A Stylometric Application of Large Language Models

Anonymous ACL submission

Abstract

In this paper 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

Stylometry is the quantitative analysis of writing style. The field of stylometry is broadly concerned with capturing the statistical properties of text that characterize an author's writing style. Stylometry generally elides the complex collection of factors that may go into an author's "voice," focusing instead on word-usage statistics derived from the text (Neal et al., 2017). Many mark the birth of the subject with the late nineteenth century work of the philologist Wincenty Lutosławski, who had an interest in finding a statistical basis for addressing a long-standing problem in Classics of estimating the temporal order of Plato's Dialogues (Howland, 1991). Toward this end, Lutosławski measured hundreds of variables to arrive at his conclusions (Lutosławski, 1897). Stylometric methods have been used at scale to understand the evolution of style in literature (Hughes et al., 2012) as well as judicial writing(Carlson et al., 2016), while also contributing to debates around uncertain authorship (Mosteller and Wallace, 1963, 1984; Binongo, 2003; Juola, 2008).

Here we introduce a new stylometric approach, *Predictive comparison testing* (PCT), based on LLMs. PCT derives from the hypothesis that training an LLM using the works of a single given author will produce a model that best captures that author's unique writing style. Given two LLMs, M_A and M_B, trained on the works of two different authors, A and B, respectively, we therefore hypothesize that a held out work by Author A will exhibit smaller predictive loss when tested with model M_A than with model M_B. Our results, using a small data set of publicly available work from Project Gutenberg, bear out this hypothesis by showing that we can reliably predict the authorship of held-out text using the predictive losses of models trained on the works of different candidate authors. We also consider a well-known example of questioned attribution of the 15th book in the *Oz* series. Our approach agrees with the current prevailing scholarly opinion about that book (Binongo, 2003), further validating our approach.

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 prediction loss on some 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 compared to those 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 Appendix 2.6.

To construct a 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. Specifically, we truncate each author's corpus so that every model is trained on an equal number of tokens. This token budget is determined by removing the longest book from each author's set and then taking the smallest remaining total token count across each author's remaining books. For our dataset, this yields a 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 (i.e., remaining books after holding out a to-be-evaluated book). The length of the sub-sequence sampled from book i is proportional to its original length, computed as:

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

The starting position of each sub-sequence is chosen uniformly at random, ensuring the sample fits within the book's bounds.

2.2 Model Architecture, Training, and Evaluation

For each author, we train GPT-2 language models (Radford et al., 2019) 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 used a training batch size of 16 and 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 construct training samples by shuffling, concatenating, and then sampling 1024-token chunks from the truncated corpus for the given author (constructed as described above, using contiguous sub-sequences selected from all but one of their books). We continue training until the cross-entropy loss fell 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 a critical component of our stylometric analyses.

We evaluate the models using the held-out book from the corresponding author. We sample 1024-token chunks from the held-out book, using a sliding window approach 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 Predictive Comparison Testing

2.3.1 Eight-Author Comparison

For each author, we compute the predictive loss of the corresponding model on a held-out book by that author, as well as on one randomly selected book from each of the other authors. We then compute the average next-token cross-entropy with sliding windows.

Figure 1B presents the evaluation results across all eight authors. Training losses are again comparable across models, ensuring a fair basis for comparison.

For each author, we compare the predictive losses of all models on that author's held-out text. For every author's held-out text, the model trained on the matching author achieves the lowest loss, indicating a clear preference for its own author's stylistic patterns. This consistent alignment provides strong evidence that the GPT-2 models have learned to encode author-specific stylometric features.

2.3.2 Baum vs. Thompson

Once again, after each training epoch, we compute the loss of every model on its corresponding held-out book as well as on one randomly selected book from the other author's corpora. For the specific case of Baum and Thompson (and following (Binongo, 2003), we include three additional evaluation texts: the contested 15^{th} *Oz* book (authorship disputed between Baum and Thompson), a non-*Oz* book authored by Thompson, and a non-*Oz* book authored by Baum. For all texts used for predictive comparison, we compute the average next-token cross-entropy loss using a sliding window approach.

Figure 1A presents the evaluation results for models trained on Baum and Thompson. The top left sub-panel (labeled "Train") confirms that both models converge to similar training loss, ensuring a fair basis for comparison. The top center sub-panel (labeled "Baum") shows that the Baum-trained model achieves lower loss on Baum's held-out book than the Thompson-trained model. Conversely, the top right sub-panel (labeled "Thompson") shows that the Thompson-trained model yields lower loss on Thompson's held-out book than the Baum-trained model.

Notably, the bottom left sub-panel (labeled "Contested") shows that the Thompson-trained models consistently achieve lower loss on the contested 15^{th} *Oz* book, aligning with the prevailing literary consensus that Thompson was its author. The bottom center and bottom right sub-panels show the models' performance on non-*Oz* books by Baum and Thompson, respectively. As expected, the Baum-trained model performs better on Baum's non-*Oz* text, while the Thompson-trained model performs better on Thompson's.

These results collectively support the conclusion that the trained GPT-2 models are able to capture distinct stylometric patterns associated with each author.

2.4 t-tests

We also conduct a t-test for the eight-author comparison. Specifically, for each author's model, we perform t-tests for (i) the loss values computed by using the author's models to predict the author's held-out text and (ii) the loss values computed by using the author's models to predict the other authors' randomly sampled texts. Table 1 shows the results of the t-tests computed for each author using the final losses. Figure 1C illustrates the distribution of loss values for each author's model across self-authored and other-authored texts. These results demonstrate that the trained GPT-2 models reliably distinguish the stylometric features of the corresponding author with high statistical significance.

Author	t**-stat**	**p-value**	**df**
Baum	16.96	5.78×10^{-9}	10.49
Thompson	21.50	6.84×10^{-12}	13.60
Dickens	18.36	6.52×10^{-17}	27.36
Melville	24.15	1.87×10^{-27}	45.15
Wells	35.17	1.16×10^{-23}	26.33
Austen	47.29	4.38×10^{-46}	54.75
Fitzgerald	26.03	2.22×10^{-18}	22.66
Twain	20.13	9.67×10^{-11}	12.22

Table 1: t-test results for each author on final losses

In addition, we perform the same paired t-test at each of the first 500 training epochs, comparing losses on the author's own held-out texts to losses on texts from other authors. Figure 1D shows the t-values as training progresses. For all authors except Twain, the t-statistic exceeds the threshold corresponding to $p < 0.001$ after just one epoch, indicating rapid acquisition of author-specific stylometry. For Twain, this threshold is crossed at epoch 47. Figure 1E plots the average t-statistic across all eight authors over training epochs, further illustrating the early emergence of stylometric differentiation in the models.

The paired t-test also provides further validation for the *Oz* authorship question. On the average cross-entropy losses of the two models evaluated on the contested 15^{th} *Oz* book, the test revealed a statistically significant difference in predictive performance,

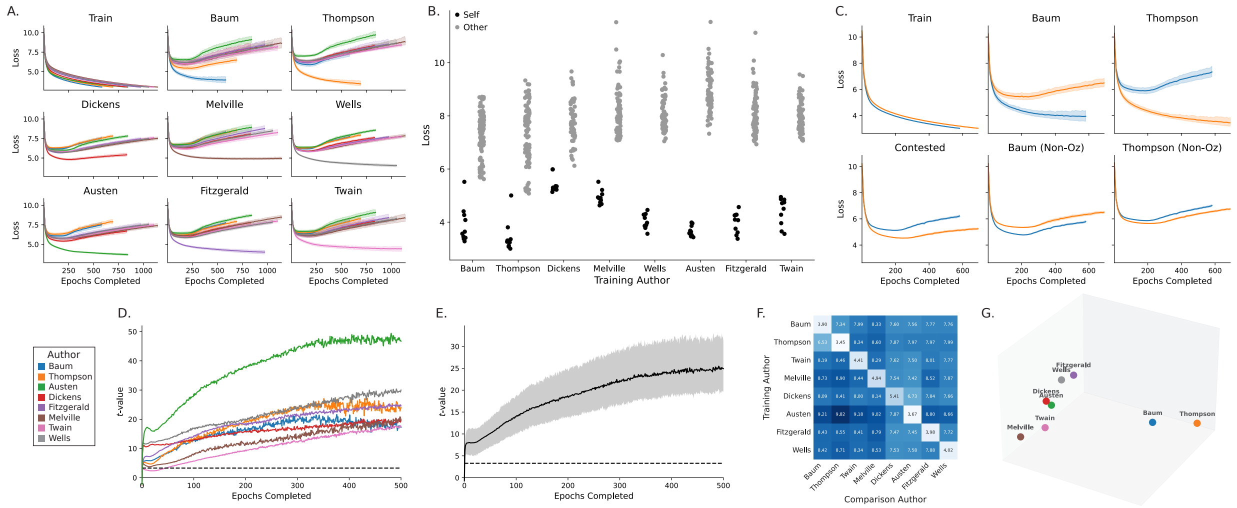

Figure 1: **A.** Average cross-entropy loss on evaluation texts for models trained on Baum and Thompson, each trained with a different random seed; error bars indicate 95% confidence interval. **B.** Average cross-entropy loss on evaluation texts across all eight authors. Error bars denote 95% confidence intervals over 10 random seeds. **C.** t-test results by author for first 500 training epochs. **D.** Averaged t-test results for first 500 training epochs.

$t(9) = 20.723, p = 6.6 \times 10^{-9}$. This provides strong evidence that the Thompson-trained models predict the contested text more accurately, aligning with prevailing literary consensus regarding its authorship.

2.5 Classification

For each author and each random seed, we evaluate models trained on that author's work by computing the loss on held-out texts from all eight authors. In every case, the model yields the lowest loss on the held-out text of its corresponding author.

2.6 Stylometric Distance

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 (the i, j-entry of the heatmap/average loss matrix in Figure 1F). Let $\overline{L_i(j)} = L_i(j) - L_i(i)$, normalizing the entries by subtracting the native author baseline. Then define the LLM-based *stylometric distance*, $d(i,j) = \frac{1}{2}\left(\overline{L_i(j)} + \overline{L_j(i)}\right)$. Figure 1G is a visualization of the relative "distances" among our author set.

Conclusions

It is possible to fine-tune large languages to write in the "style" or "voice" of a given author (see e.g., Mikros, 2025). Generally, the mathematical encoding—or "measure"—of a writer's style falls under the heading of *stylometry*.

Comparison of prediction at the level of sentences and using textual information is put forward in (Rezaei, 2025). Our work differs both in the scale and reliance purely on the loss function.

These results suggest that our approach holds promise as a new technique for machine reading approaches to text-based disciplines (Moretti, 2017, 2000; Holmes, 1998) and the practices of cultural analytics (Underwood et al., 2013). PCT also suggests a natural approach to author attribution in the case in which there is a finite set of possiblities: assign the work to the author whose PCT produces the least loss. To this we consider a well-known example of questioned attribution of the 15th book in the *Oz* series and scholarly opinion (Binongo, 2003), further validating our approach.

In this paper we introduce *predictive comparison*, a new LLM-based relative stylometric measure. It derives from a simple idea, that if an LLM can be fine-tuned to write like – i.e., in the style of – a given author by training on the work of an author, then the degree to which such a fine-tuned 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 which we produce. Lastly, this further suggest a literary authentication tool that would assign an unknown or contested work to the model which predictive comparison generates the smallest loss. We use this on a well-known and once contested book in the *Oz* series, confirming what is now accepted attribution. We believe this new idea could be of use in considering questions of authorial influence and stylistic evolution.

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. 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 fine-tuning one model for every writer of interest, a task that would require significant computational and textual resources.

References

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

Keith Carlson, Michael A. Livermore, and Daniel Rockmore. 2016. A quantitative analysis of writing style on the U.S. Supreme Court. *Wash. U. L. Rev.*, 93(6):1461.

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

Jacob Howland. 1991. Re-reading Plato: The problem of Platonic chronology. *Phoenix*, 45(3):189–214.

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

W. Lutosławski. 1897. *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of His Writings*. WC Brown Reprint Library, London.

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

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

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

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

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

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mosab Rezaei. 2025. Detecting, generating, and evaluating in the writing style of different authors. In *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.

Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu. 2013. Mapping mutable genres in structurally complex volumes. In *2013 IEEE International Conference on Big Data*, page 95–103. IEEE.

Appendix A: Authors, Books, 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**

Austen	**Tokens**	**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**

Fitzgerald	**Tokens**	**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**