# Can Moderate-Sized Language Models Imitate an Individual Academic Writing Style?

Aleksandra Hinc[1], Jan K. Krajewski[1], Leszek J. Chmielewski[1], and Arkadiusz Orłowski[1,2]

[1] Warsaw University of Life Sciences – SGGW, Institute of Information Technology, Warszawa, Poland
[2] Military University of Technology – WAT, Institute of Mathematics and Cryptology, Warszawa, Poland
arkadiusz_orlowski@sggw.edu.pl

**Abstract.** Recent advances in large language models (LLMs) have significantly improved the fluency and coherence of automatically generated text. This study investigates whether moderate-sized causal language models (up to 500 million parameters) can reproduce the individual academic writing style of a specific author through fine-tuning on a limited corpus of scientific introductions. Three open-access models (Gemma 3, GPT-2 Medium, and Qwen1.5) were evaluated before and after adaptation using six complementary metrics: perplexity, semantic similarity, a style classifier, stylometric features, stylometric distance, and idiosyncratic phrase matching. Results indicate substantial improvements in language modelling quality and moderate gains in classifier-based stylistic attribution. However, distinctive authorial traits remain weakly represented, suggesting that adaptation primarily reinforces genre-level conformity rather than robust individual style imitation.

**Keywords:** Author style imitation · Fine-tuning · Stylometry · Language models

## 1 Introduction

Recent advances in large language models (LLMs) have significantly improved the fluency and coherence of automatically generated text. Although substantial research has focused on text style transfer and controlled generation, considerably less attention has been paid to the problem of imitating the individual writing style of a specific author from scratch.

Most existing approaches address style as a controllable attribute such as sentiment, formality, or politeness. In contrast, author-specific style imitation requires capturing subtle and often implicit stylistic regularities that distinguish one individual's writing from another's, particularly within a strongly conventionalized genre such as academic prose.

This paper investigates whether moderate-sized causal language models (up to 500 million parameters) can acquire and reproduce the individual academic

writing style of a specific author through fine-tuning on a relatively small corpus. We focus on introductions to scientific papers written or co-written by the selected author and evaluate the effect of fine-tuning using both automatic and human-centred metrics [5].

Three open-access language models: Gemma 3 [2], GPT-2 Medium [10], and Qwen1.5 [12] were fine-tuned and evaluated before and after adaptation. Performance was assessed using six complementary metrics that capture different aspects of generated text: perplexity [15], semantic similarity [3], a style classifier based on word embeddings [11] and logistic regression, stylometric distance [13], matching of idiosyncratic phrases [16], and stylometric features [9].

Our results show that fine-tuning improves several quantitative metrics, particularly language modelling quality and classifier-based style attribution. However, the individual authorial style remains only weakly distinguishable, suggesting that moderate-sized models trained on limited corpora primarily capture genre-level regularities rather than highly specific stylistic traits.

## 2 Related Work

Research on stylistic control in natural language generation has primarily focused on text style transfer (TST) [8], where predefined attributes such as sentiment, politeness, or formality are modified while preserving semantic content. Existing approaches include supervised and unsupervised methods, paraphrase-based reformulations, and large-scale benchmarking frameworks [1,7,14].

A key challenge in TST research concerns evaluation [14]. Automatic metrics often measure fluency or semantic preservation, but do not necessarily capture stylistic fidelity. As a result, many studies combine embedding-based similarity measures, classifier-based attribution, and human assessment.

In contrast to attribute-based style transfer, the problem of author-specific style imitation is less frequently studied. Individual style is typically investigated in stylometry and authorship attribution research, where features such as lexical richness, sentence length, and function word frequencies are used to distinguish between authors. However, little work has explored whether generative language models can reproduce such stylistic traits when fine-tuned on a limited corpus.

The present study bridges these perspectives by combining fine-tuning of causal language models with multi-dimensional evaluation inspired by both stylometric analysis and contemporary NLP metrics.

## 3 Experimental Setup

### 3.1 Dataset

The dataset consisted of 120 introductions to scientific papers written or co-written by a single academic author (e.g. [6]). Only introductions were selected, to reduce structural variability across different sections of articles. The texts were written in English and represent a specialized variant of academic prose.

The introductions contained about 72 K words, that is, over 292 328 tokens. The average length of an introduction was 773 words, with substantial variability across samples.

After preprocessing (removal of headings, page numbers, footnotes, and formatting artifacts), the dataset was split into training, validation, and test subsets in proportions approximately 0.775 : 0.10 : 0.125 (93 : 12 : 15 texts, respectively).

For generation, the first 500 characters of each test introduction were used as prompts, and models were instructed to continue the text.

### 3.2 Models

Three open-access causal language models (CLMs) with up to 500 million parameters were selected: 1° Gemma 3 (270 M parameters) [2], 2° GPT-2 Medium (355 M parameters) [10], and 3° Qwen1.5-0.5B-Chat (500 M parameters) [12]. All models are transformer-based and were obtained from the Hugging Face repository in their publicly available pretrained versions. Only models supporting English were considered.

The selection criterion was to evaluate moderate-sized architectures that are computationally feasible to fine-tune on limited hardware resources.

### 3.3 Fine-Tuning Procedure

Fine-tuning was performed on the training subset (93 introductions) using causal language modelling. The texts were concatenated and segmented into blocks of 1024 tokens to improve contextual learning.

Each model was fine-tuned multiple times with different hyperparameter configurations, and the final configuration was selected based on validation loss. The final learning rate was set to $5 \times 10^{-6}$ for all models. The optimal number of epochs was 4 for Gemma 3 and Qwen1.5, and 6 for GPT-2 Medium.

Training and validation loss values were monitored to avoid overfitting, and the lowest loss checkpoint was used for subsequent evaluation.

### 3.4 Generation Protocol

For each of the 15 test prompts, text continuations were generated using both the original pretrained models and their fine-tuned counterparts. This enabled direct comparison between baseline and adapted versions.

All generations were produced under identical decoding settings for each model. The generated continuations were then evaluated using automatic metrics and qualitative human assessment.

## 4 Evaluation Metrics

### 4.1 Perplexity

Perplexity [15] was used as a standard measure of language modelling performance. It quantifies how well a model predicts a given sequence, with lower values indicating higher predictive confidence and improved fluency.

Although perplexity does not directly measure stylistic fidelity, it reflects improvements in internal language modelling after fine-tuning and serves as a baseline indicator of adaptation quality.

### 4.2 Semantic Similarity

Semantic similarity between generated continuations and reference introductions was computed using sentence embeddings [3]. Cosine similarity between embedding vectors was used to assess topical alignment.

This metric evaluates whether generated texts remain thematically consistent with the prompts and the reference corpus. However, high semantic similarity does not necessarily imply stylistic similarity, as semantically aligned texts may still differ stylistically.

### 4.3 Style Classifier

A simple style classifier based on TF–IDF features [11] and logistic regression was constructed using held-out author samples and a set of negative examples generated by GPT4 to estimate the probability that a text represents the target author's style.

The classifier outputs a probability score representing the degree of stylistic attribution. This approach is inspired by authorship attribution methods and provides a supervised perspective on stylistic alignment.

### 4.4 Stylometric Distance

To obtain a single aggregated stylistic measure, a stylometric feature vector was constructed for each text, and cosine distance to the mean reference-style vector was computed [13]. The resulting value lies in the interval [0, 1], where lower values indicate greater stylistic similarity to the reference corpus.

### 4.5 Idiosyncratic Phrase Matching

To assess phrase-level stylistic imitation, $n$-gram overlap between generated texts and the reference corpus was computed [16]. The number of matching n-grams was normalized relative to total n-grams in generated texts.

This metric was designed to capture author-specific lexical patterns. However, high overlap may mean memorization rather than stylistic generalization, whereas very low overlap may suggest absence of distinctive phrase usage.

### 4.6 Stylometric Features

Four stylometric features were extracted from each generated text: 1° proportion of long words (more than six letters), 2° average sentence length, 3° lexical richness, and 4° readability (Flesch index). These features capture structural and lexical regularities commonly used in stylometric analysis [9]. For each feature, values were compared to the mean values computed from the reference corpus.

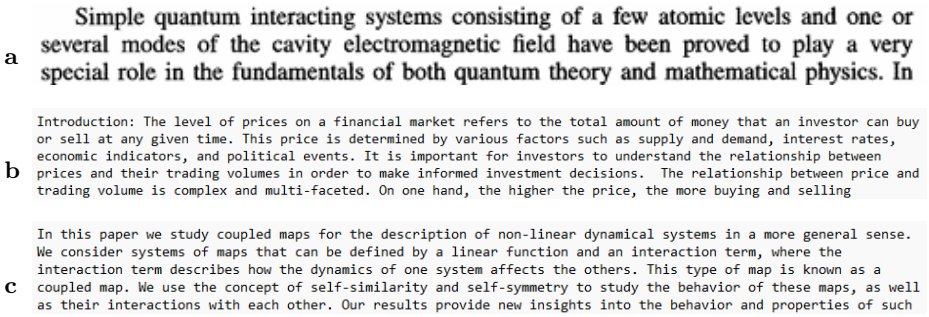


**Fig. 1.** Original introduction (**a**) and texts generated by Qwen1.5: (**b**) before fine-tuning and (**c**) after fine-tuning (fragments). For more images see [4].

## 5 Results

This section presents the quantitative and qualitative evaluation of text generations produced by the pretrained and fine-tuned models. Results are organized according to the evaluation metrics described in Section 4. For each metric, we compare baseline and fine-tuned versions of the three models. The results (except stylometric features) are collected in Table 1[3], and the values of stylometric features are shown in Table 2[3].

Images of small fragments of examples of a prompt and texts generated by the best model Qwen1.5 before and after tuning are shown in Fig. 1. For full versions of these images and images of texts generated with other models see [4].

The primary objective is not only to determine whether fine-tuning improves numerical scores but also to assess whether such improvements correspond to meaningful stylistic alignment with the target author. Therefore, particular attention is given to patterns across models and to discrepancies between different types of metrics.

### 5.1 Human Assessment (qualitative inspection)

As an initial sanity check, we conducted a qualitative human inspection by reading the outputs and comparing them to the target corpus. This step served to verify whether fine-tuning produces any perceivable change in academic tone, coherence, and overall stylistic plausibility. The inspection was not conducted as a formal user study: no predefined rating scale, rater training, or inter-annotator agreement analysis was used. Consequently, we treat this component as an illustrative qualitative check rather than a quantitative evaluation, and we rely on the automatic metrics.

[3] In mini bar graphs: reference value – narrow grey bar, positive value – blue bar (bright if larger than reference, dark otherwise), negative value – red bar.

**Table 1.** Evaluation metrics (excluding the stylometric features shown in Tab. 2). In mini bar graphs: relative value – the largest one before tuning; see also footnote 3.

| Model | Gemma 3 | GPT-2 Medium | Qwen1.5 |
|---|---|---|---|
| **a Perplexity** (smaller is better) | | | $\geq 0$ |
| Before fine-tuning | 79.66 | 61.50 | 149.71 |
| After fine-tuning | 76.98 | 51.24 | 25.79 |
| Relative change | -3.73% | -16.67% | -82.78% |
| **b Semantic similarity** (moderate is better) | | | $\in [0, 1]$ |
| Before fine-tuning | 0.575 | 0.519 | 0.579 |
| After fine-tuning | 0.567 | 0.528 | 0.620 |
| Relative change | -1.39% | +1.73% | +7.08% |
| **c Style classifier** (larger is better) | | | $\in [0, 1]$ |
| Before fine-tuning | 0.672 | 0.655 | 0.591 |
| After fine-tuning | 0.686 | 0.649 | 0.690 |
| Relative change | +2.08% | -0.92% | +16.75% |
| **d Stylometric distance** (smaller is better) | | | $\in [0, 1]$ |
| Before fine-tuning | 0.049 | 0.023 | 0.023 |
| After fine-tuning | 0.035 | 0.034 | 0.006 |
| Relative change | -28.6% | +47.8% | -73.9% |
| **e Idiosyncratic phrase matching** (larger is better) | | | $\in [0, 1]$ |
| Before fine-tuning | +0.016 | +0.013 | 0.007 |
| After fine-tuning | +0.016 | +0.013 | 0.017 |
| Relative change | 0.0% | 0.0% | 142.9% |

### 5.2 Perplexity

Perplexity values before and after fine-tuning are presented in Table 1, part a. Fine-tuning led to a reduction in perplexity for all three models, although the magnitude of improvement varied substantially.

Gemma 3 showed a modest decrease (approx. 3%), while GPT-2 Medium achieved a more pronounced improvement (about 17%). The most significant change was observed for Qwen1.5, whose perplexity decreased by more than 80%, reaching the lowest absolute value among all evaluated models.

These results demonstrate that fine-tuning substantially improved predictive confidence, particularly for Qwen1.5. However, perplexity primarily reflects internal language modelling performance and fluency rather than stylistic specificity. Therefore, while the reduction confirms successful adaptation to the training corpus, it does not by itself imply the acquisition of distinctive authorial traits.

### 5.3 Semantic Similarity

Semantic similarity results are shown in Table 1b. Overall similarity values remained within a relatively narrow range (approximately 0.52–0.62), indicating moderate thematic alignment between generated texts and the reference corpus.

Fine-tuning produced only minor changes for Gemma 3 and GPT-2 Medium. In contrast, Qwen1.5 exhibited a noticeable increase in similarity (about 7%), reaching the highest post-adaptation value among the evaluated models.

The relatively stable similarity scores suggest that even pretrained models were able to generate thematically appropriate academic continuations. Since prompts were derived from authentic introductions, semantic coherence was largely preserved regardless of fine-tuning. Consequently, improvements in semantic similarity appear less informative for assessing author-specific style and instead reflect topic continuity within the academic domain.

### 5.4 Style Classifier

The results of the TF–IDF-based style classifier are reported in Table 1c. The classifier outputs represent the estimated probability that a given text belongs to the target author's style.

Before fine-tuning, all three pretrained models already received relatively high attribution scores (approximately 0.59–0.67), suggesting that their baseline generations shared general characteristics with the reference corpus. This may reflect the strongly conventional nature of academic English.

After fine-tuning, Gemma 3 exhibited a small improvement (about 2%), while GPT-2 Medium showed a slight decrease. The most substantial change was again observed for Qwen1.5, whose attribution probability increased by nearly 17%, achieving the highest post-adaptation score among the models.

These findings indicate that fine-tuning can strengthen classifier-based stylistic alignment, particularly for Qwen1.5. However, the relatively high baseline scores and moderate final probabilities (around 0.69 at most) suggest that the classifier captures general stylistic regularities rather than highly distinctive author-specific features.

### 5.5 Stylometric Distance

Aggregated stylometric distances are reported in Table 1d. All values, both before and after fine-tuning, are relatively close to zero, indicating high similarity between the generated texts and the reference-style vector.

Before fine-tuning, GPT-2 Medium and Qwen1.5 already exhibited very low distances (approximately 0.023), while Gemma 3 showed a slightly higher value (0.049). After adaptation, stylometric distance decreased for Gemma 3 and Qwen1.5, with Qwen1.5 achieving the lowest overall value (0.006). In contrast, GPT-2 Medium showed a moderate increase in distance after fine-tuning.

The consistently low baseline distances suggest that pretrained models already generate text structurally similar to academic prose. Fine-tuning further

reduced this distance in two cases, but the absolute magnitude of change remains small. This can be interpreted as that aggregated stylometric features capture broad genre-level regularities that are largely present even without adaptation.

Therefore, while stylometric distance confirms incremental alignment after fine-tuning, it also highlights a limitation: individual authorial style may not be strongly distinguishable from general academic writing when measured using coarse structural features.

### 5.6 Idiosyncratic Phrase Matching

Results for idiosyncratic phrase matching are presented in Table 1e. Overall overlap values are low, ranging approximately between 1% and 2% of n-grams.

Fine-tuning did not change the overlap for Gemma 3 or GPT-2 Medium. In the case of Qwen1.5, phrase overlap increased substantially in relative terms, but the absolute value remained below 2%.

These findings mean that the models did not meaningfully acquire distinctive lexical patterns characteristic of the target author. The low overlap suggests that fine-tuning primarily influenced general structural and lexical tendencies rather than specific recurring expressions.

It should also be noted that academic prose may inherently contain fewer highly idiosyncratic phrases compared to more expressive genres. Nevertheless, the minimal change observed across models supports the conclusion that author-specific phrase usage was not strongly internalized during adaptation.

### 5.7 Stylometric Features

Stylometric feature values for the generated texts are presented and compared against the mean reference values in Table 2. Fine-tuning generally reduced the distance between generated and reference feature values, although the extent of improvement varied across models and features.

**Proportion of long words** Before fine-tuning, Qwen1.5 produced values closest to the reference corpus. After adaptation, all models moved closer to the target value, with GPT-2 Medium achieving the smallest post-tuning deviation. This suggests that lexical-level adjustments can be partially captured through fine-tuning.

**Average sentence length** Gemma 3 initially generated sentence lengths closest to the reference mean. Fine-tuning slightly worsened its alignment, while GPT-2 Medium improved and became the closest match. Qwen1.5 moved further away from the reference value. This indicates that sentence-level structural properties may not consistently benefit from adaptation.

**Lexical richness** GPT-2 Medium showed the closest alignment before fine-tuning. After adaptation, Qwen1.5 exhibited the largest shift toward the reference value, although in some cases the adjustment slightly overshot the target. Gemma 3 showed modest improvement. These results suggest that vocabulary diversity is moderately sensitive to fine-tuning.

**Table 2.** Comparison of stylometric features. On mini bar graphs see footnote 3.

| Model | Gemma 3 | GPT-2 Medium | Qwen1.5 |
|---|---|---|---|
| **Proportion of long words** (closer to reference is better) | | | Reference: 0.314 |
| Before fine-tuning | 0.217 | 0.238 | 0.335 |
| Absolute difference w.r.t. reference | −0.097 | −0.076 | +0.021 |
| After fine-tuning | 0.258 | 0.308 | 0.325 |
| Absolute difference w.r.t. reference | −0.056 | −0.006 | +0.011 |
| **Average sentence length** (closer to reference is better) | | | Reference: 0.707 |
| Before fine-tuning | 0.711 | 0.644 | 0.738 |
| Absolute difference w.r.t. reference | +0.004 | −0.063 | +0.031 |
| After fine-tuning | 0.666 | 0.700 | 0.753 |
| Absolute difference w.r.t. reference | −0.041 | −0.007 | +0.046 |
| **Lexical richness** (closer to reference is better) | | | Reference: 0.536 |
| Before fine-tuning | 0.327 | 0.455 | 0.639 |
| Absolute difference w.r.t. reference | −0.209 | −0.081 | +0.103 |
| After fine-tuning | 0.375 | 0.418 | 0.504 |
| Absolute difference w.r.t. reference | −0.161 | −0.118 | −0.032 |
| **Readability** (closer to reference is better) | | | Reference: 0.386 |
| Before fine-tuning | 0.535 | 0.554 | 0.288 |
| Absolute difference w.r.t. reference | +0.167 | +0.186 | −0.080 |
| After fine-tuning | 0.480 | 0.397 | 0.352 |
| Absolute difference w.r.t. reference | +0.112 | +0.029 | −0.016 |

**Readability (Flesch index)** All three models improved their readability alignment after fine-tuning. Qwen1.5 remained the closest to the reference value both before and after adaptation. From this it follows that global readability characteristics can be adjusted through training on domain-specific data.

Overall, stylometric feature analysis shows incremental improvements across several dimensions, but no consistent dominance of a single model across all features can be seen. The observed changes are relatively small in magnitude, suggesting partial stylistic adjustment rather than strong acquisition of distinctive authorial traits.

### 5.8 Summary of Results

Across all evaluated metrics, fine-tuning consistently improved language modelling performance and moderately strengthened stylistic alignment indicators. The most substantial quantitative gains were observed in perplexity, particularly for Qwen1.5, indicating effective adaptation to the training corpus.

However, improvements in stylistic measures were generally incremental. Semantic similarity remained relatively stable, classifier-based attribution increased only moderately, and stylometric feature adjustments were small in magnitude. Aggregated stylometric distance was already low before fine-tuning, suggesting strong baseline conformity to academic writing conventions. Finally, idiosyncratic phrase overlap remained minimal, indicating limited acquisition of distinctive lexical patterns.

Taken together, the results suggest that fine-tuning enhances domain adaptation more clearly than it enables imitation of highly specific authorial style.

## 6 Discussion

### 6.1 Differential Effects of Fine-Tuning

The results demonstrate that fine-tuning affected different metric categories unevenly. The strongest improvements were observed in perplexity, indicating enhanced predictive confidence and internal adaptation to the training corpus. In contrast, stylistic indicators showed only modest shifts.

This divergence suggests that fine-tuning primarily reinforced domain-specific language modelling rather than deeply restructuring stylistic behaviour. Improvements in fluency and topic coherence were more pronounced than changes in structural or lexical individuality.

### 6.2 Model Sensitivity and Parameter Scale

Among the evaluated models, Qwen1.5 exhibited the most substantial changes across several metrics. It achieved the largest reduction in perplexity and the most notable increase in classifier-based style attribution.

This pattern may suggest greater plasticity of the model under fine-tuning, potentially due to architectural or training differences. In contrast, the GPT-2 Medium model showed mixed behaviour, including a slight deterioration in stylometric distance despite improvements in perplexity. Gemma 3 demonstrated consistent but modest improvements.

These findings suggest that moderate-sized models differ in their responsiveness to small, domain-specific corpora.

### 6.3 Genre-Level Regularities vs Individual Style

A notable observation is the consistently low stylometric distance even before fine-tuning. This indicates that pretrained models already generate text structurally similar to academic prose. Consequently, distinguishing individual authorial style from broader genre conventions becomes challenging.

The limited increase in idiosyncratic phrase overlap further supports this interpretation. While global structural features can be approximated through adaptation, highly specific lexical patterns appear resistant to learning from relatively small corpora.

This suggests that, within strongly conventionalized genres such as scientific introductions, individual stylistic markers may be subtle and partially overshadowed by shared structural norms.

### 6.4 Limitations

Several limitations should be acknowledged. First, the training corpus consisted of only 93 introductions, which may be insufficient for robust acquisition of fine-grained stylistic traits. Second, collaborative authorship may reduce stylistic purity of the reference corpus. Third, evaluation relied on a limited set of stylometric features and a simple TF–IDF-based classifier, which may not capture deeper stylistic dimensions.

Finally, only moderate-sized models ($\leq$500M parameters) were considered. Larger architectures may exhibit different adaptation dynamics.

## 7 Conclusion

This study examined whether moderate-sized causal language models can reproduce the individual academic writing style of a specific author through fine-tuning on a limited corpus of scientific introductions. Three open-access models were evaluated before and after adaptation using a combination of language modelling, semantic, classifier-based, and stylometric metrics.

Fine-tuning consistently improved perplexity and moderately strengthened classifier-based stylistic attribution. However, changes in stylometric features and phrase-level overlap were limited in magnitude. In particular, idiosyncratic phrase usage remained largely unaffected, and aggregated stylometric distance was already low prior to adaptation.

These findings suggest that fine-tuning on small domain-specific corpora primarily enhances genre-level conformity rather than enabling robust imitation of distinctive authorial traits. Within highly conventionalised genres such as academic prose, individual stylistic markers may be subtle and difficult to isolate using standard evaluation metrics.

Future work may explore larger corpora, alternative stylistic representations, and larger model architectures to further investigate the limits of author-specific style adaptation.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bacco, L., Dell'Orletta, F., Lai, H., Merone, M., Nissim, M.: A text style transfer system for reducing the physician–patient expertise gap: An analysis with automatic and human evaluations. Expert Systems with Applications **233**, 120874 (2023). https://doi.org/10.1016/j.eswa.2023.120874

2. Google: gemma-3-270m. Hugging Face (2025), https://huggingface.co/google/gemma-3-270m, [Accessed: 09 Dec 2025]
3. Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., Gao, C.: A survey on the techniques, applications, and performance of short text semantic similarity. Concurrency and Computation: Practice and Experience **33**(5), e5971 (2021). https://doi.org/10.1002/cpe.5971
4. Hinc, A., Krajewski, J.K., Chmielewski, L.J., Orłowski, A.: Images for the paper (submitted to PP-RAI 2026): Can moderate-sized language models imitate an individual academic writing style? (2026), https://www.lchmiel.pl/writingstyle/, [Updated Mar 2026]
5. Hinc, A.: Analysis of the ability of artificial intelligence language models to imitate individual writing style. M.Sc. Thesis, Warsaw University of Life Sciences – SGGW, Faculty of Applied Informatics and Mathematics – WZIM (2026)
6. Janowicz, M., Orłowski, A.: A four-level "$\Xi$-system": Exact solvability and effective evolution operators via multiple scales. Reports on Mathematical Physics **54**(1), 57–69 (2004). https://doi.org/10.1016/S0034-4877(04)80005-3
7. Krishna, K., Wieting, J., Iyyer, M.: Reformulating unsupervised style transfer as paraphrase generation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP). pp. 737–762 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.55
8. Mukherjee, S., Dušek, O.: Text style transfer: An introductory overview. arXiv, arXiv.2407.14822 (2024). https://doi.org/10.48550/arXiv.2407.14822
9. Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., Woodard, D.: Surveying stylometry techniques and applications. ACM Computing Surveys **50**(6), 86 (2017). https://doi.org/10.1145/3132039
10. OpenAI: GPT-2: 1.5B release. OpenAI (2019), https://openai.com/index/gpt-2-1-5b-release/, [Accessed: 14 Dec 2025]
11. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. In: Proc. 55th Annual Meeting of the Association for Computational Linguistics. vol. 1: Long Papers, pp. 1756–1765 (2017). https://doi.org/10.18653/v1/P17-1161
12. QwenTeam: Introducing Qwen1.5. Qwen (2024), https://qwen.ai/blog?id=qwen1.5
13. Stanikūnas, D., Mandravickaitė, J., Krilavičius, T.: Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts. In: Proc. CEUR Workshop, Symp. for Young Researchers in Informatics, Mathematics and Engineering (ICYRIME) (2017), https://hdl.handle.net/20.500.12259/36092
14. Xie, Y., Gui, J., Che, Z., et al.: TSTBench: A comprehensive benchmark for text style transfer. Entropy **27**(6) (2025). https://doi.org/10.3390/e27060575
15. Xu, Z., Gupta, A., Li, T., Bentham, O., Srikumar, V.: Beyond perplexity: Multi-dimensional safety evaluation of LLM compression. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 15359–15396 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.901
16. Zhu, J., Jurgens, D.: Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In: Proc. 2021 Conf. Empirical Methods in Natural Language Processing. pp. 279–297 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.25