A Stylometric Application of Large Language Models

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

September 8, 2025

Abstract

We show that large language models (LLMs) can be used to distinguish the writings of different authors. Specifically, an individual model, trained 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 15th 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. In this paper we show, using a small set of authors and their works, that this thesis is borne out. 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 authentication 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^{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, 2019) 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

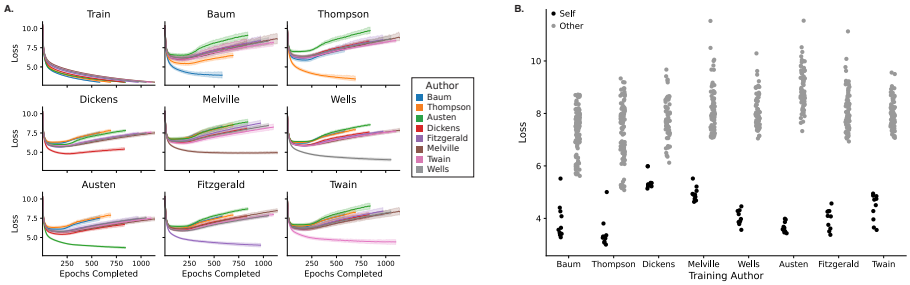

Figure 1: Cross-entropy loss across models and authors.A. Average cross-entropy loss on *Train*ing data and held-out test data from each author, plotted as a function of the number of training epochs. Each color denotes a model trained on a single author's work. Error ribbons denote bootstrap-estimated 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.

book or random sample.

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 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

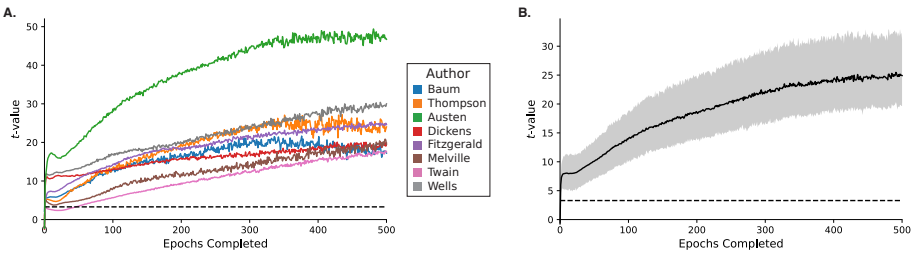

Figure 2: Same vs. other author comparisons, by model.D. Each curve denotes, as a function of the number of training epochs, the the t-statistic from a t-test comparing the distribution of losses (across random seeds) assigned to held-out texts from the given author (color) versus held-out texts from all other authors. **E.** The average t-statistic across all eight authors, as a function of the number of training epochs. Error ribbons denote bootstrap-estimated 95% confidence intervals across authors.

on another (i.e., for each author, the highest black dot is always lower than the lowest gray dot in the Panel). 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 epoch, indicating that the models rapidly acquire author-specific stylometric patterns. For Twain, this threshold is crossed at epoch 47. Figure 2B shows the average t-values across all eight authors as a function of the number of training epochs (final epoch: $t(9) = 20.723, p = 6.6 \times 10^{-9}$). 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	16.96	10.49	5.78×10^{-9}
Thompson	21.50	13.60	6.84×10^{-12}
Austen	47.29	54.75	4.38×10^{-46}
Dickens	18.36	27.36	6.52×10^{-17}
Fitzgerald	26.03	22.66	2.22×10^{-18}
Melville	24.15	45.15	1.87×10^{-27}
Twain	20.13	12.22	9.67×10^{-11}
Wells	35.17	26.33	1.16×10^{-23}

Table 1: 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.

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, excluding the diagonal entries (i.e., the losses obtained using each author's model when applied to their own held-out text). We observed (Fig. 4) that Baum and Thompson (authors of corpora largely intended to be similar) are mapped onto nearby locations, providing some evidence that the embeddings are "meaningful." 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.

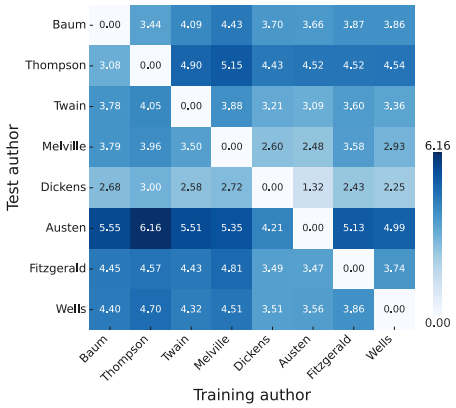

Figure 3: Confusion matrix. The matrix displays the average cross-entropy loss assigned by models trained on each author's writing (column) to held-out texts from each author (row).

3.2 Stylometric Distance

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

3.3 Predictive Attribution of the 15$^{\text{th}}$ *Oz* Book

Attribution is another application of predictive comparison. We illustrate with the well-known example of the contested authorship of the 15$^{\text{th}}$ *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 15$^{\text{th}}$ *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-

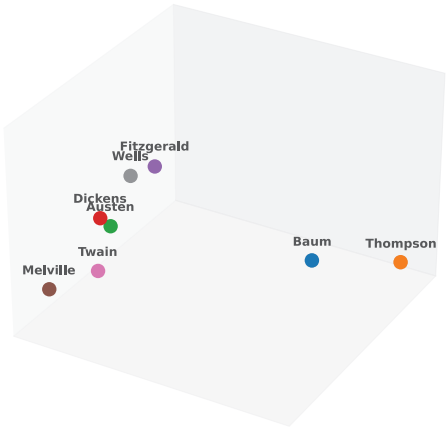

Figure 4: Multidimensional scaling plot. Three-dimensional MDS projection of the average cross-entropy loss matrix shown in Figure 3.

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

Conclusions

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 measuring the stylistic distances between different authors. We note that our approach is broadly similar to that of Rezaei (2025), who examined sentence level information using LLMs. Our work differs 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. We suggest that our approach holds promise as a new technique for machine reading approaches to text-based disciplines (Holmes, 1998;

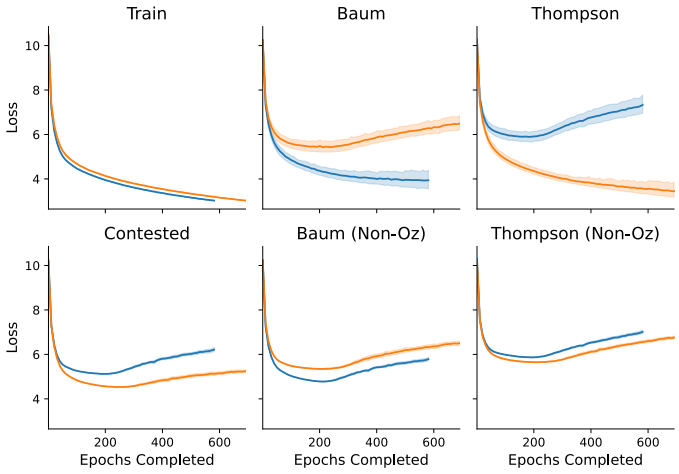

Figure 5: Cross-entropy loss across models and authors. The top sub-panels replicate the Baum (blue) and Thompson (orange) results from Figure 1. The bottom sub-panels show the cross-entropy loss assigned to a held-out text whose authorship is contested (lower left), to a held-out non-*Oz* text by Baum (lower center), and to a held-out non-*Oz* text by Thompson (lower right). Error ribbons denote bootstrap-estimated 95% confidence intervals over 10 random seeds.

Moretti, 2000, 2017) and the practices of cultural analytics (Underwood et al., 2013).

Limitations

The main limitations of this paper are (1) the lack of breadth of experiments as well as (2) the oft-acknowledged opacity of the LLM. The results in this paper serve as a proof-of-concept for the idea of using the structure of a bespoke trained LLM as a stylometric engine. Only a handful of examples have been tested, but the results on a classic stylometric test are promising. Further testing is needed to understand what kinds of writing features are being picked up by the LLM. Finally, deploying this idea at scale would require training one model for every writer of interest, a task that would require significant computational and textual resources.

References

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.

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

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.

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. (2019). Decoupled weight decay regularization.

Mikros, G. (2025). Beyond the surface: stylometric analysis of GPT-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities*, page fqaf035.

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.

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.

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.

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**