International Journal of Human Post-Edited AI Qualitative Data Analysis (IJHPAQDA)
ISSN: (Online): 3070-3433
DOI prefix: 10.65930
Journal Homepage: www.researchpublication.org

RPO
Research Publication Organization
8 The Green, Dover, DE 19901, USA

| **RESEARCH ARTICLE**

Authorship Attribution of Arabic Criminal Texts Using Large Language Models: A Comparative Evaluation of ChatGPT, DeepSeek, and Gemini

Ibrahim Alharbi

Department of English Language and Literature, College of Languages & Humanities, Qassim University, Buraydah, Saudi Arabia

Corresponding Author: Ibrahim Alharbi, E-mail: IbramimSaud202@gmail.com

| ABSTRACT

This study investigates the ability of three large language models (LLMs), ChatGPT, DeepSeek, and Gemini, to attribute authorship of Arabic criminal texts in a zero-shot setting, with no task-specific training or fine-tuning. Using a quantitative experimental design, each model attributes 24 anonymous criminal texts against reference writings from 12 Arabic authors. The results reveal limited effectiveness, with only ChatGPT achieving a statistically significant accuracy rate of 25%, above the 8.3% chance level. These findings demonstrate that current LLMs in zero-shot settings lack sufficient reliability for definitive authorship attribution (AA) of short Arabic criminal texts, highlighting a gap between their general linguistic capabilities and the specific requirements of forensic textual analysis. While LLMs show preliminary potential, their current implementation cannot replace human expertise in high-stakes forensic contexts involving Arabic texts.

| KEYWORDS

Arabic criminal texts, Authorship attribution, forensic linguistics, LLMs.|

1. Introduction

It is fascinating that language can start and end wars, unite and divide people, and create love or hatred (Everett, 2012). The word *language,* in general, may refer to multiple meanings. As Gee (2017) suggested, it may refer to the English language, a computer language, secret languages, the language of life (DNA), and other forms. However, in linguistics, when we refer to the word *language*, we mean the language that humans acquire as their native tongue. As powerful as it is, language is not just a tool for communication but also a reflection of identity and culture (Gee, 2017). Therefore, it can serve as valuable evidence in criminal contexts, where written or spoken words may play a crucial role in investigations and legal proceedings (Olsson & Luchjenbroers, 2013). According to Olsson (2009), when a text is connected to a legal or criminal case, it is classified as a forensic text.

Every piece of spoken or written language can potentially serve as forensic evidence if it is relevant to a legal or criminal investigation. Such linguistic evidence can be of great value to both law enforcement and legal professionals. The field of forensic linguistics began to gain attention in 1968, when the Swedish linguist Jan Svartvik analyzed the confession statements of Timothy John Evans, who had been executed in 1950 for a crime he did not commit. Svartvik's analysis uncovered inconsistencies in Evans's statements, indicating that parts of the confession had been fabricated. His findings played a key role in proving Evans's innocence (Grant, 2022; Olsson, 2008, 2009).

According to Huang et al. (2025), AA is essential in forensic contexts, as it helps detect suicide cases, track terrorist activities, and support criminal investigations. The importance of this task lies in its ability to help prevent wrong convictions. What if AA could be done accurately by artificial intelligence (AI)? With the huge development of technology, especially in natural language processing and machine learning, AI is becoming increasingly capable of analyzing writing styles and identifying potential authors (Atkinson-Abutridy, 2024; Thakur et al., 2024). This development opens up new possibilities for us to utilize AI in forensic linguistics, particularly in cases involving anonymous or criminal texts. According to Sousa-Silva (2024), technology has long been used for criminal purposes; nonetheless, the same technology has also been employed for positive purposes, including combating criminal activity. In the realm of AA, the effectiveness of LLMs has been demonstrated in several studies, including those by Hu et al. (2024) and Misini et al. (2024). LLMs are advanced AI systems designed for processing, understanding, and producing human language through the use of designed neural networks (Raschka, 2024).

These days, AI technologies have become an integral part of investigations. These technologies play a significant role in forensic science across various areas, including DNA sequencing and forensic document analysis (Saini et al., 2024). The reason behind the use of AI is its ability to provide large amounts of information quickly and accurately (Makei & Tokura, 2025). Therefore, this study investigated the accuracy of AI, specifically three different LLMs, ChatGPT, DeepSeek, and Gemini, that have a remarkable ability to understand, produce, and interpret human language.

2. Literature Review

A. 2.1. Forensic Linguistics

According to Coulthard et al. (2020), the field of forensic linguistics is divided into three areas: the study of legal language, the study of language in legal processes, and the role of the forensic linguist as an expert witness. The first two areas involve the analysis of the language of legal texts and the study of how people communicate throughout this process. This includes everything from the initial reports to final judgments, and everything in between. However, since the main concern of this study relates to language used in criminal contexts, our focus is limited to the third area of forensic linguistics: the role of the linguist as an expert witness. In this area, language analysis is applied to real legal cases to reveal hidden meanings, patterns, or authorship clues that may not be apparent to non-specialists. In such scenarios, this analysis can play a crucial role in solving critical cases, as it provides investigators with insights that may otherwise remain undiscovered.

It can be challenging to identify the exact origin of any practical discipline, as the practice likely took place long before it was observed, named, and formally discussed. However, when we refer to the English language, it is clear that the Swedish linguist Jan Svartvik played a key role, particularly with his 1968 study on the language of The Evans Statement, which introduced the term *forensic linguistics* in its subtitle. The case involved the murder of several people in one of London's infamous neighborhoods, which was later renamed following requests from the residents at the time. Evans and his family lived in that neighborhood, on the upper floor of a quiet man named John Christie, who was sentenced to death in July 1953 for several crimes he had committed, including the murder of Evans's wife (Grant, 2022; Olsson, 2009).

In 1949, Evans disappeared, and people began to worry about his wife and daughter. In the last quarter of that year, Evans turned himself in to the South Wales police. He was found guilty of the murder of his wife based on his statements to the police and the evidence given by John Christie and was hanged a year later. Later, Christie moved out of the neighborhood, and another tenant moved into his flat. While carrying out maintenance, the new tenant discovered that there were three dead bodies hidden in his place. The police then found evidence of several murders and tracked down Christie, who confessed to murdering Evans's wife and was executed in 1953. However, even after Christie's confession, the crime continued to be attributed to Evans for several years (Grant, 2022; Olsson, 2009).

A decade later, journalist Ludovic Kennedy began investigating Evans's case, which eventually caught the attention of Swedish professor Jan Svartvik. When the professor began examining Evans's statements, he identified several distinct language styles, with the policeman's register being the dominant one. Kennedy's interest in the case, along with Svartvik's linguistic analysis, led the Home Secretary to overturn the conviction, and Evans was eventually exonerated of the murder of his wife. As Olsson noted, this case is often regarded as the earliest instance involving linguistic analysis, and as a result, Svartvik is widely recognized as the founding figure of the discipline (Olsson, 2009).

One of the most important figures in the history of forensic linguistics is the British linguist Malcolm Coulthard. In the late 1980s, he became involved in several cases that took place in Birmingham, UK. These cases questioned the authenticity of several confession statements suspected of being fake. A special unit within the police force, the West Midlands Serious Crime Squad, was accused of corruption and fabricating evidence. One of these cases involved a man named Paul Dandy. The police claimed that Dandy had confessed to a crime that he was accused of; however, it turned out that the confession was fabricated. This was discovered through a special test called Electro-Static Detection Analysis (ESDA), which can detect marks left on paper from writing, even if the ink is not visible. The test showed that someone had added a single sentence to Paul's statement to make it look like he had confessed to the crime. Coulthard also contributed to several other cases, including the 1975 incident in which 21 individuals were killed, as well as a case involving the conviction of an individual for the murder of a child who had interrupted a burglary (Grant, 2022).

In 1994, John Olsson established the Forensic Linguistics Institute in the United Kingdom, which became a leading center for analyzing texts in legal contexts, with a focus on authorship, authenticity, meaning, and disputed language. Since then, Olsson has been involved in nearly 300 forensic investigations, ranging from analyzing alleged terrorist statements and suicide notes to assessing threats in extortion letters and evaluating police interview recordings. During this time, the discipline has shifted from a largely theoretical and marginal area of study to a recognized and applied science used in real-world legal cases. As a result, he is considered one of the key figures in the field of forensic linguistics (Olsson, 2009)

This article introduces Professional Learning Communities (PLCs) as a conceptual and practical framework aligned with contemporary understandings of effective teacher learning. PLCs provide the structures and cultures necessary for teachers to engage in sustained collaboration, reflective inquiry, and collective problem-solving.

B. 2.2. Authorship Attribution

In the previous section, we examined some cases that involved AA; however, long before that, the process of determining the likely author of a given text based on linguistic or stylistic features was already in use. Stylometry is one of the quantitative approaches that researchers have used to identify the author of a particular text. It refers to the analysis of writing style, often using linguistic features such as word choice, syntax, and orthography (Plechác, 2022). Holmes (1998) investigated the roots of stylometry, tracing them back to Augustus De Morgan's work in 1851. Augustus De Morgan was a British mathematician who attempted to determine which of the Pauline epistles were genuinely authored by St. Paul. Therefore, he analyzed the average word length as a means of distinguishing between different authors' writing styles.

The popularity of stylometry rose in the late 19th century, when Thomas Corwin Mendenhall, an American physicist, conducted a study using word length distribution curves instead of just averages. In his 1887 article, "The Characteristic Curve of Composition," he proposed the idea of analyzing the distribution of word lengths, rather than just their averages, to identify the actual author of a text. With financial support from a benefactor, August Hemenway, Mendenhall later applied this method to a real-world case involving the disputed authorship of works attributed to William Shakespeare. The results were published in Mendenhall's 1901 article *A Mechanical Solution to* *a Literary Problem,* where he compared the word length patterns in texts by Shakespeare, Francis Bacon, and Christopher Marlowe. He carefully concluded that Bacon probably did not write the disputed texts, but there were strong similarities between Shakespeare’s and Marlowe’s writing styles (Coulthard et al., 2016; Plechác, 2022).

C. 2.3. Criminal Texts and Authorship Attribution

One of the examples of how forensic linguistics can help in murder investigations is the story of Julie Turner. Julie went missing in June 2005 after leaving to meet a man named Howard Simmerson, with whom she had an affair for more than three years. After her disappearance, Julie’s husband received several text messages from an unknown number, claiming to be from Julie. Several unusual phrases, such as “sort my head out” and “sort my life out,” were mentioned in these messages, which raised doubts, as Darren knew Julie would never leave her two children without telling them. The case was handed to a forensic linguist named John Olsson to analyze. He found several linguistic markers that were unusual, including the use of full stops instead of commas and the rare co-occurrence and sequence of the phrases “head sorted out” and “sorted her life out”. These expressions were later mentioned in an interview with the man she had an affair with, suggesting he was the actual author of the messages that were sent to the husband. Further evidence, including a letter indicating suicidal and violent intentions and CCTV footage showing a drum on Simmerson’s vehicle, led to the discovery of Julie’s body inside the drum. Despite Simmerson’s claim that Julie shot herself, the jury rejected his defense. He was sentenced to life imprisonment (Olsson, 2009).

Another case that Olsson analyzed is the disappearance of Jenny Nicholl. After Jenny went missing, her parents received several text messages from her phone over the course of several days. Initially, these texts led both the police and her parents to believe that she was still alive. However, suspicions arose that someone else might have been sending the messages on her behalf. As a result, Olsson was asked to examine the texts. He discovered a significant shift in texting style after 26 June 2005. While he acknowledged that some variation could be attributed to emotional state or life changes, the overall shift was too marked to be explained by these factors alone. Sadly, Jenny was never seen again after that date. Two years later, David Hodgson was found guilty of her murder, although her body was never recovered (Olsson, 2009).

D. 2.4. Artificial Intelligence and Authorship Attribution

In the era of LLMs, Huang and his partners (2025) explored how AA has developed in response to the rise of these advanced models. They argue that, over the past few decades, there has been a noticeable shift in the field of AA from relying primarily on traditional stylometric and statistical approaches to embracing the computational capabilities of LLMs. This shift has brought further advancements to the task since LLMs are capable of automatically analyzing and extracting key linguistic features from texts, which can help identify the actual author.

Schmidt, Gorovaia, and Yamshchikov (2024) tested the performance of four LLMs, namely, GPT-4o, Claude, Gemini, and Mistral, on AA of Latin texts from the Patristic Era. The models were tested without additional training or fine-tuning on the Latin dataset. The results indicated that LLMs can successfully perform authorship tasks even for a historical, low-resource language like Latin. Regarding the comparison between the four models, GPT-4o demonstrated the best overall performance.

In a similar context to what was proposed by Schmidt and his colleagues, Huang (2024) and two other researchers conducted a comparable experiment using different LLMs and different types of texts. While Schmidt focused on Latin texts from the Patristic Era, Huang utilized modern English texts, specifically blogs and emails. The LLMs employed in Huang's study included GPT-3.5 Turbo, GPT-4 Turbo, BERT, RoBERTa, ELECTRA, and TF-IDF. Even though the two studies differed in language and dataset, both confirmed that LLMs can effectively perform authorship tasks without the need for fine-tuning (Huang et al., 2025).

In an Albanian context, Misini et al. (2024) conducted a study to investigate the ability of various machine learning technologies to identify the true author of a text by applying a wide range of linguistic features. Their datasets consisted of a large number of texts drawn from both newsroom columns and literary works. They used various features and tested several machine learning and deep learning models to determine the most effective combination for authorship attribution. The findings showed that lexical features were the most useful, and the XGBoost machine learning algorithm achieved the highest overall performance.

According to Altheneyan and Menai (2014), the Arabic language has received little attention in the field of AA compared to other languages such as English, Chinese, and Dutch. Therefore, they conducted a study in 2014 to investigate the effectiveness of four different machine learning models in AA. They collected their data from several books written by ten different well-known Arabic authors and split each book into smaller chunks. A total of 408 writing-style features were examined, including word length, punctuation, and function words. The comparative analysis revealed significant performance differences among the four Naive Bayes models. The Multivariate Bernoulli model demonstrated the highest effectiveness for Arabic AA, achieving an accuracy of 97.43 percent.

In a similar context, AlZahrani and Al-Yahya (2023) investigated the application of modern Arabic pretrained transformer-based models - namely AraBERT, AraELECTRA, ARBERT, and MARBERT- for AA of Islamic legal texts. The researchers designed and built their own dataset using Islamic law digital resources from Al-Maktaba Al-Shamela. They divided their data into four datasets, each containing 8, 16, 32, and 40 authors, to assess how the models performed with an increasing number of authors. After a series of detailed and careful experiments involving hyperparameter optimization, their results demonstrated that AraELECTRA achieved 97% accuracy on the most challenging 40-author dataset after optimization.

Wenger (1998) defines Professional Learning Communities (PLCs) as groups sharing a concern or passion, deepening their knowledge through ongoing interaction. PLCs operationalize this idea in schools by fostering shared leadership, collective inquiry, and mutual accountability. Research supports the value of PLCs in enhancing teacher learning (Avalos, 2011; Vangrieken et al., 2017). They embody Villegas-Reimers' (2003) principles of continuous, collaborative, and context-responsive professional development. Additionally, PLCs promote distributed leadership and a strategic orientation toward educational transformation (Senge, 1990; Hord & Sommers, 2008).

3. Methodology

The primary focus of this study is to investigate the ability of ChatGPT, DeepSeek, and Gemini to attribute authorship of Arabic criminal texts accurately.

The design of this study was within-subjects, as this approach is appropriate when the same subjects are exposed to all levels of a categorical variable (Seltman, 2018). In the present context, the texts served as the subjects: each of the 24 anonymous criminal texts was analyzed by all three models under identical conditions.

This study involved twelve male native speakers of Arabic. All participants provided formal, written informed consent. They completed a writing assignment consisted of three tasks. First, they hand-wrote a threatening message, followed by a blackmail letter, both designed to simulate criminal writings. Finally, each participant provided a personal reference text of at least 150 words, originally written within the past year. They were asked to complete the writing tasks based on the following prompts:

3.1. *Threatening Message Scenario*

Imagine that someone in your workplace has begun spreading false rumors about you among colleagues, and these rumors have started to harm your reputation, possibly affecting your work or relationships. Instead of confronting this person face-to-face, you decide to write them an indirect threatening message, asking them to stop immediately.

3.2. Blackmailing Letter Scenario

Imagine that you found someone’s forgotten bag in a restaurant. Upon opening it, you discovered a laptop containing private messages and photos. Instead of handing the bag over to the police, you decide to write a message to the owner demanding a sum of money in exchange for returning the bag and threatening not to return it if they refuse to pay by a specific time and place.

4. Results and Discussion

E. *4.1. Accuracy of LLMs Compared to Chance*

The descriptive results indicate modest and variable performance across the models, as summarized in 1.

Table 1

Descriptive performance of the models

Model	Correct / N	Accuracy	95% CI
ChatGPT	6 / 24	25.0%	12.0%-45.0%
DeepSeek	4 / 24	16.7%	6.7%-35.9%
Gemini	3 / 24	12.5%	4.3%-31.0%

Note. Each model produced 24 attributions; accuracy refers to the proportion of correct answers. Confidence intervals are 95% (Wilson).

F. ***4.2. Binomial Tests vs. Chance***

One-sample binomial tests against a chance level of 8.33% (1 in 12) were conducted to determine if the observed accuracies represented meaningful performance. The results of these tests are presented in Table 2. Each model's authorship predictions were compared to the actual author of the 24 criminal texts. A one-sample binomial test was conducted to determine whether each model performed significantly better than chance (8.3%, corresponding to one correct identification out of twelve possible authors).

Table 2

Binomial tests of accuracy against chance (8.33%)

Model	Correct / N	Accuracy	95% CI	Exact p (two-sided)
ChatGPT	6 / 24	25.0%	12.0%-45.0%	.012
DeepSeek	4 / 24	16.7%	6.7%-35.9%	.135
Gemini	3 / 24	12.5%	4.3%-31.0%	.447

Note. CIs are Wilson score 95% intervals for a proportion. Exact p-values test the null p=0.0833 (1/12).

Submit a text that you have written within the past 12 months. The text must contain a minimum of 150 words and may be any form of original writing, such as an email, WhatsApp message, personal post, or any other piece of writing.

Only ChatGPT's performance reached conventional statistical significance (*p* = .012). DeepSeek (*p* = .135) and Gemini (*p* = .447) did not demonstrate statistically reliable evidence of performing above chance.

4.3. ***Interpretation***

While this statistical conclusion is technically correct, it is necessary to provide context for these results. The sample size of 24 texts is a significant limitation, resulting in the wide confidence intervals visible in Tables 1 and 2 (e.g., ChatGPT's accuracy could plausibly be between 12% and 45%). This high degree of uncertainty means that:

1. ChatGPT's superiority is not yet strongly established. The lower bound of its confidence interval (12%) is only marginally above chance, and a replication with a different small sample could easily yield a non-significant result.
2. DeepSeek and Gemini's performance cannot be conclusively dismissed. The non-significant *p*-values are likely a function of low statistical power. DeepSeek's point estimate of 16.7% is double the chance rate, a difference that could become significant with a larger sample size.

While ChatGPT's 25% accuracy may appear low in an absolute sense, it is important to contextualize this figure within a practical investigative framework. In most real-world criminal investigations, AA is often used as one of the tools to prioritize leads rather than to deliver a definitive verdict. In a scenario with a dozen potential suspects, random chance would yield an 8.3% success rate. An accuracy of 25% could provide valuable guidance to human investigators.

5. Conclusion

This study investigated the potential of three LLMs as tools for authorship attribution (AA) in the specific context of Arabic criminal texts. The main conclusion is that, in this current state and under the specific conditions of this study, LLMs are not reliably effective tools for the zero-shot attribution of authorship in short Arabic criminal texts. The models' ability to accurately attribute authorship was limited. Only ChatGPT demonstrated performance statistically significant above the chance level. The other two models showed no statistically reliable evidence of performing better than random guessing. While ChatGPT's result offers a glimmer of potential, its accuracy rate and the wide confidence interval surrounding it suggest this capability is preliminary and not yet strong. This provides a critical counterpoint to the prevailing optimism surrounding LLMs, demonstrating that their proficiency in general language tasks does not guarantee reliability in high-stakes tasks, such as AA of criminal texts. The models' inconsistent performance in this setting highlights the necessity for extreme caution.

This study examined the viability of three large language models (LLMs) as instruments for authorship attribution (AA) within the highly constrained and sensitive domain of Arabic criminal texts. The findings provide compelling evidence that, under the conditions investigated, namely, zero-shot inference applied to short, domain-specific texts, current LLM architectures do not constitute reliable or robust solutions for authorship attribution. Overall attribution accuracy across the models was modest and inconsistent, underscoring substantive limitations in their capacity to extract and operationalize stable authorial signals in this linguistic and forensic context.

Notably, only ChatGPT achieved performance levels that were statistically distinguishable from chance. However, even this comparatively stronger performance must be interpreted with considerable restraint. The observed accuracy rate, coupled with a broad confidence interval, indicates substantial variability and uncertainty in the model's predictions. Such statistical dispersion suggests that ChatGPT's apparent advantage may reflect sporadic alignment with surface-level stylistic cues rather than a systematic or theoretically grounded sensitivity to deeper, author-specific linguistic patterns. Consequently, its demonstrated capability remains emergent rather than conclusive.

In contrast, the remaining two models failed to exhibit any statistically reliable evidence of outperforming random guessing. This outcome reinforces the notion that general linguistic competence, often demonstrated by LLMs across a wide range of natural language processing tasks, does not automatically translate into effectiveness in specialized applications that demand fine-grained stylistic discrimination. Authorship attribution, particularly in short texts, is fundamentally dependent on subtle, low-frequency markers of idiolect, which appear to be inadequately captured by these models when operating without task-specific adaptation.

5.2. Implications for Policy and Practice

From a theoretical perspective, the results challenge assumptions that large-scale pretraining inherently equips LLMs with robust representations of authorial style. The failure of most models to surpass chance levels suggests that authorship signals, particularly in morphologically rich and syntactically complex languages such as Arabic, may not be sufficiently encoded in the latent spaces optimized for next-token prediction. This raises important questions about the extent to which current LLM architectures model stylistic invariance versus topical or lexical salience, and calls for a re-examination of how authorial identity is conceptualized and operationalized within neural language models.

Methodologically, the findings highlight the limitations of zero-shot paradigms for authorship attribution, especially in short-text scenarios. They suggest that meaningful improvements in performance are unlikely without targeted fine-tuning, explicit stylistic feature modeling, or hybrid approaches that integrate traditional stylometric techniques with neural representations. Furthermore, the wide confidence intervals observed, even in the best-performing model, underscore the importance of robust statistical validation and uncertainty estimation in evaluating AI-based AA systems.

From a practical and forensic standpoint, the study carries significant cautionary implications. The use of LLMs for authorship attribution in criminal investigations, judicial proceedings, or intelligence contexts would be premature and potentially consequential if adopted uncritically. Given the risk of false attribution and the severe legal and ethical ramifications associated with such errors, current LLMs should not be relied upon as standalone tools for forensic authorship analysis. At most, they may serve as exploratory or supplementary instruments, subject to strict oversight and corroboration by established forensic methodologies.

5.3. Future Research Directions

The findings of the present study open several promising and necessary avenues for future research aimed at advancing the reliability and forensic validity of authorship attribution using large language models, particularly in the context of Arabic criminal texts.

First, future work should move beyond zero-shot paradigms and systematically investigate the effects of task-specific fine-tuning on authorship attribution performance. Training LLMs on curated, author-labeled Arabic corpora, especially those reflecting forensic genres such as threats, confessions, online communications, and ransom notes, may enhance their sensitivity to stable authorial markers. Such efforts should prioritize domain-relevant data and carefully control for topic leakage to ensure that improvements reflect genuine stylistic learning rather than semantic memorization.

Second, there is a clear need to explore hybrid methodological frameworks that integrate LLM-based representations with established stylometric techniques. Combining neural embeddings with interpretable linguistic features, such as function word distributions, morphological patterns, syntactic constructions, and orthographic variation, may yield more robust and explainable attribution models. This approach could bridge the gap between the predictive power of deep learning and the transparency required in forensic settings.

Third, future research should address the linguistic complexity and variability of Arabic more explicitly. Dialectal variation, code-switching, orthographic inconsistency, and morphological richness present unique challenges for authorship attribution that are insufficiently captured by models primarily trained on Modern Standard Arabic or mixed-domain data. Developing dialect-aware models and evaluating performance across varieties of Arabic would substantially enhance ecological validity.

Fourth, the impact of text length and genre warrants systematic investigation. Since criminal texts are often short and contextually constrained, future studies should assess model performance across varying text

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

ORCID ID: 0009-0006-2990-6318

References

Alsajri, A., Salman, H. A., & Steiti, A. (2024). Generative models in natural language processing: A comparative study of ChatGPT and Gemini. *Babylonian Journal of Artificial Intelligence,* 134–145.

Altheneyan, A., & Menai, M. (2014). *Naïve Bayes classifiers for authorship attribution of Arabic texts*. *Journal of King Saud University – Computer and Information Sciences, 26*(4), 473–484. https://doi.org/10.1016/j.jksuci.2014.06.006

AlZahrani, F. M., & Al-Yahya, M. (2023). *A transformer-based approach to authorship attribution in classical Arabic texts*. *Applied Sciences, 13*(12), 1–15. https://doi.org/10.3390/app13127255

Atkinson-Abutridy, J. (2024). *Large language models* (1st ed.). CRC Press.

Bissell, A. F. (1995). *Weighted cumulative sums for text analysis using word counts*. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 158*(3), 525–545. https://doi.org/10.2307/2983444

Canbay, P., Sezer, E. A., & Sever, H. (2020). *Deep combination of stylometry features in forensic authorship analysis. International Journal of Information Security Science, 9*(3), 154–163.

Coulthard, M., Johnson, A., & Wright, D. (2016). *An introduction to forensic linguistics* (2nd ed.). Routledge.

Coulthard, M., Johnson, A., & Wright, D. (2020). *The Routledge handbook of forensic linguistics* (2nd ed.). Routledge.

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. In J. F. Martínez-Trinidad, J. A. Carrasco Ochoa, & J. Kittler (Eds.), *Progress in pattern recognition, image analysis and applications* (pp. 844–853). Springer. https://doi.org/10.1007/11892755_87

Everett, D. L. (2012). *Language*. Profile Books.

Gee, J. P. (2017). *Introducing discourse analysis* (1st ed.). Routledge.

Grant, T. (2022). *The idea of progress in forensic authorship analysis*. Cambridge University Press.

Hardcastle, R. A. (1993). *Forensic linguistics: An assessment of the CUSUM method for the determination of authorship*. *Journal of the Forensic Science Society, 33*(2), 95–106.

Holmes, D. I. (1998). *The evolution of stylometry in humanities scholarship*. *Literary and Linguistic Computing, 13*(3), 111–117. https://doi.org/10.1093/llc/13.3.111

Holmes, D. I., & Forsyth, R. S. (1995). *The Federalist revisited: New directions in authorship attribution*. *Literary and Linguistic Computing, 10*(2), 111–127.

Seltman, H. J. (2018). *Experimental design and analysis*. Carnegie Mellon University.

Hu, Z., Zheng, T., & Huang, H. (2024). A Bayesian approach to harnessing the power of LLMs in authorship attribution. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 13216–13227). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.733

Huang, B., Chen, C., & Shu, K. (2024). Can large language models identify authorship? In Y. Al-Onaizan, M. Bansal, & Y. Chen (Eds.), *Findings of the Association for Computational Linguistics*: *EMNLP 2024* (pp. 445–460). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.26

Huang, B., Chen, C., & Shu, K. (2025). *Authorship attribution in the era of LLMs: Problems, methodologies, and challenges*. *ACM SIGKDD Explorations Newsletter, 26*(2), 21-43.

Makei, J., & Tokura, T. (2025). *Teaching "what" vs. teaching "why": How ChatGPT and generative AI are shaping education.* ResearchGate. https://doi.org/10.13140/RG.2.2.13559.53924

Misini, A., Canhasi, E., Kadriu, A., & Fetahi, E. (2024). *Automatic authorship attribution in Albanian texts. PLOS ONE, 19*(10), e0310057. https://doi.org/10.1371/journal.pone.0310057

Mosteller, F., & Wallace, D. L. (1963). *Inference in an authorship problem*. *Journal of the American Statistical Association, 58*(302), 275–309. https://doi.org/10.1080/01621459.1963.10500849

Olsson, J. (2008). *Forensic linguistics* (2nd ed.). Continuum.

Olsson, J. (2009). *Wordcrime* (1st ed.). Continuum.

Olsson, J., & Luchjenbroers, J. (2013). *Forensic linguistics* (1st ed.). Bloomsbury Academic.

Plechác, P. (2022) *Versification and Authorship Attribution*. Karolinum Press, Charles University.

Raschka, S. (2024). *Build a large language model (from scratch)*. Manning.

Rahman, M., Shiplu, A., Watanobe, Y., Tapader, M., Amin, M., & Peng, L. (2025). *ChatGPT and DeepSeek: Strengths, limitations, and the future of generative AI*. Journal of LATEX Class Files, 18(9), 1-19.

Saini, K., Gupta, A., Rani, S., Sethi, R., & Awasthi, P. (2024). *Artificial intelligence in forensic science* (1st ed.). CRC Press.

Sousa-Silva, R. (2024). Fighting cyber-malice: A forensic linguistics approach to detecting AI-generated malicious texts. In *Proceedings of the 1st International Conference on NLP & AI for Cyber Security* (164–174).

Gorovaia, S., Schmidt, G., & Yamshchikov, I. P. (2024). Sui generis*: Large language models for authorship attribution and verification in Latin*. In M. Hämäläinen, E. Öhman, S. Miyagawa, K. Alnajjar, & Y. Bizzoni (Eds.), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (pp. 398–412). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4dh-1.39

Thakur, K., Barker, H. & Pathan, A.-S. K. (2024). *Artificial intelligence and large language models* (1st ed.). Chapman and Hall/CRC.