A Stylometric Application of Large Language Models

Harrison F. Stropkay, Jiayi Chen, Mohammad J. Latifi,
Daniel N. Rockmore, and Jeremy R. Manning
Dartmouth College
Hanover, NH 03755, USA
{harrison.f.stropkay.25, jiayi.chen.gr, mohammad.javad.latifi.jebelli
daniel.n.rockmore, jeremy.r.manning}@dartmouth.edu

October 28, 2025

Abstract

We show that large language models (LLMs) can be used to distinguish the writings of different authors. Specifically, an individual GPT-2 model, trained from scratch on the works of one author, will predict held-out text from that author more accurately than held-out text from other authors. We suggest that, in this way, a model trained on one author's works embodies the unique writing style of that author. We first demonstrate our approach on books written by eight different (known) authors. We also use this approach to confirm R. P. Thompson's authorship of the well-studied 15$^{\text{th}}$ book of the *Oz* series, originally attributed to F. L. Baum.

1 Introduction

Herein we introduce *predictive comparison*, a new LLM-based relative stylometric measure. It derives from a simple idea, that if an LLM can be trained to write like—i.e., in the

style of—a given author by training on their work (e.g., Mikros, 2025), then the degree to which such a model can predict another author's work could be a measure of stylistic similarity. This approach builds upon a growing body of work applying language models to authorship attribution (Huang et al., 2025; Uchendu et al., 2020), extending established information-theoretic methods in stylometry (Juola and Baayen, 2005; Zhao et al., 2006).

Recent work has demonstrated the effectiveness of using perplexity and cross-entropy loss from fine-tuned language models for authorship attribution (Huang et al., 2025), achieving state-of-the-art performance on standard benchmarks. Unlike traditional stylometric approaches that rely on the direct articulation of particular features such as function word frequencies (Mosteller and Wallace, 1963) or syntactic patterns (Holmes, 1998), large language models can capture complex, hierarchical patterns in authorial style (Fabien et al., 2020). This shift from explicit feature engineering to learned representations parallels broader trends in computational literary analysis (Moretti, 2000; Underwood, 2019) and digital humanities (Hughes et al., 2012).

In this paper we show, using a small set of authors and their works, that large language models capture author-specific writing patterns. Our method differs from related approaches (Rezaei, 2025) in scale (we use entire books rather than individual sentences) and in our reliance solely on cross-entropy loss as a measure of stylometric distance. This in turn suggests a notion of stylometric distance derived from the cross-entropy loss assigned to held-out texts by models trained on known works of different authors. We believe this approach could be of use in considering questions of authorial influence and stylistic evolution (Hughes et al., 2012). Lastly, this further suggests a literary attribution tool (a common use of stylometric techniques; Binongo, 2003; Juola, 2008; Mosteller and Wallace, 1963, 1984) that would assign an unknown or contested work to the model (and author) under which predictive comparison generates the smallest loss. We illustrate this

on the well-known attribution problem of the 15$^{\text{th}}$ book in the *Oz* series, confirming what is now the accepted attribution.

2 Methods

In this section, we outline our methodology for identifying stylometric signatures using large language models. For each selected author, we train a GPT-2 model (Radford et al., 2019) on that author's corpus. We then use the trained model to compute the cross-entropy loss on held-out texts from both the target author and each of the other authors in the dataset. By comparing these losses, we assess whether the model captures author-specific stylistic patterns: a model trained on a given author should exhibit lower loss when predicting that author's own texts as compared to the texts of others.

2.1 Data and preprocessing

We consider a dataset comprising books by eight authors: Jane Austen, L. Frank Baum, Charles Dickens, F. Scott Fitzgerald, Herman Melville, Rosemary Plumly Thompson, Mark Twain, and H. G. Wells. We selected these authors because their writings are well-represented in Project Gutenberg, are all in the public domain, and are written in English—eliminating any potential confounds due to translation. For each book, we pre-process the text by stripping Project Gutenberg metadata, publisher information, illustration tags, transcriber notes, prefaces, tables of contents, and chapter headings. We standardize whitespace, remove non-ASCII characters, and lowercase all alphabetic characters. Basic statistics on token lengths and the full list of books used are provided in the Appendix.

To construct training data for each author, we randomly select one book to hold out for evaluation and train their model using the remaining books. To ensure fair comparisons across authors, we standardize the number of training tokens per author by truncating

each author's corpus. This token budget is determined by removing the longest book from each author's set and then taking the smallest of the (remaining) total token counts. For our dataset, this yields a fixed training token budget of 643,041 tokens.

To construct a truncated corpus of 643,041 tokens for each author, we sample one contiguous sub-sequence from each book in their training corpus (after holding out a to-be-evaluated book). The length of the sub-sequence sampled from book i is proportional to its original length:

$$\text{length}_i = 643{,}041 \times \frac{\text{tokens in book } i}{\text{total tokens in corpus}}.$$

The starting position of each sub-sequence is chosen uniformly at random, ensuring the sample fits within the book's bounds. Finally, we shuffle and then concatenate the sampled sub-sequences from each book, resulting in a single 643,041-token training sequence for each author. This process is repeated for each of 10 random seeds, yielding 10 different training corpora for each author.

2.2 Model architecture, training, and evaluation

For each author, we train GPT-2 language models from scratch using the `GPT2LMHeadModel` class from the Hugging Face `Transformers` library with custom architecture settings: a context window of 1024 tokens, an embedding dimension of 128, 8 transformer layers, and 8 attention heads per layer. We fit each model using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-5} to minimize the cross-entropy loss on the training data. We train models using a causal language modeling objective, whereby the model iteratively predicts the next token in the sequence given all of the previous tokens in the same training sequence.

We construct training samples by sampling 1024-token chunks from the truncated

corpus for the given author and random seed (constructed as described above, using contiguous sub-sequences selected from all but one of their books). Each training epoch consists of 40 batches, each containing 16 sequences of 1024 tokens. This results in a total of 655,360 tokens per epoch. We continue training until the cross-entropy loss falls to 3.0 or lower. (We decided on this threshold after taking random draws from the models trained on Baum's and Thompson's *Oz* books and manually inspecting the quality of the resulting samples.) Training to a fixed loss threshold (e.g., as opposed to training for a fixed number of epochs) enables us to fairly compare model performance across authors, which is the central component of our stylometric analyses.

We evaluate the models using the held-out book from the corresponding author. We partition the held-out book into 1024-token chunks to ensure that each token in the evaluation set contributes equally to the computed loss. We repeat the full process (of selecting a held-out book at random and training the model using randomly selected samples from the remaining books) using 10 different random seeds. This approach enables us to assess the robustness of our results and to ensure that the models are not overfitting to a specific book or random sample.

2.3 Investigating the contributions of function words, content words, and parts of speech

In order to investigate the contributions of different types of words to the stylometric signatures captured by our models, we carried out additional analyses using modified corpora. First, we created content-word-only corpora by replacing all function words with a special token, <FUNC>. Function words were identified using scikit-learn's list of English stop words (Pedregosa et al., 2011). Next, we created function-word-only corpora by replacing all content (i.e., non-function) words with a <CONTENT> token. Finally, we

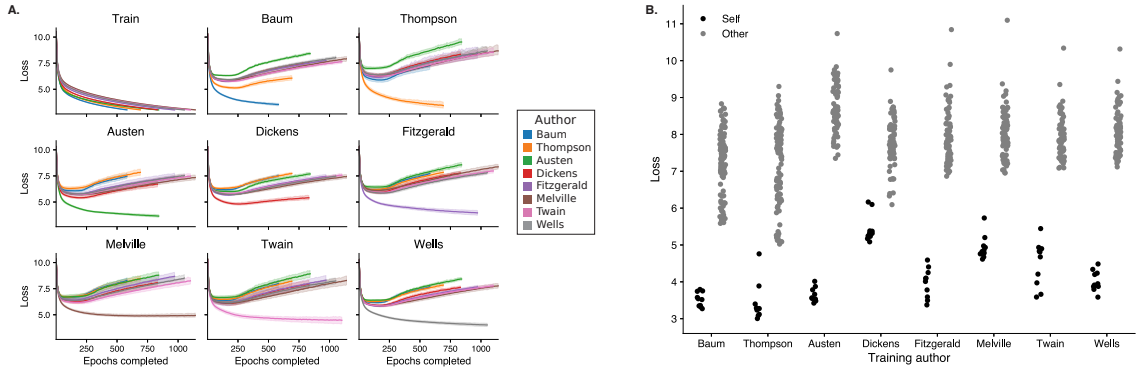


Figure 1: Cross-entropy loss across models and authors. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (x-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed. See Supplementary Materials for analogous plots using models trained on only content words (Supp. Fig. 1), only function words (Supp. Fig. 2), and only parts of speech (Supp. Fig. 3).

created part-of-speech-only corpora by using the Natural Language Toolkit (NLTK; Bird and Loper, 2004) to replace each word with its corresponding part-of-speech tag. We then re-trained our models on each of these modified corpora, following the same methodology as described above.

3 Results

3.1 Predictive comparison testing of eight classic authors

We carried out predictive comparison testing on eight classic authors (see Sec. 2.1). The top-left sub-panel of Figure 1A (labeled "Train") shows the average training loss for each author's model, computed over 10 random seeds. Training losses are comparable across models, indicating that the models are trained to similar levels of performance. The other sub-panels of Figure 1A show the average predictive (cross-entropy) loss, for each

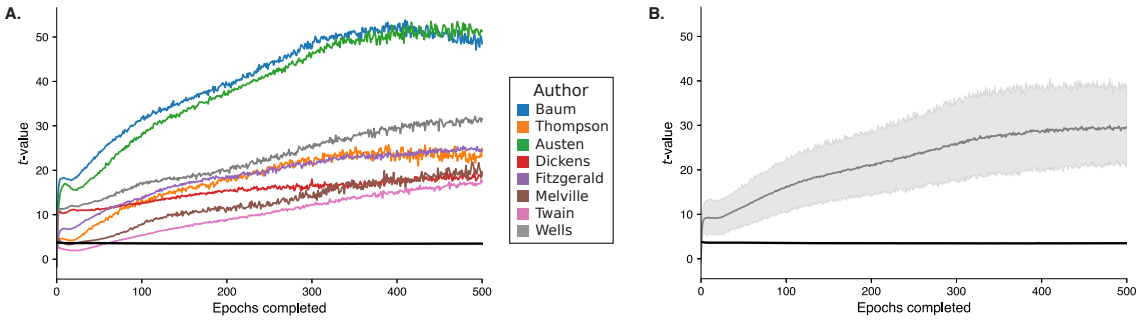


Figure 2: Same vs. other author comparisons, by model. A. Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors. See Supplementary Materials for analogous plots using models trained on only content words (Supp. Fig. 4), only function words (Supp. Fig. 5), and only parts of speech (Supp. Fig. 6).

author's model, on held-out texts from each author. For every author's held-out text, the model trained on the same author's writings produces the lowest loss, indicating a clear preference for its own author's stylistic patterns. As shown in Figure 1B, across every author we considered, and for every random seed, models trained and tested on the same author always yield smaller losses than models trained on one author and tested on another. Indeed, we achieve perfect (100%) classification accuracy when matching authors with held-out texts by labeling the held-out text according to which model produces the smallest loss.

We also wondered how many training epochs were required for the models to reliably distinguish author styles. We compared the distributions (across random seeds) of average cross-entropy losses for each author's model computed for held-out text from the *same* author versus for held-out text from *other* authors. Figure 2A displays the t-values from t-tests comparing these same versus other loss distributions for each of the first 500 training epochs. For all authors except Twain, the t-tests yielded p-values below 0.001 after just

one or two epochs, indicating that the models rapidly acquire author-specific stylometric patterns. For Twain, this threshold is crossed at epoch 77. Figure 2B shows the average t-values across all eight authors as a function of the number of training epochs (final epoch: $t(9) = 13.196, p = 3.41 \times 10^{-7}$). This latter plot provides an estimate of the performance we might expect to see in the general case (e.g., across a larger set of authors). Table 1 summarizes the results of the t-tests for each author's model after training is complete.

Model	t**-stat**	**df**	p**-value**
Baum	48.39	31.53	3.69×10^{-31}
Thompson	22.35	16.39	1.04×10^{-13}
Austen	50.64	47.38	6.48×10^{-43}
Dickens	16.37	17.84	3.46×10^{-12}
Fitzgerald	25.94	23.13	1.55×10^{-16}
Melville	23.38	23.13	1.35×10^{-17}
Twain	16.74	11.27	2.60×10^{-9}
Wells	35.73	23.68	4.15×10^{-22}

Table 1: Loss differences between same-author and other-author texts. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts. See Supplementary Materials for analogous tables using models trained on only content words (Supp. Tab. 1), only function words (Supp. Tab. 2), and only parts of speech (Supp. Tab. 3).

Despite achieving perfect classification accuracy, not all authors are equally distinctive. For example, we reasoned that authors with similar writing styles might be more confusable (i.e., yielding relatively smaller losses for models trained across different authors). We computed the average loss for each author using the models trained on the other authors' texts (Fig. 3). Authors with similar writing styles (e.g., Baum and Thompson) yield relatively small losses when evaluated using models trained on the other author's texts. In contrast, authors with more distinct writing styles (e.g., Austen and Thompson) yield relatively large losses when evaluated using each other's models. To illustrate these patterns, we also project the losses into a 3D space using multidimensional scaling (MDS; Kruskal, 1964) applied to the pairwise correlations between rows of the loss matrix, ex-

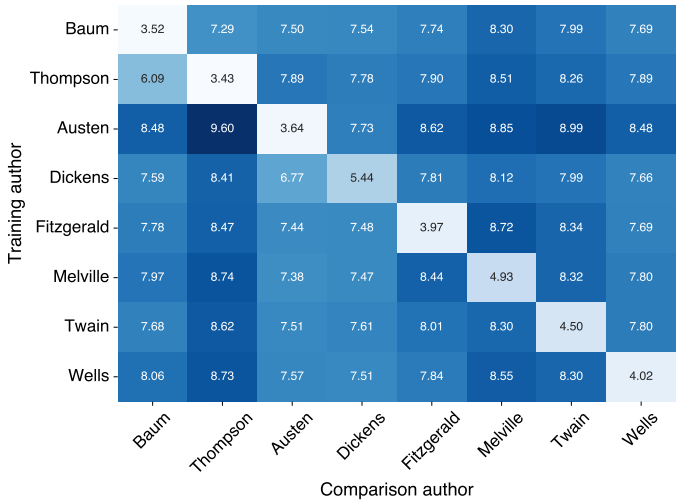


Figure 3: Confusion matrix. The matrix displays the average cross-entropy loss assigned by models trained on each author's writing (column) to held-out texts from each author (row), after subtracting the native author's baseline loss. See Supplementary Materials for analogous plots using models trained on only content words, function words, and parts of speech (Supp. Fig. 7).

cluding the diagonal entries (i.e., the losses obtained using each author's model when applied to their own held-out text; Fig. 4). We suggest that this approach might lend itself to further exploration and consideration by literature scholars, particularly if extended to a larger embedding space. For the purposes of our present work, however, we provide the plot solely as a provocative demonstration.

3.2 Stylometric distance

As indicated by Figure 4, predictive comparison suggests a natural notion of distance between authorial styles. Let $L_j(i)$ denote the average loss of a work of author i for a model trained on author j (entry i, j of the average loss matrix in Fig. 3). Let $\overline{L_j(i)} = L_j(i) - L_j(j)$, normalizing the entries by subtracting the native author's baseline loss. Then define the LLM-based *stylometric distance*, $d(i, j) = \frac{1}{2}\left(\overline{L_j(i)} + \overline{L_i(j)}\right)$. Thus, Figure 4 is a visualization of the relative "distances" among our author set.

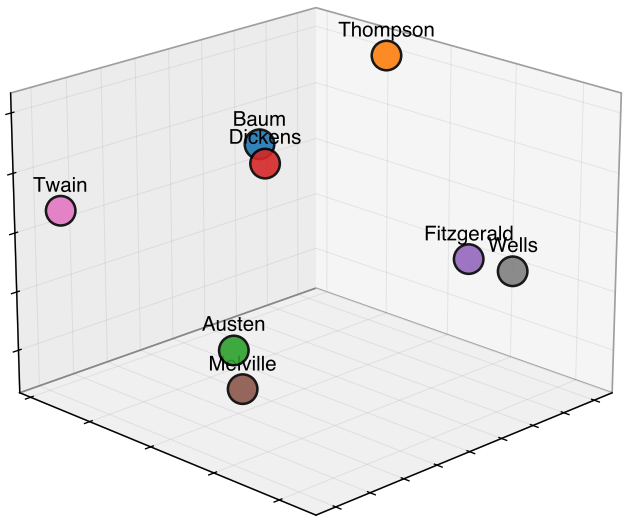


Figure 4: Multidimensional scaling plot. Three-dimensional MDS projection of the (symmetrized) average cross-entropy loss matrix shown in Figure 3. See Supplementary Materials for analogous plots using models trained on only content words, function words, and parts of speech (Supp. Fig. 8).

3.3 Predictive attribution of the 15th *Oz* book

Attribution is another application of predictive comparison. We illustrate with the well-known example of the contested authorship of the 15th *Oz* book (in a thirty-one book series), widely believed to have been written by Ruth Plumly Thompson, but originally attributed to L. Frank Baum (Binongo, 2003). We applied predictive comparison to the 15th *Oz* book, using models trained on Baum and Thompson's undisputed *Oz* books. As shown in the bottom left sub-panel of Figure 5, we find lower loss for the Thompson-trained model than for the Baum-trained model, indicating that the contested book is indeed more similar to Thompson's writing style than to Baum's. We also applied both models to a non-*Oz* book by Baum (bottom center) and Thompson (bottom right). We see

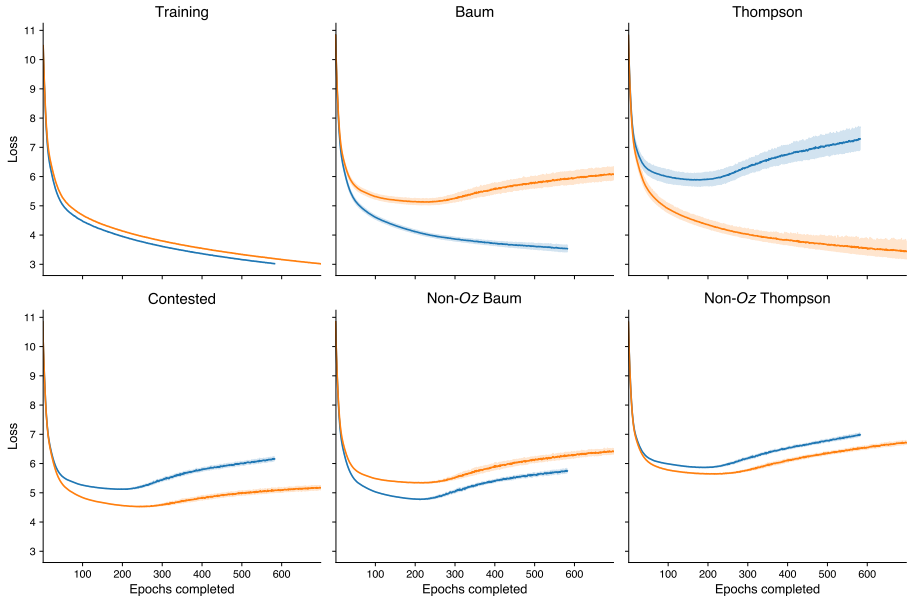


Figure 5: Cross-entropy loss across models and *Oz* authors. The top sub-panels replicate the Baum (blue) and Thompson (orange) results from Figure 1—i.e., that a given Thompson is well-distinguished from Baum and vice-versa (the two rightmost top sub-panels; error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds). The bottom sub-panels show the cross-entropy loss assigned to a held-out text whose authorship is contested (lower left), to a held-out non-*Oz* text by Baum (lower center), and to a held-out non-*Oz* text by Thompson (lower right). I.e., the contested book shows lower loss for Thompson-trained models; a non-Oz Baum book shows lower loss for Baum-trained models; and a non-Oz Thompson book shows lower loss for Thompson-trained models.

lower losses for the correct author in each case, demonstrating that predictive comparison is robust to thematic differences within the same author's writings.

3.4 Ablation studies: content words, function words, and parts of speech

The above analyses show that LLMs trained on one author's works can effectively capture the distinctive statistical patterns of that author's writing style. We carried out a series of ablation studies to investigate the contributions of different aspects of writing style. Specifically, we constructed three modified corpora for each author: (1) content-word-only corpora, in which all function words were replaced with a special token; (2) function-

word-only corpora, in which all content words were replaced with a special token; and (3) part-of-speech-only corpora, in which each word was replaced with its corresponding part-of-speech tag (see Investigating the contributions of function words, content words, and parts of speech). We then re-trained our models on each of these modified corpora and repeated the predictive comparison analyses (Supp. Figs. 1–6).

The models trained on the content-word-only corpora were intended to capture stylistic patterns related to vocabulary choice and thematic content. The models trained on the function-word-only corpora were intended to capture syntactic and grammatical patterns that transcended story-specific content. Finally, the models trained on the part-of-speech-only corpora were intended to capture higher-level syntactic patterns while abstracting away from specific word choices. Models trained on a single author's texts from each of these modified corpora all converged, achieving training losses below 3.0 well within 500 training epochs (Supp. Figs. 1, 2, and 3).

We found that models trained on content-word-only corpora reliably learned author-specific patterns for 6 of the 8 authors (Supp. Figs. 1 and 4, Supp. Tab. 1). Overall, by the final training epoch, the average t-values across all models and held-out texts were reliably greater than zero ($t(9) = 8.438, p = 1.44 \times 10^{-5}$). However, models trained only on content words were significantly less effective at distinguishing authors than models trained on the intact texts ($t(11.77) = 3.21, p = 7.68 \times 10^{-3}$).

Models trained on function-word-only corpora reliably learned author-specific patterns for 5 of the 8 authors (Supp. Figs. 2 and 5, Supp. Tab. 2). Overall, by the final training epoch, the average t-values across all models and held-out texts were reliably greater than zero ($t(9) = 4.428, p = 1.65 \times 10^{-3}$). These models were also significantly less effective at distinguishing authors than models trained on the intact texts ($t(8.36) = 4.82, p = 1.15 \times 10^{-3}$), but not significantly different from models trained on

content-word-only corpora ($t(10.29) = 1.81, p = 0.100$).

Models trained on part-of-speech-only corpora reliably learned author-specific patterns for just 3 of the 8 authors (Supp. Figs. 3 and 6, Supp. Tab. 3). By the final training epoch, the average t-values across all models and held-out texts were not reliably greater than zero ($t(9) = 1.616, p = 0.141$). These models were significantly less effective at distinguishing authors than models trained on the intact texts ($t(7.36) = 5.72, p = 6.01 \times 10^{-4}$), models trained on content-word-only corpora ($t(7.90) = 3.10, p = 1.49 \times 10^{-2}$), and models trained on function-word-only corpora ($t(10.41) = 2.11, p = 6.04 \times 10^{-2}$).

Taken together, these ablation results suggest that both content words and function words contribute to the author-unique stylometric signatures captured by our models. In contrast, grammatical structure alone, as reflected in part-of-speech sequences and captured by our methodology, appears to be more similar across authors. Distinctiveness notwithstanding, however, models trained on *all* of the corpora (intact texts, content-word-only, function-word-only, and part-of-speech-only) all rapidly converged to low training and evaluation losses. This indicates that all four corpora contain sufficient statistical regularities for GPT-2 models to learn to reliably and accurately make next-token predictions.

4 Discussion

We introduced predictive comparison, a method for stylometric analysis that leverages the predictive capabilities of language models trained on individual authors' works. Our approach rests on a straightforward principle: if a language model can learn to generate text in an author's style, then the cross-entropy loss of that model on held-out text should reflect stylistic similarity. By training separate GPT-2 models for each author and comparing their predictive performance, we aimed to develop both a measure of stylometric

distance and a practical tool for authorship attribution.

Our results demonstrate the effectiveness of this approach across multiple dimensions. Models trained and tested on the same author consistently yielded lower cross-entropy losses than models trained on different authors, achieving perfect classification accuracy across all eight authors examined. This separation emerged rapidly during training: for seven of eight authors, statistically significant discrimination was achieved after just two training epochs. The resulting stylometric distances proved meaningful, clustering authors with known stylistic similarities (e.g., Baum and Thompson) while maintaining clear separation between all author pairs. Finally, our method successfully resolved the well-studied attribution problem of the 15$^{\text{th}}$ *Oz* book, confirming Thompson's authorship in agreement with traditional stylometric analyses (Binongo, 2003).

We also conducted ablation studies to investigate the contributions of different aspects of writing style. Models trained on content-word-only and function-word-only corpora both captured author-specific patterns, though with reduced effectiveness compared to models trained on intact texts. In contrast, models trained solely on part-of-speech sequences struggled to distinguish authors, suggesting that grammatical structure alone is less distinctive. These findings highlight the importance of both lexical choice and syntactic patterns in shaping authorial style.

4.1 Relationship to prior work

Our predictive comparison approach relates closely to recent work using language model perplexity for authorship attribution (Huang et al., 2025), which independently developed a similar methodology using fine-tuned (rather than trained-from-scratch) GPT-2 models. Both approaches exploit the relationship between perplexity and cross-entropy loss, treating authorship attribution as a language modeling problem rather than a classification

task. This convergent development suggests that predictive modeling may be a natural framework for capturing authorial style.

The information-theoretic foundations of our approach connect to earlier work using cross-entropy (Juola and Baayen, 2005) and relative entropy (Zhao et al., 2006) for stylometry. These methods recognized that authorial style manifests not just in feature frequencies but in their sequential dependencies—precisely what language models are designed to capture. Our contribution extends this line of reasoning to large language models, which can learn these dependencies implicitly rather than requiring explicit feature engineering.

Compared to classification-based approaches using BERT (Fabien et al., 2020) or other transformers (Uchendu et al., 2020), predictive comparison offers conceptual simplicity: rather than training a single classifier to distinguish multiple authors, we train author-specific models that embody each writer's style. This approach naturally extends to open-set attribution problems where new authors may be introduced without retraining existing models. However, classification approaches may be more computationally efficient when dealing with fixed author sets, as they require training only a single model.

Our reliance on books as training data contrasts with most contemporary stylometry research, which typically uses shorter texts to enable larger author sets (Tyo et al., 2022). While this limits our experimental scope, it ensures that our models capture sustained stylistic patterns rather than topic-specific or context-dependent features that might dominate shorter texts (Fincke and Boschee, 2024). The success on full-length books suggests that predictive comparison can leverage the rich stylistic signal present in longer texts.

4.2 Limitations and challenges

Several limitations constrain the interpretation and application of our results. The most immediate is the limited experimental scope; we examined only eight authors writing in

English during overlapping historical periods. Whether predictive comparison maintains its effectiveness across larger author sets, different languages, or more diverse time periods remains an open question. The computational requirements of training separate models for each author may become prohibitive for attribution problems involving hundreds or thousands of candidate authors.

The opacity of large language models also presents interpretability challenges (Schuster et al., 2020). While our method successfully discriminates between authors, understanding which stylistic features drive this discrimination remains elusive. Unlike traditional stylometry, where specific features (e.g., function word frequencies) can be examined directly, the distributed representations learned by GPT-2 resist straightforward interpretation. This "black box" nature may limit adoption in domains where explanations for attribution decisions are required.

Cross-domain robustness represents another significant challenge. Prior work has shown that language model-based authorship attribution methods can struggle when training and test texts come from different genres or topics (Barlas and Stamatatos, 2020). Our experiments used books from the same genre for each author, leaving cross-domain performance unexplored. The strong performance on Baum and Thompson's *Oz* books versus their non-*Oz* works provides encouraging evidence, but systematic evaluation across diverse domains is needed.

The vulnerability of language model-based methods to adversarial attacks (Quiring et al., 2019) raises concerns about the reliability of predictive comparison in adversarial settings. Authors attempting to disguise their style or imitate others might fool language model-based attribution more easily than traditional methods that rely on subtler stylistic habits that are difficult to intentionally emulate. Evaluating robustness against both intentional obfuscation and unintentional style drift (e.g., authorial development over

time) will be crucial for practical applications.

4.3 Future directions

Several research directions could address current limitations while extending the theoretical and practical reach of predictive comparison. Understanding the theoretical relationship between cross-entropy loss and stylistic similarity would provide principled foundations for the approach. Why does minimizing cross-entropy during training lead to models that capture author-specific rather than general linguistic patterns? Connecting language model objectives to stylometric theory could yield insights for both fields.

Developing hybrid approaches that combine predictive comparison with traditional stylometric features or classification-based language-modeling methods might offset individual weaknesses. For instance, using cross-entropy loss as one feature among many in an ensemble model could improve robustness while maintaining interpretability through traditional features. Alternatively, predictive comparison could provide initial attributions that are refined using more interpretable methods.

The scalability challenge invites algorithmic innovations. Rather than training separate models from scratch for each author, could we use parameter-efficient fine-tuning methods (Houlsby et al., 2019) to adapt a single base model? Could authors be represented as vectors in a learned embedding space, with a single model conditioned on these embeddings? Such approaches might enable attribution among thousands of authors while maintaining the conceptual advantages of predictive modeling.

Finally, exploring applications beyond attribution could demonstrate the broader utility of modeling individual writing styles. For example, author-specific language models might be used to assist in literary analysis by generating counterfactual texts, such as what Austen might have written about modern themes (e.g., the impact of social media on

relationships). These approaches might also help to identify stylistic development within an author's career, or trace influence networks among authors. These applications would position predictive comparison within the broader landscape of computational literary studies.

4.4 Concluding remarks

Just as prior work has shown that it is possible to train LLMs to *write* in the "style" or "voice" of a given author (see e.g., Mikros, 2025), our work shows that LLMs may also be used to predict authorship and measure the stylistic distances between different authors. The predictive comparison method we have introduced offers a conceptually straightforward approach: models trained on individual authors' works embody their unique stylistic patterns, and the cross-entropy loss of these models on new texts provides a natural measure of stylistic similarity.

The strong empirical results—perfect attribution accuracy and meaningful stylometric distances—suggest that language models capture robust stylistic signatures, even when trained on relatively limited data. The convergence of our approach with concurrent work (Huang et al., 2025; Rezaei, 2025) indicates that the field may be moving toward predictive modeling as a unifying framework for computational stylometry. Our finding that models trained only on content words or function words can still capture author-specific patterns highlights the multifaceted nature of writing style. Both lexical choice and syntactic patterns appear to contribute to unique authorial signatures. In contrast, models trained solely on part-of-speech sequences struggled to differentiate between most authors, suggesting that grammatical structure alone is less distinctive. Overall, we suggest that our approach holds promise as a new technique for machine reading approaches to text-based disciplines (Holmes, 1998; Moretti, 2000, 2017) and the practices of cultural

analytics (Underwood et al., 2013).

Acknowledgments

We acknowledge helpful discussions with Jacob Bacus, Hung-Tu Chen, and Paxton Fitzpatrick. This research was supported in part by National Science Foundation Grant 2145172 to JRM and by a GPU cluster generously donated by the estate of Daniel J. Milstein.

Data and code availability

All code and data needed to reproduce the results in this paper are available at `https://github.com/ContextLab/llm-stylometry`.

References

Barlas, G. and Stamatatos, E. (2020). Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266. Springer.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *CHANCE*, 16(2):9–17.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Fabien, M., Villatoro-Tello, E., Motlicek, P., and Parida, S. (2020). BertAA: BERT fine-tuning

for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Fincke, S. and Boschee, E. (2024). Separating style from substance: Enhancing cross-genre authorship attribution through data selection and presentation. *arXiv preprint arXiv:2408.05192*.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111—-117.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.

Huang, W., Murakami, A., and Grieve, J. (2025). Attributing authorship via the perplexity of authorial language models. *PLOS One*, 20(7):e0327081.

Hughes, J. M., Foti, N. J., Krakauer, D. C., and Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *PNAS*, 109(20):7682–7686.

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Juola, P. and Baayen, H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl):59–67.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv*, 1711.05101.

Mikros, G. (2025). Beyond the surface: stylometric analysis of GPT-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities*, 40:587–600.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1:54–68.

Moretti, F. (2017). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso Books, Brookly, NY.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading, MA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Quiring, E., Maier, A., and Rieck, K. (2019). Misleading authorship attribution of source code using adversarial learning. In *Proceedings of the 28th USENIX Security Symposium*, pages 479–496.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rezaei, M. (2025). Detecting, generating, and evaluating in the writing style of different authors. In Ebrahimi, A., Haider, S., Liu, E., Haider, S., Leonor Pacheco, M., and Wein, S., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student*

Research Workshop), pages 485–491, Albuquerque, USA. Association for Computational Linguistics.

Schuster, T., Schuster, R., Shah, D. J., and Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Tyo, J., Dhingra, B., and Lipton, Z. C. (2022). On the state of the art in authorship attribution and authorship verification. *arXiv*, page 2209.06869.

Uchendu, A., Le, T., Shu, K., and Lee, D. (2020). Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Underwood, T., Black, M. L., Auvil, L., and Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. In *2013 IEEE International Conference on Big Data*, page 95–103. IEEE.

Zhao, Y., Zobel, J., and Vines, P. (2006). Using relative entropy for authorship attribution. In *Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006*, volume 4182 of *LNCS*, pages 92–105. Springer.

Appendix: Authors, books, and tokens

Charles Dickens	**Tokens**	**Herman Melville**	**Tokens**
A Christmas Carol	38,906	I and My Chimney	15,341
Oliver Twist	216,100	Bartleby, the Scrivener	19,112
The Old Curiosity Shop	285,895	Israel Potter	88,570
Bleak House	471,630	Omoo	134,628
Dombey and Son	482,161	Mardi, Vol. II	150,347
David Copperfield	479,387	The Confidence-Man	129,059
A Tale of Two Cities	181,593	White Jacket	190,577
Nicholas Nickleby	446,457	Mardi, Vol. I	132,358
American Notes	129,214	Moby-Dick	285,066
The Pickwick Papers	432,546	Typee	114,239
Great Expectations	244,897		
Martin Chuzzlewit	455,995		
Little Dorrit	449,230		
Hard Times	142,759		
Total	**4,456,770**	**Total**	**1,259,297**

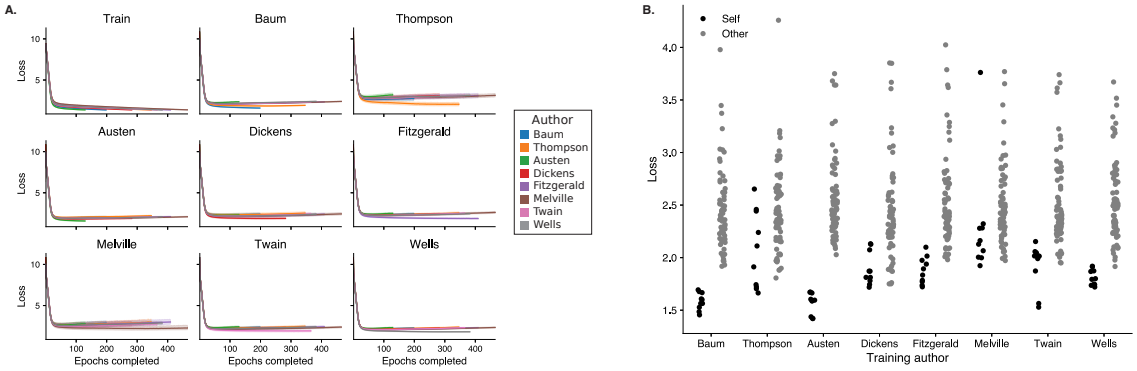
L. Frank Baum	Tokens	Ruth Plumly Thompson	Tokens
Ozma of Oz	52,039	The Giant Horse of Oz	51,036
Dorothy and the Wizard in Oz	53,849	The Cowardly Lion of Oz	61,666
Tik-Tok of Oz	63,781	Handy Mandy in Oz	44,778
The Road to Oz	52,866	The Gnome King of Oz	51,687
The Magic of Oz	51,166	Grampa in Oz	55,169
The Patchwork Girl of Oz	75,703	Captain Salt in Oz	61,797
The Wonderful Wizard of Oz	49,686	Ozoplaning with the Wizard of Oz	50,660
The Lost Princess of Oz	60,418	The Wishing Horse of Oz	59,490
The Emerald City of Oz	70,781	The Lost King of Oz	58,105
The Tin Woodman of Oz	57,338	The Hungry Tiger of Oz	53,543
Rinkitink in Oz	62,241	The Silver Princess in Oz	47,964
The Marvelous Land of Oz	54,733	Kabumpo in Oz	62,693
Glinda of Oz	51,218	Jack Pumpkinhead of Oz	49,661
The Scarecrow of Oz	59,593		
Total	**815,412**	**Total**	**708,249**

Jane Austen	**Tokens**	**Mark Twain**	**Tokens**
Sense And Sensibility	153,718	Adventures Of Huckleberry Finn	147,655
Mansfield Park	201,611	A Connecticut Yankee In King Arthur'S Court	150,327
Lady Susan	29,043	Roughing It	208,545
Northanger Abbey	98,090	The Innocents Abroad	246,321
Emma	207,830	The Adventures Of Tom Sawyer, Complete	95,059
Pride And Prejudice	157,777	The Prince And The Pauper	88,409
Persuasion	106,027		
Total	**954,096**	**Total**	**936,316**

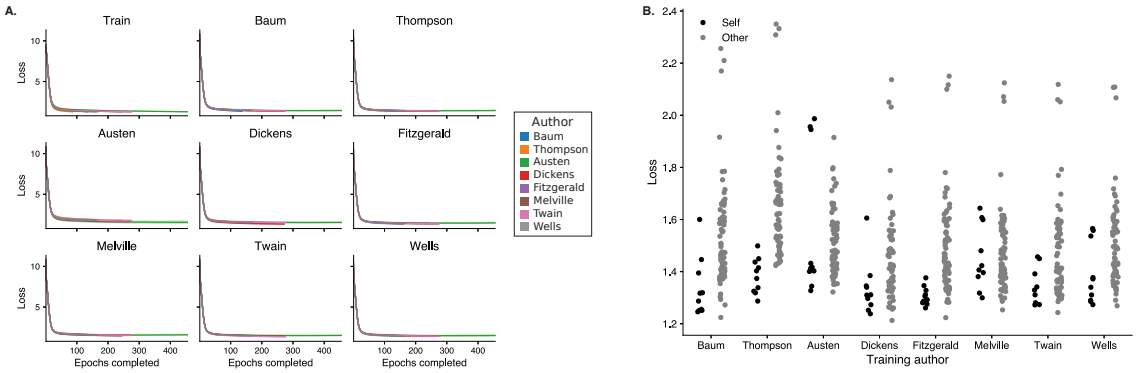
F. Scott Fitzgerald	**Tokens**	**H. G. Wells**	**Tokens**
The Beautiful And Damned	168,147	The Red Room	4,944
Flappers And Philosophers	84,707	The First Men In The Moon	87,615
This Side Of Paradise	100,796	The Island Of Doctor Moreau	55,967
All The Sad Young Men	85,411	The Open Conspiracy	40,271
Tales Of The Jazz Age	109,997	A Modern Utopia	105,810
The Pat Hobby Stories	51,069	The Sleeper Awakes	98,228
The Great Gatsby	65,136	The New Machiavelli	185,158
Tender Is The Night	145,925	The War Of The Worlds	75,727
		Tales Of Space And Time	94,711
		The Invisible Man: A Grotesque Romance	65,584
		The Time Machine	40,184
		The World Set Free	80,518
Total	**811,188**	**Total**	**934,717**

Supplementary materials for: A Stylometric Application of Large Language Models

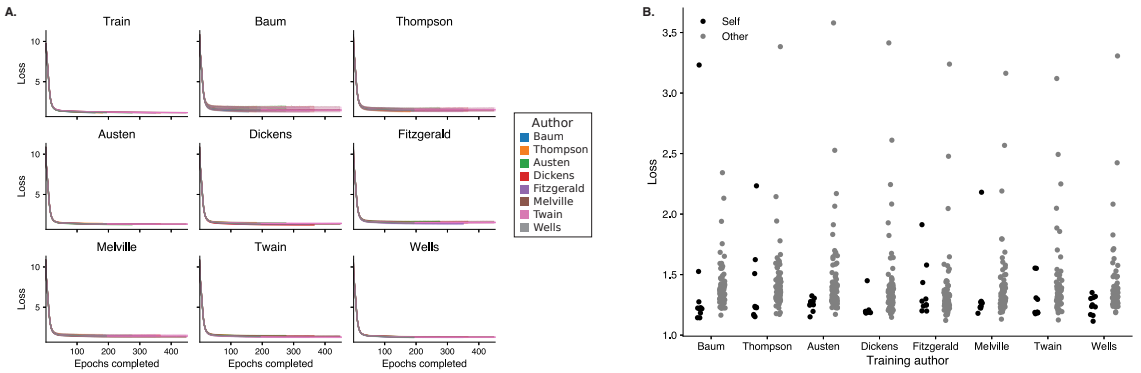
Harrison F. Stropkay, Jiayi Chen, Mohammad J. Latifi,
Daniel N. Rockmore, and Jeremy R. Manning
Dartmouth College
Hanover, NH 03755, USA
{harrison.f.stropkay.25, jiayi.chen.gr, mohammad.javad.latifi.jebelli
daniel.n.rockmore, jeremy.r.manning}@dartmouth.edu



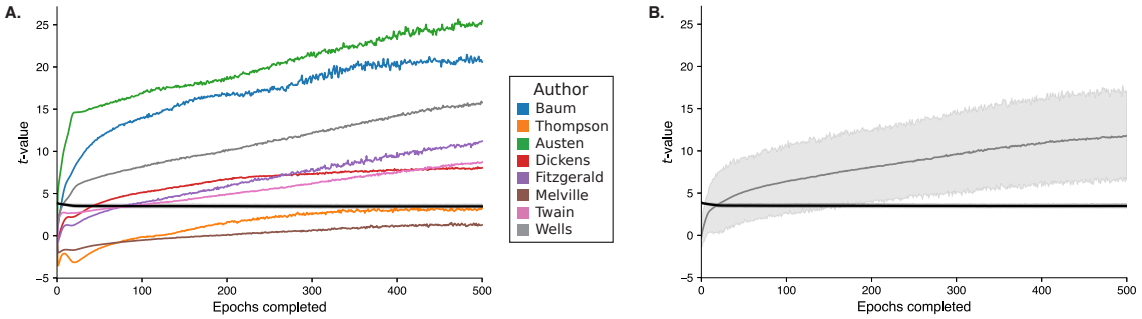
Supplementary Figure 1: Cross-entropy loss across models and authors using only content words. Follows the general format of Figure 1 in the main text, but uses models trained on only content words. All function words are masked out using <FUNC>. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (x-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



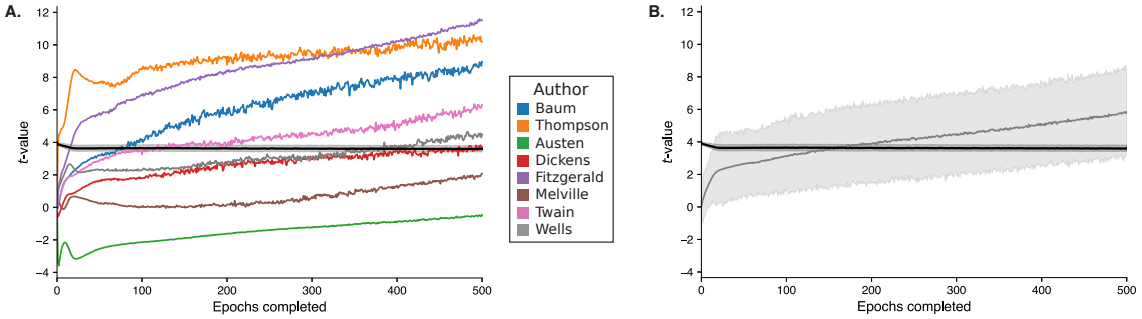
Supplementary Figure 2: Cross-entropy loss across models and authors using only function words. Follows the general format of Figure 1 in the main text, but uses models trained on only function words. All content words are masked out using <CONTENT>. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (*x*-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



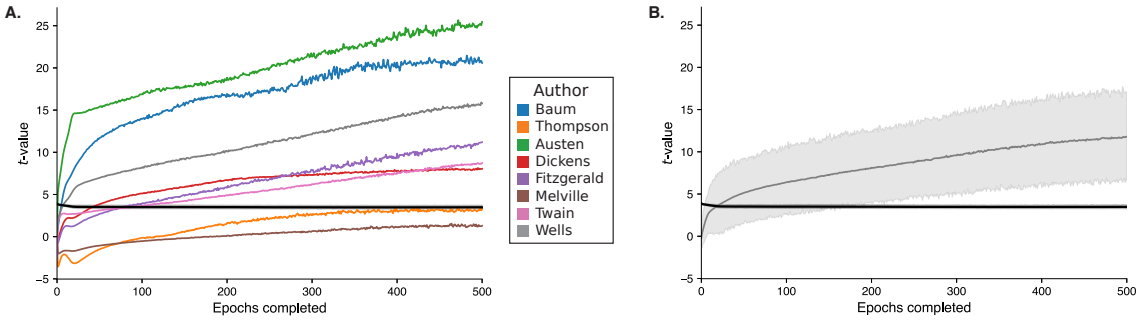
Supplementary Figure 3: Cross-entropy loss across models and authors using only parts of speech. Follows the general format of Figure 1 in the main text, but uses models trained on only parts of speech. All words are replaced with their corresponding part of speech tag. **A.** Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds. **B.** Cross-entropy loss assigned to held-out test data by each author's model (*x*-axis). Held-out test data is either from the *same* author (black) or from *other* authors (gray). Each dot denotes the average loss (across all 1024-token chunks) for a single random seed.



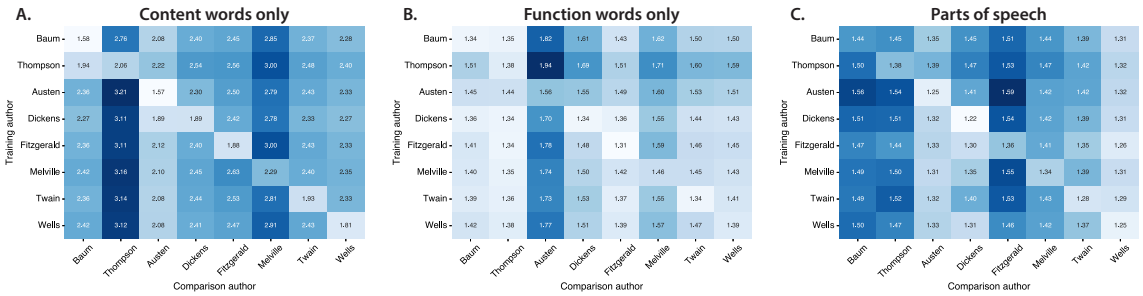
Supplementary Figure 4: Same vs. other author comparisons, by model, using only content words. Follows the general format of Figure 2 in the main text, but uses models trained on only content words. All function words are masked out using <FUNC>. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



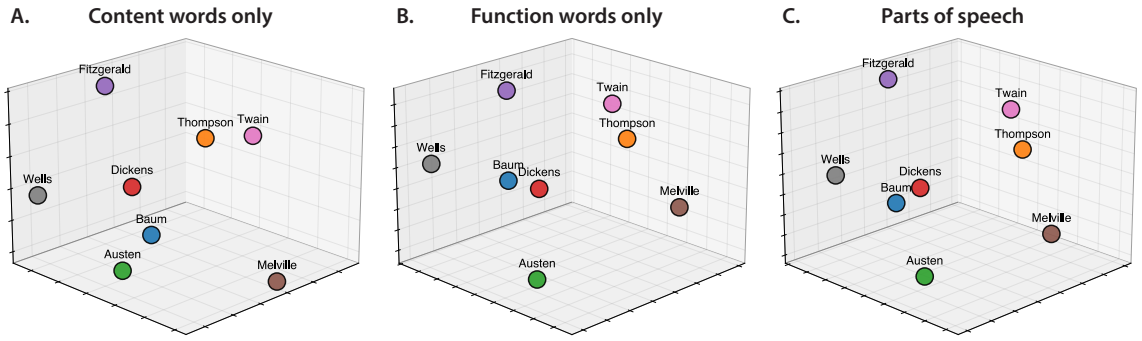
Supplementary Figure 5: Same vs. other author comparisons, by model, using only function words. Follows the general format of Figure 2 in the main text, but uses models trained on only function words. All content words are masked out using <CONTENT>. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



Supplementary Figure 6: Same vs. other author comparisons, by model, using only parts of speech. Follows the general format of Figure 2 in the main text, but uses models trained on only parts of speech. All words are replaced with their corresponding part of speech tag. **A.** Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **B.** The average t-statistic across all eight authors, as a function of the number of training epochs. The black curves in both panels indicates the average t-value corresponding to $p = 0.001$, for each epoch. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.



Supplementary Figure 7: Confusion matrices. Follows the general format of Figure 3 in the main text, but shows confusion matrices for models trained on only content words (A), only function words (B), and only parts of speech (C). Within each panel, the matrix displays the average cross-entropy loss assigned by models trained on each author's writing (column) to held-out texts from each author (row), after subtracting the native author's baseline loss.



Supplementary Figure 8: Multidimensional scaling plots. Follows the general format of Figure 4 in the main text, but shows MDS projections of the (symmetrized) average cross entropy loss matrices shown in Figure 7, for models trained on only content words (A), only function words (B), and only parts of speech (C).

Model	t-stat	df	p-value
Baum	20.58	68.36	5.27×10^{-31}
Thompson	3.29	11.33	6.97×10^{-3}
Austen	25.46	70.33	3.20×10^{-37}
Dickens	8.04	37.39	1.13×10^{-9}
Fitzgerald	11.21	49.02	3.97×10^{-15}
Melville	1.28	10.28	0.2274
Twain	8.73	22.50	1.12×10^{-8}
Wells	15.79	71.87	4.53×10^{-25}

Supplementary Table 1: Loss differences between same-author and other-author texts using only content words. Follows the general format of Table 1 in the main text, but uses models trained on only content words. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.

Model	t-stat	df	p-value
Baum	8.97	20.85	1.34×10^{-8}
Thompson	10.20	22.96	5.39×10^{-10}
Austen	-0.46	9.52	0.6581
Dickens	3.69	17.46	1.73×10^{-3}
Fitzgerald	11.52	77.98	1.70×10^{-18}
Melville	2.08	17.29	0.0529
Twain	6.31	34.66	3.14×10^{-7}
Wells	4.49	15.94	3.76×10^{-4}

Supplementary Table 2: Loss differences between same-author and other-author texts using only function words. Follows the general format of Table 1 in the main text, but uses models trained on only function words. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.

Model	t-stat	df	p-value
Baum	0.04	9.22	0.9695
Thompson	1.05	10.91	0.3179
Austen	5.72	77.97	1.89×10^{-7}
Dickens	4.41	62.85	4.14×10^{-5}
Fitzgerald	0.43	14.25	0.6704
Melville	0.81	11.98	0.4337
Twain	2.43	20.88	0.0240
Wells	4.05	56.90	1.55×10^{-4}

Supplementary Table 3: Loss differences between same-author and other-author texts using only parts of speech. Follows the general format of Table 1 in the main text, but uses models trained on only parts of speech. Each row displays the results of a t-test comparing the average loss values assigned by each author's model (after training is complete) to the author's held-out text and to the other authors' randomly sampled texts.