

Detecting the Aura of an Article – Was the Original Manuscript Seen?

Thomas Asselborn*, Magnus Bender^{§¶}, Ralf Möller[†], Sylvia Melzer^{*‡}

* University of Lübeck, Institute for Information Systems, Ratzeburger Allee 160, 23562 Lübeck, Germany
Email: {thomas.asselborn, sylvia.melzer}@uni-luebeck.de
0009-0005-3011-7626, 0000-0002-0144-5429

§ Aarhus University, Department of Management, Universitetsbyen 61, 8000 Aarhus C, Denmark
Email: magnus@mgmt.au.dk, 0000-0002-1854-225X

¶ Aarhus University, Center for Contemporary Cultures of Text, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark

† University of Hamburg, Institute for Humanities-Centered AI, Warburgstraße 28, 20354 Hamburg, Germany
Email: ralf.moeller@uni-hamburg.de, 0000-0002-1174-3323

‡ University of Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany

Abstract—This paper investigates whether scholarly texts reveal, through linguistic patterns, whether authors engaged directly with original artefacts or relied on facsimiles, digital or analogue. Drawing on debates about reproduction and materiality, it frames this question as a classification task and evaluates machine learning approaches using both synthetic and real-world data. While initial experiments under controlled conditions suggest that certain linguistic features may correlate with different modes of access, these patterns prove unstable in practice. Key challenges include the lack of reliable ground truth labels, the influence of domain-specific writing conventions, and multilingual variation. In particular, highly standardised descriptive language in specialised fields limits the discriminative power of textual features. The findings indicate that the task is fundamentally ill-posed under standard assumptions, highlighting the limits of data-driven methods in capturing epistemic conditions that are only partially encoded in text.

I. INTRODUCTION

THE increasing availability of digitised cultural artefacts has transformed scholarly research practices in the humanities. Large-scale digital collections promise new forms of access, enabling researchers to engage with materials that would otherwise remain geographically or materially inaccessible. At the same time, this shift raises a fundamental question: to what extent does direct engagement with an original artefact differ from interaction through digital reproductions, and can this distinction be detected in scholarly writing?

This question resonates with longstanding theoretical debates, most prominently articulated by Walter Benjamin, who argued that mechanical reproduction alters the “aura” of the artwork [3], and more recently revisited by Bruno Latour and Adam Lowe, who suggest that such qualities may instead be redistributed across networks of originals and reproductions. In parallel, current work on “digital twins” of artefacts emphasizes the epistemic role of digital representations in shaping how knowledge is produced and accessed [9].

From a computational perspective, these debates invite a seemingly straightforward question: can the mode of access to an artefact, direct or mediated, be inferred from the language

of scholarly texts? Standard supervised learning approaches, where texts are classified based on linguistic features, seem well suited to this task. However, as this case study demonstrates, this formulation rests on a set of assumptions that do not hold in practice.

This paper reports on an exploratory attempt to model this distinction using a combination of synthetic data generation and machine learning methods, including both linear classifiers and transformer-based architectures. Initial results obtained under controlled conditions suggested that certain linguistic patterns, particularly those associated with material description versus abstract or argumentative discourse, may correlate with different modes of engagement. However, these findings proved highly unstable when applied to real-world data.

Closer inspection revealed several fundamental challenges. First, the key variable of interest, whether an author had direct access to an original artefact, is typically not explicitly documented and cannot be reliably reconstructed from textual data alone. Second, domain-specific writing conventions, particularly in fields such as codicology, introduce highly standardised forms of material description that blur the distinction between direct and mediated observation. Third, the presence of unaccounted multilingual variation, including segments in languages such as Georgian, introduces additional sources of variability that can dominate model behaviour.

Together, these factors suggest that the task is not merely difficult but fundamentally ill-posed under standard machine learning assumptions. This case study provides an analysis of where such approaches work and where they break. In doing so, it highlights the limits of purely data-driven methods for capturing epistemic distinctions that are only partially encoded in textual form and depend on contextual and domain-specific knowledge.

The contribution of this paper is therefore twofold. Empirically, it demonstrates how different classes of models respond to domain-specific and multilingual signals, often relying on surface-level regularities rather than meaningful epistemic

distinctions. Conceptually, it argues that the difficulty of the task reflects a more in-depth issue: the conditions under which scholarly knowledge is produced are not fully transparent in the textual record and cannot be inferred without external annotation or domain expertise.

II. RELATED WORK

Past work has explored whether copies of original manuscripts influence how scholars perceive and analyse them. Building on this, D. F. McKenzie [11] reconceptualizes texts as material and social constructs, arguing that meaning is shaped not only by content but also by form, production, and use. From this perspective, the distinction between original and facsimile is not merely technical but interpretive, as each instantiation of a text generates its own conditions of meaning.

Within this broader framework, questions of originality and authenticity have been further examined in library and information science, where the relationship between original and reproduction is treated as context-dependent, not fixed [6].

A foundational account of this relationship is provided by Walter Benjamin, who introduces the concept of “aura” to describe the unique existence of an artwork in time and space, arguing that technologies such as photography and film erode this quality [3]. This position is subsequently challenged by Bruno Latour and Adam Lowe, who argue that in the context of digital facsimiles, aura is not simply lost but redistributed across networks of originals and reproductions [9].

The relevance of these questions is further reflected in current research agendas. For instance, the second phase of the Excellence Cluster “Understanding Written Artefacts” introduces a Concepts & Methods Unit titled “Digital Twins”, which explicitly addresses the epistemic implications of working with digital surrogates such as images or 3D scans instead of original manuscripts [5]. This development indicates that the shift toward digital representations is not merely technical but raises fundamental questions about interpretation, materiality, and scholarly practice.

III. METHOD

The principle of the task, to classify whether a researcher has seen the original artefact or not, seems to lead to rather simple solutions. This paper assesses two possibilities. First, we treat it as a binary classification: the researcher has either seen the original or the facsimile. Thus, a binary classifier can be trained in a supervised way. To do so, the training data need to have enough examples from both classes, namely “original seen” and “facsimile only”. What counts as enough depends on the dataset and its diversity in representing the classes of interest as well as the task of interest.

One of the problems when trying to verify the hypothesis that humanities scholars write differently about a manuscript if they saw the original in contrast to only having seen the (digital) facsimile is that there is not enough training data available to train classifiers. Thus, a way of generating synthetic data needs to be used. We experimented with using generative LLMs (Large Language Models) like ChatGPT to

generate data for us. The process is outlined in Fig. 1. See Appendix¹ for a more detailed description.

IV. APPLICATION AND RESULTS

This section focuses on the process of how we generated synthetic datasets, both for training and testing, and how both were used to train and test classifiers. Finally, the best two models were tested on a small dataset of real scientific articles.

A. Synthetic Data Generation

Because no sufficiently large labelled corpus was available, synthetic training data were generated using LLMs. To derive plausible distinctions between direct and mediated artefact access, several contemporary LLMs (full list¹) were asked whether scholars would write differently about artefacts examined in person versus through digital reproductions. Across models, recurring themes included materiality, sensory engagement, authenticity, contextual understanding, and descriptive richness.

These recurring themes were consolidated into a set of generation guidelines and used to create a balanced synthetic corpus. Using Llama 3.1 70B, 4,200 synthetic research articles were generated, comprising 2,100 examples for the class *original seen* and 2,100 examples for the class *digital only*. The prompts differed only in the assumed mode of access to the artefact, while explicitly instructing the model not to reveal this distinction directly. Instead, any differences were intended to emerge through subtle linguistic variation.

Generated articles contained approximately 1,000 words on average and were manually screened to remove malformed outputs before training.

B. Training the Classifiers

Having compiled the dataset, different classification models were trained, all with the following identical parameters:

- Total number of papers: 4200
 - Papers of class “original seen”: 2100
 - Papers of class “digital only”: 2100
- Papers held back for testing: 100 per class (200 in total)
- Of the 4000 papers for training, the train/validation split is 80%/20%, i.e.,
 - Papers for training: 1600 per class
 - Papers for validation during training: 400 per class

All training examples were pre-processed: n-grams like “digital”, “scan” or “reproduction”, among others, were removed to avoid the models relying on obvious features. Additionally, standard stopwords like “the”, “this”, or “and” were removed.

The classical machine learning models, i.e., decision tree, random forest, logistic regression, naïve Bayesian networks and SVM, were implemented using scikit-learn [12]. Note that we used the standard tokenizer for the models, i.e., a counting-based tokenizer for the naïve Bayesian network and a TF-IDF tokenizer for the other models. In terms of token lengths, both unigrams and bigrams were considered. Initial experiments

¹Paper Appendix: <https://doi.org/10.25592/uhhfdm.18819>

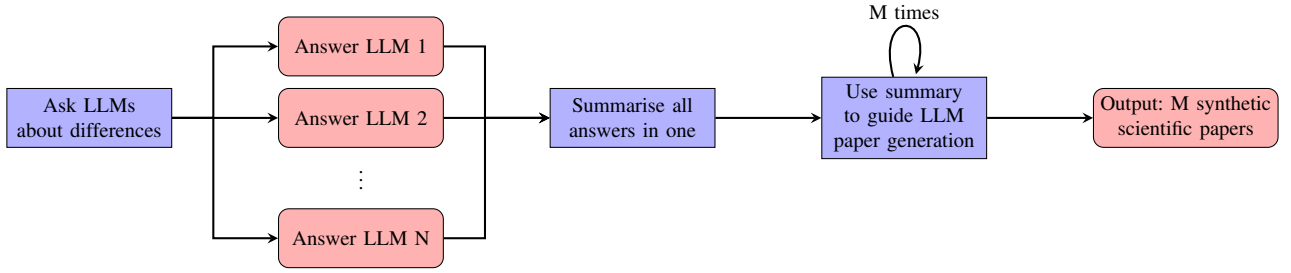


Fig. 1: Process to generate synthetic data for training classifiers by using N different LLMs.

also included trigrams to potentially capture more context, but the performance overall did not improve compared to using unigrams and bigrams alone.

For the transformer-based approaches, we have used a standard BERT [7] and a Longformer model [2]. The reason behind that is that while the classical BERT model² is faster during training and inference, its context window is much smaller than the total length of the synthetic research papers. One solution to that problem is to truncate the input to the maximum possible context window of the BERT model. Another solution is to use models with longer context windows, like the Longformer. This, however, significantly increases the hardware requirements for training and inference. To find a good balance, both models were tested here.

As for the Longformer, we fine-tuned the model `allenai/longformer-base-4096`³ and will, for brevity, now simply refer to it as Longformer. For fine-tuning a learning rate of 2×10^{-6} and 10 epochs were used for both transformer-based models. All other hyperparameters have been kept at default. Please keep in mind that the hyperparameters are not necessarily the best ones for this task but just a starting point for further analysis.

C. Results on the Synthetic Dataset

At first, quantitative measurements were taken to verify how well the models were trained on clues found in the synthetic dataset. To do so, the performance on the synthetic testing dataset was measured first. Table I shows the metrics for precision, recall, F1-score, and accuracy for all the models tested when reporting class “original”. Note that the test set is balanced, i.e., the number of examples per class is identical and thus, performance for the second class, “digital only”, is symmetric and omitted here. See [10] for information about the metrics.

The best overall precision is given by the Longformer model followed by the classical BERT with truncated inputs, i.e., the transformer-based models. The classical ML models are, in general, not far behind though, with logistic regression and random forest having the same precision score. The decision tree is the worst by far, which is not surprising given that the feature space for texts is typically huge and

sparse. Looking at the recall, the SVM is the best model with the Longformer being a close second. BERT, logistic regression and the random forest provide identical recall, with the naïve Bayesian network being second to last. As with precision, the decision tree comes in last again. For recall, one classical ML model, the SVM, was best overall, but the transformer-based models are still in the top three. Having a closer look at the F1-score and the accuracy, the overall order is similar with the transformer-based models providing a good overall score, the classical ML models being close behind with the decision tree being worse overall. Confusion matrices for the best transformer-based model, the Longformer, and the best overall classical ML model, the SVM, were generated. Fig. 2 shows the confusion matrix for the SVM (right) on the 200 testing samples. The confusion matrix indicates a fairly balanced but moderate error distribution, with a slight preference for the class “original seen” over “digital only”, and both false positives and false negatives occurring at non-negligible rates, suggesting no strong class-specific bias but limited discriminative performance overall. This requires more tuning in the future to truly find the best overall SVM classifier, which was not the focus of this paper. Fig. 2 also shows the confusion matrix for the Longformer (left). Compared to the SVM, the Longformer is more balanced overall, with discriminative performance still leaving room for improvement.

While the transformer-based models provided better results on the synthetic testing dataset, the classical ML models provide the benefit of being interpretable. We gathered the top twelve features for logistic regression, the naïve Bayesian network and SVM to get more qualitative results about which uni- and bigrams push the models into which class. Table II shows the top features for logistic regression. Looking at the features for the class “digital only”, n-grams like “this study”, “linguistic”, “medieval”, “facsimiles” indicate that the classifier classifies texts using analytical and academic writing as being closer to class “digital only”. In contrast, the features that push the model towards the class “original” are more about materiality, sensory experiences and physical descriptors like “physical characteristics”, “ink”, “parchment”, “binding”, “faint”, or “worn”. However, both top features also provide a strong indication that the model is taking shortcuts when classifying, because it boils down to “physical properties indicate original seen” and “academic writing indicates access

²<https://huggingface.co/google-bert/bert-base-uncased>

³<https://huggingface.co/allenai/longformer-base-4096>

TABLE I: Classification performance on the balanced test set (Original class reported).

Model	Precision	Recall	F1-score	Accuracy
BERT (Truncation)	0.753	0.730	0.741	0.745
Longformer	0.768	0.760	0.764	0.765
Decision Tree	0.608	0.590	0.599	0.605
Logistic Regression	0.730	0.730	0.730	0.730
Naïve Bayes (Count)	0.655	0.720	0.686	0.670
Random Forest	0.730	0.730	0.730	0.730
SVM	0.713	0.770	0.740	0.730

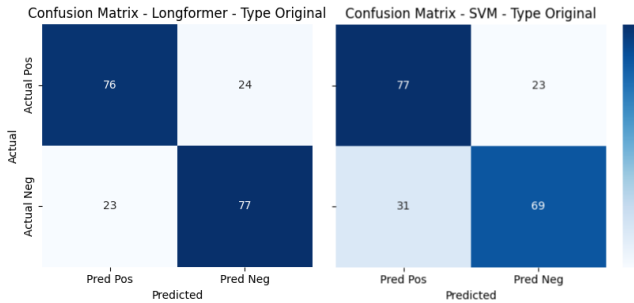


Fig. 2: Confusion matrices for Longformer (right) and SVM (left), both for the Original class.

to digital facsimiles only”. In a way, this is expected to a certain degree, though, because the way the LLM wrote the synthetic papers emphasises exactly those features.

The naïve Bayesian network’s top twelve features are shown in Table III. Compared to the logistic regression models, its learned features per class are mostly useless. The network trains on specific, rare tokens and names and assigns a huge weight to them. This will most likely lead to this model not generalising well to real-world data.

Finally, the top features for the SVM are considered in Table IV. As can be seen, the results are almost a carbon copy of what the logistic regression model already showed. The SVM learned that linguistic patterns describing physical properties indicate that the author has seen the original, while more distant, academic writing indicates that the author has only had access to a digital facsimile.

TABLE II: Top features for Logistic Regression. Positive weights push predictions toward class 1 (*original*); negative weights push predictions toward class 0 (*digital_only*).

Class 0: digital_only		Class 1: original	
Feature	Weight	Feature	Weight
while the	-1.2860	physical characteristics	+1.3954
despite	-0.9730	faint	+1.2237
despite its	-0.9467	the ink	+1.1557
while	-0.9257	worn	+1.1513
linguistic	-0.8755	almost	+1.1462
medieval	-0.8095	slightly	+1.1332

D. Results on the Catalogue of the Georgian Manuscripts

For our next experiments, we applied the classifiers to real-world texts. Namely, we tested them on texts from the

TABLE III: Top features for naïve Bayes. Positive log-odds push predictions toward class 1 (*original*); negative log-odds push predictions toward class 0 (*digital_only*).

Class 0: digital_only		Class 1: original	
Feature	Log-odds	Feature	Log-odds
marlowe manuscript	-4.2124	black book	+4.4771
the marlowe	-4.1971	avenant	+4.4199
physical access	-4.0658	wilde	+4.4081
athanasian	-4.0301	bos	+4.0548
the athanasian	-3.9348	the bos	+4.0026
ancrene	-3.9348	redwood	+3.9091

TABLE IV: Top features for linear SVM. Positive weights push predictions toward class 1 (*original*); negative weights push predictions toward class 0 (*digital_only*).

Class 0: digital_only		Class 1: original	
Feature	Weight	Feature	Weight
while the	-1.2791	physical characteristics	+1.2059
though	-0.9676	almost	+1.1231
despite	-0.8923	delicate	+1.0775
despite its	-0.8647	faint	+1.0732
while	-0.8488	subtle	+1.0694
form	-0.8351	worn	+1.0444

book “Catalogue of the Georgian Manuscripts” by Jikurashvili, Gippert, Outtier et al. [8]. This book contains articles about 93 Georgian codices, one scroll and 14 fragments published by the Iviron Monastery on Mount Athos. Overall, 95 articles are present in this book. In preparing the book, the authors travelled to the Iviron Monastery and thus had access to the original manuscripts.

Similar to the results for the synthetic dataset, we gathered results for this catalogue book. Because the book only contains texts from the class “original”, we report only the recall for this class as a measure of label retention rather than classification accuracy. Table V shows the results in decreasing order. It can be seen that all classical ML models, except the decision tree, perform very well on this dataset. One surprising result, however, is that the BERT model with the small context window and truncation of the input shows poor performance. This suggests that the beginning of most examples contains cues that are present in synthetic articles from authors having only had access to digital facsimiles. More work needs to be done to verify that.

To quantify uncertainty, we report 95% confidence intervals for proportions using the Wilson score interval [14], [4], which provides a reliable estimate of the true underlying probability under finite sample sizes. It is defined as

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}},$$

where $\hat{p}$ denotes the observed proportion, n the number of samples and $z = 1.96$ for a 95% confidence level.

Looking at the upper and lower bounds of the 95% Wilson confidence intervals, the differences between logistic regres-

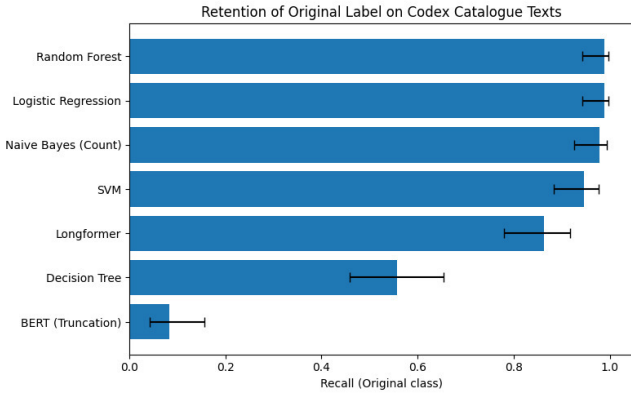


Fig. 3: Retention of the Original class on Codex catalogue texts (Original-only corpus, $N=95$). Bars indicate recall; error bars show 95% Wilson confidence intervals[14], [4]. Linear models largely retain the Original label, while contextual transformer models systematically reclassify catalogue prose as Digital

TABLE V: Retention of the Original class on Codex catalogue texts (Original-only corpus, $N = 95$).

Model	TP	Recall _{Orig}	FN rate	95% CI
Logistic Regression	94	0.989	0.011	[0.943, 0.998]
Random Forest	94	0.989	0.011	[0.943, 0.998]
Naïve Bayes (Count)	93	0.979	0.021	[0.926, 0.994]
SVM	90	0.947	0.053	[0.883, 0.977]
Longformer	82	0.863	0.137	[0.780, 0.918]
Decision Tree	53	0.558	0.442	[0.458, 0.654]
BERT (Truncation)	8	0.084	0.916	[0.043, 0.157]

sion, random forest, naïve Bayes and SVM are not statistically meaningful. The Longformer is genuinely worse as the logistic regression, random forest and naïve Bayes because its upper bound is still below the lower bounds of those models. The decision tree shows high uncertainty, as that range is rather large, yet it is clearly worse compared to the other models. The standard BERT model is not comparable to the others.

TABLE VI: SVM predictions for two excerpts from the Catalogue of the Georgian Manuscripts [8]. Contributions are given as $w_j x_j$ (feature weight $\times$ TF-IDF value). Both examples show the words that led the SVM the most to misclassify or correctly classify the examples.

Misclassified (Pred: digital_only)		Correct (Pred: original)	
Token	Contribution	Token	Contribution
des	+0.0297	st	+0.0278
holy	-0.0294	after	+0.0232
styles	+0.0252	of st	+0.0191
forever	-0.0200	without	-0.0186
veneration	-0.0177	brown	+0.0176
the holy	-0.0177	12	+0.0175

To get a better understanding of what the models have actually learned, Table VI shows the features that led the SVM to correctly classify and misclassify a text from the codicological book, together with their contribution to the

overall score. Both examples were chosen because their overall score was the strongest in the respective classes over the complete set of texts from the book. The misclassified example shows a mixture of weakly informative and ambiguous features. Several tokens associated with religious or abstract language (e.g., *holy*, *veneration*) push the prediction toward the *digital_only* class, overriding weaker material cues such as *ink* or *letters*. In contrast, the correctly classified example contains more consistent material and descriptive indicators (e.g., *brown*, *ink*, numerical and positional references), which align with the learned profile of *original* manuscripts. This suggests that the model is sensitive to lexical signals of physicality but can be misled by semantically dominant yet less discriminative vocabulary.

E. Challenges During Classification

What initially appears as a straightforward classification task turns out to be fundamentally ill-posed due to the interaction of unobservable labels, domain-specific language conventions, and hidden multilingual variation.

After examining the content of the book “Catalogue of the Georgian Manuscripts”, two previously unforeseen problems emerged. The first problem concerns the genre of the text. The texts in this book are all from the genre of codicological texts. [1] provides a checklist of how researchers should write codicological texts. A formalisation of codicological texts is also shown in the TEI standards[13], with texts printed in books being similar in style to the structure shown in the standard. Both sources highlight one problem with the real-world texts we used for our experiments, namely that texts about codices follow a specific structure or “formula”. Thus, this text genre is different from other types of scientific articles and also from the texts we generated for our synthetic dataset.

The second issue is that the book contains many passages in the Georgian language. Those words are a problem for all the models tested, as it is unlikely that they were seen during the training phase.

One fundamental problem when trying to perform the classification in a simple way is the following: the class labels “original seen” and “facsimile only” are not as clear as they may seem at first. They depend on the language used during writing and on genre-specific language conventions. Additionally, for some real-world texts with multiple authors, one author may have seen the original while co-authors only saw pictures taken by that author. In this case, it is not well defined which class such an article should belong to.

V. CONCLUSION

This study set out to investigate whether linguistic patterns can be used to distinguish between scholarly texts produced with direct access to original artefacts and those based on mediated sources. While initial experiments using synthetically generated data suggested that such a distinction may be learnable under controlled conditions, the results do not generalize reliably to real-world data.

A central limitation of this study lies in the absence of ground truth labels for the key variable of interest: namely, whether an author had direct access to the original artefact. Despite attempts to obtain such annotations from domain experts, no consistent or sufficiently large labelled dataset could be established. This limitation is not merely practical but reveals a deeper issue: often, scholarly texts do not explicitly encode the conditions under which knowledge was produced. As a result, the distinction between direct and mediated engagement with an object is often epistemically opaque and cannot be reconstructed post hoc with high confidence.

To address data scarcity, synthetic text generation was employed as a form of augmentation. While this enabled controlled experimentation, it also introduced a significant methodological constraint. Models trained and evaluated on synthetic data performed well within that domain but failed to transfer robustly to authentic scholarly texts, particularly in highly specialized genres such as codicological writing. This domain shift suggests that the classifier may partially rely on artefacts of the synthetic data rather than stable linguistic markers of epistemic stance.

The breakdown of model performance in codicological texts is particularly instructive. As a discipline focused on the material properties of manuscripts, codicology employs a highly standardised and materially grounded descriptive language. In such contexts, both direct and indirect forms of engagement may produce similar linguistic patterns, thereby reducing the discriminative power of surface-level textual features. This finding underscores the importance of genre and disciplinary conventions as confounding factors in computational analyses of scholarly discourse.

Taken together, these results suggest that the expectation of automatically inferring scholars' access to original artefacts from textual data alone may be misguided. Rather than demonstrating a robust classification capability, this study highlights the limits of such an approach under conditions of limited and weakly supervised data. At the same time, the exploratory findings indicate that the mode of access to an object may still leave subtle linguistic traces, though these are highly sensitive to data quality, genre, and modelling assumptions.

Future work should prioritize the creation of expert-annotated corpora and explore alternative methodological approaches, including the use of richer linguistic features and the incorporation of contextual metadata. More broadly, the difficulty of this task points to a fundamental tension in the study of scholarly practices: while computational methods promise large-scale analysis, the epistemic conditions of knowledge production often remain only partially visible in the textual record.

ACKNOWLEDGMENT

The research for this paper was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC

2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796. Parts of this research were conducted within the scope of the Centre for the Study of Manuscript Cultures (CSMC) at University of Hamburg.

This contribution was partially funded by the Danish National Research Foundation (DNRF193) through TEXT: Centre for Contemporary Cultures of Text at Aarhus University.

During the preparation of this manuscript, the authors used generative AI tools (DeepL, DeepL Write, GitHub Copilot) for: spelling and grammar check and suggestions for wording. The authors reviewed and edited all generated content and took full responsibility for the final text.

Special thanks to Jost Gippert for allowing us to use his book for our experiments.

REFERENCES

- [1] Beadle, R., Hanna, R.: Describing and cataloguing medieval english manuscripts: A checklist. In: Da Rold, O., Treharne, E. (eds.) *The Cambridge Companion to Medieval British Manuscripts*, p. 13–38. Cambridge Companions to Literature, Cambridge University Press (2020)
- [2] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer (2020), <https://arxiv.org/abs/2004.05150>
- [3] Benjamin, W.: The work of art in the age of mechanical reproduction. In: Arendt, H. (ed.) *Illuminations*, pp. 217–251. Schocken Books, New York (1968)
- [4] Brown, L.D., Cai, T.T., DasGupta, A.: Interval estimation for a binomial proportion. *Statistical Science* **16**(2), 101–133 (2001)
- [5] Centre for the Study of Manuscript Cultures: Digital twins (cmu 6). <https://www.csmc.uni-hamburg.de/research/concepts-and-methods-units/cmu-6.html> (2026), last updated: 2026-03-10; accessed: 2026-04-09; archived at: <https://web.archive.org/web/20260409153215/https://www.csmc.uni-hamburg.de/research/concepts-and-methods-units/cmu-6.html>
- [6] Dahlström, M.: Copies and facsimiles. *International Journal of Digital Humanities* **1**(2), 195–208 (2019). <https://doi.org/10.1007/s42803-019-00017-5>
- [7] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* **abs/1810.04805** (2018)
- [8] Jikurashvili, T., Gippert, J., Outtier, B., Kim, S., Asatiani, K., Dolakidze, M., Jghamaia, L., Karanadze, M., Kavtaria, M., Khevsuriani, L., Kiknadze, L., Tseradze, T., Guledani, T., Kvirkevelia, E., Lomsadze, G., Sarjveladze, S., Shanidze, M., Tvaltvadze, D., Boudalis, G.: *Catalogue of the Georgian Manuscripts, Holy Monastery of Iviron. Holy Monastery of Iviron* (01 2022)
- [9] Latour, B., Lowe, A.: The migration of the aura, or how to explore the original through its facsimiles. In: Bartscherer, T., Coover, R. (eds.) *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*, pp. 275–297. University of Chicago Press, Chicago (2011)
- [10] Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK (2008)
- [11] McKenzie, D.F.: *Bibliography and the Sociology of Texts*. Cambridge University Press, Cambridge (1999)
- [12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**(null), 2825–2830 (Nov 2011)
- [13] Text Encoding Initiative Consortium: Tei p5: Guidelines for electronic text encoding and interchange: 11. manuscript description (2026), <https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>, version 4.11.0. Accessed 2026-04-14
- [14] Wilson, E.B.: Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**(158), 209–212 (1927)