T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models

Rebecca M. M. Hicke[1,*], David Mimno[2]

[1]*Department of Computer Science, Cornell University, USA*
[2]*Department of Information Science, Cornell University, USA*

Abstract
Large language models have shown breakthrough potential in many NLP domains. Here we consider their use for stylometry, specifically authorship identification in Early Modern English drama. We find both promising and concerning results; LLMs are able to accurately predict the author of surprisingly short passages but are also prone to confidently misattribute texts to specific authors. A fine-tuned `t5-large` model outperforms all tested baselines, including logistic regression, SVM with a linear kernel, and cosine delta, at attributing small passages. However, we see indications that the presence of certain authors in the model's pre-training data affects predictive results in ways that are difficult to assess.

Keywords
stylometry, large language models, Early Modern English drama

1. Introduction

Stylometry is a key tool for computational humanities research. Author identification provides a clear test case for methods that seek to identify "style," which in turn can be used to answer many questions of interest to humanists. However, current attribution methods require substantial amounts of known-author text for training as well as large amounts of text for identification. Large language models (LLMs) are powerful and now widely used. They develop a statistical model of language through training on a large, unorganized corpus. By encoding information from large amounts of contextual data in their parameters, they are often able to extract subtle, complex patterns from relatively short text segments. LLMs have proven useful for tasks such as detecting scenes in German dime novels [29], predicting TEI/XML annotations for plain-text editions of plays [18], and understanding ancient Korean documents [28].

In this work we consider whether LLMs can be applied to authorship identification and whether they might allow us to stretch the boundaries of stylometry to increasingly short passages. To evaluate these questions, we consider a deliberately difficult setting: Early Modern English drama. The language of Early Modern drama is sufficiently far from contemporary

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

*Corresponding author.

rmh327@cornell.edu (R. M. M. Hicke); mimno@cornell.edu (D. Mimno)
https://rmatouschekh.github.io (R. M. M. Hicke); https://mimno.infosci.cornell.edu (D. Mimno)
0009-0006-2074-8376 (R. M. M. Hicke); 0000-0001-7510-9404 (D. Mimno)

CEUR Workshop Proceedings (CEUR-WS.org)

English that it may be challenging for LLMs primarily trained on modern text to parse. Additionally, the culture of co-authorship and collaboration among writers during the Early Modern era often makes it difficult to distinguish stylistic delineations between individuals.

Despite its challenges, the attribution of Early Modern drama is a well-studied field, and techniques like cosine delta [24] achieve high accuracy at identifying plays. Yet, these methods still struggle to attribute short passages of text. We are specifically interested in determining whether fine-tuned LLMs can improve performance in this area. To this end, we provide the LLM with 5 to 450 word speaker utterances for both fine-tuning and testing. The average length of utterances in our test dataset is only 28.2 words.

We have three primary findings. First, for short texts the fine-tuned LLM outperforms all tested baselines, including logistic regression, a support vector machine (SVM) with a linear kernel, and cosine delta. Accuracy varies by author and is not fully explained by the number of plays by the author in the fine-tuning set. Second, LLMs are more prone than cosine delta to confidently misattribute texts to specific authors. These "scapegoat" authors often have large vocabularies and word use similar to the corpus average. Third, trained LLMs may be able to quantify "style". When we apply the model trained on Early Modern drama to "attribute" excerpts of plays written between the 1500s and 1900s, we see an increasing proportion attributed to Shakespeare, possibly suggesting a quantification of his lasting influence.

2. Related Work

Many different methods have been used to perform authorship attribution tasks with Early Modern drama. These include function word adjacency networks [8], multi-view learning [7], clustering algorithms [1, 23, 14], and SVMs with rolling attribution [20]. All of these studies attempt to attribute complete plays except [20], which attributes scenes with more than 100 lines. We are not aware of any use of large language models for Early Modern attribution.

Attempts to attribute shorter passages in Early Modern drama have been controversial. These studies include the attribution of 63 words from *Macbeth* [25] and samples of 173 words from *Henry VI, Part 1* [26]. They have been critiqued [9, 22] in part because the sections of text studied were so short. While we attempt short text attribution, we select samples broadly from many plays rather than focusing on specific passages.

Work has also been done on the attribution of short texts in different fields. Cosine similarity is effective at attributing 500 word excerpts from blogs [11]. Similarly, topic models are able to attribute email and blog snippets with average length 39 and 57 words [30] and the Source Code Authorship Profile (SCAP) method attributes tweets of 140 characters or shorter with high accuracy [12, 3]. None of these studies use LLMs, and all use modern datasets.

Some researchers have begun testing the feasibility of using LLMs for attribution. These studies used the embedding output of LLMs to train custom attribution models using LSTMs [10] or CNNs [15]. Our work uses a simpler LLM method, in which we fine-tune the original model to directly generate author names, without the need for any additional coding or customization. In addition, we use a corpus with less clear delineation.

Table 1
Example of input-output pair used during fine-tuning. We include the author label as *masked* text to be generated.

Input	Output
AUTHOR: <extra_id_0> | All the damnable degrees Of drinkings have you, you staggered through one Citizen. Is Lord of two fair Manors, called you master Only for Caviar.	AUTHOR: John Webster | All the damnable degrees Of drinkings have you, you staggered through one Citizen. Is Lord of two fair Manors, called you master Only for Caviar.

3. Data & Methods

We use a collection of Early Modern English drama—plays written in the 1500s and 1600s—gathered from two sources: the Folger Digital Anthology of Early Modern English Drama (EMED) [5] and the Shakespeare His Contemporaries corpus (SHC) [13]. We first gathered 367 plays from the EMED corpus and then added the 181 remaining plays from SHC.[1] In order to remove features that may distinguish files from different corpora, we stripped all non-accent non-ASCII characters from the play texts and replaced them with standardized alternatives where appropriate. Each XML file offered regularized spellings of non-standard words in the play. In creating our corpus, we used the regularized spellings from the EMED corpus that agreed with the greatest number of other sources when possible and the SHC regularizations otherwise. We chose to use the regularized text for two reasons. First, we did not want the model to be able to distinguish between authors based on spelling choices. Although differences in spelling may help the model identify authors, they are not indicative of the kinds of stylistic difference we are interested in studying. Second, we hypothesized that standardizing the play texts would make them appear more similar to modern text and thus improve the model's ability to accurately tokenize the input. Finally, we removed all line breaks from the texts as the different corpora do not consistently mark them.

We then split each play into speaker utterances to create a challenging but coherent identification problem. We separated any utterance longer than 450 words into multiple samples by splitting directly after every 450th word, regardless of sentence or line breaks. We then removed utterances with fewer than 5 words. Because authors sometimes develop distinctive speaker voices within a play, we hypothesize that separating the texts by speaker utterance adds an extra layer of difficulty to the attribution task.

We further reduced the training and testing corpora to maximize validity and statistical reliability. We removed all plays with fewer than 300 remaining utterances, plays by multiple authors, and plays by authors with fewer than three works in the corpus. Plays that were mislabeled as by a single author, but were actually of disputed (co-)authorship were placed into a separate subcorpus. We were thus left with 253 plays by 23 authors in the primary corpus and 23 plays in the subcorpus. Further details about the corpora are listed in the appendix.

We used these corpora to assess the capability of several different authorship attribution methods to label short texts. Specifically, we tested logistic regression, SVMs with a linear ker-

[1] Because the original Shakespeare His Contemporaries corpus is no longer publicly available, we have drawn these sources from a port of the original Github linked in the citation.

nal, cosine delta [24], Pythia [4], Falcon [19], and several fine-tuned T5 models [21] of varying sizes. T5 is a generative large language model and the pre-trained T5 models are optimized with a masked language modeling objective. Thus, in order to fine-tune T5 to perform authorship attribution, we created a series of input and output pairs where the inputs are formatted as an utterance with the author's name masked and the corresponding outputs are the same utterances with the author's name revealed (i.e. Table 1). The tag `<extra_id_0>` was used to mask the author's name because it follows the format of tags used during T5's pre-training regime. Initial experimentation found that using this tag provided good accuracy. It is important to note that the model could emit any string, but in practice the fine-tuned model only generated author names present in our corpus except during a later application to a comparative dataset (Section 7).

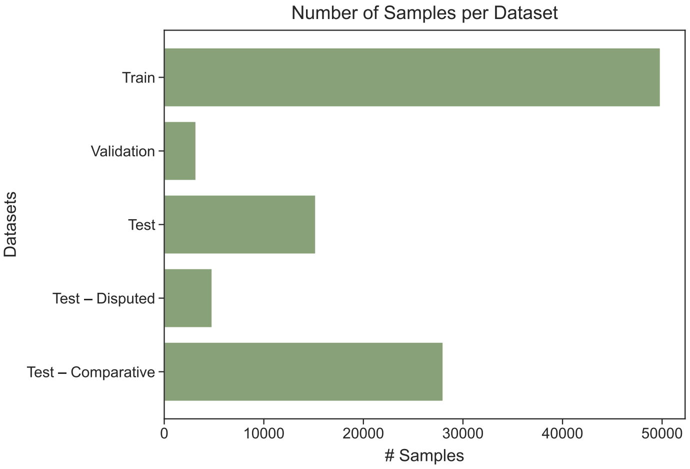

Figure 1: The size of each dataset used in the experiment by number of samples.

One play from each author in our corpus was withheld from the training dataset. From the remaining $n - 1$ plays by each author, we included 235 random samples in the training dataset and 15 samples in the validation dataset used for parameter tuning. We included another 50 samples from each of these plays in the final test dataset for which we report results. Thus, we draw 300 distinct samples from each play withheld from the training dataset. To the test dataset, we added 200 randomly selected samples from each of the plays withheld from training. We then created a separate test dataset containing 200 samples from each of the 23 plays in the disputed authorship corpus. Figure 1 shows the relative size of the training, validation, and test datasets, as well as the held-out disputed set and a set of post-Early Modern plays used in Section 7.

We fine-tuned a small, base, and large version of T5 on the train and validation datasets, using batch sizes of 16, 8, and 4 respectively and running for 10 epochs. The additional fine-tuning hyperparameters are reported in Section A of the appendix. We then asked each model to predict labels for every sample in the primary test dataset. Finally, we used the best performing model, `t5-large`, to predict labels for the test dataset of disputed authorship plays. We also experimented with fine-tuning two comparable decoder-only generative LLMs: Pythia with 1 billion parameters and Falcon with 1 billion parameters. The input and output strings described above were edited for these experiments so that the AUTHOR tag was placed at the end of each string. However, both models hallucinated extensively; Pythia produced 10,221 unique strings as author names and Falcon produced 9,180. Even when the first two words of each produced string, stripped of punctuation, were used as the predicted author name Pythia and Falcon still performed considerably worse than `t5-large`. We thus omit a further analysis of these models from the paper.

For our baseline comparisons we used only the original quotation without the AUTHOR prefix and T5 tags. The correct authors were included as labels. We ran two logistic regression models and two SVM models with linear kernels: one version of each used TF-IDF weighted word counts as features and the other used plain word counts. Each of these baselines was implemented using the `sklearn` package. Cosine delta [24] is a popular improvement on Burrows delta [6] that represents texts using z-score weighted word frequencies for the *n* most frequent words and compares sample texts to the training corpus using cosine similarity. To run cosine delta, we used an adapted version of the `faststylometry` package with a vocabulary size of 5,000 unigrams [27]. We chose a vocabulary size of 5,000 because we found it optimized performance on the plays in the training data without over-fitting and decreasing performance on the withheld plays. Each sample was assigned to the author with the highest cosine similarity value. All baseline models were evaluated on the same test/train splits as the T5 models and the TF-IDF, z-score, and word count values were fit on only the training dataset. For every model, we experimented with using combinations of unigrams, bigrams, and trigrams but found that using only unigrams resulted in the highest performance.

4. Comparing Models

Results are shown in Table 2 for the per-sample accuracy of each attribution method and the accuracy of the "majority vote" predicted author of each play. In order to display the effect of play-specific language such as character names and settings, we show predictive results for both held-out *sections* of plays and fully held-out plays.

We begin by establishing that accurate author attribution is possible for this dataset using only the available information. It is known that authorship attribution is more reliable for longer samples. To establish an upper bound on expected performance, we thus apply a cosine delta model to the full held-out text of each play rather than the short samples we use for all other experiments. This setting increases the length of attributed samples by a factor of 50 for plays in the training set and 200 for fully held-out plays. Cosine delta accurately attributes 94.9% of the long samples, performing better on plays in the training set than those fully held-out.

The fine-tuned `t5-large` model correctly attributes more short samples than any other method tested. It accurately labels 52.7% of held-out samples from plays included in the training dataset and 33.2% of samples from plays fully withheld from training. `t5-large` performs substantially worse on the individual sample level than the cosine delta upper bound, but it only falls seven plays short of the upper bound when attributing plays to the most-predicted author. Although it is not surprising that results are better for partially-seen plays, the accuracy of both subsets exceeded our expectations. Because the text excerpts we use are very short, they frequently contain no named entities, and we thus conclude that attribution was not performed solely using this information.

Longer samples were more accurately attributed. The average length of correctly attributed

Table 2
LLMs have the highest predictive accuracy for short texts. The ± values represent 95% confidence intervals. For context, we show both a lower bound (random guessing based on author frequency) and an upper bound (cosine delta on large text segments, all other rows are evaluated on short texts).

Method	% Correct (In)	% Correct (Out)	% Correct by Play
Upper Bound			
Cosine Delta (Long Texts)	95.8	87.0	94.9
LLMs			
Fine-tuned `t5-large`	52.7 ± 0.3	33.2 ± 0.4	91.9
Fine-tuned `t5-base`	45.9 ± 0.3	27.6 ± 0.4	74.9
Fine-tuned `t5-small`	22.4 ± 0.3	11.8 ± 0.3	28.5
Baselines			
Linear SVM (TF-IDF)	48.0 ± 0.3	23.3 ± 0.4	90.2
Logistic Regression (Word counts)	45.0 ± 0.3	23.5 ± 0.4	86.0
Linear SVM (Word counts)	43.6 ± 0.3	21.7 ± 0.4	88.1
Logistic Regression (TF-IDF)	42.6 ± 0.3	19.3 ± 0.4	66.4
Cosine Delta (Short Texts)	24.1	18.2	89.8
Most Prominent Author	14.2	4.3	13.2
Random	7.5	4.3	14.5

samples in our primary test dataset was 36.7 words whereas the average length of misattributed samples was 20.7 words. Figure 2 shows the distribution of sample lengths and the accuracy for each range. Accuracy exceeds 50% with only 20 words (random is ≈5%). Model scale also effects accuracy. The `t5-large` model performed better than the smaller models we compare it to, `t5-base` and `t5-small`. We observe that `t5-large` does 30.3% better on samples from plays included in training, 21.4% better on samples from plays withheld from training, and 63.4% better at attributing plays by majority vote than `t5-small`. This effect may be due to the larger model's greater capacity to fit the particulars of author-specific language in fine-tuning, a greater capacity to represent linguistic variation in pre-training, or some combination of both.

It appears that the reason for the large improvement in play attribution accuracy with model size is a significant reduction in the assignment of large numbers of samples to 2–3 specific (incorrect) authors, which we call *scapegoating*. `t5-small` assigns 60.5% of misattributed segments to the top two scapegoated authors (Thomas Heywood and William Shakespeare), `t5-base` assigns 32.8% of misattributed samples to two authors (Heywood and James Shirley), and `t5-large` only attributes 25.6% of misattributed samples to two authors (also Heywood and Shirley). Because misattributions both occur less frequently and are spread more evenly between authors in the larger models, it is more likely that the author of a play will have the majority of samples assigned to them.

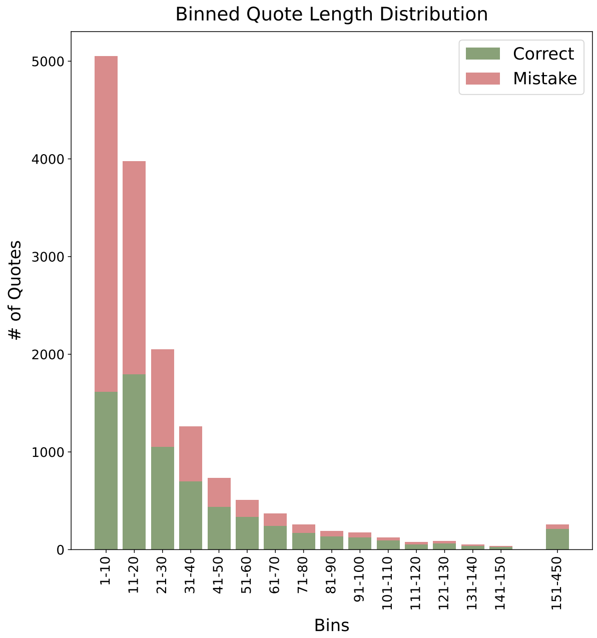

Figure 2: Length distribution of samples in the test dataset. Sample lengths are binned by 10s, and all quotes longer than 150 words are placed in one bin.

Logistic regression and linear SVM prove to be strong baselines. However, `t5-large` performs 4.7% better on samples from plays included in training and 9.9% better on samples from withheld plays than linear SVM with TF-IDF values, the highest performing of these baselines. Since these models have access to the same data, the difference must either come from the LLM's ability to use arbitrary combinations of non-sequitive words or its access to patterns from pre-training. It is important to note that we do not know what data T5 saw during pre-training. However, because the `t5-small` model performs worse than logistic regression, it is unlikely that this is the sole source of improvement. Logistic regression and linear SVM are also prone to scapegoating: all models assign over 35% of misattributed samples to two primary authors (Shakespeare and Shirley). Linear regression with TF-IDF values is a particularly egregious scapegoater, assigning over 50% of samples to Shakespeare and Shirley, over double the number that `t5-large` assigns to Shirley and Heywood.

In addition to the "merged samples" upper bound, we apply cosine delta to the short samples. This approach performs worse than all methods but `t5-small` and the simple baselines. However, cosine delta achieves high performance at the play level. Even `t5-large` only attributes 6 more of the 253 plays in the original corpus correctly. This, again, appears to be related to scapegoating. Cosine delta assigns the samples it misattributes relatively evenly between authors, only assigning 12.4% to the two most scapegoated authors (Richard Brome and Thomas Middleton). Thus, while cosine delta may be less accurate overall than T5, the way in which it fails is less skewed. Compared to T5, SVMs, and logistic regression, it is less likely to confidently misattribute a play.

Table 3
Percentage of samples correctly attributed by `t5-large` for plays withheld from and included in the training dataset and # of plays in the corpus by author. The authors are ordered by accuracy on samples from plays included in training.

Author	% Correct (In)	% Correct (Out)	# Plays
William Shakespeare	79.0	72.0	30
Margaret Cavendish	74.9	68.5	12
James Shirley	61.7	58.5	31
John Lyly	60.6	60.0	8
Thomas May	58.0	0.5	3
John Fletcher	54.7	56.5	15
Christopher Marlowe	53.6	55.5	6
Thomas Killigrew	52.0	50.0	4
Robert Greene	51.0	2.0	3
Ben Jonson	49.7	59.5	14
Philip Massinger	49.0	46.5	13
Thomas Heywood	49.0	36.0	19
Richard Brome	42.4	32.0	15
Thomas Nabbes	40.5	17.5	5
George Chapman	38.4	13.0	11
William Davenant	36.7	24.5	4
Thomas Middleton	36.0	33.0	13
John Marston	34.7	22.5	7
John Ford	25.3	15.5	7
Thomas Dekker	22.0	11.5	6
Henry Glapthorne	22.0	5.0	3
Robert Wilson	20.0	19.0	3
John Webster	9.0	4.5	3

5. Accuracy by Author

For the best-performing model, `t5-large`, accuracy varies considerably by author for both withheld plays and those included in training (Table 3). Authors with more plays in the training set are more accurately predicted for the held-out set; the Pearson correlation coefficient between these values is 0.65, with $p < 10^{-3}$.

The `t5-large` model performs well on samples from many of the well-represented authors in our corpus. For 9 of the 23 authors, the model accurately attributes more than 50% of samples from plays included in training, well above random. The four authors for whom the model performs best on samples from included plays are Shakespeare (79.0%), Margaret Cavendish (74.9%), Shirley (61.7%), and John Lyly (60.6%). The model also accurately attributes many samples from the withheld plays by these authors: 72.0% of samples from Shakespeare[2] are correctly attributed, 68.5% of samples from Cavendish[3], 58.5% of samples from Shirley[4], and

[2] *Antony and Cleopatra*
[3] *The Wooers*
[4] *The Sisters*

60.0% of samples from Lyly[5].

The reasons that the model attributes samples from these four authors with such high accuracy differ. Shirley and Shakespeare are the authors with the most and second-most plays in the dataset, with 31 and 30 plays in the corpus respectively. But Cavendish (8) and Lyly (12) are close to the average. Authors comparable to Cavendish in representation, such as George Chapman (11), Philip Massinger (13), and Thomas Middleton (13), all have accuracies below 50% for plays included in training. Similarly, authors comparable to Lyly such as John Ford (7) and John Marston (7) both have accuracies below 35% on included plays. Therefore, there is likely something distinctive about these two authors that makes them easier for the model to identify. Note that Cavendish is the only female author in the corpus (we were unable to include others), so we are not able to determine if her plays are distinctive because she has an individual style or if women authors of the period wrote differently from men.

To further explore the cause of Cavendish and Lyly's distinctiveness, we compare each author's usage of the 100 most frequent words in the corpus. We first calculate z-scores comparing the frequency with which an author used each word to the mean frequency of that word's usage for all authors in the dataset. The frequencies are normalized by author so that no single author skews the distribution and we ensure that the set of 100 most frequent words contains no named entities. We then sum the absolute values of each author's z-scores to create a 'uniqueness' metric. For a further exploration and validation of this metric, please see Section B of the appendix. The summed z-scores ranged from 47.5 to 138.2. The author with the most unique usage of common words by this metric is Cavendish, with a score of 138.2. The authors with comparable play counts to Cavendish each have considerably lower scores (Chapman: 50.57, Massinger: 69.9, Middleton: 67.7). The second most distinctive author is Thomas Killigrew, with a summed z-score of 108.7. Indeed, Killigrew has a very high accuracy on samples from included plays (52.0%) considering only four of his works are in the corpus. Lyly also has a relatively high summed z-score of 99.4, which is the fourth largest in the dataset. Again, this is higher than the scores of comparably represented authors (Ford: 68.3, Marston: 70.8), but not by as much. Notably, both Shirley and Shakespeare have low uniqueness scores by this metric. Shakespeare's is the lowest (47.5) and Shirley's is the 16th lowest (67.8). In addition, both authors have large vocabularies; Shakespeare has the largest vocabulary and Shirley the third-largest of all authors in the dataset. Both of these trends are likely related to their prominence within the training dataset, but they may still be meaningful. It is possible that Shakespeare and Shirley's uniqueness comes from using words that the other authors do not, instead of using common words uniquely. Overall, it seems that an author's usage of common words does affect how well the model can identify their writing. But it does not explain all of the variation seen in the dataset.

5.1. Quote Misattribution

There is also considerable variation in how the fine-tuned `t5-large` model misattributes quotes (Figure 3). Instead of assigning the misattributed quotes to authors randomly, it scapegoats two primary authors, Heywood and Shirley, and assigns them a disproportionate number. A

[5] *Sappho and Phao*

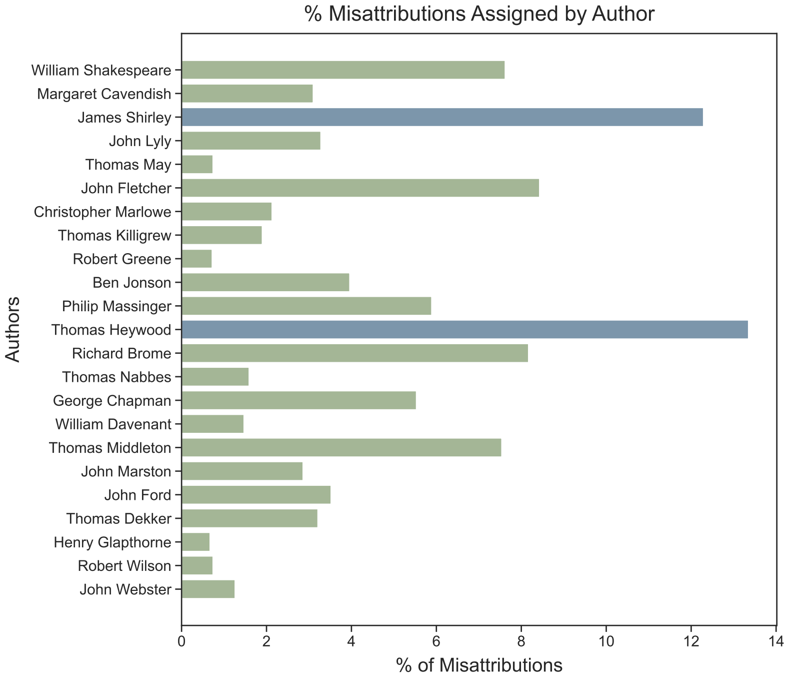

Figure 3: Percentage of misattributed samples assigned to each author.

confusion matrix depicting who quotes are misattributed to by original author demonstrates that the scapegoating phenomenon is not caused by confusion between specific pairs of authors (Figure 4). Instead, the misattributions to Heywood and Shirley are spread throughout the dataset. Again, it appears that contribution to the corpus is one factor that affects who samples are misattributed to. The Pearson's R correlation between the number of plays by an author in the dataset and the percentage of misattributed samples assigned to them is 0.86 with $p < 10^{-6}$. The outliers from this relationship appear to be Heywood, Shakespeare, Cavendish, and Ben Jonson (Figure 5).

Examining authors' scores for the summed z-score metric again provides an indication of why some are scapegoated. Cavendish's high uniqueness score likely means it is more difficult for the model to mistake a given quote for hers. In contrast, Heywood, who has the most samples misattributed to him, has the second-lowest uniqueness score in the corpus, 49.4. He also has the second-largest vocabulary. The combination of these factors may help explain why he is so frequently scapegoated. Given a random quote from the test dataset, Heywood is more likely than most authors to have all of the words in the sample in his vocabulary. Even if he doesn't, the model could have learned that he is more likely to use a broad range of words than other authors. In addition, common word usage in the average corpus sample is likely to resemble Heywood's. Shirley, who has the second-most misattributed quotes assigned to him, has the third-largest vocabulary and the 16th lowest uniqueness score. Thus, it appears that vocabulary size and common word usage are factors that affect to whom the model's misattributes quotes.

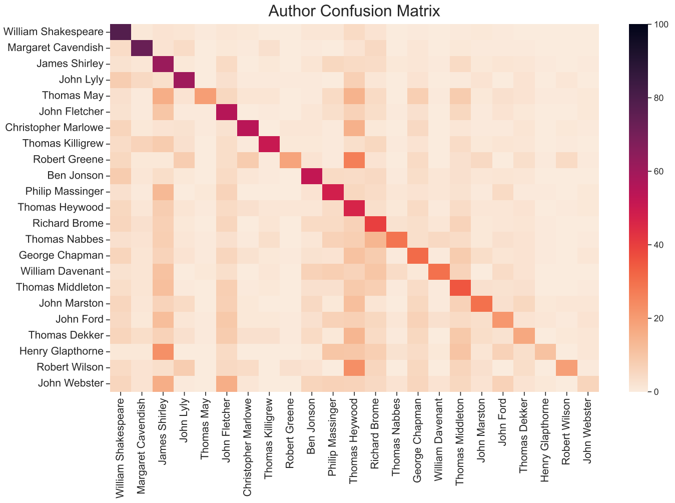

Figure 4: Confusion matrix demonstrating how frequently samples from row authors were misattributed to column authors. Each matrix row sums to 100%. Prolific authors Heywood, Shakespeare, and Shirley are most commonly guessed.

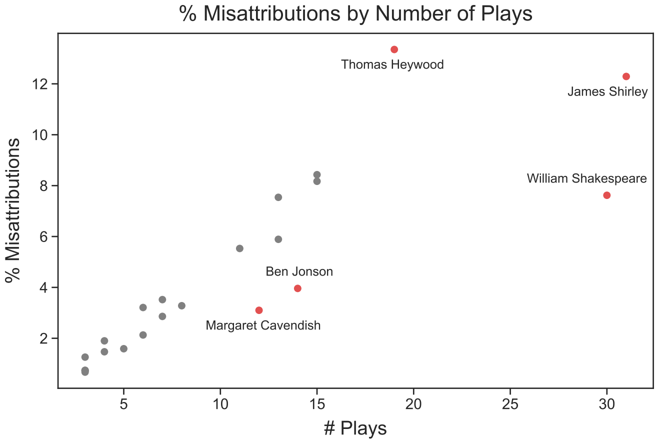

Figure 5: The relationship between the number of plays by an author and the % of misattributed samples assigned to them. Visual outliers are highlighted and labeled.

However, there are two major outliers which indicate that these three factors—number of plays, common word usage, and vocabulary size—cannot be the only ones affecting scapegoating. These are Jonson and Shakespeare. Shakespeare has both the largest vocabulary and lowest uniqueness score of any author in the corpus, and yet samples are less likely to be misattributed to him than would be expected given his contribution to the dataset. Similarly, Jonson has the fourth-largest vocabulary and the 19th lowest uniqueness score, yet he also stands out as an outlier to whom fewer samples are misattributed than expected. We hypothesize that these outliers are caused by the model's pre-training. Of the authors included in our corpus, Shakespeare and Jonson are among the best-known today. The model is likely to have seen the writing of these authors during pre-training, and may therefore be more likely to correctly label data from these authors than would be expected given only the fine-tuning process.

6. Accuracy by Play

Interesting patterns and outliers emerge when we examine the model's play-by-play accuracy at attributing samples. There are several authors, like Cavendish, for whom the proportion of correctly attributed samples is largely consistent across all plays and some, like Brome, for whom there is considerable variation in the play-level accuracy but for whom there are no noticeable outliers. When there is an outlier among an author's plays, there is usually (though not always) an identifiable reason for why that play stylistically differs from the rest of the author's work.

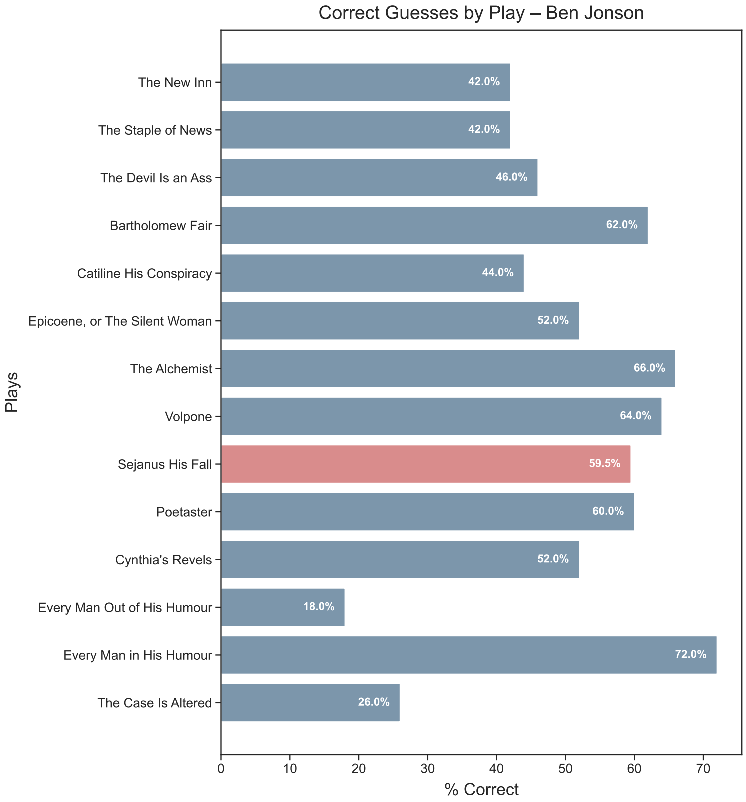

Figure 6: Percentage of samples correctly attributed for each play by Ben Jonson. The pink bar represents the play withheld from training.

A representative example of this can be seen in Ben Jonson's plays. Samples from all but two of Jonson's plays are correctly attributed more than 40% of the time, including those from the withheld play (Figure 6). However, only 18% of samples from *Every Man Out of His Humour* and 26% of samples from *The Case is Altered* are attributed to Jonson, causing *Every Man Out of His Humour* to be misattributed by the model. Both of these plays differ from Jonson's typical work. Although *Every Man Out of His Humour* was advertised as a sequel to the well-received *Every Man in His Humour*, it is very different from the original play [17]. It was the longest play written for a public theater performance during the Elizabethan era and was very poorly received. After its failure, Jonson began writing for private theaters instead [2]. Thus, it is likely that the play marks a stylistic experiment within Jonson's work. Interestingly, this play is still correctly attributed by sample-level cosine delta with 16% of samples. *The Case is Altered* is unique because it is the earliest surviving of Jonson's plays. Jonson excluded it from his collected works when they were first published and, even when it was eventually published in 1609, his name was only included in some copies [16]. *The Case is Altered* therefore likely represents an early work which the author was not proud of, and from whose style he matured.

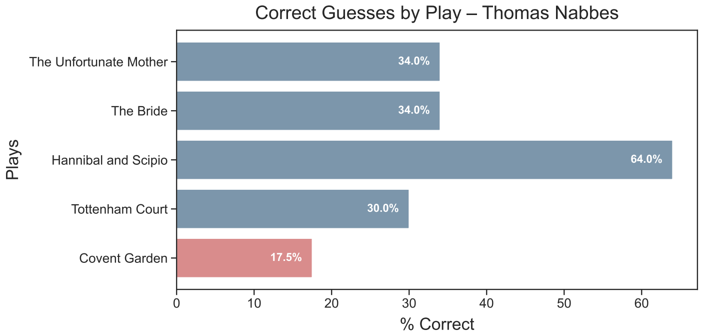

Figure 7: Percentage of samples correctly attributed for each play by Thomas Nabbes. The pink bar represents the play withheld from training.

Another outlier is *Covent Garden* by Thomas Nabbes. Although there is some play-level variation in Nabbes' attribution accuracy, *Covent Garden* is the only play for whom the model correctly attributes less than 20% of samples and the only one it misattributes, assigning Brome 23% of samples (Figure 7). While this variation may be in part because *Covent Garden* was withheld from training, the underlying reason for the model's confusion is likely that Nabbes' *Covent Garden* was written as a direct response to Richard Brome's play *The Weeding of Covent Garden*, which is also included in the dataset. There are likely named entities that cross-over between these two works and there may even be stylistic similarities. Sample-level cosine delta correctly attributes *Covent Garden* to Nabbes, but with only 13% of samples. It assigns 10% of samples to Brome.

We also see that the model performs poorly on the withheld plays of almost all authors with only three works in the corpus. For four of these five authors, less than 5% of samples from withheld plays are correctly attributed. The only deviation from this pattern is Robert Wilson; 19% of samples from the withheld Wilson play are correctly attributed. However, the withheld Wilson play is a prequel to one included in the training set. Thus, the model has more knowledge of this play than it would otherwise. It appears that including two plays by an author, or 470 samples, in the training data is not sufficient for the model to learn to extrapolate an author's style to an unseen text. It thus suggests a boundary for how much data may be needed for LLMs to be used for authorship attribution.

6.1. Disputed and Co-Authorship

We also asked the `t5-large` model predict the author of samples from 23 plays which are of disputed authorship or which are believed to be co-authored, although they were labeled as written by a single author in the corpora we drew from. We determined which plays were co-authored or of disputed authorship using the Oxford National Dictionary of Biography, which provides a detailed biography for each author in this corpus.

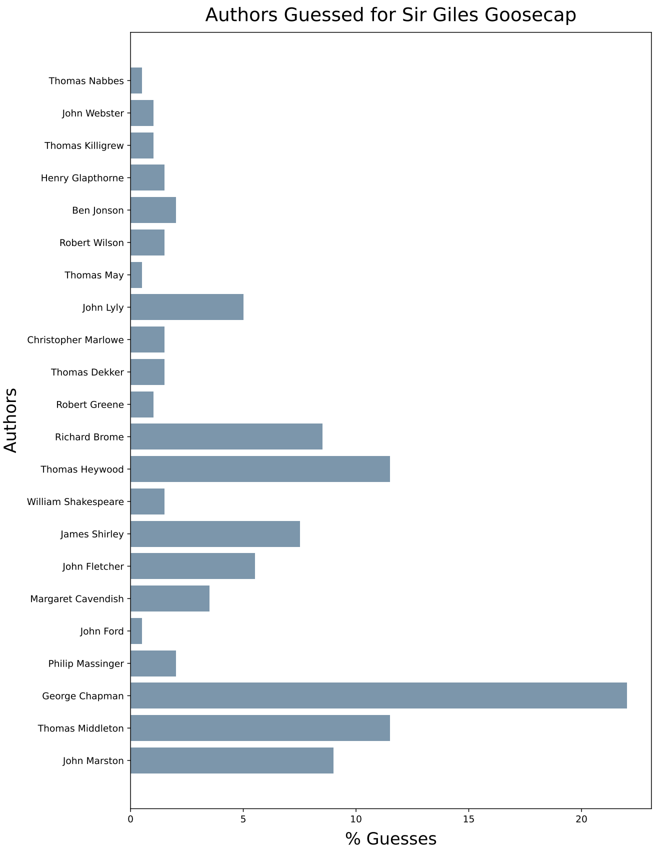

Figure 8: Percentage of samples attributed to authors for *Sir Giles Goosecap*.

Overall, we found that the model greatly struggles to make clear attributions for plays that were co-authored or of disputed authorship unless Shakespeare was a contributor. The only exception is *Sir Giles Goosecap*, which is hypothesized to have been written by George Chapman. The model attributes 22% of samples from *Sir Giles Goosecap* to Chapman (Figure 8). This is comparable to two other plays by Chapman in the original dataset: *All Fools*, which was withheld from training and from which 13% of samples are correctly attributed, and *The Widow's Tears*, from which 24% of samples are correctly attributed. Thus, the model results support the overall attribution of this play to Chapman. Sample-level cosine delta does not support this attribution, assigning only 5.5% of samples to Chapman. For no other non-Shakespearian play in this subcorpus is there enough evidence to argue for an attribution or co-attribution to authors in the dataset. It is particularly difficult to make assumptions about plays that are suspected to have been written by authors for whom the model's performance on the initial corpus is low. Even if these authors are only attributed a small proportion of samples from a play, these results are often comparable to those for their plays in the original dataset, meaning no conclusion can be reached.

Co-authorship also confused the model, particularly plays that were co-written with authors outside of the original corpus. The model often attributed a large proportion of quotes from these plays to Heywood. However, since there is no evidence that Heywood helped to author these plays, it is likely that this is an artifact of scapegoating. This trend also means that it is difficult to attribute plays to Heywood. 22% of samples from *The Fair Maid of the Exchange*, which Heywood is suspected to have co-authored, are attributed to him. However, a comparable proportion of samples are attributed to Heywood for multiple other plays in this dataset, meaning that we cannot use this as evidence for his authorship. This is a clear example of a case in which the model's misattribution patterns detrimentally affect its usability. The results are confusing even for plays from whom all of the suspected contributors are in the dataset, such as *The Laws of Candy*.

Thus, the results for non-Shakespearian plays provide little evidence for or against certain writers' authorship. While sample-level cosine delta appears to have no clear advantage over `t5-large` in attributing samples from these plays, the two methods attribute samples in very different ways. In some cases, `t5-large` more strongly attributes a play to its suspected author, and in others sample-level cosine delta does. Often the models attributed samples to different subsets of authors.

A very interesting pattern emerges when we look at the plays co-authored by Shakespeare in this test corpus. Over 50% of samples from each of the eight plays that Shakespeare contributed to are attributed to him, with little to no samples attributed to those who he supposedly co-authored the plays with, even if they are in the dataset. The most significant indication we see of another author's contribution to one of these plays is for *The Two Noble Kinsmen*. Here, only 54% of samples are attributed to Shakespeare and 8.5% are attributed to Fletcher, with whom he wrote the play. However, this is still not a strong signal of Fletcher's involvement. This pattern again suggests that the `t5-large` model recognizes Shakespeare from pre-training. If the model had seen these plays attributed solely to Shakespeare during pre-training, as is likely, it may help explain why it assigns them so confidently to Shakespeare despite the influence of other authors. In contrast, sample-level cosine delta never assigns more than 25% of samples from any of these plays to Shakespeare, and the presence of his theorized co-authors is much more prominent in the results.

7. Stylistic Development Over Time

In addition to the narrower task of author attribution, a measure of stylometric similarity can also be used to quantify authors' influence. To study shifts in dramatic style over time, we created a comparative corpus of plays written between the 14th and 18th centuries. In this

Table 4
The number of plays and authors by century in the comparison dataset.

Century	# Plays	# Authors
1500s	14	13
1600s	61	53
1700s	18	16
1800s	18	14
1900s	29	13

corpus, we included 74 plays gathered from the EMED and SHC corpora not written by authors in our training dataset. To these, we added 67 additional plays from Project Gutenberg (see Table 4). We performed the same utterance separation and splitting with these plays as with the original corpus and formatted the input and output pairs identically. Further details can be found in the appendix. The `t5-large` model fine-tuned on the original dataset was asked to predict authors for 200 samples from each of these plays. The percentages we report in Figure 9 are averaged by the original text author instead of by play; for example, we calculate the percentage of samples attributed to Heywood from each author writing in the 1500s and then average those percentages to reach the depicted value. This was to prevent any single writer whose style may somehow mimic that of an author in our original dataset from skewing the results.

In the 1500s and 1600s, the greatest proportion of samples are assigned to Thomas Heywood. This aligns with the scapegoating trends we saw in the original corpus. However, starting in the 1700s the greatest proportion of samples are assigned to Shakespeare, and this value increases in the 1800s and 1900s (Figure 9), for which nearly half of the samples from each author were attributed to Shakespeare. In addition, if we attribute plays to an author by majority vote, no plays in the 1500s are assigned to Shakespeare, but 97% of plays are attributed to him by the 1900s. This result does not imply that 20th century plays are similar to Shakespeare, only that of the Early Modern authors known to the model, Shakespeare is both distinct and increasingly more similar to more recent plays than any other Early Modern author.

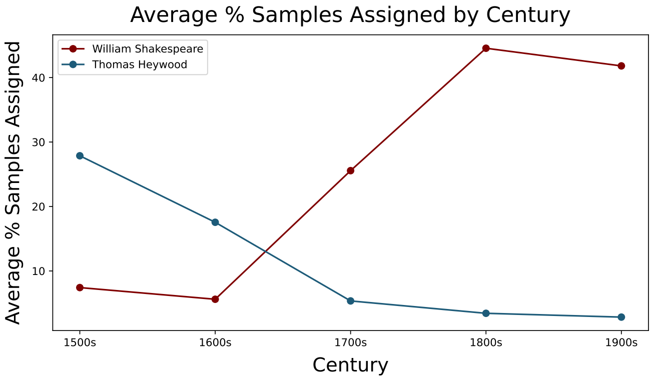

Figure 9: The average % of samples assigned to Thomas Heywood and William Shakespeare for samples from each century in the comparative corpus.

8. Conclusion

Generative large language models provide a promising tool for stylometry. While simpler methods such as cosine delta remain more accurate for larger text segments, we find that LLMs, particularly at larger scales, are remarkably effective at predicting the author of a difficult corpus of short 5–450 word text segments, which are more aligned with LLMs' shorter input windows.

In addition to quantitative power, LLM-based stylometric analysis provides evidence for a range of interpretive arguments both when it succeeds (such as with Margaret Cavendish) as well as when it fails (both in scapegoating and in the stylistic differences in the work of Ben Jonson). Because T5 demonstrates an ability to recognize style, it may prove useful in other situations where recognizing implicit signals is key such as tracking genre differences and stylistic movements. There are also substantial practical advantages to using fine-tuned LLMs: despite their complexity and computational intensity, generative LLMs provide a remarkably simple text-in/text-out user interaction that requires no specialized software.

However, there are several disadvantages to using pre-trained LLMs for authorship attribution. They are more computationally intensive than more traditional methods of authorship attribution and the content and effect of pre-training corpora are difficult to assess. In addition, the ways in which the model confidently misattributes texts means that it is more likely to produce misleading results than traditional attribution methods. Given the differences that emerged between the performance of cosine delta and the fine-tuned LLM, using the two methods in conjunction may provide more accurate results than using either method separately. Due to the weaknesses we have observed, however, we recommend against using LLMs for authorship attribution in forensic or legal settings.

Acknowledgments

We would like to thank Federica Bologna, Katherine Lee, Noam Ringach, Rosamond Thalken, Andrea Wang, Matthew Wilkens, and Gregory Yauney for their thoughtful feedback. This work was supported by the NEH project AI for Humanists and Cornell University's Hopcroft Fellowship.

References

[1] A. S. Arefin, R. Vimieiro, C. Riveros, H. Craig, and P. Moscato. "An Information Theoretic Clustering Approach for Unveiling Authorship Affinities in Shakespearean Era Plays and Poems". In: *PLoS ONE* 9.10 (2018), e111445. doi: 10.1371/journal.pone.0111445.

[2] A. Augustyn. *Authors of the Medieval and Renaissance Eras, 1100 to 1660.* New York, New York: Encyclopaedia Britannica, Inc, 2014.

[3] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps. "Time-aware authorship attribution for short text streams". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* Santiago, Chile, 2015. doi: 10.1145/2766462.2767799.

[4] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. "Pythia: A suite for analyzing large language models across training and scaling". In: *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, Hawaii, USA, 2023, pp. 2397–2430. doi: 10.48550/arXiv.2304.01373.

[5] M. Brown, M. Poston, and E. Williamson. *A Digital Anthology of Early Modern English Drama*. Online corpus. N/a. URL: https://emed.folger.edu.

[6] J. Burrows. "'Delta': a measure of stylistic difference and a guide to likely authorship". In: *Literary and linguistic computing* 17.3 (2002), pp. 267–287. DOI: 10.1093/llc/17.3.267.

[7] A. B. Duque, F. d. A. T. de Carvalho, and R. Vimieiro. "A Multiview Clustering Approach for Mining Authorial Affinities in Literary Texts". In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. Salvador, Bahia, Brazil, 2019. DOI: 10.1109/bracis.2019.00146.

[8] M. Eisen, A. Ribeiro, S. Segarra, and G. Egan. "Stylometric analysis of Early Modern period English plays". In: *Digital Scholarship in the Humanities* 33.3 (2018), pp. 500–528. DOI: 10.1093/llc/fqx059.

[9] D. Freebury-Jones. "Unsound deductions in early modern attribution: the case of Thomas Watson". In: *ANQ: A Quarterly Journal of Short Articles, Notes and Reviews* 33.2-3 (2020), pp. 164–171. DOI: 10.1080/0895769x.2019.1612231.

[10] J. Huertas-Tato, A. Huertas-Garcia, A. Martin, and D. Camacho. *PART: Pre-trained Authorship Representation Transformer*. arXiv paper. 2022. DOI: 10.48550/arxiv.2209.15373.

[11] M. Koppel, J. Schler, and S. Argamon. "Authorship attribution in the wild". In: *Language Resources and Evaluation* 45 (2011), pp. 83–94. DOI: 10.1007/s10579-009-9111-2.

[12] R. Layton, P. Watters, and R. Dazeley. "Authorship attribution for twitter in 140 characters or less". In: *2010 Second Cybercrime and Trustworthy Computing Workshop*. Ballarat, Victoria, Australia, 2010. DOI: 10.1109/ctc.2010.17.

[13] M. Mueller. *Shakespeare His Contemporaries: a corpus of Early Modern Drama 1550-1650*. Online corpus. 2015. URL: https://github.com/JonathanReeve/corpus-SHC.

[14] L. M. Naeni, H. Craig, R. Berretta, and P. Moscato. "A novel clustering methodology based on modularity optimisation for detecting authorship affinities in Shakespearean era plays". In: *PloS ONE* 11.8 (2016), e0157988. DOI: 10.1371/journal.pone.0157988.

[15] M. Najafi and E. Tavan. "Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis". In: *CLEF 2022 – Conference and Labs of the Evaluation Forum*. Belmeloro University Complex, Bologna, Italy, 2022. DOI: 10.18653/v1/w17-4914.

[16] E. H. C. Oliphant. "Problems of Authorship in Elizabethan Dramatic Literature". In: *Modern Philology* 8.3 (1911), pp. 411–459. DOI: 10.1086/386843.

[17] *Oxford Dictionary of National Biography*. Oxford, England: Oxford University Press, 2004.

[18] J. Pagel, N. Sihag, and N. Reiter. "Predicting Structural Elements in German Drama". In: *Proceedings of the Second Conference on Computational Humanities Research*. Monastery of the Grauwzusters, Antwerp, Belgium, 2021. DOI: 10.2307/1145292.

[19] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only". In: (2023).

[20] P. Plecháč. "Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns". In: *Digital Scholarship in the Humanities* 36.2 (2021), pp. 430–438. DOI: 10.1093/llc/fqaa032.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning* 21.1 (2020), pp. 5485–5551. DOI: 10.48550/arxiv.1910.10683.

[22] P. Rizvi. "The problem of microattribution". In: *Digital Scholarship in the Humanities* 34.3 (2019), pp. 606–615. DOI: 10.1093/digitalsh/fqy066.

[23] O. A. Rosso, H. Craig, and P. Moscato. "Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers". In: *Physica A: Statistical Mechanics and its Applications* 388.6 (2009), pp. 916–926. DOI: 10.1016/j.physa.2008.11.018.

[24] P. W. Smith and W. Aldridge. "Improving authorship attribution: optimizing Burrows' Delta method". In: *Journal of Quantitative Linguistics* 18.1 (2011), pp. 63–88. DOI: 10.1080/09296174.2011.533591.

[25] G. Taylor. "Empirical Middleton: Macbeth, Adaptation, and Microauthorship". In: *Shakespeare Quarterly* 65.3 (2014), pp. 239–272. DOI: 10.1353/shq.2014.0030.

[26] G. Taylor and J. V. Nance. "Imitation or collaboration? Marlowe and the early Shakespeare canon". In: *Shakespeare, Origins and Originality* 68 (2015), pp. 32–47. DOI: 10.1017/cbo9781316258736.003.

[27] T. Wood. *faststylometry: Burrows Delta*. 2021. URL: https://github.com/fastdatascience/faststylometry.git.

[28] H. Yoo, J. Jin, J. Son, J. Bak, K. Cho, and A. Oh. "HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States, 2022. DOI: 10.18653/v1/2022.findings-naacl.140.

[29] A. Zehe, L. Konle, L. K. Dümpelmann, E. Gius, A. Hotho, F. Jannidis, L. Kaufmann, M. Krug, F. Puppe, N. Reiter, et al. "Detecting scenes in fiction: A new segmentation task". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. online, 2021. DOI: 10.18653/v1/2021.eacl-main.276.

[30] H. Zhang, P. Nie, Y. Wen, and X. Yuan. "Authorship attribution for short texts with author-document topic model". In: *Knowledge Science, Engineering and Management: 11th International Conference, KSEM 2018*. Changchun, China, 2018. DOI: 10.1007/978-3-319-99365-2_3.

A. T5 Fine-Tuning Hyperparameters

Parameter	**Value**
Evaluation Strategy	Epoch
Learning Rate	2x10^{-5}
Weight Decay	0.01
Save Total Limit	3

B. Examination of Z-Score Uniqueness

To explore the validity of our uniqueness metric, we ran 1,000 synthesized trials to examine what the expected correlation between dataset contribution and the uniqueness metric would be given randomly assigned plays. Concretely, in each trial we randomly assigned plays to synthetic authors in the same proportions they are assigned to authors in our true dataset. We then calculated the Spearman's rho correlation between number of plays and uniqueness values for each trial. We plot the binned synthetic correlations and the true correlation from our dataset in Figure 10. The true correlation from our dataset, depicted with the vertical red line, is 0.2 away from any value reached in our synthesized trial. Thus, it seems that there are some notable deviations from the expected trend in our dataset.

To further explore this relationship, we averaged the uniqueness values for each synthetic author over all trials and plotted these values as well as the true values in Figure 11. It is

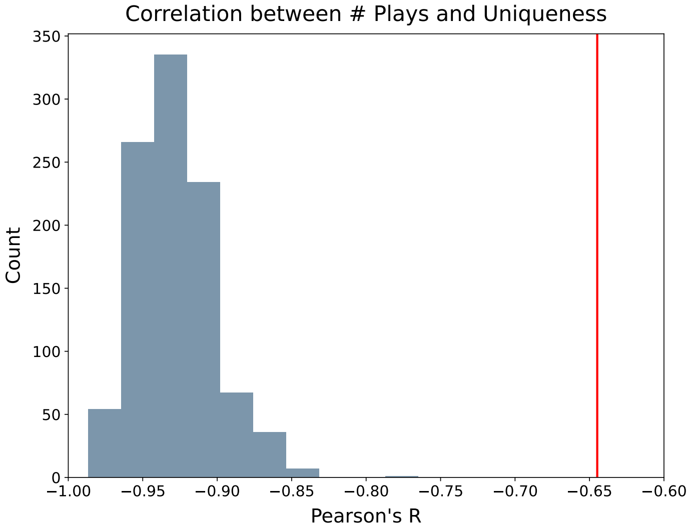

Figure 10: Binned correlations between uniqueness scores and number of plays from each synthesized trial. The red line represents the correlation from our true dataset.

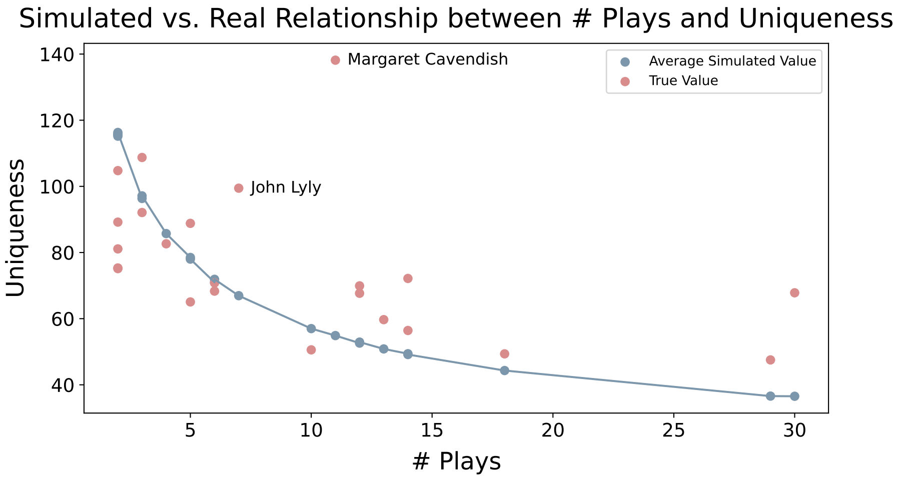

Figure 11: The relationship between the averaged uniqueness values for each synthetic author (blue) and the number of plays and the true uniqueness values from the corpus (red) and the number of plays.

clear that the true uniqueness values frequently deviate from the expected relationship between uniqueness and number of plays. In particular, Margaret Cavendish and John Lyly have much higher uniqueness values than expected given the number of plays they contribute to the training dataset. Because of this, we believe that this metric represents a valuable measure of uniqueness and does not simply reemphasize the impact of contribution to the training corpus.

C. Original Corpus Contents

All plays in our original training and test corpora by author. The withheld plays are bolded and italicized.

Author	**Plays**
Richard Brome	The Northern Lass, The City Wit or The Woman Wears the Breeches, The Queen's Exchange (The Royal Exchange), The Weeding of Covent Garden or The Middlesex Justice of Peace, The Novella, The Queen and Concubine, The New Academy or The New Exchange, The Sparagus Garden (Tom Hoydon o' Tanton Deane), ***The English Moor or The Mock Marriage***, The Antipodes, The Damoiselle or The New Ordinary, A Mad Couple Well Matched, The Lovesick Court or The Ambitious Politic, The Court Beggar, A Jovial Crew or The Merry Beggars
Margaret Cavendish	The Lady — Part 1, The Lady — Part 2, The Unnatural Tragedy, Wit's Cabal — Part 1, Wit's Cabal — Part 2, Love's Adventures — Part 1, Love's Adventures — Part 2, Several Wits, The Matrimonial Trouble — Part 1, The Matrimonial Trouble — Part 2, The Religious, ***The Wooers***

George Chapman	The Blind Beggar of Alexandria, A Humorous Day's Mirth, ***All Fools***, The Gentleman Usher, May Day, The Widow's Tears, Bussy D'Ambois, Monsieur D'Olive, Caesar and Pompey (The Wars of Caesar and Pompey), The Tragedy of Charles Duke of Byron, The Revenge of Bussy D'Ambois
William Davenant	The Cruel Brother, Albovine King of the Lombards, The Just Italian, ***The Wits***
Thomas Dekker	Old Fortunatus, Satiromastix or The Untrussing of the Humorous Poet, ***The Honest Whore — Part 2***, Match Me in London, If It Be Not Good the Devil Is in It
John Fletcher	The Faithful Shepherdess, The Woman's Prize or The Tamer Tamed, Bonduca, Valentinian, The Mad Lover, The Chances, The Loyal Subject, The Humorous Lieutenant (Generous Enemies, Demetrius and Enanthe), Women Pleased, ***The Island Princess***, The Wild Goose Chase, The Pilgrim, Rule a Wife and Have a Wife, A Wife for a Month
John Ford	The Lover's Melancholy, The Broken Heart, 'Tis Pity She's a Whore, ***Love's Sacrifice***, Perkin Warbeck, The Fancies Chaste and Noble
Henry Glapthorne	The Hollander, Ladies' Privilege, ***Wit in a Constable***
Robert Greene	Friar Bacon and Friar Bongay, ***The Scottish History of James the Fourth***, Orlando Furioso
Thomas Heywood	The Four Prentices of London, Edward IV — Part 1, Edward I — Part 2, ***The Royal King and the Loyal Subject***, How a Man May Choose a Good Wife from a Bad, A Woman Killed with Kindness, If You Know Me Not You Know Nobody or The Troubles of Queen Elizabeth — Part 1, If You Know Me Not You Know Nobody or The Troubles of Queen Elizabeth — Part 2, The Fair Maid of the West or A Girl Worth Gold — Part 1, The Wise Woman of Hogsdon, The Rape of Lucrece, The Golden Age or The Lives of Jupiter and Saturn, The Brazen Age, The Iron Age — Part 1, The Iron Age — Part 2, The English Traveller, Love's Mistress, A Challenge for Beauty
Ben Jonson	The Case is Altered, Every Man in His Humour, Every Man Out of His Humour, Cynthia's Revels, Poetaster, ***Sejanus His Fall***, Volpone, The Alchemist, Epicoene or The Silent Women, Catiline His Conspiracy, Bartholomew Fair, The Devil is an Ass, The Staple of News, The New Inn
Thomas Killigrew	The Prisoners, The Princess, The Parson's Wedding, ***Claricilla***

John Lyly	***Sappho and Phao***, Campaspe (Alexander, Campaspe, and Diogenes), Gallathea, Endymion, Midas, Love's Metamorphosis, Mother Bombie, The Woman in the Moon
Christopher Marlowe	***Tamburlaine the Great — Part 1***, Tamburlaine the Great — Part 2, The Jew of Malta, Doctor Faustus, Edward the Second, THe Massacre at Paris
John Marston	Antonio and Mellida, Antonio's Revenge, Jack Drum's Entertainment, What You Will, ***The Malcontent***, Parasitaster or The Fawn, The Dutch Courtesan
Philip Massinger	The City Madam, ***The Duke of Milan***, The Maid of Honour, The Bondman, The Unnatural Combat, The Renegado or The Gentleman of Venice, A New Way to Pay Old Debts, The Roman Actor, The Great Duke of Florence, The Picture, The Emperor of the East, The Guardian, The Bashful Lover
Thomas May	***The Heir***, Cleopatra — Queen of Egypt, Julia Agrippina — Empress of Rome
Thomas Middleton	The Phoenix, Michaelmas Term, A Trick to Catch the Old One, A Mad World My Masters, The Puritain or The Widow of Watling Street, ***Your Five Gallants***, The Widow, The Mayor of Quinborough, A Chaste Maid in Cheapside, More Dissemblers Beside Women, Women Beware Women, A Game at Chess
Thomas Nabbes	***Covent Garden***, Tottenham Court, Hannibal and Scipio, The Bride, The Unfortunate Mother
William Shakespeare	The Comedy of Errors, Richard III, The Taming of the Shrew, The Two Gentlemen of Verona, Romeo and Juliet, Richard II, King John, The Merchant of Venice, Henry IV — Part 1, Henry IV — Part 2, Much Ado About Nothing, Henry V, Julius Caesar, As You Like It, Twelfth Night, Hamlet, Merry Wives of Windsor, Troilus and Cressida, Othello, Measure for Measure, Macbeth, King Lear, ***Antony and Cleopatra***, Coriolanus, Cymbeline, The Tempest

James Shirley	The School of Compliment, The Maid's Revenge, The Wedding, The Witty Fair One, The Grateful Servant, The Humorous Courtier, Love's Cruelty, The Ball, The Traitor, Hyde Park, Changes or Love in a Maze, The Bird in a Cage (The Beauties), The Young Admiral, The Gamester, The Opportunity, The Example, The Lady of Pleasure, The Coronation, The Duke's Mistress, The Royal Master, The Doubtful Heir, The Constant Maid, The Gentleman of Venice, Saint Patrick for Ireland — Part 1, The Politician, The Arcadia, The Imposter, ***The Sisters***, The Cardinal, The Brothers, The Court Secret
John Webster	The White Devil (Vittoria Corombona), ***The Duchess of Malfi***, The Devil's Law Case (When Women Go to Law the Devil is Full of Business)
Robert Wilson	***The Three Ladies of London***, The Three Ladies of London, The Cobbler's Prophecy

D. Disputed and Co-Authored Corpus Contents

All plays in the disputed and co-authored corpus by the author they were attributed to in the original corpora.

Labeled Author	**Plays**
George Chapman	Sir Giles Goosecap, Two Wise Men and All the Rest Fools
Thomas Dekker	Patient Grissel, The Wonder of a Kingdom
John Ford	The Laws of Candy, The Queen
Henry Glapthorne	Revenge for Honor (The Parricide)
Robert Greene	George a Green the Pinner of Wakefield
Thomas Heywood	The Fair Maid of the Exchange
John Marston	Histriomastix or The Player Whipped, The Insatiate Countess
Thomas Middleton	Anything for a Quiet Life, The Family of Love
William Shakespeare	Henry VI — Part 1, Henry VI — Part 2, Henry VI — Part 3, Henry VIII, Pericles — Prince of Tyre, Timon of Athens, Titus Andronicus, The Two Noble Kinsmen
John Webster	Appius and Virginia, The Thracian Wonder

E. Comparison Corpus Contents

All plays in the comparative corpus by author. Plays that were attributed to Shakespeare by the model are bolded and italicized.

Author	Plays
Robert Armin	The Two Maids of More-Clacke
Thomas Baker	The Fine Lady's Airs
James Nelson Barker	***The Indian Princess***
J. M. Barrie	***Dear Brutus***, ***Peter Pan***
Lording Barry	Ram Alley
Barnabe Barnes	The Devil's Charter
Clifford Bax	***Square Pegs***
Francis Beaumont	The Knight of the Burning Pestle
Dabridgecourt Belchier	Hans Beer-Pot (See Me and See Me Not)
Arnold Bennett	***The Great Adventure***
William Berkeley	The Lost Lady
Hugh Henry Brackenridge	The Battle of Bunkers Hill
Alexander Brome	The Cunning Lovers
Robert Browning	***A Blot in the Scutcheon***
Henry Burnell	Landgartha
Lodowick Carlell	The Deserving Favorite
Richard Claude Carton	***Lady Huntworth's Experiment***
William Cartwright	The Royal Slave
William Cavendish	The Country Captain, The Variety
Susanna Centlivre	The Busie Body, The Perjur'd Husband

Robert Chamberlain	The Swaggering Damsel
George Coleman	***John Bull***
Abraham Cowley	Love’s Riddle
Aleister Crowley	***Household Gods***
Robert Daborne	A Christian Turned Turk
John Denham	The Sophy
Thomas Drue	The Duchess of Suffolk
William Dunlap	***Andre***
Lord Dusany	***If***
Nathan Field	Amends for Ladies, A Woman is a Weathercock
Jasper Fisher	Fuimus Troes (The True Trojans)
Phineas Fletcher	Sicelides
Ralph Freeman	Imperiale
John Galsworthy	***A Bit O’ Love***, ***The Eldest Son***, ***A Family Man***, ***The First and the Last***, ***The Foundations***, ***The Fugitive***, ***Joy***, ***Justice***, ***The Little Dream***, ***The Little Man***, ***Loyalties***, ***The Mob***, ***The Skin Game***, ***Strife***
Thomas Godfrey	The Prince of Parthia
Johann Wolfgang von Goethe	***Faust***
John Gough	The Strange Discovery
Fulke Greville	Alaham
John Johns	Adrasta
William Kemp	A Knack to Know a Knave
Henry Killigrew	The Conspiracy
John Kirke	The Seven Champions of Christendom
James Sheridan Knowles	***The Love Chase***
Thomas Kyd	Soliman and Perseda, The Spanish Tragedy (Hieronimo is Mad Again)
Maurice Kyffin	Andria

William Habington	The Queen of Aragon
Samuel Harding	Sicily and Naples
Joseph Harris	The City Bride
William Haughton	Englishmen for My Money
Peter Hausted	The Rival Friends
William Hawkins	Apollo Shroving
Gorges Edmond Howard	***The Female Famester***
Henrik Ibsen	***A Doll's House***, Hedda Gabler
Elizabeth Inchbald	***Such Things Are***, ***The Widow's Vow***
Jerome K. Jerome	***Fanny and the Servant Problem***, ***Woodbarrow Farm***
Henry Arthur Jones	***Dolly Reforming Herself***, ***Michael and His Lost Angel***
D. H. Lawrencee	***Touch and Go***
John Leacock	***The Fall of British Tyranny***
Thomas Lodge	The Wounds of Civil War
Samuel Low	***The Politician Out-Witted***
Sir William Lower	The Phoenix in Her Flames
Thomas Lupton	All for Money
James Mabbe	The Spanish Bawd (Calisto and Meliboea)
Charles Macklin	The Covent Garden Theatre
Gervase Markham	The Dumb Knight, Herod and Antipater
Shakerley Marmion	A Fine Companion, Holland's Leaguer
John Mason	The Turk
Jasper Mayne	The City Match
Edward Moore	***The Gamester***
Thomas Morton	***Speed the Plough***
Arthur Murphy	***The Grecian Daughter***

Thomas Newman	The Andrian Woman (Andria), The Eunuch
Mordecai Manuel Noah	***She Would Be a Soldier***
John O’Keeffe	***Wild Oats***
Henry Nevil Payne	The Fatal Jealousie
Arthur Pinero	***The Big Drum***, The ’Mind the Paint’ Girl
Henry Porter	The Two Angry Women of Abingdon
Thomas Randolph	***The Jealous Lovers***
Thomas Rawlins	The Rebellion
Nathaniel Richards	Messalina — The Roman Empress
Robert Rogers	Ponteach: The Savages of America
Edmond Rostand	***Cyrano de Bergerac***
Samuel Rowley	The Noble Spanish Soldier (The Noble Soldier or A Contract Broken Justly Revenged), When You See Me You Know Me (Henry the Eighth)
Joseph Rutter	The Shepherds’ Holiday
S. S.	The Honest Lawyer
W. S.	Thomas Lord Cromwell
Edward Sharpham	Cupid’s Whirligig, The Fleer
George Bernard Shaw	***Arms***, ***The Devil’s Disciple***, ***Fanny’s First Play***, ***Man and Superman***
John Stephens	Cynthia’s Revenge
William Stevenson	Gammer Gurton’s Needle
Algernon Charles Swinburne	***The Duke of Gandia***, ***Erechtheus***, ***Rosamund***
Robert Tailor	The Hog Hath Lost His Pearl
Brandon Thomas	***Charley’s Aunt***
Thomas Tomkis	Albumazar, Lingua or The Combat of the Tongue and the Five Senses of Superiority
Cyril Tourneur	The Atheist’s Tragedy

Royall Tyler	***The Contrast***
Nicolas Udall	Ralph Roister Doister
George Wapull	The Tide Tarrieth No Man
Oscar Wilde	***Vera***, ***A Woman of No Importance***
George Wilkins	The Miseries of Enforced Marriage
Nathaniel Woodes	The Conflict of Conscience
Robert Yarington	Two Lamentable Tragedies
Richard Zouch	The Sophister