`TaskType.CAUSAL_LM`), with LoRA rank $r = 16$, $\alpha = 32$, and dropout $= 0.1$.

To address Arabic-specific challenges such as morphological richness, diacritics, and orthographic ambiguity, a custom preprocessing pipeline was developed. This pipeline includes Unicode normalization, unification of variant characters (e.g., different forms of Alef and Yeh), and cleaning of punctuation, diacritics, and Latin script artifacts. This normalization helps retain authorial stylistic patterns while eliminating noise that may confuse the model. During inference, a similar prompt without the target output guides the model to generate stylized text, using $\text{top\_}p = 0.9$, $\text{temperature} = 0.7$, and repetition penalties to balance creativity and fluency. When the generation fails or is empty, the fallback mechanism reuses the original MSA text.

Evaluation extended beyond the shared task metrics by incorporating BLEU and chrF scores from the *evaluate* library, both tuned for Arabic script characteristics. Although only Qwen2.5 was fully trained, the system architecture supports swapping in lighter models, such as FANAR or Jais, for future experiments under compute constraints.

4.2 Task 2: Authorship Identification

Our Arabic author classification leverages the discriminative capabilities of the AraBERT-v2 transformer, specifically optimized for authorship attribution. We fine-tuned the AraBERTv2 model [2] using the HuggingFace Transformers framework with a sequence classification head. Texts were preprocessed using a lightweight Arabic-aware pipeline

[2] aubmindlab/bert-base-arabertv2

that removed non-Arabic noise while preserving stylistic cues. Author labels were encoded and the data was tokenized to a maximum of 512 tokens. Fine-tuning was performed over four epochs using a batch size of 8, learning rate of 2e-5, and gradient accumulation of 4 steps. Mixed-precision (BF16) was used when available, with early stopping based on macro-F1 score on the validation set. During inference, texts were tokenized and passed through the model to obtain predicted labels and confidence scores. Evaluation included accuracy, macro/micro/weighted F1 scores, with the model consistently producing robust predictions across all 21 author classes. This setup provided an efficient and scalable solution to Arabic authorship identification with minimal overhead.

Configuration A (QWEN2.5L-LoRA) uses generative pre-training with sequence-to-sequence objectives, 4-bit quantization, LoRA rank-8 adaptation, batch size 16-32, max length 256, training time ∼8-12 hours, memory usage ∼16-22GB VRAM, achieving macro-F1 ∼0.82-0.87;

Configuration B (AraBERT-Full) employs discriminative pre-training with masked language modeling, full parameter fine-tuning, FP32 precision for stability, batch size 8-16, max length 512, training time ∼2-3 hours, memory usage ∼6-8GB VRAM, achieving macro-F1 ∼0.85-0.92.

5 Results

5.1 Task 1: Authorship Style Transfer Results

Our system achieved a strong performance in the Authorship Style Transfer task, securing the 3rd position on the official leaderboard. The best-performing model, Qwen2.5-7B-Instruct, achieved a BLEU score of 20.30 and a chrF score of 52.56, which were competitive compared to the top scorer's 24.58 BLEU and 59.01 chrF. Despite its smaller size relative to other models like Fanar-1.9B and Jais-13B, Qwen2.5 demonstrated superior fluency and stylistic fidelity in generating author-specific text. Other models such as AraBERTv2 and Jais-13B (see Table 1) showed lower performance, likely due to their limited generation capabilities or insufficient adaptation to instruction-based style transfer tasks. These results highlight the effectiveness of instruction-tuned LLMs, such as Qwen2.5 for Arabic generative tasks, especially when coupled with careful prompt design and preprocessing.

Model	BLEU	chrF
Jais-13B	15.17	47.32
AraBERTv2	17.78	46.72
Fanar-1.9B	18.39	48.32
Qwen2.5-7B	20.30	52.56

Table 1: Performance of our models on Task 1 (Leaderboard Position: 3rd)

Model	Accuracy	Precision	Recall	Macro F1
AraBERTv2	0.865	0.865	0.785	0.828
MARBERT	0.762	0.722	0.691	0.727
Qwen2.5-7B	0.745	0.789	0.732	0.701

Table 2: Comparison of the performance of our Models on Task 2

5.2 Task 2: Authorship Identification Results

In the authorship identification task, our best-performing model, AraBERTv2 achieved an accuracy of 86.49% and a macro F1 score of 82.82%, approaching the top system's performance of 92.42% accuracy and 89.89% macro F1. AraBERTv2 outperformed other tested models such as MARBERT and Qwen2.5, as shown in Table 2. This indicates the suitability of AraBERTv2 for fine-grained classification tasks in Arabic. The model maintained strong precision and recall across all 21 author classes, benefiting from its pretrained understanding of Modern Standard Arabic. In contrast, Qwen2.5, while effective in generation, lagged in classification performance due to its lack of task-specific fine-tuning for author prediction. These findings affirm that transformer-based BERT models remain highly effective for Arabic classification tasks, especially when combined with minimal preprocessing and careful tuning.

6 Conclusion

Despite the promising results, several limitations remain. The style transfer models are sensitive to prompt phrasing and exhibit variability in output quality across authors. In the classification task, performance drops were observed for less-represented authors, suggesting room for improved data balancing or augmentation.

Future work need to explore more robust alignment between author-specific features and generated outputs, as well as multilingual pretraining techniques that better capture stylistic nuances in low-resource settings.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI). F. Alam's work on this publication was also partially funded by the project "AI-EDAPT: Artificial Intelligence for Educational Adaptation, Personalization, and Transformation" (HBKU-OVPR-SRG-02-2), awarded by the Office of the Vice President for Research at Hamad Bin Khalifa University. The findings presented herein are solely the responsibility of the author(s).

References

Muhammad Abdul-Mageed, Abdelrahim Elmadany, and 1 others. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah, and Hamzah Luqman. 2025. The arageneval shared task on arabic authorship style transfer and ai-generated text detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

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.

Hamed Alshammari and Khaled Elleithy. 2024. Toward robust arabic AI-generated text detection: Tackling diacritics challenges. *Information*, 15(7):419.

Yousef Altaher, Ali Fadel, Mazen Alotaibi, Mazen Alyazidi, Mishari Al-Mutairi, Mutlaq Aldhbuiub, Abdulrahman Mosaibah, Abdelrahman Rezk, Abdulrazzaq Alhendi, Mazen Abo Shal, and 1 others. 2022. Masader plus: A new interface for exploring+ 500 arabic nlp datasets. *arXiv preprint arXiv:2208.00932*.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2021. Masader: Metadata sourcing for arabic text and speech data resources. *Preprint*, arXiv:2110.06744.

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

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

Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4):431–447.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

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. *arXiv preprint arXiv:2406.15586*.

Shuai Liu, Shantanu Agarwal, and Jonathan May. 2024. Authorship style transfer with policy optimization. *arXiv preprint arXiv:2403.08043*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, and 1 others. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Zhonghui Shao, Jing Zhang, Haoyang Li, Xinmei Huang, Chao Zhou, Yuanchun Wang, Jibing Gong, Cuiping Li, and Hong Chen. 2024. Authorship style transfer with inverse transfer data augmentation. *AI Open*, 5:94–103.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. Fanar: An arabic-centric multimodal generative ai platform.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.