Who Wrote This Review?
Evaluating Authorial Voice in Real vs. GPT-2 Generated Music Reviews

Doruk Efe Kanber
Student ID: 3163051
Milan, Italy

Tulga Kağan Temel
Student ID: 3168178
Milan, Italy

Eren Karsavuranoğlu
Student ID: 3164647
Milan, Italy

Abstract

Generative language models are now used in all kinds of writing. Tools such as ChatGPT and Gemini have become common in everyday communication. This trend raises concerns about whether they can preserve individual writing styles. As their use grows, it is unclear whether these models capture deeper stylistic traits. Fine-tuning on individual authors enables the model to learn and reflect deeper linguistic patterns. We test this approach by fine-tuning GPT-2 models on music reviews written by different Pitchfork users and generating synthetic texts. We train authorship classifiers on the original reviews and use them to evaluate whether the generated texts preserve the author's voice. While classifiers perform well on original reviews, they often predict the same author for generated reviews. The results suggest that the synthetic reviews sound similar across authors despite fine-tuning. This finding highlights a limitation in GPT-2 model.

1 Introduction

Personal voice defines good writing. Readers can easily spot stylometric differences between a WhatsApp chat and an academic paper. Large language models now produce thousands of articles, reviews, and social-media posts every day (Reuters Institute, 2024). Editors fine-tune ChatGPT to "match" a house style. Machine-assisted texts mixes the boundary between an author's own style and a model's default voice, especially in terms of word choice.

This loss of personal voice is not just an aesthetic issue. It has measurable linguistic and social consequences. Hovy (2015) showed that demographic factors like age and gender significantly affect writing style, influencing word choice, syntax, and even semantics. Models trained without considering these differences performed worse across sentiment, topic, and author classification tasks. When models try to generalize, they often lose the stylistic details that make writing feel personal and socially distinct. As large language models become central to written communication, understanding how much of that individuality they preserve becomes essential.

To address this, we ask a clear question: can a language model reflect a writer's style after it has been fine-tuned on their past work? Answering this gives us a way to measure how much personal voice survives in generated text. It also helps us understand where and how style changes. If style stays strong, we can trust models to assist writing without taking over the author's voice. If style fades, we know to be more careful. Either way, it gives editors, developers, and researchers a better view of how generative tools affect writing.

Our Contribution

We introduce a pipeline to test whether large language models preserve author style after fine-tuning. Our approach employs an authorship classifier as a specific functional test to evaluate the stylistic fidelity of generated texts, drawing inspiration from broader methodologies for targeted model evaluation in NLP (e.g., Röttger et al., 2021).

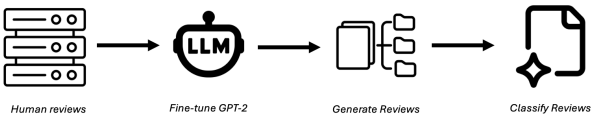

Figure 1: Pipeline overview

We provide a dataset of 5,054 music reviews written by ten different *Pitchfork* authors. We choose GPT-2 as our base model because it is

lightweight, publicly available, and easy to fine-tune with limited computation time. We fine-tune GPT-2 separately for each author and generate new reviews to mimic their style. To evaluate stylistic similarity, we train a linear Support Vector Machine using word-level features extracted with a TF-IDF model. This model reaches 98.2% accuracy on held-out test data, confirming that word-level signals are strong indicators of author identity. However, when tested on generated reviews, accuracy drops to 32.5%. Most misclassifications cluster around a few prominent authorial styles, with one style (that of Marc Hogan) being particularly overrepresented in the generated texts. These results suggest that our fine-tuned GPT-2 models generate fluent content, but they fail to keep many of the stylistic features found in human-written reviews.

Our results show the limits of small models in keeping personal style. The setup we use can be reused to test bigger models.

2 Experiments

2.1 Dataset

We use the Pitchfork Music Reviews dataset (a publicly available collection of editorial reviews written by music critics). We examined the number of reviews per author to identify those with the most contributions. The dataset initially includes reviews from hundreds of authors with a highly skewed distribution, totaling 18,393 reviews. To build a balanced classification task, we selected the top 10 authors with most reviews. This reduced our dataset to 5,054 reviews.

Preprocessing

We applied a minimal cleaning pipeline: (i) HTML entities are un-escaped, (ii) runs of whitespace collapsed to a single space, and (iii) curly quotes normalised to straight ASCII quotes. We deliberately keep case, punctuation, and numbers intact because these surface cues are crucial for stylometric classification. No stemming, lemmatisation, or stop-word removal is applied. After filtering the 10 target authors, the dataset was partitioned into an 80% training set and a 20% held-out test set, stratified by author. Hyperparameter tuning for the classifier was performed using 5-fold stratified cross-validation on the training set.

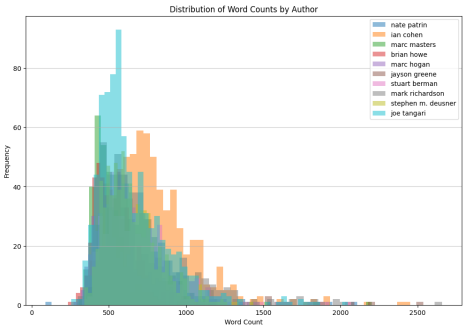

Figure 2: Distribution of review word counts by author. Most reviews fall between 400 and 800 words, but styles vary across authors.

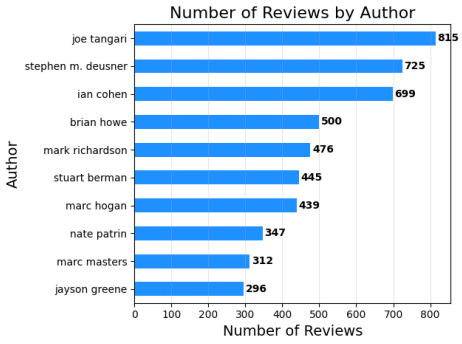

Figure 3: Number of reviews per author. The dataset is curated to include the top 10 authors by volume.

Descriptive Statistics

To better understand the characteristics of our dataset, we conducted exploratory analysis. Figure 3 shows the number of reviews written by each selected author. While the dataset was selected to be balanced, natural variation remains, with Joe Tangari contributing the most reviews (815) and Jayson Greene the fewest (296). Figure 2 displays the distribution of word counts per review for each author. Most reviews fall between 400 and 800 words, but the distribution reveals clear stylistic differences. For example, reviews by Ian Cohen tend to be longer on average, whereas Marc Masters' reviews are typically shorter.

The final dataset contains 5,054 reviews across 10 authors. On average, each review contains 661 words, yielding an estimated 3.2 million tokens overall. After cleaning, the vocabulary consists of approximately 29,000 unique tokens.

2.2 Synthetic Review Generation

To evaluate whether GPT-2 can capture individual writing styles, we fine-tuned a separate GPT-2

model for each of the 10 selected authors.

We used the `transformers` library from Hugging Face for both training and generation. Each model is based on the small "gpt2" checkpoint and fine-tuned using that author's review texts. We included author-specific tokens (e.g., <|AUTHOR_0|>) and a <|REVIEW_START|> marker in the tokenizer vocabulary to condition generation, aiming to help the model distinguish between authors and align generations with the intended writing voice. We also experimented with one-shot anchoring by priming the model with a real review before generation, but this approach resulted in less coherent and stylistically inconsistent outputs.

Each model was fine-tuned on a GPU-enabled environment with early stopping to prevent overfitting. For every author, we generated 20 synthetic reviews using nucleus sampling with the following decoding parameters: `top_k=50`, `top_p=0.92`, `temperature=0.7`, and a maximum length of 512 tokens. The input prompt for generation combined the author token and the review start marker.

This process yielded a total of 200 synthetic reviews (20 per author), which we later used to assess whether stylistic signals are preserved and distinguishable from real human-written reviews.

2.3 Author Attribution

We approach author identification as a multi-class classification problem, where each review is assigned to one of ten authors based on its writing style. The field of stylometry has long established that authors can be distinguished by patterns in their textual features (e.g., Argamon et al., 2007). To represent these stylistic patterns, we experiment with several feature types:

- Word-level TF-IDF (1–2 grams)
- Character-level n-gram counts (3–5 grams)
- Character-level TF-IDF (3–5 grams)
- Sentence embeddings from MiniLM

These features serve as input to multiple classifiers, including `LinearSVM`, `LogisticRegression`, and `NaiveBayes`. Across all setups, we intentionally avoid semantic or metadata features. Our goal is to isolate and model stylistic differences visible at the surface level of the text.

The dataset is split into training, validation, and test sets. All models are trained exclusively on real reviews written by human authors. Evaluation on synthetic reviews is deferred to a later section. For development and tuning, we use 5-fold stratified cross-validation to ensure class balance and a fair assessment of each author's representation.

Among all configurations, the best performance is achieved by a `LinearSVM` trained on TF-IDF word features, reaching an average accuracy and F1-Macro score of 98.2% on a held out test data with a 20% split.

To further explore the learned stylistic patterns, we extract the top 10 most predictive n-grams for each author (Figure 4). These terms reveal the lexical signatures that differentiate writing styles from structural choices like "there is" to genre-linked phrases like "the album" or "dubstep."

Figure 4: Top 10 predictive word n-grams per author based on feature importance in the linear SVM.

We also visualize the TF-IDF representation of reviews using t-SNE (Figure 5). Distinct clusters emerge per author, suggesting strong stylistic separability even without contextual modeling.

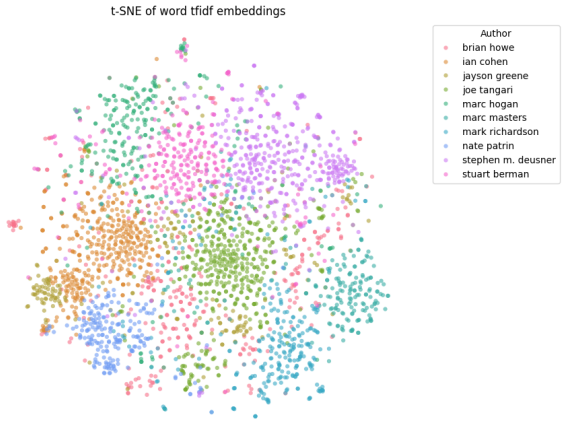

Figure 5: t-SNE visualization of review embeddings using word-level TF-IDF vectors. Each point represents one review.

2.4 Evaluation Metrics

We use two main metrics to evaluate author attribution: accuracy and macro-averaged F1-score. Accuracy measures how often the model predicts the correct author. Macro F1 gives equal weight to each author by averaging F1-scores across all classes, helping account for slight differences in review counts.

We report these metrics on both real and synthetic reviews. The model is trained only on real reviews, and tested separately on both sets.

To better understand model behavior, we include a confusion matrix (Figure 6) that shows how often the model confuses one author with another. These patterns help us interpret stylistic overlap and classification boundaries.

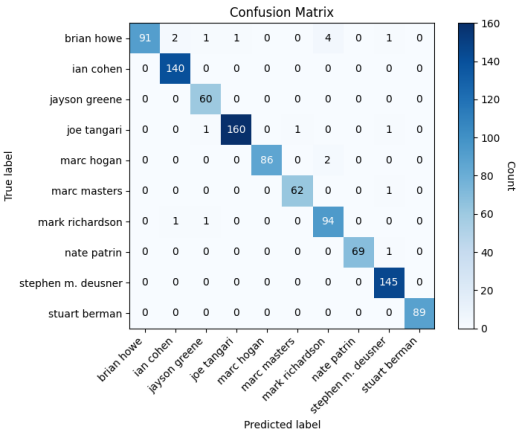

Figure 6: Confusion matrix for author classification on real reviews using n-grams.

3 Results

The classifier reaches an accuracy of 32.5% and a macro-averaged F1 score of 25.3% on GPT-2-generated reviews. Compared to the performance on original reviews, this marks a clear decline, indicating that much of the original writing style is lost during generation. (Figure 7) shows that predictions are heavily skewed toward a few authors. Marc Hogan, in particular, accounts for a large share of the classifications, receiving misattributions from Brian Howe (17), Jayson Greene (15), and Stuart Berman (14), among others. While some individual voices—like Ian Cohen with 17 correct predictions—still show traces of distinctiveness, many others are frequently confused or absorbed into dominant clusters. For instance, reviews generated for Nate Patrin and Stuart Berman are never identified correctly. Overall, these results suggest that the fine-tuned model captures surface-level fluency but fails to reproduce the full stylistic range of each author.

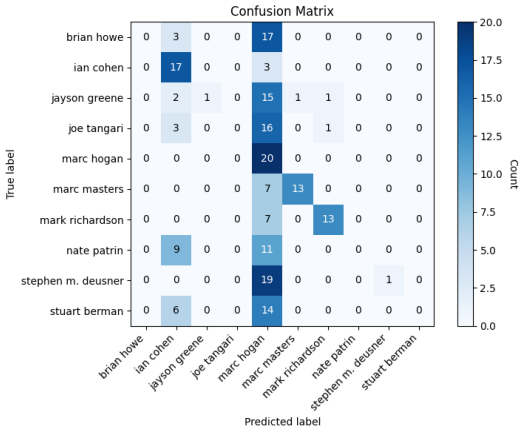

Figure 7: Confusion matrix for author classification on generated reviews using n-grams.

4 Discussions

Our results show that while GPT-2 can generate fluent and coherent reviews after per-author fine-tuning, it often struggles to preserve each writer's unique style. The classifier accuracy on generated reviews drops to 32.5%, down from 98.2% on real ones, and macro F1 decreases from 98.2% to 25.3%. These shifts suggest that although the model captures genre and topic well, finer stylistic details are more difficult to retain.

This outcome is informative. It reflects the capacity limits of small-scale generative models in

stylometric tasks and highlights the shift in distribution between real and synthetic prose. At the same time, it provides a foundation for evaluating and improving stylistic control in future text generation work.

A large portion of the generated reviews were classified as written by Marc Hogan. This consistent misattribution suggests that the generated texts resemble Hogan's original writing more than others. One possible explanation is that GPT-2's pretraining data may already align more closely with Hogan's style. As an American journalist and Northwestern graduate, Hogan's writing may reflect a more "standardized" or widely accepted form of English. If his linguistic patterns are closer to what the model has seen during pretraining, the generated outputs may naturally converge toward that style, even after fine-tuning on other authors. This observation raises broader questions about stylistic bias in generative models and which voices are more easily reproduced.

These patterns point to interesting directions for future work. Our setup was limited to 10 authors with balanced review counts. Further experiments could explore whether these findings generalize to larger and more diverse sets, and whether results hold under different generation parameters or fine-tuning techniques.

Despite the limitations of using GPT-2 (124M parameters) and standard decoding strategies, our work introduces a reusable evaluation pipeline that pairs generation with stylometric classification. This framework can be applied to larger models, such as GPT-3 or LLaMA,. It also opens the door to decoding strategies that better preserve stylistic identity, such as classifier-guided generation or style-constrained sampling.

In short, per-author fine-tuning of GPT-2 produces readable and thematically appropriate text, but stylistic consistency remains a challenge. Addressing this will likely require stronger models, clearer style objectives, and more targeted decoding methods. Our study helps define that challenge and provides tools for tackling it.

5 Related Works

Our work is part of a broader effort to explore how language models can reflect the writing style of individual authors. Syed et al. (2020) fine-tune models using denoising autoencoders to rewrite input in a target author's voice, showing that stylistic transfer is possible without parallel data. While their work focuses on rewriting, we study style reproduction in text generated from scratch. Liu et al. (2024) apply parameter-efficient fine-tuning to LLaMA-2 and find that even small updates can shift generation toward a writer's idiolect. Our classification-based evaluation offers a way to measure whether such shifts actually preserve authorial identity. Khan et al. (2023) take a more controlled approach by learning continuous style embeddings for long-range consistency. In contrast, we test how far basic fine-tuning can go without style-specific constraints. Alvero et al. (2024) show that LLMs tend to produce writing aligned with socially dominant groups, raising concerns about style homogenization. We observe a similar collapse in our confusion matrix, where generated reviews cluster around just a few author labels. Finally, McCoy et al. (2023) find that GPT-2 often reuses local patterns from training data, even when global structure is novel. Our findings build on this by showing that such reuse may lead to repeated stylistic signals across authors.

6 Conclusion

This paper tested whether GPT-2 can generate text that preserves the writing style of individual authors. We fine-tuned separate models for ten Pitchfork reviewers and evaluated the output using a classifier trained on real reviews. The classifier performed well on human-written text but struggled with the generated reviews. Most generated texts were misclassified as coming from the same few authors.

These results show that while GPT-2 can produce fluent and coherent text, it often fails to retain the distinct stylistic features of each author. One possible reason is that GPT-2's original training data consists of general internet text. Some writing styles may be overrepresented. This could explain why certain authors like Marc Hogan dominate the predictions.

Our main contribution is a simple and reusable pipeline for testing stylistic preservation in generated text. It combines author-specific generation with classification-based evaluation. Future work could apply this setup to larger models like GPT-3 or LLaMA, or explore fine-tuning methods such as LoRA.

Acknowledgments

We thank Dirk Hovy for his guidance throughout the course and Paul Röttger for his valuable feedback and support during this project

References

[Argamon et al., 2007] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822. doi:10.1002/asi.20553.

[Liu et al., 2024] Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024. Customizing large language model generation style using parameter-efficient finetuning. In *Proceedings of the 17th International Natural Language Generation Conference (INLG)*, pages 412–426. Association for Computational Linguistics.

[McCoy et al., 2023] R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670. doi:10.1162/tacl_a_00567.

[Khan et al., 2023] Aleem Khan, Andrew Wang, Sophia Hager, and Nicholas Andrews. 2023. Learning to generate text in arbitrary writing styles. *arXiv preprint* arXiv:2312.17242.

[Alvero et al., 2024] A. J. Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F. Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text. *Journal of Big Data*, 11:138. doi:10.1186/s40537-024-00986-7.

[Syed et al., 2020] Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9008–9015.

[Röttger et al., 2021] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 41–58.

[Hovy, 2015] Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762. https://doi.org/10.3115/v1/P15-1073.

A Supplementary Figures

This appendix contains supplementary figures referenced in the main text or providing additional details about the experimental setup and results.

A.1 Author Attribution Model Selection

Table 8 shows the performance of different classifier and feature combinations explored for the author attribution task on the original human-written reviews.

Model	Vectorizer	Acc_Mean	Acc_Std	F1_Mean	F1_Std	Val_Acc	Val_F1
Linear SVM	Word TF-IDF (1-2)	0.989769	0.005463	0.988613	0.007063	0.987129	0.987059
Linear SVM	Char Count (3-5)	0.971947	0.007879	0.971520	0.008437	0.961386	0.959572
Logistic Regression	Word TF-IDF (1-2)	0.973267	0.005362	0.971883	0.006779	0.960396	0.958597
Logistic Regression	Char Count (3-5)	0.966007	0.010571	0.964826	0.011973	0.948515	0.946816
Naive Bayes	Char Count (3-5)	0.955446	0.009956	0.954343	0.010859	0.938614	0.935518
Logistic Regression	Char TF-IDF (3-5)	0.949175	0.013341	0.948478	0.013697	0.920792	0.920354
Logistic Regression	Mini-LM Embeddings	0.456436	0.026824	0.451974	0.025541	0.439604	0.427619

Figure 8: Performance comparison of various classifiers and feature sets for author attribution on real reviews. Best performance (Accuracy: 0.982, F1-Macro: 0.982) was achieved with Linear SVM using Word TF-IDF (1-2 grams).

A.2 Inter-Author Stylistic Distances

Figure 9 visualizes the mean pairwise distances between author clusters in the t-SNE projection of the word TF-IDF embeddings (as shown in Figure 5 in the main text).

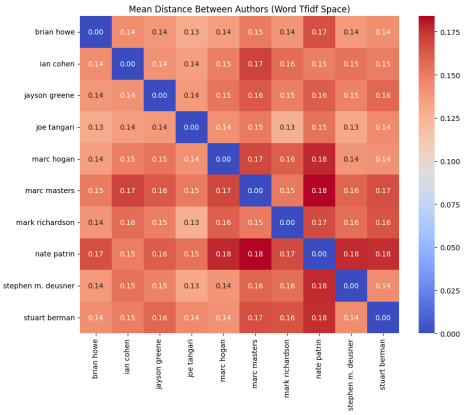

Figure 9: Heatmap of mean pairwise Euclidean distances between author centroids in the 2D t-SNE projection of word TF-IDF embeddings. The color scale ranges from blue (distance $\approx$ 0.00, higher similarity) to red (distance $\approx$ 0.175, lower similarity).