The AraGenEval Shared Task on Arabic Authorship Style Transfer and AI Generated Text Detection

Shadi Abudalfa[1], Saad Ezzini[1], Ahmed Abdelali[2], Hamza Alami[3], Abdessamad Benlahbib[3], Salmane Chafik[4], Mo El-Haj[5,8], Abdelkader El Mahdaouy[4], Mustafa Jarrar[6,9], Salima Lamsiyah[7], Hamzah Luqman[1]

[1]King Fahd University of Petroleum & Minerals, [2]Humain,
[3]Sidi Mohamed Ben Abdellah University, [4]Mohammed VI Polytechnic University, [5]VinUniversity, [6]Hamad Bin Khalifa University,
[7]University of Luxembourg, [8]Lancaster University, [9]Birzeit University

Abstract

We present an overview of the AraGenEval shared task, organized as part of the ArabicNLP 2025 conference. This task introduced the first benchmark suite for Arabic authorship analysis, featuring three subtasks: Authorship Style Transfer, Authorship Identification, and AI-Generated Text Detection. We curated high-quality datasets, including over 47,000 paragraphs from 21 authors and a balanced corpus of human- and AI-generated texts. The task attracted significant global participation, with 72 registered teams from 16 countries. The results highlight the effectiveness of transformer-based models, with top systems leveraging prompt engineering for style transfer, model ensembling for authorship identification, and a mix of multilingual and Arabic-specific models for AI text detection. This paper details the task design, datasets, participant systems, and key findings, establishing a foundation for future research in Arabic stylistics and trustworthy NLP.

1 Introduction

The rise of user- and machine-generated Arabic content across social media platforms, digital journalism, literary archives, and online educational resources has created an urgent demand for advanced NLP tools capable of analysing, transforming (Abudalfa et al., 2024; Abdu et al., 2025), and verifying text style (El-Haj et al., 2024; El-Haj and Ezzini, 2024). Unlike general stylistic analysis, which seeks to characterise an author's linguistic footprint, Authorship Style Transfer (AST) aims to actively modify a given text to reflect the stylistic features of a target author while preserving its semantics. Meanwhile, the proliferation of Arabic content generated by large language models (LLMs) has raised the stakes for AI-generated text detection systems (Zmandar et al., 2023). As the line between human and synthetic writing becomes increasingly blurred, particularly in Arabic with its orthographic and dialectal variability, it is critical to establish robust benchmarks and methodologies for style manipulation (Mughaus et al., 2025) and content authenticity assessment. Prior efforts in Arabic readability modelling (El-Haj and Rayson, 2016) and corpus development (El-Haj and Koulali, 2013) have laid essential groundwork for Arabic linguistic resource creation, but there remains a significant gap in structured evaluations targeting stylistic transformation and AI-authored text detection.

To address this need, we launched the AraGenEval Shared Task, hosted at the ArabicNLP 2025 conference (co-located with EMNLP 2025). AraGenEval complements prior Arabic NLP shared tasks (Malaysha et al., 2024) and aims to fill a critical gap in Arabic style transfer and authorship detection, where no dedicated benchmark has previously been released. AraGenEval features three subtasks designed to advance research in Arabic authorial style processing:

1. **Authorship Style Transfer (AST)**: Given a formal Arabic input, generate a stylistically faithful version in the voice of a specific author from a curated set of 21 classical and modern writers.

2. **Authorship Identification**: Determine the most likely author for a given text segment using multiclass classification.

3. **ARATECT (Arabic AI-Generated Text Detection)**: Distinguish between human- and LLM-generated Arabic texts across news and literary genres.

Motivation

The motivation behind AraGenEval is both linguistic and socio-technical. Authorship style transfer (AST) offers valuable insights into how stylistic signals operate in Arabic, supporting applications

Proceedings of The Third Arabic Natural Language Processing Conference, pages 1–13

such as educational feedback, personalisation, and literary imitation, while addressing the typological and orthographic characteristics of the language (Alqahtani and Yannakoudakis, 2022). At the same time, Arabic authorship identification and AI-generated text detection have become increasingly important for digital forensics, media verification, and preserving cultural authenticity, as demonstrated by recent work on stylometric detection of LLM-generated Arabic text (Al-Shaibani and Ahmed, 2025) and competitive system development in shared evaluation tasks (Chowdhury et al., 2024; AL-Smadi, 2025). Furthermore, studies show that models trained on English frequently fail to generalise to Arabic due to differences in script, morphology, and dialectal variation, underscoring the need for dedicated Arabic-specific evaluation frameworks (Al-Shaibani and Ahmed, 2025).

Challenges

Arabic presents unique challenges for AST and detection:

- **Stylistic Variation**: Arabic exhibits a continuum of registers from Modern Standard Arabic to regional dialects, with authorial voice often tied to historical, literary, or journalistic contexts (Habash, 2010).
- **Data Sparsity**: Compared to English, there are far fewer large-scale, author-labelled Arabic corpora (El-Haj and Koulali, 2013; El-Haj and Ezzini, 2024).
- **Morphological Richness**: Arabic's complex morphology makes it harder to isolate stylistic features from lexical ones (El-Haj et al., 2018).

AraGenEval

AraGenEval[1] offers a unified framework and high-quality datasets to benchmark models on these challenges. We collected over 47,000 human-written paragraphs from 21 classical and modern Arabic authors, and curated a balanced corpus of human- and AI-generated news and literary texts. Submissions were evaluated via BLEU and chrF for generation, macro-F1 for multiclass classification, and accuracy/F1 for binary classification.

The task received strong engagement from the global NLP community:

- 72 teams registered (115 participants in total).
- 37 unique submissions to the leaderboard across the three subtasks.
- 16 countries, including: India, Pakistan, Saudi Arabia, Qatar, Tunisia, Egypt, Palestine, Algeria, Morocco, Japan, Vietnam, UAE, Spain, UK, US, and France.

AraGenEval contributes the first benchmark suite tailored for Arabic authorship manipulation and AI-authorship detection, and sets the foundation for future research in Arabic stylistics, forensic linguistics, and trustworthy NLP.

2 Related Work

Authorship Style Transfer (AST) is a specialized task in natural language generation that modifies the stylistic elements of a text, such as lexical choice, syntactic patterns, and rhetorical flourishes, to mimic a target author's voice while preserving the original content. Unlike broader Text Style Transfer (TST), AST specifically targets writer-specific traits, including narrative tone, sentence complexity, and idiosyncratic phrasing. The focus of TST was to modifies stylistic attributes (e.g., politeness, formality, sentiment) of text while preserving its core content.

Recent advances in deep learning and LLMs have significantly advanced TST research, enabling more nuanced and convincing stylistic adaptations. The researchers use different methods and approaches to solve this challenge. Supervised approaches use parallel data with encoder-decoder models (e.g., sequence-to-sequence) (Hu et al., 2022; Gong et al., 2019) that models the problem as a translation task. Other approaches include copy mechanism (Pan et al., 2024; Chawla and Yang, 2020) proposed to better support sections of text which should not be changed (e.g., some proper nouns and rare words) (Merity et al., 2016). (Hu et al., 2017) exploited deep learning methods like Variational Autoencoders (VAE) and Denoising Autoencoders (DAE) to modify textual styles while preserving the original content. They utilize the VAE framework to learn the latent representation of text and employ a style classifier to discern the style attribute vector.

Authorship Identification is the task of determining the author of a text from a set of known candidates (Mosteller and Wallace, 1963). The

[1]AraGenEval URL: https://ezzini.github.io/AraGenEval

field is historically rooted in **stylometry**, the quantitative study of literary style, which operates on the premise that authors have unique linguistic "fingerprints" (Mosteller and Wallace, 1963; Lagutina et al., 2019). Traditional approaches involved manually engineering a wide array of lexical and syntactic features, including word frequencies, sentence lengths, and punctuation usage, and using them to train classical machine learning classifiers, including logistic regression, Naive Bayes, and support vector machines (SVM) (Aborisade and Anwar, 2018; Bacciu et al., 2019). However, the advent of deep learning marked a paradigm shift, moving the field from manual feature engineering to automated feature extraction (Bauersfeld et al., 2023; Huang et al., 2025). Recently, machine learning methods have explored recurrent neural networks (RNNs) (Bagnall, 2015), long short-term memory networks (LSTMs) (Qian et al., 2017), convolutional neural networks (CNNs) at character and word levels (Ruder et al., 2016; Shrestha et al., 2017), and hybrid Siamese or attention-based networks (Boenninghoff et al., 2019; Saedi and Dras, 2021). With the rise of pre-trained language models, BERT and its variants (Devlin et al., 2019; Fabien et al., 2020; Huertas-Tato et al., 2022) have become the dominant paradigm, often enhanced by supervised contrastive learning (Khosla et al., 2020). While effective, they remain challenged by cross-domain generalization and explainability (Rivera-Soto et al., 2021). More recently, LLMs have been applied for feature extraction, annotation, and even end-to-end attribution, showing promise in domain transfer and interpretability (Brown et al., 2020; Huang et al., 2024, 2025).

Within Arabic NLP, authorship identification has been investigated across diverse genres, from classical literature and poetry to modern social media. Shared tasks such as PAN/CLEF (Rosso, 2017) on author profiling and AraPlagDet (Bensalem et al., 2015) on plagiarism detection provided early benchmarks, though neither directly addressed multi-author attribution in Arabic. A recent survey of 27 Arabic studies highlights large performance variability, driven by differences in genre, feature design, and dataset size, and emphasizes the difficulty posed by morphology and diglossia (Alqahtani and Dohler, 2023). More recent advances demonstrate the advantage of Arabic-specific pre-trained models such as AraBERT (Antoun et al., 2020a), AraELECTRA (Antoun et al., 2020b), and CAMeLBERT, which consistently outperform multilingual baselines on tasks including attribution of classical poetry and Islamic legal texts (AlZahrani and Al-Yahya, 2023; Alqurashi et al., 2025). Nevertheless, cross-domain transfer remains a persistent challenge, as models trained on social media rarely generalize to literary or journalistic prose. The lack of unified, large-scale Arabic benchmarks makes systematic evaluation difficult, a gap that AraGenEval seeks to fill by providing a multi-genre, multi-author benchmark for Arabic authorship identification.

Arabic AI-Generated Text Detection is framed as a binary classification task, aiming to determine whether a given text was authored by a human or produced by a machine. Approaches applied to this task are typically grouped into four main categories (Wu et al., 2025): (i) *statistics-based methods*, which exploit entropy or n-gram distributions to capture distributional irregularities in machine text (Shen et al., 2023; Mitchell et al., 2023); (ii) *neural-based methods*, including fine-tuned transformers such as BERT and RoBERTa, which achieve strong performance but face robustness challenges under adversarial conditions (Ippolito et al., 2020; Li et al., 2025); (iii) *watermarking approaches*, embedding token-level or hidden-space signals to enable proactive detection (Kirchenbauer et al., 2023; Zhao et al., 2023); and (iv) *LLM-as-detector frameworks*, where large models themselves are used to classify or explain text origins (Wang et al., 2024b; Su et al., 2025).

Recent work has also explored leveraging Arabic-specific transformer architectures for generative text detection, highlighting both linguistic and orthographic challenges in low-resource settings (Alshammari and Elleithy, 2024). To standardize evaluation, recent benchmarks such as MultiSocial (Macko et al., 2025), XDAC (Go et al., 2025), and M4GT-Bench (Wang et al., 2024b) test cross-domain generalization, while shared tasks like SemEval-2024 Task 8 (Wang et al., 2024a), the GenAI Content Detection Task on academic essay authenticity (Chowdhury et al., 2024), and the M-DAIGT challenge (Lamsiyah et al., 2025) and , and the GenAI Content Detection Task 3, which focused on detector performance in a setting with a large but fixed set of known domains and models (Dugan et al., 2025). However, the field still lacks large-scale, standardized benchmarks and shared tasks for Arabic. Addressing this gap, recent evaluation on the AIRABIC dataset

demonstrates that current detectors like GPTZero and OpenAI's Text Classifier struggle with Arabic, especially in the presence of diacritics, revealing detection accuracy as low as 30% and underscoring design limitations in Semitic language contexts (Alshammari and Ahmed, 2023). Motivated by this gap, AraGenEval's ARATECT subtask proposes the first multi-genre evaluation framework dedicated to Arabic AI-generated text detection.

3 Data Collection and Selection

3.1 Authorship Style Transfer

We began by gathering works from 21 distinct authors with all sources publicly accessible. For each author, a selection of 10 books was made. The texts were then divided into coherent paragraphs using the Natural Language Toolkit (NLTK)[2]. In particular, this tool was employed to partition the material into segments of 2048 characters, ensuring no overlap between sections. Furthermore, the word_tokenize function from NLTK was applied to tokenize the paragraphs, after which any segment exceeding 2048 tokens was excluded. We then employed the GPT-4o mini LLM to convert the selected paragraphs into a more formalized standard style. The prompt utilized for this process is presented in Listing 1.

Listing 1: Prompt Applied in Building the Arabic Style Transfer Dataset

```
{"role": "system",
"content": "You are a helpful assistant."},
{
"role": "user",
"content": f"Rewrite the following text in
    Modern Standard Arabic (MSA) while
    maintaining its original meaning but
    changing the style to be more formal,
    neutral, and consistent with modern
    writing standards. Ensure the language is
    polished and does not reflect the
    author's original stylistic features:
    {text}"}
```

We selected parallel source–target pairs that could be accommodated within the context length restrictions of the LLMs under evaluation, as the generated texts were relatively long. For tokenization, the jais-family-13b-chat model was employed to process these pairs. Only instances in which the total number of tokens across both source and target texts was under 1900 were preserved. We divided the collected dataset into three sets: training, validation, and testing. A statistical overview is provided in Table 1.

Author	Train	Test	Val
A. Amin	2892	594	246
A. T. Pasha	804	142	53
A. Shawqi	596	46	58
A. Rihani	1557	624	142
T. Abaza	755	191	90
G. K. Gibran	748	240	30
J. Zaydan	2762	562	326
H. Hanafi	3735	1002	548
R. Barr	2680	512	82
S. Moussa	984	282	119
T. Hussein	2371	534	253
A. M. Al-Aqqad	1820	499	267
A. G. Makawi	1520	464	396
G. Le Bon	1515	358	150
F. Zakaria	1771	294	125
K. Kilani	399	109	25
M. H. Heikal	2627	492	260
N. Mahfouz	1630	343	327
N. El Saadawi	1415	382	295
W. Shakespeare	1236	358	238
Y. Idris	1140	349	120

Table 1: Authorship style transfer dataset statistics by author and data split.

3.2 Authorship Identification

For this task, we employed the same dataset described in Section 3.1. However, rather than using the ground truth text as the target text, we assign the author's name as the label, since this task involves multiclass classification rather than text generation.

3.3 Arabic AI-Generated Text Detection (ARATECT)

To support the ARATECT subtask, we created a dataset specifically designed to train and evaluate systems for detecting AI-generated news articles in Arabic.

The first step involved collecting 2,900 news articles from multiple categories from two Arabic news websites, Al Jazeera[3] and Hespress[4], to represent human-written samples across a variety of categories. To generate AI-written counterparts, we extracted the titles from these human-written articles and used them as input prompts. The content of the original articles was used to guide the AI in mimicking human writing style. After filtering and qualitative analysis, we selected a subset of 2,400 total articles to move forward with. Several high-performing reasoning and non-reasoning

[2] https://www.nltk.org

[3] https://www.aljazeera.net

[4] http://hespress.ma

language models were employed to generate the AI-written news content, including variants of Gemini (Gemini-2.5-pro) and GPT (gpt-3.5, gpt-4o-mini, gpt-4o, gpt-o4-mini). Each model was prompted using a standardized prompt shown in Listing 2.

Using this prompt on the 2,400 human-written articles, we generated 2,400 AI-generated counterparts using different LLMs, resulting in a training set of 4,800 samples. This training set was used to fine-tune a baseline model for detecting AI-generated news articles in Arabic.

For the test and development sets generation, we used an agent-based approach incorporating the aforementioned fine-tuned detection model into the pipeline illustrated in Figure 1. In this pipeline, we engage in an iterative interaction with the LLM:

- The model is first prompted to generate a news article based on a given title and writing style.
- The generated text is then evaluated by the baseline model.
- If the text is flagged as AI-generated, we inform the LLM that its previous output was detected as such, and request a new version.
- This process is repeated until the generated text is either classified as human-written (it is included in the dataset) or a predefined iteration threshold n_i is reached (we move to the next example).

As a final result, we obtained a balanced dataset of 5,800 news article samples, containing both human-written and AI-generated texts, split into 4,800 for training, 500 for development, and 500 for testing to support comprehensive model evaluation.

Listing 2: Prompt's Key Components for Generating News Articles

```
-- Each time this prompt is used, a role is
    randomly selected to influence the
    assistant writing style.

-- Randomly select one of the following
    journalist roles:

Role Definition:
       - "You are Tarik Mekouar, an expert
           Arabic journalist. Here is an
           example of how Tarik wrote:
           {first_paragraph}".
       - "You are Amal Kanin, a professional
           Arabic news writer with a focus on
           clear, unbiased reporting. Here is
           an example of how Amal wrote:
           {first_paragraph}".
       - "You are Youssef Yaakoubi, a
           friendly and engaging Arabic
           journalist, writing in an
           easy-to-understand style. Here is
           an example of how Youssef wrote:
           {first_paragraph}".
       - "You are Manal Lotfi, an opinion
           Arabic writer, focusing on
           offering personal insights on
           current news. Here is an example
           of how Manal wrote:
           {first_paragraph}".

-- Instructions:

Write a '{article_length}'-word news article
    about the following topic : '{title}'.

Focus only on the article content. Do not
    include a title.
```

4 Subtasks with Evaluation Tracks

We ran three subtasks via CodaBench platform with two main phases, development and testing phases.

4.1 Authorship Style Transfer

This subtask challenges participants to develop systems that can rewrite a given formal Arabic text to emulate the distinct style of a specific author, while ensuring the original meaning of the text is preserved. The evaluation of the generated text is based on its closeness to the target author's style. The primary metric for this task is the *BLEU* score, which measures the correspondence between the machine-generated output and high-quality reference translations. Additionally, the *chrF* score is used as a secondary metric, which evaluates character n-gram precision and recall, providing a more granular assessment of stylistic similarity.

4.2 Authorship Identification

The goal of this subtask is to identify the author of a given Arabic text from a set of 21 possible authors. This is a multiclass classification problem where systems are expected to analyze the stylistic features of the text to make an accurate prediction. The primary evaluation metric is the *Macro-F1 score*, which calculates the F1 score for each author independently and then averages them, treating all classes equally. *Accuracy*, the proportion of correctly identified authors, serves as the secondary metric.

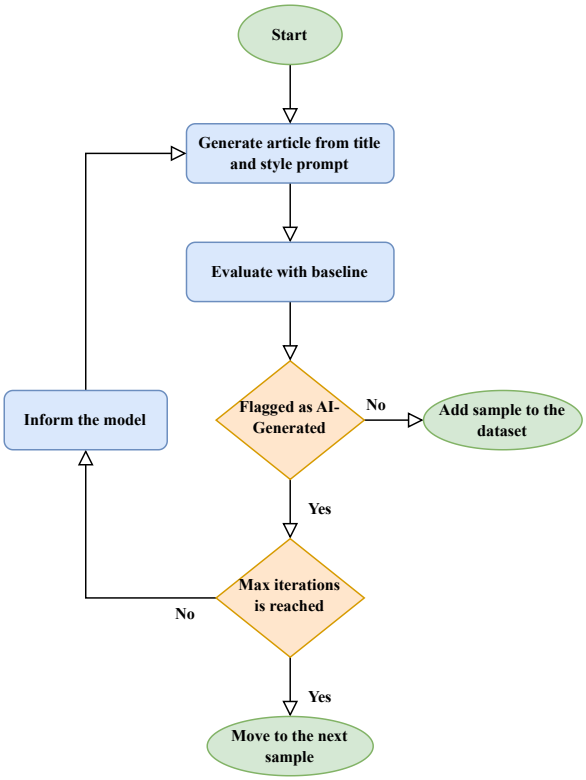

Figure 1: News generation pipeline for subtask 3

4.3 Arabic AI-Generated Text Detection

This subtask, also known as ARATECT, focuses on distinguishing between human-written and AI-generated Arabic texts. Participants are tasked with building a binary classification model to detect AI-generated content within the domain of Arabic news. The performance of the systems is evaluated primarily based on the F1-Score, which provides a balance between precision and recall. Accuracy is used as a secondary metric to measure the overall correctness of the classification.

4.4 Participants Systems

4.4.1 Subtask 1: Authorship Style Transfer

For the Authorship Style Transfer task, participants explored a range of generative models and fine-tuning strategies. The winning team, **ANLPers** (Nacar et al., 2025), achieved top performance by employing prompt engineering with AraT5, framing the task as an explicit natural language instruction in Arabic. This was followed by **Nojoom.AI** (KARA ACHIRA et al., 2025), who fine-tuned several pre-trained Seq2Seq models, including mBART and AraT5, and incorporated LoRA for efficient adaptation. The third-place team, **MarsadLab** (Biswas et al., 2025b), also leveraged parameter-efficient fine-tuning, applying LoRA to instruction-following Arabic LLMs like Qwen2.5-7B-Instruct. Other teams, such as **Osint** (Agrahari et al., 2025), fine-tuned an AraT5-based encoder-decoder model with author conditioning.

4.4.2 Subtask 2: Authorship Identification

The Authorship Identification task saw a variety of approaches, from complex ensembles to traditional machine learning. The winning team, **Sebaweh** (Helmy et al., 2025), developed a robust ensemble model that combined four fine-tuned transformer-based models: AraBERT, CAMELBERT, Arabic XLM-ROBERTa, and GATE-AraBERT. The third-place team, **Athership** (Samir et al., 2025), also used an ensemble approach with a dual-model logit fusion of AraBERT and AraELECTRA. The fourth-place team, **MISSION** (ALHARBI, 2025), fine-tuned the ALLaM-7B-Instruct-preview model using prompt engineering. In contrast, the eighth-place team, **Amr&MohamedSabaa** (Sabaa and Sabaa, 2025), demonstrated the effectiveness of traditional methods by combining word-level and character-level TF-IDF features with a Logistic Regression classifier. Other participants, such as **NLP_wizard** (Hany, 2025), used a lightweight approach with pre-trained XLM-ROBERTa embeddings fed into classical classifiers like LinearSVC. **Jenin** (Malhis et al., 2025) team conducted a layer-wise analysis of the fine-tuned BERT model to locate where author-discriminative signals emerge and how the model encodes style.

4.4.3 Subtask 3: Arabic AI-Generated Text Detection

For the ARATECT task, participants employed a diverse set of models and techniques. The winning team, **LMSA** (Zita et al., 2025), used an ensemble-based framework that integrated multilingual and Arabic-specific models, namely Fanar, AraBERT, and XLM-RoBERTa, with a majority voting strategy. The third-place team, **MISSION** (ALHARBI, 2025), fine-tuned AraModernBERT on a combination of the official dataset and an external dataset. The fourth-place team, **PTUK-HULAT** (Duridi et al., 2025), fine-tuned multilingual transformer models based on XLM-ROBERTa. The fifth-place team, **BUSTED** (Zain et al., 2025), conducted a comparative study of AraELECTRA, CAMELBERT, and XLM-ROBERTa, finding that the multilingual XLM-ROBERTa performed best. Other notable approaches included

CUET-NLP_Team_SS306's use of a chunking strategy with AraBERT to handle long input sequences (Nath et al., 2025) and **REGLAT**'s morphology-aware AraBERT model (Labib et al., 2025).

4.5 Results

This section presents the results for each of the three subtasks. A total of 37 unique submissions were made to the leaderboard across all tasks.

4.5.1 Subtask 1: Authorship Style Transfer

The results for the authorship style transfer task are shown in Table 2. The top-performing systems achieved BLEU scores around 24.5. Team **ANLPers** secured the first place with a BLEU score of 24.58, closely followed by team **Nojoom.AI** with a score of 24.46.

4.5.2 Subtask 2: Authorship Identification

The authorship identification task was highly competitive. As shown in Table 3, the top 11 participants achieved high performance, with only a 10% difference in their Macro-F1 scores. Team **Sebaweh** ranked first with a Macro-F1 of 0.8989, followed by team **batoolnajeh** with 0.8716.

4.5.3 Subtask 3: Arabic AI-Generated Text Detection

The results for the ARATECT subtask are presented in Table 4. The top participant, **LMSA**, achieved an F1-Score of 0.8641. It is worth noting that some users deleted their accounts after the submission phase, which may indicate that they belonged to the same team as other participants.

5 Discussion

The results from the AraGenEval shared task offer several key insights into the state of Arabic authorship analysis. Across all three subtasks, transformer-based models were the dominant approach, demonstrating their strong capabilities in capturing the nuances of Arabic. In the AST task, the success of prompt-engineered and LoRA-adapted models like AraT5 (Agrahari et al., 2025) and Qwen (Biswas et al., 2025a) highlights a trend towards more explicit and efficient methods for controlling generative style. The top systems showed that framing the task as a natural language instruction allows models to better leverage their pre-trained knowledge.

The Authorship Identification task was highly competitive, with ensemble methods proving particularly effective. The winning system's combination of four different transformer models (Helmy et al., 2025) and the third-place system's logit fusion (Samir et al., 2025) approach underscore the value of model diversity to capture complementary stylistic features. Notably, a traditional approach using TF-IDF features also achieved a top-10 rank, indicating that well-crafted feature engineering remains a viable strategy, especially when computational resources are limited.

Challenges such as handling long documents were addressed by some teams through chunking strategies, showing the importance of data processing in addition to model selection (Helmy et al., 2025).

For AI-Generated Text Detection, the results were more varied. The success of the winning ensemble, which included both Arabic-specific and multilingual models, suggests that a combination of specialized and broad linguistic knowledge is beneficial. The strong performance of systems based solely on multilingual models like XLM-ROBERTa (Zita et al., 2025) was a key finding, indicating their robust generalization capabilities for detecting stylistic artifacts of AI generation, even when not specifically pre-trained on large Arabic corpora.

6 Conclusion and Future Work

The AraGenEval shared task successfully established the first comprehensive benchmark for Arabic authorship style transfer, identification, and AI-generated text detection. The strong participation and the variety of systems submitted underscore the growing interest and need for research in this area. The results confirm the effectiveness of transformer-based architectures across all three subtasks, with specific strategies like prompt engineering, model ensembling, and the use of multilingual models leading to top performances. The task also highlighted the continued relevance of traditional feature-based methods and the importance of robust data handling techniques.

Future work should build on the foundation laid by this shared task. For style transfer, research could explore more advanced controllable generation techniques and develop more nuanced evaluation metrics that go beyond surface-level similarity. For authorship identification, expanding the dataset to include more authors, genres, and dialects would

Rank	Team	Participant	BLEU	chrF	Paper Submitted	System Used
1	ANLPers	omarnj	24.58	59.01	Yes	Prompt Engineering with AraT5
2	Nojoom.AI	nojoom	24.46	59.33	Yes	Fine-tuned mBART and AraT5
3	MarsadLab	rafiulbiswas	20.30	52.56	Yes	LoRA with Qwen2.5-7B-Instruct
4	Osint	shifali	19.87	54.97	Yes	Fine-tuned AraT5
5	PSAU-Wadi	moh55mm5	0.13	26.60	No	-
6	-	syedsaba	0.00	0.27	No	-
7	-	tejasree	0.00	0.18	No	-
8	Neuiry_st	baoflowin502	0.00	0.01	No	-

Table 2: Leaderboard for Subtask 1: Authorship Style Transfer. The ranking is based on the primary metric, BLEU.

Rank	Team	Participant	F1-Score	Accuracy	Paper Submitted	System Used
1	Sebaweh	muhammad-helmy	0.8989	0.9242	Yes	Ensemble of 4 Transformers
2	-	batoolnajeh	0.8716	0.9086	No	-
3	Athership	moamin007	0.8597	0.8952	Yes	Logit Fusion of AraBERT & AraELECTRA
4	MISSION	7h4m3r	0.8375	0.8905	Yes	Fine-tuned ALLaM-7B-Instruct
5	Jenin	jenin	0.8347	0.8738	Yes	Fine-tuned AraBERT
6	ANLPers	omarnj	0.8314	0.8752	Yes	Fine-tuned CAMEL-BERT
7	MarsadLab	rafiulbiswas	0.8282	0.8650	Yes	Fine-tuned AraBERTv2
8	Amr& MohamedSabaa	mohamedsabaa	0.8274	0.8890	Yes	TF-IDF with Logistic Regression
9	CIOL	tasnim_meem	0.8267	0.8641	Yes	Fine-tuned CAMEL-BERT
10	NLP_wizard	nlp_wizard	0.8130	0.8528	Yes	XLM-R Embeddings + LinearSVC
11	Osint	shifali	0.7967	0.8334	Yes	Fine-tuned AraBERTv2
12	Couger AI	sabarinathan1	0.3676	0.6707	No	-
13	-	syedsaba	0.0078	0.0317	No	-

Table 3: Leaderboard for Subtask 2: Authorship Identification. The ranking is based on the primary metric, Macro-F1 Score.

enable the development of more generalizable models. For AI text detection, future tasks should incorporate text generated by newer and more diverse LLMs, as well as adversarial examples, to test the robustness of detection systems. Finally, fostering the development of more high-quality, large-scale Arabic datasets will be crucial for advancing research in all aspects of Arabic NLP.

Limitations

While the AraGenEval shared task provides a valuable contribution, several limitations should be acknowledged. The authorship transfer dataset, though carefully curated, is confined to a specific set of 21 authors and primarily covers the literary domain. This may limit the generalizability of the developed systems to other genres, such as social media or scientific writing. For the AI-generated text detection subtask, the training data was produced by a finite set of LLMs available at the time of dataset creation; detection models may not be robust against newer, more advanced generative models. Furthermore, the evaluation metrics, while standard, have known limitations. BLEU and chrF for style transfer do not fully capture stylistic fidelity or semantic preservation, and F1-score for classification tasks does not account for the subtlety of errors. Finally, the competitive nature of a shared task, with its inherent time and computational constraints, may have prevented teams from exploring more complex or resource-intensive approaches.

References

Fahad J Abdu, Raed Mughaus, Shadi Abudalfa, Moataz Ahmed, and Ahmed Abdelali. 2025. An empirical evaluation of arabic text formality transfer: a comparative study. *Language Resources and Evaluation*, pages 1–61.

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 269–276. IEEE.

Shadi I Abudalfa, Fahad J Abdu, and Maad M Alowaifeer. 2024. Arabic text formality modifica-

Rank	Team	Participant	F1-Score	Accuracy	Paper Submitted	System Used
1	LMSA	kaoutar	0.8641	0.8660	Yes	Ensemble of Fanar, AraBERT, XLM-R
2	-	deleted_user_25186	0.8065	0.7860	No	-
3	MISSION	7h4m3r	0.8044	0.7860	Yes	Fine-tuned AraModernBERT
4	PTUK-HULAT	tasneemduridi	0.7823	0.7640	Yes	Fine-tuned XLM-ROBERTa
5	BUSTED	alizain157	0.7701	0.7600	Yes	Fine-tuned XLM-ROBERTa
6	ANLPers	omarnj	0.7617	0.7860	Yes	Fine-tuned XLM-ROBERTa
7	-	deleted_user_27804	0.7583	0.7680	No	-
8	Osint	shifali	0.7522	0.7180	Yes	mBERT with linguistic features
9	PalNLP	mutazay	0.7443	0.7060	No	-
10	NLP_wizard	nlp_wizard	0.7423	0.7000	Yes	XLM-R Embeddings + RidgeClassifier
11	Jenin	jenin	0.6845	0.5520	Yes	Fine-tuned AraBERT
12	CUET-NLP_ Team_SS306	sowravnath	0.6722	0.5280	Yes	AraBERT with chunking
13	CIOL	tasnim_meem	0.6574	0.7040	Yes	Fine-tuned AraBERTv2
14	Hedi	seifbenayed	0.6541	0.4860	No	-
15	REGLAT	mariamlabib	0.6289	0.6460	Yes	Morphology-aware AraBERT
16	Couger AI	sabarinathan1	0.6238	0.5320	No	-

Table 4: Leaderboard for Subtask 3: Arabic AI-Generated Text Detection (ARATECT). The ranking is based on the primary metric, F1-Score.

tion: A review and future research directions. *IEEE Access*.

Shifali Agrahari, Hemanth Simhadri, Ashutosh Verma, and Ranbir Sanasam. 2025. Osint at arageneval shared task: Fine-tuned modeling for tracking style signatures and ai generation in arabic texts. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Maged S Al-Shaibani and Moataz Ahmed. 2025. The arabic ai fingerprint: Stylometric analysis and detection of large language models text. *arXiv preprint arXiv:2505.23276*.

Mohammad AL-Smadi. 2025. Integrityai at genai detection task 2: Detecting machine-generated academic essays in english and arabic using electra and stylometry. *arXiv preprint arXiv:2501.05476*.

HAMER ALHARBI. 2025. Mission at arageneval shared task: Enhanced arabic authority classification. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

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.

Fatimah Alqahtani and Helen Yannakoudakis. 2022. Authorship verification for arabic short texts using arabic knowledge-base model (arakb). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 205–213.

Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. 2025. Bert-based classical arabic poetry authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119.

Hamed Alshammari and EI-Sayed Ahmed. 2023. Airabic: Arabic dataset for performance evaluation of ai detectors. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 864–870. IEEE.

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

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. 2020a. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, Julinda Stefa, and 1 others. 2019. Cross-domain authorship attribution combining instance-based and profile-based features notebook for pan at clef 2019. In *CEUR WORKSHOP PROCEEDINGS*, volume 2380. CEUR-WS.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

Leonard Bauersfeld, Angel Romero, Manasi Muglikar, and Davide Scaramuzza. 2023. Cracking double-blind review: authorship attribution with deep learning. *Plos one*, 18(6):e0287611.

Imene Bensalem, Imene Boukhalfa, Paolo Rosso, Lahsen Abouenour, Kareem Darwish, and Salim Chikhi. 2015. Overview of the araplagdet pan@ fire2015 shared task on arabic plagiarism detection. In *FIRE workshops*, pages 111–122.

Md. Rafiul Biswas, Mabrouka Bessghaier, Firoj Alam, and Wajdi Zaghouani. 2025a. Marsadlab at arageneval shared task: Llm-based approaches to arabic authorship style transfer and identification. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Md. Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Firoj Alam, and Wajdi Zaghouani. 2025b. MarsadLab at AraGenEval: Arabic Authorship Style Transfer and AI Generated Text Detection. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. *Preprint*, arXiv:2010.05090.

Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keles, Fatema Ahmad, Tasnim Mohiuddin, George Mikros, and Firoj Alam. 2024. Genai content detection task 2: Ai vs. human–academic essay authenticity challenge. *arXiv preprint arXiv:2412.18274*.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. Genai content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388.

Tasneem Duridi, Areej Jaber, and Paloma Martínez. 2025. Ptuk-hulat at arageneval shared task: Fine-tuning xlm-roberta for ai-generated arabic news detection. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. *Culture*, 2:1–359.

Mahmoud El-Haj and Paul Rayson. 2016. Osman—a novel arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association.

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. Dares: Dataset for arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 103–113.

Mo El-Haj and Saad Ezzini. 2024. The multilingual corpus of world's constitutions (mcwc). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 57–66.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. 2025. XDAC: XAI-driven detection and attribution of LLM-generated news comments in Korean. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22728–22750, Vienna, Austria. Association for Computational Linguistics.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Mena Hany. 2025. Nlp_wizard at arageneval shared task: Embedding-based classification for ai detection and authorship attribution. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Muhammad Helmy, Batool Najeh Balah, Ahmed Mohamed Sallam, and Ammar Sherif. 2025. Sebaweh at arageneval shared task: Berense - bert based ensembler for arabic authorship identification. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. 24(1):14–45.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1587–1596. JMLR.org.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

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.

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. Part: Pre-trained authorship representation transformer. *arXiv preprint arXiv:2209.15373*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Hafsa KARA ACHIRA, Mourad Bouache, and Mourad Dahmane. 2025. Nojoom.ai at AraGenEval shared task: Advancing authorship style transfer for arabic text. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Mariam Labib, Nsrin Ashraf, Mohammed Aldawsari, and Hamada Nayel. 2025. Reglat at arageneval shared task: Morphology-aware arabert for detecting arabic ai-generated text. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. 2025. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3091–3113, Vienna, Austria. Association for Computational Linguistics.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Almujaiwel, Ismail Berrada, and Houda Bouamor. 2024. Arafinnlp 2024: The first arabic financial nlp shared task. *arXiv preprint arXiv:2407.09818*.

Huthayfa Malhis, Mohammad Tami, and Huthaifa I. Ashqar. 2025. Jenin at arageneval shared task: Parameter-efficient fine-tuning and layer-wise analysis of arabic llms for authorship style transfer and classification. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Raed Mughaus, Shadi Abudalfa, Hamzah Luqman, Fahad Abdu, Mohammed AlAli, Nawaf Al-Dowayan, and Ahmed Abdelali. 2025. Ma'aks: manually-curated parallel dataset for arabic text sentiment swap. *Language Resources and Evaluation*.

Omer Nacar, Serry Sibaee, Mahmoud Reda, Adel Al-Habashi, Yasser Ammar, and Wadii Boulila. 2025. Anlpers at arageneval shared task: Descriptive author tokens for transparent arabic authorship style transfer. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Sowrav Nath, Shadman Saleh, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2025. Cuet-nlp_team_ss306 at arageneval shared task: A transformer-based framework for detecting ai-generated arabic text. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. Unsupervised text style transfer via llms and attention masking with multi-way interactions. *Preprint*, arXiv:2402.13647.

Chen Qian, Tianchang He, and Rao Zhang. 2017. Deep learning based authorship identification. *Report, Stanford University*, pages 1–9.

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 the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919.

Paolo Rosso. 2017. Author profiling at PAN: from age and gender identification to language variety identification (invited talk). In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, page 46, Valencia, Spain. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*.

Amr Sabaa and Mohamed Sabaa. 2025. Amr&mohamedsabaa at AraGenEval shared task: Arabic authorship identification using term frequency – inverse document frequency features with supervised machine learning. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*, 70:101241.

Eman Samir, Mahmoud Rady, Maria Bassem, Mariam Hossam, Amin Mohamed, Nisreen Hisham, and Sara Gaballa. 2025. Athership at arageneval shared task: Identifying arabic authorship with a dual-model logit fusion. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge, Xing Yang, and Yanghe Feng. 2023. Textdefense: Adversarial text detection based on word importance entropy. *arXiv preprint arXiv:2302.05892*.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22015–22036, Vienna, Austria. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Ali Zain, Sareem Farooqui, and Muhammad Rafi. 2025. Busted at arageneval shared task: A comparative study of transformer-based models for arabic ai-generated text detection. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

Kaoutar Zita, Attia Nehar, Abdelkader Khelil, Slimane Bellaouar, and Hadda Cherroun. 2025. Lmsa at arageneval shared task: Ensemble-based detection of ai-generated arabic text using multilingual and arabic-specific models. In *Proceedings of the third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics (ACL).

Nadhem Zmandar, Mo El-Haj, and Paul Rayson. 2023. FinAraT5: A text to text model for financial Arabic text understanding and generation. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 262–273, Vienna, Austria. NOVA CLUNL, Portugal.