ANLPers at AraGenEval Shared Task: Descriptive Author Tokens for Transparent Arabic Authorship Style Transfer

Omer Nacar[*]
Tuwaiq Academy
o.najar@tuwaiq.edu.sa

Serry Sibaee
Prince Sultan University
ssibaee@psu.edu.sa

Mahmoud Reda
Zagazig University
redamahmoud722@gmail.com

Yasser Al-Habashi
Prince Sultan University
yalhabashi@psu.edu.sa

Adel Ammar
Prince Sultan University
aammar@psu.edu.sa

Wadii Boulila
Prince Sultan University
wboulila@psu.edu.sa

Abstract

Authorship style transfer enables the generation of text that imitates a specific writer's linguistic and stylistic patterns, a challenging task in morphologically rich languages like Arabic. We tackle this problem in the AraGenEval 2025 shared task, exploring conditioning strategies to guide a fine-tuned `UBC-NLP/AraT5v2-base-1024` model in producing text aligned with target authors' styles. Our investigation compares implicit modeling, numeric and descriptive author tokens, and explicit prompt engineering in Arabic. Explicit natural language instructions proved most effective, achieving the highest competition scores with BLEU of 24.58 and chrF of 59.01, securing first place, while demonstrating that interpretable approaches can rival or surpass more opaque methods.

1 Introduction

The task of Text Style Transfer (TST) aims to modify stylistic properties of a text while preserving its semantic content (Hu et al., 2022). A challenging sub-field is authorship style transfer, which involves rewriting a text to match the unique style of a specific author (Shao et al., 2024). Arabic authorship style transfer presents unique challenges due to the language's rich morphological structure and diverse writing styles. The task, as defined in the AraGenEval 2025 shared task (Organizers, 2024), requires transforming Modern Standard Arabic (MSA) text to match the distinctive style of specific Arabic authors.

We conduct a systematic investigation of different conditioning strategies using the `UBC-NLP/AraT5v2-base-1024` model (Elmadany et al., 2022). Our work explores four main methodologies: (1) standard fine-tuning without special conditioning, (2) numeric author tokens for explicit author identification, (3) descriptive author tokens for human-readable conditioning, and (4) prompt engineering with explicit Arabic instructions.

Extensive experiments show that explicit prompt engineering delivers the best results, outperforming non-interpretable numeric tokens by leveraging the model's language understanding through clear, natural prompts. This approach secured first place in the AraGenEval Shared Task (Abudalfa et al., 2025) and offers insights for building effective, interpretable Arabic style transfer systems.

2 Background

Text style transfer has become a prominent area of research (Hu et al., 2022). Early work focused on disentangling style from content, whereas recent trends have shifted towards end-to-end transfer without explicit disentanglement.

Authorship style transfer, specifically, has been tackled with various methods. Some approaches focus on data augmentation to create paired corpora for training compact models, a technique shown to be highly effective (Shao et al., 2024). The challenge is often compounded in low-resource scenarios, where only a few examples of a target author's style are available (Patel et al., 2022). Recent work has introduced lightweight and efficient models like TinyStyler (Horvitz et al., 2024), which leverage pre-trained authorship embeddings to achieve strong performance in few-shot settings, even outperforming large models like GPT-4. Our work contributes to this area by systematically evaluating different conditioning methods for a T5-based model on Arabic, a morphologically rich language that remains under-explored in this domain.

The detection of AI-generated content is another related field of study, with recent work focusing on distinguishing between human and GenAI-generated Arabic text on social media platforms using machine learning models (Alghamdi et al.,

[*]Corresponding author: o.najar@tuwaiq.edu.sa

Proceedings of The Third Arabic Natural Language Processing Conference, pages 49–53

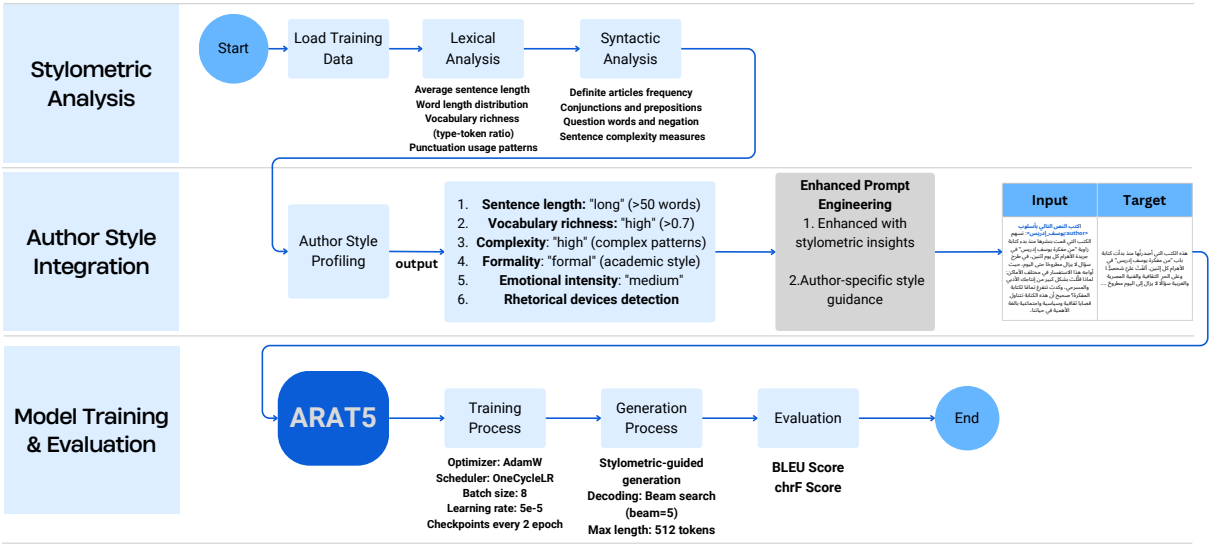

Figure 1: Pipeline overview for the proposed authorship style transfer approach.

2024). This is relevant to our participation in Subtask 3 of AraGenEval.

The AraGenEval 2025 shared task on Authorship Style Transfer provides a dataset containing text from 21 Arabic authors. The goal is to take an input text in Modern Standard Arabic (MSA) and transform it into the style of a target author. We participated in all three subtasks offered: Authorship Style Transfer (Subtask 1), Authorship Identification (Subtask 2), and ARATECT for AI-generated text detection (Subtask 3). This paper focuses primarily on our work for Subtask 1.

3 System Overview

Our approach is centered on fine-tuning the `UBC-NLP/AraT5v2-base-1024` model, a T5-based encoder–decoder architecture pre-trained on a large corpus of Arabic text (Elmadany et al., 2022). The core of our investigation involved systematically testing four different methods for conditioning the model on the target author's style, each employing a distinct input format to guide the model. Figure 1 presents an overview of the complete pipeline, which is organized into three main stages.

The first stage, *Stylometric Analysis*, extracts lexical and syntactic features from the training corpus, including sentence length, vocabulary richness, syntactic complexity, formality, emotional intensity, and rhetorical device usage (Gómez-Adorno et al., 2018). In the second stage, *Author Style Integration*, these stylistic attributes are distilled into a profile that informs two conditioning strategies: (1) enhanced prompts augmented with stylometric insights and (2) author-specific style guidance. The third stage, *Model Training & Evaluation*, applies these conditioning strategies in fine-tuning AraT5, followed by generation and evaluation against baseline and alternative approaches.

Table 1 outlines the shift from implicit style modeling to explicit, instruction-based conditioning. The baseline relies solely on input–output pairs, leaving style inference to the model. Token-based methods introduce minimal explicit signals, while prompt engineering—framing style transfer as direct, human-readable instructions—proves most effective by leveraging the model's pre-trained stylistic knowledge.

4 Experimental Setup

4.1 Dataset & Preprocessing

The shared task dataset contains writings from 21 authors split into training, validation, and test sets as provided by the shared task. The training set contains 35,122 samples, the validation set contains 4,157 samples, and the test set contains 8,413 samples, proportionally distributed per author. All texts were normalized by removing extraneous whitespace, unifying punctuation forms, and standardizing Arabic diacritics. Special tokens were inserted according to the conditioning method described in Table 1.

4.2 Hyperparameters

Experiments were implemented in `PyTorch 2.1.0` and `Hugging Face Transformers 4.38.1`, with training managed via `Accelerate` and `Datasets`.

Approach	Conditioning Method	Input Format with Speical Tokens
Standard Fine-Tuning (Baseline)	No explicit conditioning signal. The model learns the mapping implicitly from paired data.	النص الأصلي باللغة العربية الفصحى
Numeric Author Tokens with FT	A unique numeric token (e.g., `author_0`) is prepended to the input to specify the target author.	`<author_id>`: النص الأصلي باللغة العربية الفصحى
Descriptive Author Tokens with FT	Human-readable tokens (e.g., `<author:Yusuf_Idris>`) are used instead of numeric ones to improve interpretability.	`<author:name>`: النص الأصلي
Prompt Engineering with FT (Our Best System)	The task is framed as an explicit natural language instruction in Arabic, prepended to the input.	اكتب النص التالي بأسلوب `<author:name>: [source_text]`

Table 1: Overview of the four experimental approaches for authorship style transfer.

Approach	BLEU	chrF
Standard Fine-tuning	20.50	58.50
Numeric Tokens	24.04	59.15
Descriptive Tokens	24.00	59.00
Prompt Engineering	**24.58**	**59.01**

Table 2: Official results on the AraGenEval 2025 test set. Our prompt engineering system ranked first.

All code was executed on NVIDIA A100 GPUs (80GB VRAM) under CUDA 12.2. Models were fine-tuned with a batch size of 64 (across 4 GPUs), AdamW optimizer (weight decay 0.01), a 5×10^{-5} learning rate, and OneCycleLR scheduling with 1000 warmup steps. Training ran for up to 10 epochs with early stopping based on validation loss.

4.3 Generation Settings

Generation used beam search with sampling (`num_beams`=2, temperature=0.6, `top_k`=20, `top_p`=0.8, repetition penalty=1.05, length penalty=0.6) and a 512-token output cap to balance quality and diversity.

4.4 Evaluation Metrics

Evaluation was conducted using two metrics: **BLEU**, the primary measure of n-gram precision between generated and reference texts (Papineni et al., 2002), and **chrF**, a character n-gram F-score metric (Popović, 2015) often better suited for morphologically rich languages such as Arabic.

5 Results and Analysis

Our experimental results on the official test set clearly show the progression in performance across the four conditioning strategies. The prompt engineering approach achieved the highest scores, securing first place in the competition. The results are summarized in Table 2.

As shown in Table 2, explicit author conditioning was essential, with all conditioned methods outperforming the baseline. Human-readable tokens proved as effective as numeric ones, showing that interpretability does not reduce performance. Prompt engineering achieved the strongest results, enabling the model to leverage its pre-trained understanding of Arabic, while also reducing common errors such as semantic drift, incomplete style transfer, and repetition by better preserving entities and semantic fidelity.

5.1 Dataset Stylometric Analysis

We conducted a post-hoc stylometric analysis of the 21 authors using a custom `StylometricAnalyzer`, extracting lexical, syntactic, and statistically categorized features to create individual stylistic profiles. The resulting heatmap (Figure 2) revealed strong stylistic homogeneity, with minimal variation in core features like sentence length, vocabulary richness, complexity, and formality. Punctuation-based cues offered little discrimination, and the only notable outlier was ثروت أباظة, who showed lower emotional intensity—highlighting the challenge of style transfer in this dataset.

This observation provides a compelling explanation for the superior performance of our prompt engineering approach. Methods relying on implicit signals or simple author tokens must learn these subtle distinctions from the data alone. In contrast, the explicit instruction اكتب النص التالي بأسلوب leverages the vast, latent knowledge of the pre-trained AraT5 model. It effectively commands the

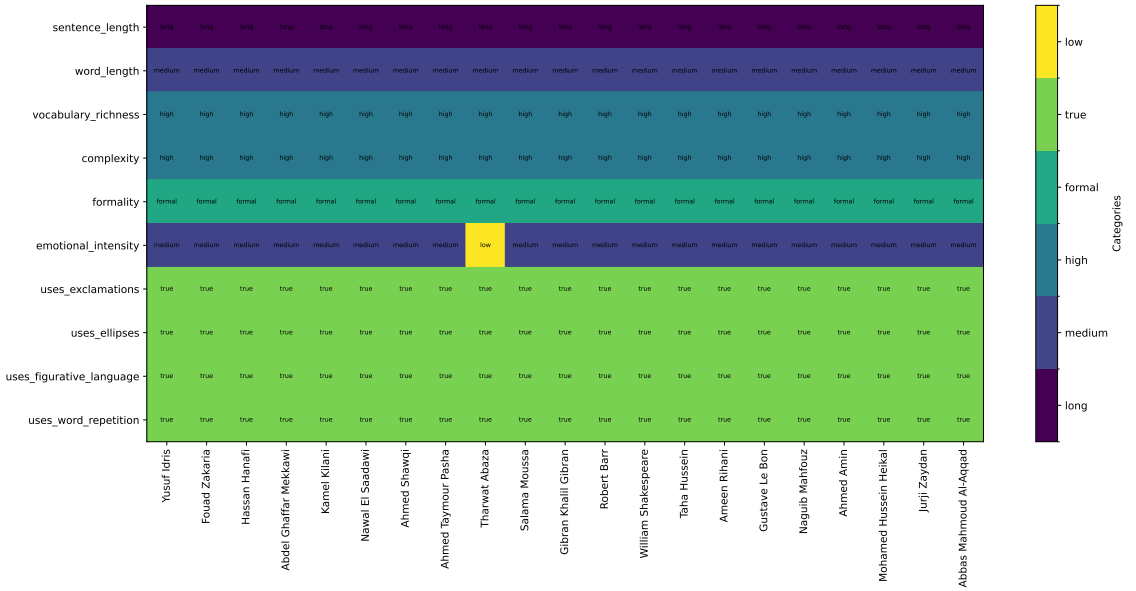

Figure 2: Stylometric characteristics heatmap.

model to access its deep understanding of authorial voice, which goes far beyond what our statistical metrics can measure. This allows it to capture the unique, nuanced characteristics of each author, leading to its first-place performance.

6 Results in Additional Shared Tasks

6.1 Subtask 2: Authorship Identification

We addressed class imbalance through weighted loss during training. After preprocessing and tokenization, several Arabic-specific BERT-based models were fine-tuned. The best-performing configuration, `bert-base-arabic-camelbert-mix-sentiment` (Inoue et al.), trained for 10 epochs with early stopping, reached an accuracy of 95.3% and a macro F1-score of 95.1% on the development set. Our system ranked 6th, achieving an F1-score of 0.83138 and an accuracy of 87.52%, which is only 6.7 percentage points lower in F1-score compared to the top-ranked system (0.89886). The official leaderboard results for both subtasks are summarized in Table 3.

6.2 Subtask 3: ARATECT (Arabic AI-Generated Text Detection)

For AI-generated text detection, the dataset was already balanced. After minimal cleaning and tokenization, transformer-based models converged in just 3 epochs. Our top model, `XLM-RobertaForSequenceClassification` (Ruder et al., 2019), achieved an accuracy of 99.36% and a macro F1-score of 99.3% on the development set. Our system ranked 6th, achieving an F1-score of 0.76 and an accuracy of 79%, which is only 10 percentage points lower in F1-score compared to the top-ranked system (0.86).

Task	Accuracy	F1
Authorship ID (Dev)	0.95	0.95
Authorship ID (Test)	0.87	0.83
ARATECT (Dev)	0.99	0.99
ARATECT (Test)	0.79	0.76

Table 3: Performance metrics for Subtasks 2 and 3.

7 Conclusion

In this paper, we presented our winning system for the AraGenEval 2025 Arabic Authorship Style Transfer task. Our systematic investigation demonstrates that explicit prompt engineering with natural Arabic instructions is a highly effective method for conditioning a T5 model. We found that simpler, interpretable conditioning methods are potent and that leveraging a model's linguistic capabilities through clear prompts yields superior results compared to merely adding special tokens. Future work could explore integrating stylometric features directly into the prompt, extending the framework to multi-author style transfer, and developing real-time applications. Our findings underscore the value of prompt engineering as a powerful and interpretable technique for controllable text generation in Arabic.

Acknowledgments

The authors gratefully acknowledge the support provided by Tuwaiq Academy and the computational resources by Prince Sultan University.

References

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)*, Association for Computational Linguistics.

Reem Alghamdi, Areej Al-Wabil, and Muna Al-Razgan. 2024. Distinguishing arabic genai-generated tweets and human tweets utilizing machine learning. *Engineering, Technology & Applied Science Research*, 14(5):16720–16726.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, and 1 others. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

Helena Gómez-Adorno, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, and Gerardo Sierra. 2018. Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas*, 22(1):47–53.

Sol Horvitz, Luis Ortiz, Anjali Sharan, and Alexandra Getman. 2024. Tinystyler: Efficient few-shot text style transfer with authorship embeddings. *arXiv preprint arXiv:2406.15586*.

Zhumin Hu, Zhaofeng Tu, Zur G-BETH, Jdn Lrec, and Vıctor T-BI. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):1–21.

G Inoue, B Alhafni, N Baimukan, H Bouamor, and N Habash. The interplay of variant, size, and task type in arabic pre-trained language models. arxiv 2021. *arXiv preprint arXiv:2103.06678*.

AraGenEval Organizers. 2024. Overview of the arageneval 2024 shared task. Shared Task Website. (Placeholder citation).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Krish Patel, Saizhu Zong, Vicky Shao, Ao Peng, and He He. 2022. Low-resource authorship style transfer: Can non-famous authors be imitated? *arXiv preprint arXiv:2212.08986*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Vicky Shao, Xinyi Chen, Saizhu Zong, Yuerou Yang, Ao Peng, and He He. 2024. Authorship style transfer with inverse transfer data augmentation. *AI Open*, 5:14–22.