Jenin at AraGenEval Shared Task: Parameter-Efficient Fine-Tuning and Layer-Wise Analysis of Arabic LLMs for Authorship Style Transfer and Classification

Huthayfa Malhis
Independent Researcher
huthayfa.malhis@gmail.com

Mohammad Tami
Arab American University
mabutame@gmail.com

Huthaifa I. Ashqar
Arab American University
huthaifa.ashqar@aaup.edu

Abstract

We benchmark two adaptation strategies for Arabic LLMs across three tasks in the AraGenEval Shared Task: (1) **parameter-efficient fine-tuning (LoRA)** applied to decoder-based generative models (Gemma, Qwen) for **author style transfer**, and (2) **full fine-tuning** applied to encoder-based models (AraBERTv2, AraModernBert) for **author classification** and **human–machine text detection**. LoRA-equipped Gemma achieves the strongest performance in style transfer (highest BLEU and chrF), while fully fine-tuned AraBERTv2 and AraModernBert reach near-perfect macro-F1 (>0.99) in classification and detection. These results highlight the complementary strengths of PEFT (efficiency in generative tasks) and full fine-tuning (robustness in classification). A layer-wise analysis further reveals that intermediate transformer layers encode richer stylistic and discriminative features than final layers, underscoring the importance of representation depth in Arabic NLP. All code and models are available at: https://github.com/mtami/AraGenEval2025.

1 Introduction

Large language models (LLMs) have transformed natural language processing (NLP) in recent years, enabling impressive progress in tasks ranging from machine translation to text generation (Ashqar & Tami, 2025). However, Arabic remains underexplored compared to English and other high-resource languages, despite being one of the most widely spoken languages worldwide, with over 400 million speakers across diverse dialects and stylistic registers (Al-Sarem et al., 2020). The morphological richness, diglossia, and wide stylistic variability of Arabic present unique challenges for adapting LLMs to downstream tasks. Prior benchmarks for Arabic LLMs are limited in scope, typically focusing on sentiment analysis or question answering, leaving important areas such as style transfer, author classification, and AI-generated text detection largely under-studied (A. Najjar et al., 2025; A. A. Najjar et al., 2025).

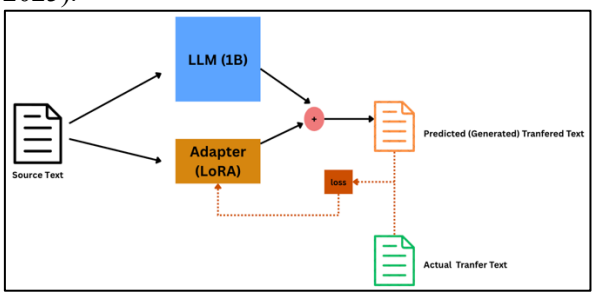

Figure 1: Parameter-efficient fine-tuning applied to Arabic LLMs for generative tasks.

In this paper, we address these gaps by providing a multi-task evaluation of Arabic LLMs, targeting three representative tasks, which is part of a AraGenEval Shared Task (Abudalfa et al., 2025):

[1] Author Style Transfer (AST): rephrasing Modern Standard Arabic into the stylistic voice of prominent Arabic authors.

[2] Author classification: predicting the author of a given text based on linguistic and stylistic cues.

[3] Human vs. machine text detection: distinguishing between human-written and AI-generated Arabic text, a growing concern with the rise of generative AI.

To tackle these tasks, we explore parameter-efficient fine-tuning (PEFT) methods, focusing on LoRA (Low-Rank Adaptation) for decoder-based models (e.g., Gemma, Qwen) as shown in Figure 1,

Proceedings of The Third Arabic Natural Language Processing Conference, pages 99–106

and full fine-tuning for encoder-based BERT variants (AraBERTv2, AraModernBert). We further introduce a layer-wise analysis framework to probe which layers in transformer models best capture stylistic and discriminative signals for Arabic, offering interpretability alongside performance.

Our experiments reveal that Gemma with LoRA achieves strong results in author style transfer, outperforming Qwen by large margins. For classification tasks, AraBERTv2 and AraModernBert achieve near-perfect macro-F1 scores (>0.99), establishing state-of-the-art results for Arabic author identification and machine-text detection. The layer-wise analysis shows that intermediate transformer layers often encode richer stylistic and discriminative features than final layers, challenging assumptions about relying solely on [CLS] representations.

The contributions of this paper are threefold:

- A benchmark-style evaluation of Arabic LLMs across diverse stylistic and discriminative tasks.
- Empirical evidence of the effectiveness of parameter-efficient fine-tuning for Arabic LLMs.
- A novel layer-wise interpretability analysis revealing how Arabic stylistic cues are encoded across model depths.

2 Tasks and Background

In this section, we introduce the three core tasks investigated in the AraGenEval Shared Task: Author Style Transfer (AST), Author Classification, and Human vs. Machine Text Detection. Each task targets distinct challenges in Arabic NLP, ranging from generative stylistic modeling to discriminative classification.

2.1 Author Style Transfer (AST)

Definition. Author Style Transfer involves rewriting an input passage in Modern Standard Arabic (MSA) into the stylistic voice of a target author while preserving semantic meaning. For example, a neutral MSA passage such as *"القول بالعموم وحده ورفض الخصوص، يعبر عن تحويل.."* may be restyled into Hassan Hanafi's philosophical rhetoric as *"والقول بالعموم وحده وإنكار الخصوص هو تحويل.."*.

Motivation. This task is essential for studying how Arabic stylistic variation can be captured and reproduced by large language models. Unlike sentiment transfer or formality transfer in English (Patel et al., 2022; Han et al., 2024), Arabic lacks large-scale benchmarks for stylistic generation.

Related Work. Prior Arabic NLP efforts have concentrated mainly on sentiment analysis, named entity recognition, and QA/reading comprehension, supported by resources such as AraBench and ArabicGLUE (Almanea, 2021; Alqahtani & Dohler, 2023; Masri et al., 2024; Sammoudi et al., 2024; Tami et al., 2024). Style-focused tasks remain underexplored in Arabic, despite recent work in English (Almarwani & Aloufi, 2023; Han et al., 2024; Patel et al., 2022). Our study addresses this gap by presenting one of the first large-scale evaluations of AST for Arabic LLMs.

2.2 Author Classification

Definition. Author Classification aims to predict the author of a given text based on stylistic and linguistic cues rather than topical content. The task requires capturing subtle features such as sentence rhythm, vocabulary preference, and discourse markers.

Motivation. Authorship identification is critical for applications in literary studies, plagiarism detection, and digital forensics (Al-Sarem et al., 2020; Alqahtani & Dohler, 2023). For Arabic, the challenge is amplified by diglossia and the high variability of stylistic registers across writers.

Related Work. While AraBERT and AraELECTRA have been widely applied to sentiment and topic classification tasks, studies on stylistic authorship attribution in Arabic are rare (Joshi et al., 2024; Khoboko et al., 2025; Lv et al., 2023). Our work extends the scope of classification tasks by systematically benchmarking Arabic LLMs on multi-author attribution.

2.3 Human vs. Machine Text Detection

Definition. Human vs. Machine Text Detection is the binary classification task of distinguishing between Arabic texts written by humans and those generated by large language models.

Motivation. The rise of generative AI has intensified concerns about misinformation, academic integrity, and authorship verification (Najjar et al., 2025; Najjar A.A. et al., 2025). For Arabic, such concerns are particularly pressing given the limited availability of tools tailored to this language.

Related Work. AI-generated text detection has been studied in English using tools such as GLTR and DetectGPT, but Arabic benchmarks remain scarce. Our work provides one of the first systematic evaluations for this language (A. Najjar et al., 2025; A. A. Najjar et al., 2025).

3 Datasets

All datasets used in this work were released as part of the AraGenEval Shared Task (Abudalfa et al., 2025). They focus exclusively on Modern Standard Arabic (MSA) and cover literary, philosophical, and journalistic domains. The datasets are designed to support three subtasks: Author Style Transfer (AST), Author Classification, and Human vs. Machine Text Detection.
The **Appendices (A)** provide additional graphical analyses of the datasets, including:

- Distribution of samples across authors (Figure 4),
- Distribution of text lengths (Figure 5),
- Word clouds highlighting lexical fingerprints of authors (Figure 6),
- t-SNE visualizations of author clustering based on AraBERT embeddings (Figure 7).

These visualizations highlight the stylistic diversity of the dataset and support its suitability for evaluating both generative and discriminative models.

3.1 Author Style Transfer (AST) Dataset

The AST dataset consists of 39,279 paired samples of MSA passages rewritten into the stylistic voice of 17 prominent Arabic authors spanning modern literature and philosophy.

- **Average length**: ~335 words per sample.
- **Range**: short phrases to long essays, up to 1,843 words.
- **Total size**: ~13.1M words.

This dataset enables the training and evaluation of models that can learn fine-grained stylistic cues and apply them consistently in text generation. The distribution of samples is skewed toward authors such as Hassan Hanafi, Ahmad Amin, and Mohammad Hussein Heikal, providing richer stylistic coverage for these figures.

3.2 Author Classification Dataset

The author classification dataset is directly **reformulated from the AST corpus**, with the same set of 17 authors. Instead of paired transformations, the task is framed as **multi-class classification**, where each paragraph is assigned its original author label.
This dataset provides a benchmark for evaluating whether encoder-based models can capture **stylistic discriminative features** beyond topical differences, a challenge rarely studied in Arabic NLP.

3.3 Human vs. Machine Text Detection Dataset

The detection dataset, named ARATECT, was newly created within the shared task to address the growing need for Arabic resources in AI-generated text detection. The construction followed these steps:

- **Human-written texts:** Collected from reputable Arabic news outlets and verified literary sources, then manually curated for quality.
- **Machine-generated texts:** Produced by Arabic-capable LLMs (e.g., GPT-4, Mistral, LLaMA) under diverse prompting strategies.
- **Annotation:** Assigned binary labels (Human vs. AI), with balanced domain coverage across news and literature.

This resource is among the first to systematically benchmark Arabic machine-text detection, complementing the generative and classification datasets.

4 System Overview

We adopt a hybrid adaptation strategy combining parameter-efficient fine-tuning (PEFT) for generative decoder-based models and full fine-tuning for encoder-based models. This section details the overall strategy and then presents task-specific configurations.

4.1 Overall Strategy

Our approach combines PEFT for decoder-based models (Gemma, Qwen) and full fine-tuning for

encoder-based BERT variants (AraBERTv2, AraModernBert). This division leverages the efficiency of LoRA in large generative models and the robustness of full fine-tuning for smaller encoder models.

4.2 Task-Specific Configurations

For **AST**, we used Gemma3-1B and Qwen2.5-1.5B fine-tuned using LoRA. The algorithm includes conditional generation. While input is concatenation of source text and target author name as a control token, output is a rewritten passage. The loss function is a standard cross-entropy on next-token prediction. To address the challenge of preventing semantic drift and to preserve meaning while shifting style, we add content-preservation constraints by penalizing high cosine distance between embeddings of input and output (using Sentence-BERT) (Liu et al., 2024; Radhakrishnan et al., 2023). This is shown in Figure 1.

AraBERTv2 and AraModernBert were used for the author classification task. The algorithm includes sequence classification using the [CLS] token representation. We fully fine-tuned with cross-entropy loss over 17 author classes. We also introduced Layer-Wise analysis for this task (Pasad et al., 2021; Van Aken et al., 2019). Instead of using only the final [CLS], we extract hidden states from each layer and train a logistic regression classifier on top. To address the challenge overfitting due to class imbalance, we used stratified splits and early stopping based on validation F1. This equation shows the Layer-Wise analysis:

$$h^l = BERT_l(x), \hat{y}^l = (Wh^l + b)$$

where we report F1 across layers $l = 1..12$ to identify the most informative depth.

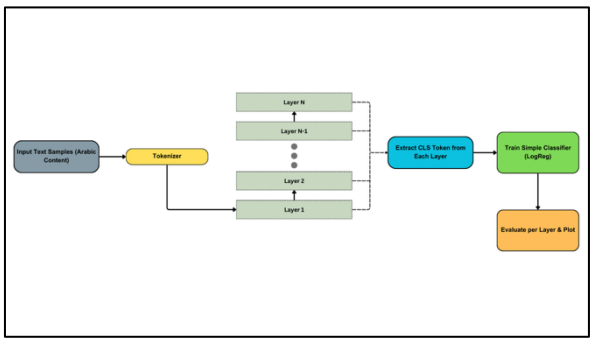

Figure 2: Layer-wise analysis**.**

For **Human vs. Machine Detection,** we also used fine-tuned AraBERTv2 and AraModernBert for binary classification with labels are {Human, AI}. We addressed the challenge of high lexical overlap between human and machine texts by applying data augmentation by paraphrasing human samples to expand stylistic variance and make the classifier robust.

4.3 Distinguishing Configurations

LoRA vs. Full Tine-Tuning: LoRA was used only for decoder models (Gemma, Qwen) due to efficiency in large generative models. Encoder models (AraBERTv2, AraModernBert) were fully fine-tuned since they are relatively small.

Intermediate vs. Final Layers: For classification, we explicitly compared performance across layers to uncover interpretability insights using layer-wise analysis.

5 Experimental Setup

For all tasks, data was split into training, development, and test sets (70/15/15 for style transfer and author classification; 80/10/10 for human vs. machine detection), stratified by class to preserve distribution. Preprocessing included standard Arabic normalization (removing diacritics, unifying punctuation, and normalizing character variants) and model-specific tokenization with a maximum sequence length of 512. Results are summarized in Table 1.

Encoder-based models (AraBERTv2, AraModernBert) were fully fine-tuned using AdamW ($lr = 2e-5$, batch size$= 4$, epochs$= 3$, 5% warmup). Decoder-based models (Gemma, Qwen) employed LoRA adapters ($r \in \{16,32,64\}$, dropout $= 0.05$, $lr = 1e-4$), applied to attention and projection modules.

Implementation used Hugging Face Transformers (v4.41.2), PEFT (v0.11.1), PyTorch (v2.3.0), and scikit-learn (v1.5.0). Evaluation metrics varied by task: BLEU/chrF for style transfer, accuracy and macro-F1 for classification, and accuracy/F1 for machine-text detection.

Task	Split	Models	Metrics
[1]	70/15/15	Gemma, Qwen	BLEU, chrF
[2]	70/15/15	AraBERTv2, AraModernBert	Accuracy, Macro-F1
[3]	80/10/10	All	Accuracy, Macro-F1

Table 1: Experimental Setup Summary.

6 Results

In this section, we present results separately for each sub-task: Author Style Transfer (AST),

Author Classification, and Human vs. Machine Detection. This structure highlights the comparative strengths of parameter-efficient fine-tuning (LoRA) and full fine-tuning across tasks.

6.1 Author Style Transfer (AST)

Table 2 reports BLEU and chrF scores for Gemma and Qwen models fine-tuned with LoRA adapters of varying ranks. The results indicate that Gemma consistently outperforms Qwen across both metrics. The best configuration is Gemma with rank r=32r=32r=32, which achieves a BLEU score of 19.04 and a chrF score of 55.14. In contrast, Qwen at rank r=16r=16r=16 performs considerably worse, obtaining a BLEU of 10.18 and chrF of 44.42.

Table 2: Results on 100 unseen Arabic articles.

Model Variant	BLEU Score	chrF Score
Gemma (r=64)	18.85	55.00
Gemma (r=32)	19.04	55.14
Gemma (r=16)	18.13	54.75
Qwen (r=16)	10.18	44.42

6.2 Author Classification

The results for author classification are presented in Table 3. AraBERTv2 achieved the highest performance, with an accuracy of 89.7% and a macro-F1 score of 0.89. AraModernBert followed with an accuracy of 87.1% and a macro-F1 score of 0.87. The layer-wise analysis provides additional insights: AraBERTv2 shows peak discriminative performance in intermediate layers (7–10), while AraModernBert encodes stylistic information more evenly across deeper layers. These findings highlight that intermediate transformer layers carry stronger stylistic signals than final layers, suggesting that representation depth plays a critical role in modeling stylistic variation in Arabic text

Table 3: Results for author classification.

Model	Accuracy	F1	Best Layer
AraBERTv2	89.71%	0.89	7
AraModernBert	87.1%	0.87	20

6.3 Human vs. Machine Detection

The binary classification results for distinguishing human- from AI-generated text are shown in Table 4. Both models reached near-ceiling performance, with AraModernBert achieving the highest accuracy of 99.4% and AraBERTv2 achieving the best macro-F1 of 0.9932.

Table 4: Results for human vs. machine detection.

Model	Accuracy	F1
AraBERTv2	99.3%	0.9932
AraModernBert	99.4%	0.9923

6.4 Comparative Insights

The comparison between full fine-tuning (for classification tasks) and LoRA (for generative tasks) highlighted clear trade-offs. Full fine-tuning enabled stable convergence and higher robustness under limited data, while LoRA delivered strong performance with fewer trainable parameters, making it attractive for scaling across multiple tasks.

To improve interpretability, we conducted a **layer-wise probing analysis**. Instead of relying only on the final [CLS] token, we extracted hidden states from each transformer layer (l = 1..12) and trained lightweight classifiers on them. Results show that **mid-level layers (7–10 in AraBERTv2)** captured the strongest stylistic and discriminative cues, while final layers tended to compress information and reduce distinctiveness. This suggests that intermediate layers preserve stylistic richness, consistent with findings in English models (Pasad et al., 2021; Van Aken et al., 2019). Figure 3 illustrates this trend for AraBERTv2 vs. AraModernBert.

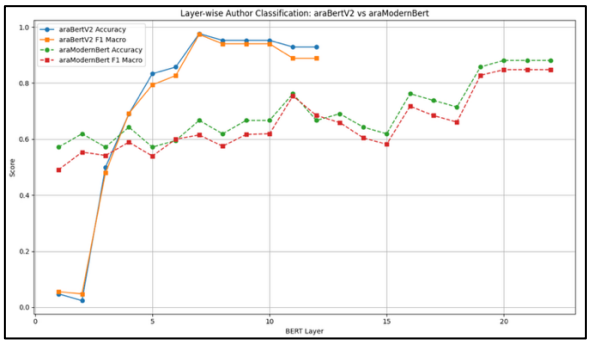

Figure 3: Layer-wise performance comparison between AraBERTv2 and AraModernBert for the author classification task. Both accuracy and

macro-F1 scores are shown across transformer layers**.**

Figure 3 also illustrates how performance evolves across layers of AraBERTv2 and AraModernBert. AraBERTv2 reaches peak accuracy and F1 around the middle layers (7–10), stabilizing near 0.99, while AraModernBert shows steadier gains across layers, with slightly lower but more consistent performance. This suggests AraBERTv2 encodes discriminative stylistic features earlier in its hierarchy, while AraModernBert distributes them more evenly, indicating differences in representational depth and efficiency.

6.5 Error Analysis

For author classification, common confusions occurred between authors with overlapping stylistic traits (e.g., similar sentence lengths or frequent religious expressions). For AST, errors often manifested as partial rewrites where the system retained source author lexical choices rather than fully adapting to the target style. For AI-generated text detection, misclassifications were rare but notable: in a few cases, highly fluent ChatGPT-like generations were labeled human, while noisy user-generated social media text was mislabeled as machine, showing the limits of surface-level stylistic cues.

7 Conclusion

We benchmarked Arabic LLMs on three challenging tasks including AST, author classification, and AI-generated text detection: comparing full-tuning and PEFT. Results showed that Arabic-specialized models, particularly AraBERTv2, achieve strong performance, with layer-wise analysis revealing where task-relevant features emerge. While domain sensitivity and limited benchmark resources remain challenges, this work offers one of the first multi-task evaluations of Arabic LLMs, establishing a replicable foundation and pointing toward broader dialectal coverage, cross-lingual transfer, and improved interpretability as key directions for future research.

This work highlights that PEFT, combined with careful layer-wise analysis, can unlock the full potential of Arabic LLMs, which brings stylistic shade, discriminative power, and robustness against AI-generated text detection into closer reach for underrepresented languages.

References

Abudalfa, S., Ezzini, S., Abdelali, A., Alami, H., Benlahbib, A., Chafik, S., El-Haj, M., Mahdaouy, A. El, Jarrar, M., Lamsiyah, S., & Luqman, H. (2025). The AraGenEval Shared Task on Arabic Authorship Style Transfer and AI-Generated Text Detection. *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*.

Almanea, M. M. (2021). Automatic methods and neural networks in Arabic texts diacritization: a comprehensive survey. *IEEE Access*, *9*, 145012–145032.

Almarwani, N., & Aloufi, S. (2023). SANA at NADI 2023 shared task: Ensemble of Layer-Wise BERT-based models for Dialectal Arabic Identification. *Proceedings of ArabicNLP 2023*, 625–630.

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

Al-Sarem, M., Saeed, F., Alsaeedi, A., Boulila, W., & Al-Hadhrami, T. (2020). Ensemble methods for instance-based Arabic language authorship attribution. *IEEE Access*, *8*, 17331–17345.

Ashqar, H. I., & Tami, M. (2025). Translation with LLMs through Prompting with Long-Form Context. *Authorea Preprints*.

Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *ArXiv Preprint ArXiv:2403.14608*.

Joshi, S., Khan, M. S., Dafe, A., Singh, K., Zope, V., & Jhamtani, T. (2024). Fine tuning LLMs for low resource languages. *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, 511–519.

Khoboko, P. W., Marivate, V., & Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, *20*, 100649.

Liu, S., Agarwal, S., & May, J. (2024). Authorship style transfer with policy optimization. *ArXiv Preprint ArXiv:2403.08043*.

Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., & Qiu, X. (2023). Full parameter fine-tuning for large language models with limited resources. *ArXiv Preprint ArXiv:2306.09782*.

Masri, S., Raddad, Y., Khandaqji, F., Ashqar, H. I., & Elhenawy, M. (2024). Transformer Models in Education: Summarizing Science Textbooks with AraBART, MT5, AraT5, and mBART. *ArXiv Preprint ArXiv:2406.07692*.

Najjar, A. A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2025). Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity. *ArXiv Preprint ArXiv:2501.03203*.

Najjar, A., Ashqar, H. I., Darwish, O., & Hammad, E. (2025). Leveraging Explainable AI for LLM Text Attribution: Differentiating Human-Written and Multiple LLMs-Generated Text. *ArXiv Preprint ArXiv:2501.03212*.

Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914–921.

Patel, A., Andrews, N., & Callison-Burch, C. (2022). Low-resource authorship style transfer: Can non-famous authors be imitated? *ArXiv Preprint ArXiv:2212.08986*.

Radhakrishnan, S., Yang, C.-H. H., Khan, S. A., Kiani, N. A., Gomez-Cabrero, D., & Tegner, J. N. (2023). A parameter-efficient learning approach to arabic dialect identification with pre-trained general-purpose speech model. *ArXiv Preprint ArXiv:2305.11244*.

Sammoudi, M., Habaybeh, A., Ashqar, H. I., & Elhenawy, M. (2024). Question-Answering (QA) Model for a Personalized Learning Assistant for Arabic Language. *ArXiv Preprint ArXiv:2406.08519*.

Tami, M., Ashqar, H. I., & Elhenawy, M. (2024). Automated Question Generation for Science Tests in Arabic Language Using NLP Techniques. *ArXiv Preprint ArXiv:2406.08520*.

Van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How does bert answer questions? a layer-wise analysis of transformer representations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1823–1832.

A Appendices

For AST dataset, Figure 4 illustrates the distribution of text samples collected for various authors in a dataset used to fine-tune a LLM for Arabic author style transfer. The dataset includes prominent Arabic literary and philosophical figures, with Hassan Hanafi, Ahmad Amin, and Mohammad Hussein Heikal having the highest number of samples, indicating a richer representation of their stylistic patterns for training the model. The horizontal bars visualize the number of samples per author, supporting tasks like stylistic imitation and authorship transformation.

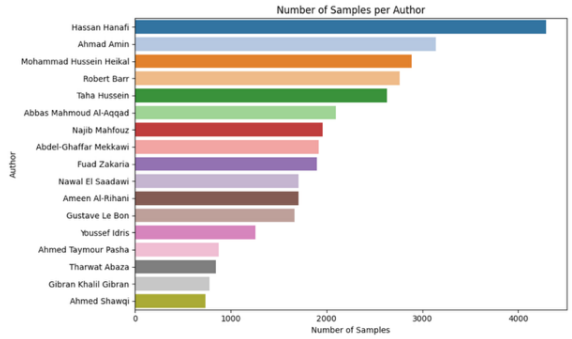

Figure 4: Number of Samples per Author in Arabic Author Style Transfer Dataset.

Moreover, Figure 5 displays the distribution of MSA text lengths, measured in number of words, across the dataset used for fine-tuning the author style transfer model. The distribution is highly concentrated around 350–400 words, with a sharp peak indicating that most samples fall within this range. The presence of a kernel density estimate (KDE) overlay highlights the unimodal and right-skewed nature of the data, where very few samples exceed 600 words. This suggests a consistent and controlled sample length throughout the dataset, which is beneficial for stable training and style learning in LLMs.

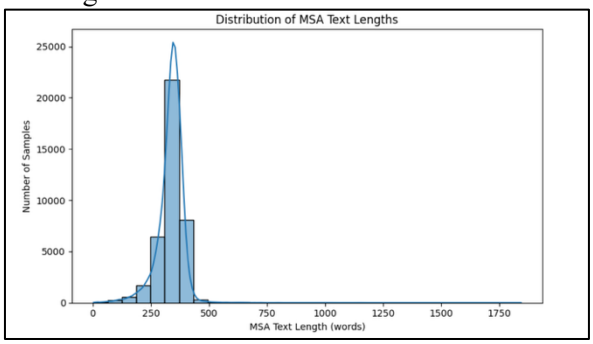

Figure 5: Distribution of MSA Text Lengths in the Arabic Author Style Transfer Dataset.

Figure 6 shows 17-word cloud subplots visualizing the most frequent and prominent words in the writings of each author from the Arabic AST dataset. The diversity of themes is evident: authors like Nawal El Saadawi and Abbas Al-Aqqad focus on gender and humanism, while Taha Hussein and Ahmad Amin emphasize thought and knowledge. Poets like Ahmed Shawqi and Gibran Khalil Gibran favor expressive and emotional lexicons, whereas philosophers such as Fuad Zakaria and Hassan Hanafi employ rational and abstract terminology.

These visualizations highlight the unique lexical fingerprints of each author, showcasing their stylistic identity. Such distinctions are foundational for fine-tuning language models to perform accurate author style transfer, as the model must learn to emulate not just surface-level vocabulary, but the deeper thematic and stylistic choices each author consistently demonstrates.

Figure 6: Word Clouds of Most Frequent Words Across 17 Arabic Authors. (a–q) show the most frequent words used by different authors in the dataset: (a) Youssef Idris, (b) Tharwat Abaza, (c) Taha Hussein, (d) Robert Barr, (e) Nawal El Saadawi, (f) Najib Mahfouz, (g) Hassan Hanafi, (h) Mohammad Hussein Heikal, (i) Gustave Le Bon, (j) Gibran Khalil Gibran, (k) Fuad Zakaria, (l) Ahmed Taymour Pasha, (m) Ameen Al-Rihani, (n) Ahmed Shawqi, (o) Ahmad Amin, (p) Abbas Mahmoud Al-Aqqad, and (q) Abdel-Ghaffar

Mekkawi. Each subplot highlights the author's dominant vocabulary, providing insight into their unique lexical and thematic style.

For Author Classification, the t-SNE visualization shown in Figure 7 represents the clustering of Arabic text samples based on [CLS] token embeddings produced by a fine-tuned AraBERTv2 model, trained for the task of author classification. Each point represents a text sample, and colors correspond to different authors. The embeddings were projected into 2D space using t-SNE for visualization purposes.

Figure 7 illustrates how well the fine-tuned AraBERTv2 model captures the distinct stylistic and semantic features of different authors in the dataset. Clear and well-separated clusters, such as those for Nawal El Saadawi, Taha Hussein, and Robert Barr, suggest that the model has successfully learned author-specific linguistic patterns, enabling high confidence in distinguishing between them.

Some clusters are positioned close to others (e.g., Ahmad Amin and Mohammad Hussein Heikal), indicating potential stylistic or thematic similarities between those authors' writing. Meanwhile, others like William Shakespeare (likely translated texts) or George Zaidan show strong separation, hinting at distinct lexical or structural traits.

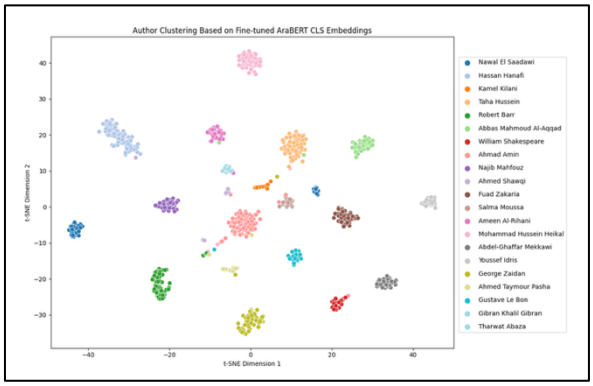

Figure 7: Author Clustering Based on Fine-Tuned AraBERT CLS Embeddings.