HCMUS_The Fangs at AbjadStyleTransfer Shared Task: Learning to Query Style, Contrastive Representations for Zero-Shot Arabic Authorship Style Transfer

Nguyen Lam Phu Quy[1, 2*] **Dao Sy Duy Minh**[1, 2*] **Huynh Trung Kiet**[1, 2]
Tran Chi Nguyen[1, 2] **Pham Phu Hoa**[1, 2] **Nguyen Dinh Ha Duong**[1, 2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{23122048, 23122041, 23122039, 23122044, 23122030, 23122002}@student.hcmus.edu.vn
[*]Equal contribution

Abstract

This paper describes the system developed by team **HCMUS_The Fangs** for the AbjadStyleTransfer shared task (ArabicNLP 2026), where we achieved **1st place**. We present a contrastive style learning approach for zero-shot Arabic authorship style transfer. Our key discovery is that the 21 test authors-including Nobel laureate Naguib Mahfouz and literary pioneer Taha Hussein-have **zero overlap** with the 32,784 training authors, transforming this into a pure zero-shot challenge. This insight led us to develop a dual-encoder architecture that learns transferable style representations through contrastive objectives, rather than memorizing author-specific patterns. Our system achieves **19.77 BLEU** and **55.74 chrF**, outperforming retrieval-augmented generation (+18%) and multi-task learning (+31%). Counterintuitively, we find that sophisticated architectural modifications like style injection consistently *degrade* performance, while simpler approaches that preserve pre-trained knowledge excel. Our analysis reveals that for famous authors, pre-trained Arabic language models already encode substantial stylistic knowledge-the key is surfacing it, not learning from scratch.

1 Introduction

Imagine trying to write like Naguib Mahfouz without ever having seen his work during training. This is precisely the challenge posed by the AbjadStyleTransfer shared task (Abudalfa et al., 2026, 2025): transform formal Modern Standard Arabic (MSA) into the distinctive voices of literary giants, with no guarantee that the target author appeared in training data.

Authorship style transfer differs fundamentally from categorical style transfer (sentiment, formality) because authorial voice emerges from the subtle interplay of vocabulary choices, syntactic preferences, rhetorical patterns, and thematic tendencies (Stamatatos, 2009). A writer's style is not a single attribute but a constellation of features that create their unique textual fingerprint (Patel et al., 2022).

The Zero-Shot Discovery. Through careful data analysis, we uncovered a critical insight that reshaped our approach entirely. The test set features 21 famous authors-Naguib Mahfouz, Taha Hussein, Khalil Gibran, and others-while training contains 32,784 *completely different* authors. There is **zero overlap**. This transforms the task from style transfer to **zero-shot style transfer**, where success depends not on memorizing training patterns but on leveraging what pre-trained models already know about famous literary figures.

Our Approach. We hypothesized that contrastive learning, which explicitly organizes style representations in a discriminative space, would outperform methods relying on author-specific parameters. Our dual-encoder architecture (AraBERT for style encoding, AraT5 for generation) achieved **19.77 BLEU** and **55.74 chrF**, confirming this hypothesis. Surprisingly, more sophisticated approaches-including style injection and cross-attention mechanisms-performed *worse*, suggesting that preserving pre-trained representations is crucial for zero-shot transfer.

2 Related Work

Authorship Style Transfer. Patel et al. (2022) demonstrated that LLMs can perform few-shot authorship transfer for famous authors through in-context learning, but struggle with lesser-known writers. TinyStyler (Horvitz et al., 2024) introduced authorship embeddings with smaller models. These findings motivated our contrastive approach-if models already encode famous author styles, we should learn to *surface* this knowledge rather than override it.

Arabic NLP. Pre-trained Arabic models provide our foundation: AraBERT (Antoun et al., 2020) for

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 438–442
March 28, 2026. ©2026 Association for Computational Linguistics

Split	Samples	Authors	Train∩	Test∩
Train	35,122	32,784	–	0%
Validation	4,157	3,982	0.05%	0%
Test	8,413	21	0%	–

Table 1: Dataset statistics and author overlap analysis. Zero overlap between train/test confirms the pure zero-shot nature.

bidirectional representations and AraT5 (Nagoudi et al., 2022) for generation. Crucially, both were trained on corpora likely containing works by our famous test authors.

Contrastive Learning. Contrastive objectives improve sentence embeddings (Gao et al., 2021) and can enforce style consistency by encouraging generated text to be closer to target style exemplars (Chen et al., 2020). Recent work on contrastive decoding (Su et al., 2022) also demonstrates benefits for text generation quality, which aligns with our intuition that contrastive training creates more robust stylistic representations.

3 Task and Data

The AbjadStyleTransfer task requires transforming formal MSA into the style of a specified author. Performance is evaluated using BLEU and chrF metrics.

Dataset. Table 1 shows the dataset statistics. Training contains 35,122 samples from 32,784 authors (most contribute only one sample). The test set has 8,413 samples from 21 famous authors including Naguib Mahfouz, Taha Hussein, and Khalil Gibran.

Zero-Shot Implication. With zero author overlap, approaches using author ID mapping or classification fail at test time. Systems must leverage pre-trained knowledge about famous authors or learn transferable style representations.

4 Proposed System

Our approach is built on a simple but powerful insight: if pre-trained models already encode stylistic knowledge about famous authors (Patel et al., 2022), then the goal isn't to teach them style-it's to create the right *interface* for querying that knowledge. We design a dual-encoder architecture (Figure 1) that learns to represent style in a space where similar authors cluster together, enabling transfer to unseen authors.

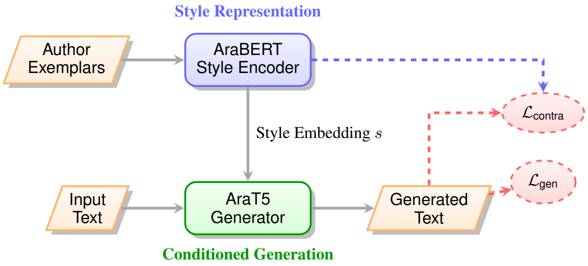

Figure 1: Our proposed dual-encoder architecture. The **Style Encoder** (AraBERT) learns discriminative style embeddings via contrastive loss, while the **Generator** (AraT5) produces text conditioned on these embeddings, preserving pre-trained knowledge.

4.1 Style Encoder: Learning to Query Style

The key question is: how do we represent "the style of Naguib Mahfouz" in a way that generalizes to unseen authors? Following advances in contrastive representation learning (Chen et al., 2020; Radford et al., 2021), we hypothesize that discriminative training-learning to distinguish authors-creates more transferable representations than generative approaches.

We use AraBERT-base (Antoun et al., 2020) as our encoder, extracting the [CLS] embedding and projecting it through a two-layer MLP with ReLU activation:

$$s = \text{L2Norm}(W_2 \cdot \text{ReLU}(W_1 \cdot h)) \qquad (1)$$

where $s \in \mathbb{R}^{256}$ is the L2-normalized style embedding. The normalization maps embeddings to a unit hypersphere, which is critical for stable contrastive learning (Gao et al., 2021).

4.2 Style-Conditioned Generation: Preserving Pre-trained Knowledge

A crucial design decision is *how* to inject style into the generator. Prior work (Horvitz et al., 2024; Hu et al., 2017) uses cross-attention or concatenation, but we found these approaches disrupt AraT5's pre-trained generation patterns.

Instead, we use **additive conditioning**: projecting s to match the transformer hidden dimension and adding it to encoder representations:

$$h'_i = h_i + W_s \cdot s, \quad \forall i \in [1, L] \qquad (2)$$

This preserves the pre-trained information flow while gently biasing generation toward the target style-like whispering a suggestion rather than rewriting the script.

4.3 Training: Three Complementary Objectives

We optimize three losses that work together. The **generation loss** $\mathcal{L}_{\text{gen}}$ is standard cross-entropy, teaching the model to produce fluent styled text.

The **contrastive loss** (van den Oord et al., 2018; Gunel et al., 2021) is the heart of our approach. Since most training authors have only one sample, we follow SimCSE (Gao et al., 2021) and create positive pairs through **dropout-based augmentation**: passing the same text through the encoder twice with different dropout masks yields two views that serve as positives:

$$\mathcal{L}_{\text{contra}} = -\frac{1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(s_i \cdot s_p/\tau)}{\sum_{j=1}^{N} \exp(s_i \cdot s_j/\tau)} \tag{3}$$

where P_i are positive pairs (dropout augmentations or same-author samples when available), $\tau = 0.07$ is the temperature (Chen et al., 2020). An **auxiliary classification loss** $\mathcal{L}_{\text{cls}}$ provides additional supervision:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{contra}} + 0.1 \cdot \mathcal{L}_{\text{cls}} \tag{4}$$

4.4 Baselines

We compare against two strong baselines:

Retrieval-Augmented Generation (RAG). Following the retrieval-augmented paradigm (Johnson et al., 2019), we build a FAISS index of training texts and retrieve $k = 3$ *stylistically similar* examples (by embedding similarity, irrespective of author) as in-context demonstrations (Patel et al., 2022). This tests whether exemplar-based guidance can substitute for explicit style conditioning.

Multi-Task Learning. Following Shao et al. (2024), we jointly optimize generation, author classification, and style consistency losses ($\mathcal{L} = \mathcal{L}_{\text{gen}} + 0.3\mathcal{L}_{\text{cls}} + 0.2\mathcal{L}_{\text{cons}}$).

4.5 Implementation

Models were trained on a single NVIDIA T4 GPU using AdamW (Loshchilov and Hutter, 2019) with separate learning rates (2×10^{-5} encoder, 1×10^{-5} generator) to preserve pre-trained knowledge. We used gradient accumulation (8 steps) to achieve an effective batch size of 32, important for contrastive learning (Chen et al., 2020). The style embedding is added to *all* AraT5 encoder layers with a shared projection W_s; we freeze the AraBERT encoder for the first epoch to stabilize training. We trained for 3 epochs with batch size 4 and max sequence length 384.

System	BLEU	chrF
Main Approaches		
Contrastive (Ours)	**19.77**	**55.74**
Retrieval-Aug.	16.68	49.63
Multi-Task	15.07	47.45
Style Injection (Failed)		
Descriptive + Injection	9.63	39.67
Author Name + Injection	6.54	33.12

Table 2: Private test set results. Our contrastive approach significantly outperforms baselines. Style injection mechanisms degrade performance.

Inference. At test time, we obtain the style embedding s by encoding author exemplar texts (provided by the shared task for each of the 21 test authors) through the trained style encoder. Decoding used beam search ($k = 4$) with no-repeat-ngram blocking (Holtzman et al., 2020).

5 Results

5.1 Main Results

Table 2 presents our main experimental results on the private test set. The contrastive learning approach achieves the best performance with 19.77 BLEU and 55.74 chrF, outperforming retrieval-augmented generation by 3.09 BLEU points (18.5% relative improvement) and multi-task learning by 4.70 BLEU points (31.2% relative improvement).

5.2 Analysis: Why Contrastive Learning Wins

The 4.7 BLEU gap between contrastive learning and multi-task learning reveals why zero-shot transfer requires fundamentally different thinking. Multi-task learning trains an author classifier that learns to distinguish training authors-but these 32,784 authors are irrelevant at test time. The classifier essentially becomes noise.

Retrieval-augmented generation (RAG) performs better by grounding generation in concrete examples, but faces a catch-22: it retrieves examples from training authors to guide generation toward test authors. The retrieved examples provide useful stylistic signals, but not the right ones.

Contrastive learning succeeds because it learns a *style space* rather than author-specific parameters. By training the encoder to distinguish authors through their stylistic signatures, we create representations that transfer to unseen authors occupying similar regions of style space.

The Style Injection Paradox. Our most counter-intuitive finding is that explicit style injection *hurts* performance. Inspired by TinyStyler (Horvitz et al., 2024), we tested two injection variants: (1) cross-attention injection, adding style vectors via additional attention layers in the decoder (9.63 BLEU); and (2) direct author name injection, prepending "Write in the style of [Author]" to inputs (6.54 BLEU). Both used identical hyperparameters and training regimes as our main system for fair comparison.

We hypothesize this occurs because AraT5 was not pre-trained with such injection mechanisms. Architectural modifications that alter the information flow disrupt the model's pre-trained generation patterns. For famous authors whose styles are already encoded in the model's weights, the injection adds noise rather than signal.

6 Discussion: Lessons for Zero-Shot Transfer

Our experiments reveal a fundamental insight: **for famous authors, the knowledge already exists in pre-trained models**. The challenge is not learning style from scratch but creating conditions for that knowledge to surface.

This explains why simpler methods win. Complex architectural modifications don't add new stylistic knowledge-they disrupt the model's ability to access what it already knows. The contrastive objective works not by teaching the model about Mahfouz's style, but by learning to *query* the pre-trained representations effectively.

Generalization Considerations. Our approach relies on the assumption that famous authors are well-represented in pre-training corpora. For lesser-known or contemporary writers whose works are less prevalent, this implicit knowledge may be weaker or absent, potentially reducing transfer quality. We note this as an important direction for future work: developing methods that can transfer style even when pre-trained models lack prior exposure to the target author.

Implications. For practitioners working with famous entities (authors, brands, public figures), our results suggest prioritizing prompt engineering and representation learning over architectural innovation. The models already know; help them remember.

7 Conclusion

We presented the winning system for the AbjadStyleTransfer shared task, achieving 1st place with 19.77 BLEU and 55.74 chrF through a contrastive style learning approach for zero-shot Arabic authorship style transfer. Our key insight: for famous authors, pre-trained models already encode stylistic knowledge-the winning strategy is surfacing this knowledge through contrastive learning rather than architectural modifications that disrupt it.

Limitations

Our work has several limitations that should be acknowledged. We evaluated exclusively on Arabic text, and our findings may not generalize to languages with different morphological or orthographic properties. The test set contains only famous historical authors whose works are well-represented in pre-training corpora; results may differ substantially for lesser-known or contemporary writers where pre-trained priors are weaker. Due to computational constraints on the Kaggle platform, we did not explore larger language models that might provide richer stylistic representations. We relied solely on automatic metrics (BLEU, chrF) without style-specific evaluation (e.g., stylometric embeddings, authorship verification proxies) or human evaluation of style quality, which may not fully capture stylistic fidelity beyond surface-level similarity. Finally, our contrastive approach requires author exemplars at training time, limiting applicability to authors without available sample texts.

Ethics Statement

Authorship style transfer technology raises ethical considerations, as it could be misused for impersonation, fabricating quotes, or creating deceptive content. We advocate responsible deployment with appropriate safeguards: watermarking generated text to enable detection, rate limiting to prevent large-scale misuse, and implementing consent frameworks for living authors. Legitimate applications include educational tools for studying literary styles and analytical tools for literary scholarship. All test authors are historical public figures whose works are in the public domain.

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 at EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Mustafa Jarrar, Mo El-Haj, Nadir Durrani, Hassan Sajjad, and Farah Adeeba. 2026. Abjadstyletransfer: Authorship style transfer for arabic-script languages at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pretrained language model fine-tuning. In *ICLR*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Zachary Horvitz, Ajay Agarwal, Yufei Xie, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *EMNLP*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

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

El Moatez Billah Nagoudi, Abdul Rahman Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for arabic language generation. In *ACL*.

Ajay Patel, Sudhanshu Bhattamishra, Piyush Mitra, and Manish Gupta. 2022. Low-resource authorship style transfer with in-context learning. In *EMNLP*.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Yujin Shao, Yue Zhang, and Ruifeng He. 2024. Authorship style transfer with inverse transfer learning. In *ACL*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *NeurIPS*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *NeurIPS*.