A Comparative Pattern Analysis of Qwen 2.5 and Gemma 3 Text Generation

KALIN KOPANOV, TATIANA ATANASOVA
Institute of Information and Communication Technologies,
Bulgarian Academy of Science,
Acad. G. Bonchev Str., Block 2, 1113 Sofia,
BULGARIA

Abstract: - This study examines whether two instruction-tuned language models, Qwen 2.5 (32B) and Gemma 3 (27B), exhibit distinct linguistic patterns for accurate, automated text attribution. We created a dataset of 6,000 LLM-generated text outputs (3,000 per model) to 300 prompts across ten categories for diverse contextual analysis. Afterward, we trained four classifiers: Logistic Regression; Support Vector Machine (SVM); Random Forest; and Gradient Boosting, using Term Frequency-Inverse Document Frequency (TF-IDF) features and various syntactic and stylistic cues. Findings show that TF-IDF features alone are effective for text attribution and that SVM is the most accurate attribution tool, achieving a 99% success rate. At the same time, the ensemble methods of Random Forest and Gradient Boosting were improved by the addition of syntactic and stylistic markers. However, lexical frequency patterns remained the primary predictor, which indicates that simple methods can effectively categorize text. Further analysis also revealed that Qwen 2.5 typically produces structured, formal outputs, while Gemma 3 favors a more expressive, narrative style. Our final results show that all classifiers can effectively identify AI-generated text, which may have future implications for academic integrity, content moderation, and automated plagiarism detection. Considering the constant evolution of Large Language Models (LLMs), better benchmarking methods and additional features are required to precisely attribute AI-generated text across different scenarios.

Key-Words: - Qwen 2.5, Gemma 3, Large Language Models (LLMs), Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, AI-generated content, pattern analysis.

Received: August 17, 2024. Revised: May 15, 2025. Accepted: June 13, 2025. Published: July 29, 2025.

1 Introduction

Large Language Models (LLMs) have rapidly developed over the past few years, often making it difficult to distinguish between the outputs of different systems. Furthermore, open-source LLMs can now produce fluent and contextually appropriate text that resembles human writing, which raises fundamental questions about their internal generative mechanisms. This research aims to examine whether these models produce unique writing styles or linguistic signatures while delivering comparable performance results. To successfully discriminate between LLM-generated outputs, machine-learning text attribution systems need to understand these traits.

Recent research has found that different LLMs, such as GPT-3.5, GPT-4, and Bard, display significant linguistic variations that enable high-accuracy attribution to the correct model, [1]. However, prior work has primarily focused on a few popular models or distinguishing AI-generated text from human text, leaving a gap in systematic comparisons between alternative LLMs. It remains unclear whether two state-of-the-art models from different development lineages, such as Qwen 2.5 (developed by Alibaba's AI research division) and Gemma 3 (designed by Google's DeepMind team), consistently exhibit distinct styles across diverse content domains, [2], [3]. Examination of this question will advance our understanding of LLMs' behavior and will improve the detection of AI-generated content.

Research studies currently use limited detection tools that depend on public data and pre-trained models, raising concerns about the accuracy and reproducibility of the research results. Furthermore, the translation of AI-generated content and manual editing or paraphrasing frequently reduces detection accuracy to approximately 50%. Even minor prompt adjustments can lower the performance of state-of-the-art detectors by more than 50%, [4], [5]. Research [6], has confirmed this trend and found that the detection of LLM-generated text depends heavily on the choice of model and tool and that

paraphrasing can decrease detection accuracy by as much as 40%. The study [7], has also found that Turnitin and other advanced detection systems identify AI-generated submissions at a rate of only 55%, which challenges the effectiveness of current detection methods. Such research shows a need to study specific models in controlled assessments.

This study is a comparative pattern analysis of the text outputs produced by Qwen 2.5 and Gemma 3 across a broad range of categories. We analyzed texts of sufficient length to capture each model's stylistic tendencies that allow for attribution while remaining concise enough to permit controlled comparisons. Our method blends a classifier-based attribution test with direct linguistic inspection, which measures lexical diversity, sentence structure complexity, and domain-specific language to confirm and explain the observed patterns. Whereas most detection research contrasts AI-generated text with human writing, we test whether automated classifiers can distinguish between two state-of-the-art LLMs. We show that Qwen 2.5 and Gemma 3 leave separable linguistic signatures that allow for the identification of model-specific training and design choices, showing that AI-generated text is multi-model and more complex than initially understood.

2 Related Works

The development of LLMs has transformed text generation and can now produce high-quality content for a wide range of topics. These advanced models have created additional avenues for academic research, but they also pose problems regarding unethical use and plagiarism. The essential role of these LLMs in scholarly communication requires the development of robust methods to distinguish AI-generated content from human-generated content. At the same time, open-source developments have driven technological progress while intensifying debates about academic integrity, author authenticity, and proper AI text labeling practices. Studies [6], [8], identify a need for a theoretical framework that transcends basic deterministic or instrumental views of technology and advocates instead for a perspective that acknowledges the dynamic and collaborative interplay between generative AI, individuals, and the scientific system.

Ongoing research also highlights significant shortcomings in existing AI-generated text detection methods. For instance, the study [9], introduces AI-Catcher. This multimodal deep learning framework combines a multilayer perceptron for extracting linguistic and statistical features with a convolutional neural network (CNN) that captures sequential patterns in text. This integrated approach demonstrates that using enriched and diverse feature representations can improve detection accuracy. However, the development and evaluation of AI-Catcher were based solely on OpenAI's ChatGPT-generated scientific text, which may limit its effectiveness when applied to content produced by other language models or in non-scientific domains.

A notable concern is the interaction between real-world assessment practices and automated detection systems. The study [7], found that many instructors either overlooked or misinterpreted these flags even when detection software flagged AI-generated passages in student essays with moderate accuracy. This oversight resulted in significant underreporting of potential academic misconduct. Their findings underscore a need for comprehensive training for academic staff that allows them to critically evaluate detection outcomes rather than relying solely on software-generated metrics.

In addition, empirical studies indicate that publicly available detectors frequently misclassify text, erroneously labeling AI-generated content as human-authored. The research [4], evaluated several free and commercial detection systems using short text samples and observed that most tools are consistently inclined to identify text as human-authored, particularly when the AI-generated content was paraphrased or underwent human editing. This systematic bias increases the incidence of false negatives, which impedes the reliable detection of AI-assisted plagiarism and compromises academic integrity standards.

While many detection efforts have focused on identifying texts generated by OpenAI's GPT variants, few studies have systematically examined newer models such as Qwen 2.5 (32B-instruct) and Gemma 3 (27B-it; introduced on 12 March 2025). The study [6], examines general performance comparisons of Qwen 2.5 Max and Gemma but does not systematically address how current detectors handle these newer models. Some recent work has demonstrated the value of comparative analysis between different AI systems. Research [10], analyzed multiple NLP models and LLMs for multilingual geo-entity detection to show how different models and languages affect performance, which could help develop attribution methods. Due to the different developmental paths of Qwen 2.5 and Gemma 3, both models generate distinctive text. As such, detection tools need customized training data to achieve reliable performance. The literature lacks specific research on style analysis or detection

benchmarking for these models. The detection and attribution of AI-generated text become increasingly complex because domain-specific and open-source models continue to mature and expand.

3 Problem Formulation

Advancements in LLMs have resulted in a wide range of synthetic text generation techniques, yet gaps remain in the literature regarding systematic, head-to-head comparisons of different models' outputs. We selected Qwen 2.5 (32B) and Gemma 3 (27B) because mid-range LLMs, which have approximately 30B parameters, have reliable performance and can generate rich, nuanced content.

This study aims to fill research gaps by detailing how synthetic text is generated, particularly in LLMs that produce diverse and context-specific outputs. Our research investigates consistency, lexical patterns, and stylistic indicators inherent in their generated texts by examining the instruction-tuned versions of the Qwen 2.5 32B and Gemma 3 27B models.

In doing so, we address the following key research questions:

RQ1: How can different Machine Learning classifiers distinguish between text generated by Qwen 2.5 and Gemma 3?

RQ2: What roles do lexical, syntactic, and stylistic features play in identifying the source model of a given text?

RQ3: To what extent do additional syntactic and stylistic cues enhance classification beyond TF-IDF alone?

By answering these questions, we aim to identify distinctive characteristics in the text generated by those two LLMs and establish effective methods for model attribution, thus contributing to greater transparency in AI-generated content.

4 Problem Solution

Our approach pursues two lines of effort: a classifier that assigns texts to the correct model and a comparison of lexical diversity, rhetorical structures, and domain-specific expressions. This strategy allows us to better attribute AI content and offers practical insights into academic integrity policies.

Recent research reveals that advanced LLMs produce unique "fingerprints" in text output that are influenced by differences in model architecture, training data, and fine-tuning objectives, [1]. Detailed examinations of vocabulary, syntax, and specific phrases show that these linguistic characteristics can effectively act as indicators for attributing models. We hypothesize that Qwen 2.5 and Gemma 3 will exhibit unique stylistic characteristics as well. As AI is increasingly responsible for a growing amount of content, intellectual property, and transparency issues become more prominent when AI acts as a co-author or principal author, [11]. The ability to identify whether Qwen 2.5 or Gemma 3 generated a text will ultimately increase trust in scholarly communication. The following subsections describe the experimental workflow, including data preparation, model training, classification accuracy, and confidence analysis.

4.1 Synthetic Data Generation Workflow and Experimental Setup

Our four-stage experimental process is summarized in Figure 1:

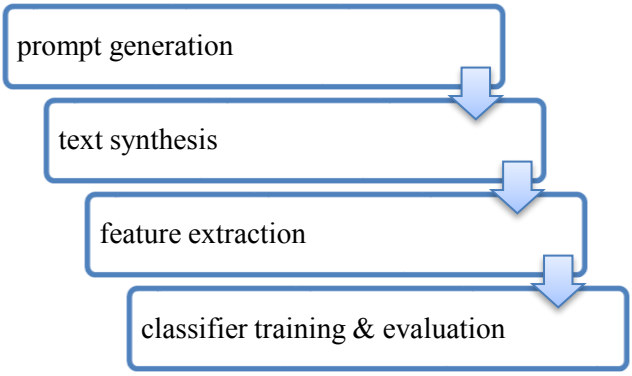

Fig. 1: Experimental workflow overview

4.1.1 Synthetic Data Generation

During our experimental phase, we utilized Claude 3.7 Sonnet to create 300 prompts in ten domain categories, [12]. These categories included creative storytelling, technical explanations, data-analysis reports, casual conversation, marketing copy, news reporting, instructional tutorials, philosophical discourse, legal and contractual writing, and entertainment reviews. Each prompt was crafted to generate synthetic text in English with a target length of 200 to 250 words. To reduce variability, we structured every prompt in the same way, starting with a *system message* ("You are a helpful creative-writing assistant.") and followed by a *user* directive ("<domain prompt> Provide only the answer. The result should be a creative, well-written text of 200–250 words."). This approach ensured standardized formatting across all unique 3,000 prompts, which eliminates structural variations that

might occur in the generated texts. Our use of ten different domains forced the LLMs to alter their style, and vocabulary, and consider different contexts. This design allowed us to assess how consistently each model's characteristics appeared in varied content.

4.1.2 Model Setup and Computational Pipeline

After generating synthetic data, we created Python scripts that ran on a GPU server of the HEMUS supercomputer (Heterogeneous Multi-functional Supercomputer) of the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. While the HEMUS system includes several GPU servers, we used only one for our experiment, which used eight NVIDIA A100-SXM4-40GB GPUs. We deployed two instruction-tuned open-source language models from Hugging Face:

- **google/gemma-3-27b-it**, [13], processed the 3,000 prompts in 1,939 minutes using 6 GPUs (running three instances simultaneously, with 2 GPUs per instance), generating 644,680 tokens.
- **Qwen/Qwen2.5-32B-Instruct**, [14], processed the same prompts in 968 minutes using 8 GPUs (running three instances simultaneously, unevenly distributed across instances), generating 807,161 tokens.

When normalized to a single GPU, the total compute time reached nearly 50 hours.

These experiments did not aim to optimize speed or token-generating efficiency, but only to produce synthetic texts that accurately corresponded to each prompt. Both models operated under the same generation settings: temperature=0.6, top-p=0.95, top-k=30, repetition_penalty=1.1, and a limit of 512 new tokens. We established a fixed random seed for each batch, discarding and regenerating any responses that were outside the desired length. This ensured consistent length and formatting across all outputs, isolating each model's intrinsic style as the primary source of variation.

Finally, we developed a Python script that trained and evaluated classifiers on the dataset. The script performed data loading, feature extraction, model training, and performance assessment that enabled systematic parameter tuning and ensured reproducibility. The detailed configurations of the Machine Learning classification models and the corresponding experimental results are outlined in the following sections.

4.2 Machine Learning Classification Model Training

Our approach to differentiate between Qwen 2.5 from Gemma 3 produced content required us to develop a complete Machine Learning system.

4.2.1 Model Setup and Computational Pipeline

We used four Machine Learning classification models to evaluate the distinguishing features between outputs:

- **Logistic Regression** is the baseline model as it can be trained quickly, its ease of interpretation, and capacity to generate probability estimates. This characteristic makes it especially effective for high-dimensional, sparse textual features.
- **Support Vector Machine (SVM)** is selected for its effectiveness when managing high-dimensional text data and its capacity to capture complex nonlinear relationships through kernel transformations, which ensures strong generalization performance on unseen examples.
- **Random Forest** offers an ensemble approach that reduces overfitting by aggregating multiple decision trees and provides built-in feature importance metrics that reveal key lexical and stylistic markers distinguishing the two models.
- **Gradient Boosting** is included for its sequential learning approach that iteratively improves upon previous predictions, effectively capturing subtle patterns in text generation that other algorithms might overlook.

Together, these complementary algorithms create a comprehensive evaluation framework for identifying linear and nonlinear patterns in LLM-generated text, enabling robust cross-validation of our findings across various classification paradigms.

4.2.2 Organization of Training

We developed two feature sets for our classification experiments:

1. TF-IDF Only: We extracted lexical features using TF-IDF vectorization, [15], limiting the selection to 5,000 n-grams (unigrams and bigrams) based on established heuristics for mid-sized corpora (min_df = 5, max_df = 0.7).

2. All Features: We combined TF-IDF vectorization with additional syntactic and stylistic features:

- *Computational Efficiency*: Sentence-length statistics, part-of-speech ratios, and punctuation profiles can be extracted quickly and efficiently, maintaining peak memory usage below two gigabytes.

- *Linguistic Coverage*: These features include surface forms, clause-level organization, and grammatical preferences. Preliminary checks indicate that the use of conjunctions and medium-length sentences effectively differentiate Qwen 2.5 from Gemma 3, suggesting variations in training corpora or decoding routines.
- *Complementarity*: These features offer insights that extend beyond what word frequency alone can provide, particularly benefiting tree-based models.

We then used the *FeatureExtractor* class to boost potential dependency-relation counts, sentence-embedding cosine distances, and discourse markers. Our pilot run on a 10% subset shows a further 0.4% gain when universal-dependency bigrams are included.

4.2.3 Organization of Training

After labeling the outputs from the Qwen 2.5 and Gemma 3 models with 3,000 samples each, we divided the total dataset of 6,000 samples using an 80:20 ratio for training and validation, resulting in 4,640 training samples and 1,160 validation samples. The remaining 200 samples (100 from each model) were set aside as a held-out test set. Each text instance underwent processing through our feature-extraction pipeline, which implemented the functions described above.

4.2.4 Hyperparameter Tuning

Hyperparameter selection was conducted through iterative experimentation rather than exhaustive grid or randomized search. We set the random seed to 42 for all operations to ensure reproducibility. Specific configurations included:

- **Logistic Regression**: C = 1.0 with L2 penalty, liblinear solver, balanced class weights, maximum 1,000 iterations, tolerance = 1×10^{-4}.
- **SVM**: LinearSVC with C = 1.0, L2 penalty, dual = False, maximum 2,000 iterations, tolerance = 1×10^{-4}, balanced class weights, with isotonic calibration via CalibratedClassifierCV.
- **Random Forest**: 100 trees, max_depth = 10, min_samples_split = 10, min_samples_leaf = 4, max_features = "sqrt", balanced class weights, n_jobs = -1.
- **Gradient Boosting**: 50 estimators, learning_rate = 0.1, max_depth = 3, subsample = 0.6, max_features = 0.3, min_samples_split = 20, min_samples_leaf = 10.

4.2.5 Model Performance and Feature Importance

All classifiers demonstrated strong performance during their development stage (training and validation phases) in distinguishing between Qwen 2.5 and Gemma 3:

- **Logistic Regression**: Achieved 98.45% accuracy with TF-IDF only, with a slight drop to 96.64% using all features.
- **SVM**: Reached 98.79% accuracy with TF-IDF, maintaining 98.53% with all features.
- **Random Forest**: Improved from 94.22% with TF-IDF to 95.69% with all features.
- **Gradient Boosting**: Showed the biggest improvement, from 93.28% to 96.21% when adding non-lexical features.

Feature importance analysis (detailed in Appendix in Table 1 and Table 2) reveals two distinct patterns: the contraction "isn" (from "isn't") serves as the most significant single discriminator across all models, while conjunction patterns such as "but also" and "not only," along with structural cues like average sentence length and conjunction ratio, consistently appear across models.

Tree-based classifiers benefit from stylistic signals, including uppercase ratio and punctuation density, especially when word frequencies alone are insufficient for decisive classification.

4.3 Experimental Classification Results

Classification accuracy by all four models was high, although with differences in confidence distribution between classifiers and feature sets. SVM achieved the highest overall accuracy at 99.0%, outperforming the other methods. Both SVM and Logistic Regression excelled in distinguishing Qwen 2.5 from Gemma 3 text, while Random Forest and Gradient Boosting trailed slightly in accuracy. Notably, even the lowest-performing configuration (Gradient Boosting with TF-IDF only) still had 91.0% accuracy, indicating that differentiating Qwen 2.5 vs. Gemma 3 outputs is largely successful across all tested classifiers.

4.3.1 Adding Syntactic and Stylistic Features to TF-IDF

Comparing feature sets shows that adding syntactic and stylistic features to TF-IDF sometimes provided marginal improvements to accuracy but was not universally beneficial. The SVM's performance reached a threshold of 99.0% accuracy with both feature sets, suggesting that lexical TF-IDF features alone captured nearly all the information needed for this classifier. Logistic Regression performed

slightly better with TF-IDF only (98.5%) than with "all features" (97.0%), suggesting that the additional features may have introduced minor noise or overfitting for this linear model. In contrast, the ensemble models benefited from the richer feature set: Gradient Boosting saw a substantial accuracy jump from 91.0% (TF-IDF only) to 96.0% with all features included. Random Forest improved modestly from 94.5% to 95.5%. These patterns suggest that while simple lexical features are powerful discriminators for linear classifiers, more complex models can use supplementary syntactic/stylistic cues to improve the classification of Qwen 2.5 vs. Gemma 3 texts.

4.3.2 Receiver Operating Characteristic Curve Analysis

We conducted Receiver Operating Characteristic (ROC) curve analysis on all four classifiers to further evaluate classifier performance beyond accuracy metrics. ROC curves provide a graphical representation of a classifier's discriminative ability by plotting the True Positive Rate against the False Positive Rate at various threshold settings, while the Area Under the Curve (AUC) offers a single scalar value summarizing overall classifier performance.

As shown in Figure 2, all four classifiers demonstrate strong performance in distinguishing between Qwen 2.5 and Gemma 3 outputs, with AUC values significantly exceeding the random classifier baseline (AUC = 0.5). The SVM classifier achieved particularly impressive results, with perfect classification (AUC = 1.00) for Gemma 3 with all features and Qwen 2.5 with TF-IDF only. This exceptional performance aligns with the SVM's superior accuracy metrics noted in Section 4.3.

Logistic Regression similarly performed well, especially when classifying Gemma 3 outputs (AUC = 0.99 with all features, AUC = 0.98 with TF-IDF only). The ensemble methods showed more variability, with Random Forest performing better on Gemma 3 outputs (AUC = 0.97) than Qwen 2.5 (AUC = 0.83), while Gradient Boosting displayed more balanced performance across both models.

The impact of feature selection varies across classifiers. For SVM, TF-IDF features alone were sufficient to achieve optimal performance when classifying Qwen 2.5 outputs (AUC = 1.00), while for Random Forest, the addition of syntactic and stylistic features substantially improved discrimination for both models.

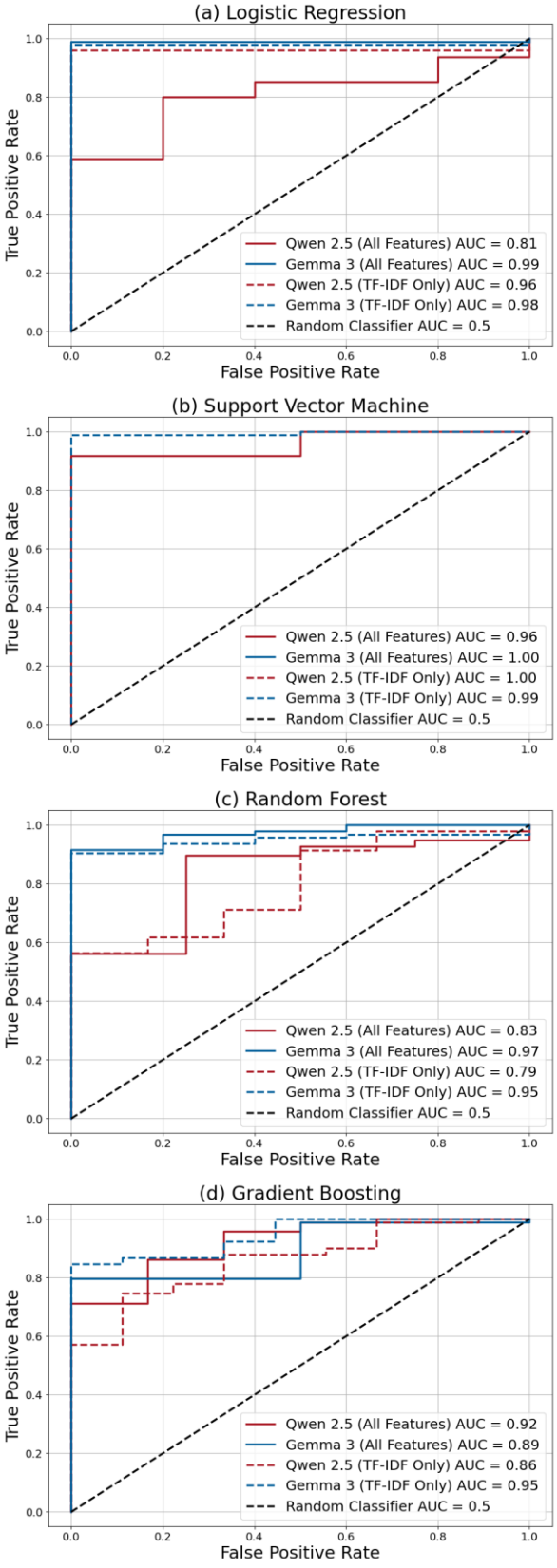

Fig. 2: ROC curves for distinguishing between Qwen 2.5 and Gemma 3 outputs: (a) Logistic Regression, (b) Support Vector Machine, (c) Random Forest, and (d) Gradient Boosting

These patterns support our conclusions in Section 4.3.1 that lexical features provide strong signals for linear classifiers, but ensemble methods benefit from additional feature representations.

The consistently high AUC values across all classifiers confirm that Qwen 2.5 and Gemma 3 produce outputs with distinctive characteristics that can be reliably detected and differentiated using Machine Learning techniques, regardless of the specific classification approach employed.

4.3.3 Calculation of Confidence Values

Beyond overall accuracy, an analysis of prediction confidence highlights essential differences in how reliably each model distinguishes between text sources. Based on the model's output score, we categorized each prediction's confidence as high, medium, or low.

Confidence levels are categorized as follows:

- **High Confidence**: Predictions with a probability greater than 0.9 (90%)
- **Medium Confidence**: Predictions with probability between 0.7 and 0.9 (70-90%)
- **Low Confidence**: Predictions with probabilities between 0.5 (50%) and 0.7 (70%)

These thresholds are applied uniformly across all four models (Figure 3), with high confidence typically reflecting a strong posterior probability for the predicted class.

Logistic Regression is relatively confident, as illustrated in Figure 3(a). Using all features, it labeled about two-thirds of each class with high confidence (60 Qwen and 66 Gemma cases) and very few as low confidence. When Logistic Regression was limited to TF-IDF features, the model became more conservative, with high-confidence predictions dropping to approximately 41% per class for all classifiers. In this TF-IDF-only scenario, the model was still highly accurate (even slightly more so), but it signaled uncertainty more often, suggesting improved calibration (fewer overconfident errors).

SVM attains near-perfect accuracy and makes most of its predictions with high confidence. As shown in Figure 3(b), for both Qwen 2.5 and Gemma 3 outputs, SVM (in either feature setting) assigned high confidence (93-97%) to its classifications, leaving almost no low-confidence cases. This indicates a clear separation between Qwen 2.5 and Gemma 3 patterns that allow for higher confidence calls by SVM.

Random Forest stands out for the opposite reason, as seen in Figure 3(c), as most of its predictions were not high confidence. With all features,

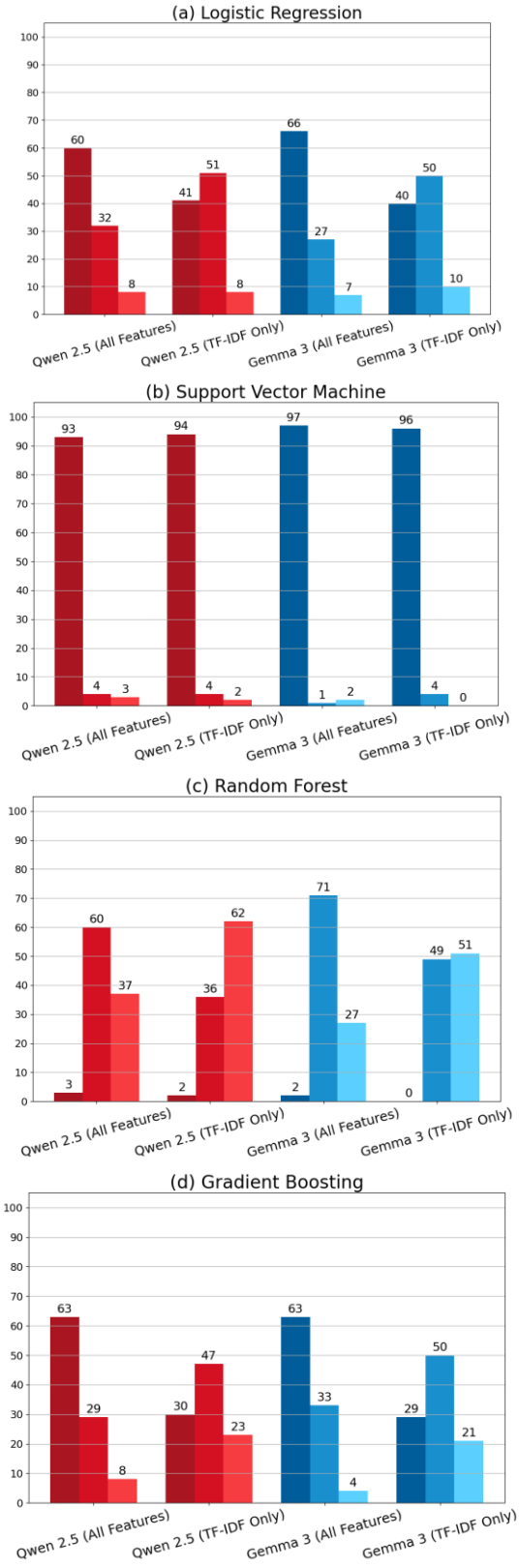

Fig. 3: Confidence level distribution across different classifiers: (a) Logistic Regression, (b) Support Vector Machine, (c) Random Forest, and (d) Gradient Boosting.

Each group of three bars represents predictions for a model configuration, with bars showing (from darkest to lightest shade) high/medium/low confidence

Random Forest made only three high-confidence classifications for Qwen outputs and 2 for Gemma, while making 37 and 27 low-confidence cases, respectively. With TF-IDF only, it produced virtually no high-confidence decisions. This suggests that the Random Forest model assigned marginal probabilities of a difference between Qwen 2.5 and Gemma 3 outputs, even when correct.

Gradient Boosting exhibited a similar trend to Logistic Regression, as depicted in Figure 3(d). With all the features set, it made 63 high confidence and only 8 and 4 low confidence calls for Qwen and Gemma, respectively. In contrast, with TF-IDF only, its high-confidence count roughly halved (30 Qwen, 29 Gemma), and low-confidence predictions increased.

5 Discussion

The experimental results show that raw accuracy and prediction confidence influence classifier reliability. For instance, SVM achieved near-perfect accuracy and produced a high proportion of high-confidence predictions, which suggests that the feature space was well-separated. On the other hand, Logistic Regression, when limited to TF-IDF features, showed a slight improvement in accuracy and a shift towards more medium-confidence predictions, which suggests that the removal of additional features reduced overconfidence and improved calibration. The ensemble models behaved differently in our experiments. Gradient Boosting benefited from the inclusion of syntactic and stylistic features, which resulted in a substantial increase in high-confidence predictions and accuracy. Although the Random Forest classifier reached approximately 95% accuracy, it frequently produced low-confidence calls, so it often found only a small distinction between Qwen 2.5 and Gemma 3 outputs, even when accurately classifying outputs.

Our method shows promising results, but it comes with certain limitations. The main difficulty stems from the expected decline in attribution accuracy when texts experience extensive editing or paraphrasing. Human editing processes can eliminate or modify the distinctive linguistic patterns of Qwen 2.5 and Gemma 3, which would decrease the classifier's performance. Furthermore, the classifiers received training data only from the current model outputs. Taking this into account, we acknowledge that the linguistic features of the attribution systems require continuous updates because subsequent fine-tuning, task-specific retraining, and domain adaptation would significantly modify them.

Several strategies have been discussed to address these limitations, such as the regular retraining or calibration of the classifiers with updated data from the latest LLM versions, which could help maintain higher detection accuracy. Additionally, using hybrid detection methods that combine lexical and stylistic classifiers could boost accuracy. Another effective approach involves implementing confidence-thresholding mechanisms that prompt further human review for classifications with lower confidence, thereby reducing the chances of misclassification. Advanced linguistic features such as syntactic dependencies and semantic embedding may enhance detection accuracy as generative models advance.

The examined models demonstrate accurate text classification abilities, but their confidence distributions show substantial variation, which needs evaluation during real-world use. Furthermore, our assessment of linguistic features shows that Qwen 2.5 and Gemma 3 produce different patterns in their text generation when executing different tasks. The models differ in their shared phrases and entity patterns, vocabulary selection, syntactical structure, readability metrics, and final output style.

We conducted a basic analysis of the LLMs' stylistic traits. The result was that Qwen's outputs display a structured and deliberate style. The sentences in Qwen tend to be longer and more intricate, characterized by formal and precise diction that maintains a clear expository voice. This approach results in organized and professional text, whether the model describes an ancient temple or outlines contractual terms.

By contrast, Gemma's outputs reflect narrative richness and expressiveness. It uses a mix of sentence lengths, often adding short sentences for emphasis, employs a diverse vocabulary, and uses expressive punctuation for dramatic effect. Gemma also infuses personality and creativity into factual or instructional writing. These characteristics make Gemma's texts captivating and dynamic, with vivid scenes and lively dialogues that engage the reader.

The stylistic differences between Qwen 2.5 and Gemma 3 reflect their developers' different philosophies and training methods. The documentation shows Qwen 2.5 received training on 18 trillion tokens while its corpus contained

primarily technical and code-related content. The technical focus of Alibaba on enterprise reliability and structured outputs through alignment techniques results in a formal, methodical tone that matches the emphasis on precision and factuality. Accordingly, Qwen's output often consists of complex, well-organized paragraphs with methodical progressions and structured arguments. This is optimal for structured analytical tasks requiring professional text with a consistent voice throughout.

Gemma 3, according to public information, was trained on approximately 14 trillion tokens drawn from a multilingual corpus that spans more than 140 languages. The training method follows DeepMind's mission to make AI accessible to users. The combination of Gemma's architecture and post-training optimization for conversational fluency leads to an expressive narrative style that uses vivid sensory details, rhythmic sentence structures, and stylistic techniques such as sentence fragments and conversational asides. The model demonstrates exceptional performance in creative and dialog-based tasks because its outputs exhibit wide emotional ranges and dynamic narrative voices. The architectural and training differences between these models explain the distinct patterns observed in our comparative analysis.

The results show that syntactic and stylistic cues occasionally enhance classification accuracy, but these additional features fail to outperform well-optimized TF-IDF features. The tested models provided essential practical insights, especially when feature interpretability or specific analyses were needed. The classification results show that the two language models produce distinct lexical patterns, which suggests that future research could investigate how combining more complex feature sets or hybrid approaches would enhance attribution accuracy.

Despite our observations of their differences, both LLMs generate coherent and appropriate text for the prompts. The models demonstrate comparable readability and grammatical precision, yet they maintain distinct writing styles. The comparison shows that both models achieve outstanding results while preserving their linguistic features in the generated outputs.

6 Conclusion

Our findings indicate that Machine Learning classifiers can successfully distinguish between texts produced by Qwen 2.5 and Gemma 3, even when using basic lexical features such as TF-IDF. Both linear and ensemble classifiers utilize unique lexical patterns specific to each model's outputs. While adding more syntactic and stylistic features can occasionally improve classification accuracy, these richer features do not consistently surpass the effective baseline established by well-tuned frequency-based markers. However, including more advanced linguistic cues might be advantageous for applications that require higher interpretability.

While our classifiers show strong performance in controlled experiments, further investigation is needed to assess their reliability in realistic situations that involve extensive text editing or paraphrasing. Future attribution systems will probably need ongoing recalibration, the incorporation of more extensive linguistic feature sets, and hybrid approaches to maintain their effectiveness, particularly as models like Qwen 2.5 and Gemma 3 progress and advance. Adopting these strategies will be crucial for securing accurate attribution in practical settings.

This study highlights the practical use of automated attribution techniques for identifying LLM-generated content in real-world scenarios like content moderation, plagiarism detection, and document authentication. The unique stylistic fingerprints identified here can be applied in multiple ways: (i) immediate flagging of questionable, AI-produced sections in learning-management systems; (ii) forensic authorship verification for journal or grant submissions, where section-level attribution reports could aid in maintaining academic integrity; and (iii) combining "watermark + style" detectors, which merge lexical signatures with cryptographic watermarks, to enhance plagiarism prevention methods against paraphrasing.

A TF-IDF-only SVM can perform these checks in milliseconds on conventional hardware. Conversely, more complex ensemble models may be reserved for critical cases where additional latency is permissible. The classifier demonstrates strong accuracy across different prompts, indicating resilience to varying input styles. Nonetheless, the decision between a simple lexical baseline and a detailed ensemble depends on resource limitations, the need for transparency, and the tolerable margin of error.

While the results are encouraging, several limitations exist. First, it remains uncertain how well the distinctive patterns we observed would play out in other genres or applications. Although we attempted to cover a broad range of prompts, this set cannot exhaustively represent all possible inputs. There may be prompt formulations or content areas where Qwen 2.5 and Gemma 3 outputs converge

stylistically, reducing the efficacy of our feature-based attribution. Another important caveat is that the human evaluation in this study was not comprehensive. Our analysis is based on computational metrics and differences, and we have not verified whether human readers notice or care about these same differences, or whether one model's output might be preferred in terms of readability or quality. In addition, our investigation was confined to comparing two specific model versions. We did not test how our approach performs with other LLMs or models with smaller or larger parameter counts, so the uniqueness of the identified features to this particular pair remains an open question. These limitations highlight that as strong as our classifiers performed here, their real-world application requires further validation in broader conditions.

Looking ahead, the insights gained here open several avenues for future research. A possible next step is to extend this comparative pattern analysis to multilingual settings, examining whether the attribution techniques remain accurate in other languages or if new language-specific features emerge when distinguishing model outputs in other languages. Similarly, cross-domain studies would be valuable: applying the developed classifiers to content far outside the training domain would test the robustness of the identified stylometric cues under substantial shifts in topic and style. Beyond standalone analysis, there is practical merit in integrating these detection methods with plagiarism detection and authorship verification systems. Such integration could help flag AI-generated sections in academic or professional writing and verify content origins, thereby growing the number of tools available for ensuring the integrity of academic publications or education. Another promising direction is to develop detectors capable of handling documents with mixed human and AI authorship. As AI assistants become common co-writers, future attribution methods must discern partial AI contributions embedded within human-written text, a significantly more challenging scenario than the binary classifications we tackled. Exploring these directions will not only address the current gaps in generalization and applicability but will also move the field closer to effective solutions for AI text attribution.

Taken together, these observations highlight the practical feasibility of distinguishing between Qwen 2.5 and Gemma 3 outputs and underscore the importance of calibration and feature selection in real-world deployment. While each model excels at generating coherent and contextually appropriate text, their unique stylistic traits remain detectable through a variety of computational approaches. By refining these methods in subsequent research, it may be possible to adapt classification systems to an even broader spectrum of generative models while preserving consistency, interpretability, and high classification accuracy.

Acknowledgement:
We acknowledge the provided access to the e-infrastructure of the Centre for Advanced Computing and Data Processing, with financial support from Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020) and co-financed by the European Union through the European Structural and Investment funds.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors wrote, reviewed and edited the content as needed and they have not utilised artificial intelligence (AI) tools. The authors take full responsibility for the content of the publication.

References:

[1] Rosenfeld, A., & Lazebnik, T. (2024). Whose LLM is it anyway? Linguistic comparison and LLM attribution for GPT-3.5, GPT-4 and Bard. https://doi.org/10.48550/arXiv.2402.14533.

[2] Qwen Team (2024, September). Qwen2.5: A party of foundation models, [Online]. https://qwenlm.github.io/blog/qwen2.5/ (Accessed Date: October 5, 2024).

[3] Gemma Team. (2025). Gemma 3 technical report. Kaggle, [Online]. https://goo.gle/Gemma3Report (Accessed Date: October 5, 2024).

[4] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z.

[5] Zhang, Y., Ma, Y., Liu, J., Liu, X., Wang, X., & Lu, W. (2024). Detection vs. Anti-detection: Is Text Generated by AI Detectable? In Sserwanga, I., Joho, H., Ma, J., Hansen, P., Wu, D., Koizumi, M., & Gilliland, A. J. (Eds.), *Wisdom, Well-Being,*

Win-Win. iConference 2024. Lecture Notes in Computer Science, Vol. 14596, pp.209–222. Springer, Cham. https://doi.org/10.1007/978-3-031-57850-2_16.

[6] Aydin, O., Karaarslan, E., Erenay, F. S., & Bacanin, N. (2025). Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma. https://doi.org/10.48550/arXiv.2503.04765.

[7] Perkins, M., Roe, J., Postma, D., & McGaughran, J. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6.

[8] Watson, S., Brezovec, E., & Romic, J. (2025). The role of generative AI in academic and scientific authorship: An autopoietic perspective. *AI & Society*. https://doi.org/10.1007/s00146-024-02174-w.

[9] Alhijawi, B., Jarrar, R., AbuAlRub, A., & Bader, A. (2025). Deep learning detection method for large language models-generated scientific content. *Neural Computing and Applications, 37*(1), 91–104. https://doi.org/10.1007/s00521-024-10538-y.

[10] Kopanov, K. (2024). Comparative Performance of Advanced NLP Models and LLMs in Multilingual Geo-Entity Detection. *Cognitive Models and Artificial Intelligence Conference (AICCONF '24)*, pp.106-110. ACM. https://doi.org/10.1145/3660853.3660878.

[11] Formosa, P., Bankins, S., Matulionyte, R., & Ghasemi, O. (2024). Can ChatGPT be an author? Generative AI creative writing assistance and perceptions of authorship, creatorship, responsibility, and disclosure. *AI & Society*. https://doi.org/10.1007/s00146-024-02081-0.

[12] Anthropic. (2025). Claude 3.7 Sonnet, [Online]. https://www.anthropic.com/claude/sonnet (Accessed Date: March 5, 2025).

[13] Google. (n.d.). *Gemma 3 27B instruction-tuned model*, [Online]. https://huggingface.co/google/gemma-3-27b-it (Accessed Date: March 5, 2025).

[14] Qwen Team. (n.d.). *Qwen 2.5 32B instruction-tuned model*, [Online]. https://huggingface.co/Qwen/Qwen2.5-32B-Instruct (Accessed Date: March 5, 2025).

[15] Senbel, S. (2021). Fast and memory-efficient TFIDF calculation for text analysis of large datasets. In H. Fujita, A. Selamat, J. CW. Lin, & M. Ali (Eds.), Advances and trends in artificial intelligence: Artificial intelligence practices (pp. 557-563). Springer. https://doi.org/10.1007/978-3-030-79457-6_47.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Tatiana Atanasova conceptualized the study, established the theoretical framework, supervised the research process, and reviewed and enhanced the manuscript.
- Kalin Kopanov designed the methodology, developed the experimental framework, conducted the data collection and analysis, and drafted the initial manuscript.

Both authors jointly interpreted the results and formulated the conclusions. All authors have read and agreed to the published version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The presented study is supported by the National Science Fund at the Ministry of Education and Science of Bulgaria, with Contract № КП-06-Н85/16.

Conflict of Interest

The authors have no conflicts of interest to declare.

APPENDIX

Table 1. Overall top ten features. List the features that receive the highest raw importance scores in at least one of the four classifiers

Rank	Feature (description)	Logistic Regression	SVM	Random Forest	Gradient Boosting
1	isn (contraction “isn’t”)	4.122	2.450	0.037	0.119
2	conjunction ratio (share of conjunction tokens)	3.615	2.356	0.036	0.090
3	function word ratio (common closed-class words)	4.074	2.597	0.016	0.057
4	punctuation, comma (,)	0.366	0.262	–	–
5	uppercase ratio (share of capital letters)	–	–	0.049	0.096
6	punctuation, period (.)	–	–	0.023	0.053
7	sentence count (number of sentences)	–	–	0.041	0.094
8	average sentence length (words per sentence)	–	–	0.035	0.109
9	punctuation, colon (:)	–	–	0.025	0.036
10	such (word “such”)	2.203	1.351	0.014	–

Note: Higher values within each column indicate features with greater discriminative power for that specific model. Cross-column comparisons are meaningful only between linear-model pairs (Logistic Regression and SVM) or between tree-based pairs (Random Forest and Gradient Boosting) because the two families use different scoring scales

Table 2. Leading lexical, syntactic, and stylistic cues for each classifier. The layout clarifies that the linear models rely mostly on lexical markers, whereas the tree ensembles heavily rely on sentence-level and stylistic statistics

Family	Feature (description)	Logistic Regression	SVM	Random Forest	Gradient Boosting
Lexical	isn (contraction)	4.122	2.450	0.037	0.119
	such	2.203	1.351	0.014	–
	ensure	2.381	1.333	0.009	0.012
Syntactic	conjunction ratio	3.615	2.356	0.036	0.090
	average sentence length	0.232	0.038	0.035	0.109
	sentence count	0.143	0.041	0.041	0.094
Stylistic	function word ratio	4.074	2.597	0.016	0.057
	uppercase ratio	1.262	0.856	0.049	0.096
	punctuation density (all marks per 100 tokens)	0.662	0.408	0.025	0.040

Note. Higher values within each column indicate features with greater discriminative power for that specific model. Cross-column comparisons are meaningful only between linear-model pairs (Logistic Regression and SVM) or between tree-based pairs (Random Forest and Gradient Boosting) because the two families use different scoring scales