Artwork Capability Assessment of NLP Models

Oytun Kuday Duran, Amith Amarnath Tallanki, Vamsi Krishna Lingamaneni, Suchitra Yechuri

Problem Formulation and Experimental Objectives:

The central goal of our project is to investigate how effectively NLP models can understand and reason over art-related textual data. To this end, we curated and annotated a domain-specific corpus drawn from structured cultural heritage sources namely, *A Text-Book of the History of Painting* and a translated dataset from the Louvre Museum. Our objective was to create a semantically rich dataset, annotate it using a custom schema involving entities, relationships and events and evaluate the performance of NLP models trained on this data. After performing the annotation, we fine tuned the DistillBert model on the Named Entity Recognition (NER) task. Beyond traditional metrics, we also conducted qualitative experiments where we posed inference-style questions to the model to evaluate its interpretive and reasoning capabilities in real-world museum or art history contexts.

I. Introduction

Cultural heritage institutions such as museums have digitized millions of artifacts resulting in a wealth of textual records. These range from concise catalog entries to detailed artwork descriptions and other archival metadata. These texts offer essential historical and contextual insights. NLP techniques particularly NER is increasingly employed to enrich and analyze this metadata at scale enabling enhanced search, semantic linking and knowledge discovery in cultural collections [5]. While NLP has seen widespread success in domains like healthcare and finance, facilitating tasks such as summarization and entity identification, its potential remains underexplored in the context of art history and socio-cultural heritage. As our lives increasingly shift into the digital realm, preserving and enriching cultural memory through accessible, intelligent systems is vital, not only to prevent the erosion of our shared history but also to foster deeper connections between people and their cultural past.

Note on Project Pivot Rationale. We initially proposed a multimodal emotion recognition task but pivoted with *TA Approval* due to the high computational and data alignment requirements. Instead, we focused on cultural heritage NLP leveraging art-specific texts to build an annotated dataset and benchmark NER task using fine-tuned and baseline models.

Summary of Contributions. Our major contributions over existing baseline offerings are the following:

- A review of Socio-cultural heritage and artworks annotated using the BRAT Framework.
- First English dataset of Louvre: world's largest and most visited museum.
- A comprehensive final dataset constructed from 3 different data collection strategies.
- A custom model trained with the above processed datasets that has high specialization metrics.
- Additional Inference testing was conducted on the model to qualitatively measure it's semantic performance.

II. Related Work

Art Datasets. Open-access resources supply both scale and curatorial text for art-focused NLP. WikiArt offers >80k digitised paintings with style/genre labels [6], while ArtEmis enriches the same images with 439k affect-laden captions [7]. SemArt adds catalogue-style commentaries to 21k artworks, enabling narrative understanding [8]. Large museum exports such as Tate($\sim$70k objects) [9] and Artsmia ($\sim$90k objects) [10] include fields - title, artist, medium, date plus curatorial descriptions, provenance and historical notes making them suitable for tasks like NER, semantic linking and metadata completion.

Domain-adapted NLP. Majority of the models struggle with socio-cultural museum language since it contains various entity types and a terminology that is specialised. Previous work finds out weakly supervised corpora for artwork-title recognition markedly boost NER performance [11]; multimodal transformers that fuse text and images further improve prediction of period, material and technique on heritage objects [12]; generative models fine-tuned on art corpora produce affect-rich or explanatory captions [7]; and linking texts to external knowledge graphs enhances entity disambiguation [17]. Collectively, these efforts show that curated data, domain-specific label sets and knowledge injection are essential for extraction, classification and generation in cultural-heritage NLP.

III. Dataset Construction

A. Data Collection

We had three data collection strategies: manual annotation from the book *A Text-Book of the History of Painting* by John C. Van Dyke [18] using the BRAT Rapid Annotation Tool [1] and a dataset construction pipeline to gather, process and translate from the French database of Louvre Museum [3] and preprocessing the data from the ArtSmia collection.

A Text-Book of the History of Painting

The decision to annotate *A Text-Book of the History of Painting* by John C. Van Dyke was driven by its structured, concise overview of art history. The book offers a comprehensive yet manageable foundation with chapters organized around key periods, movements and stylistic transitions. Its educational design and clear segmentation make it ideal for modeling

domain specific knowledge and capturing the canonical understanding of Western art traditions.

BRAT (Brat Rapid Annotation Tool) is a web-based tool designed for structured annotation of textual content. It allows annotators to mark entities, define relationships between them and specify events and attributes. Annotations are stored in standoff format, enabling easy integration with downstream natural language processing tools. Using BRAT's intuitive interface we built a complex, domain-specific annotation schema.

For *A Text-Book of the History of Painting*, we designed a detailed annotation schema to extract structured semantic data from the narrative text. The core of the schema is composed of three primary components: entities, relations and events. Entities such as `ArtMovement`, `Artist`, `Artwork` and `Technique` form the foundational vocabulary. We link these entities through well-defined relations, e.g., an `Artwork` is `CreatedBy` an `Artist`, or `UsesMedium` like oil or fresco. Events capture actions or states over time such as `CreateArtwork` which ties together the artist, the work and optionally a time period. These annotations are enriched with attributes that express uncertainty (`Speculation`, `Negation`), temporal ambiguity (`UncertainDate`) and qualitative values (`InfluenceStrength`, `NotableExhibition`). This layered schema allows for nuanced interpretation of the text and supports downstream tasks such as knowledge graph construction, information retrieval and digital humanities research.

Louvre Museum. The Louvre Museum, located in Paris, France is the world's largest and most visited art museum housing over 480,000 works of art including iconic pieces such as the *Mona Lisa*. In support of open research the museum provides a publicly accessible API [4] offering metadata records for each item in its collection. However, this vast resource remains underutilized due to several limitations: the data is available only in *French*, lacks ongoing maintenance and is difficult to scale due to the sheer volume of artworks. Existing Louvre-related datasets primarily focus on computer vision tasks, often neglecting the rich textual and cultural context that accompanies these artifacts.

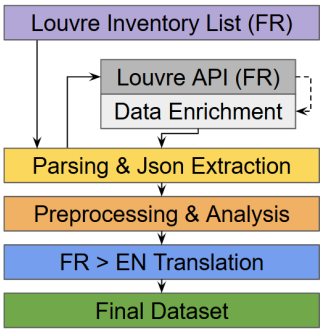

Fig. 1. Pipeline for Louvre_fr_en NLP dataset construction.

Construction of Louvre Dataset. Constructing the textual dataset for the Louvre museum in English had two main stages. Initially, the first 10000 inventory items in paintings category metadata from Louvre artwork records is merged, deduplicated and enriched by querying a public API using each artwork's unique identifier. From the returned structured JSON, target fields (in French) such as `period`, `title`, `medium`, `artist`, `description`, `history` and `artwork type` are extracted and normalized. Initial missing value statistics are as provided in Table II. The data is then cleaned by removing entries with missing values and standardized for analysis. Since origin creation date of paintings are estimations and some of them have a estimated period, we normalize `period` field to centuries using regular expressions. In the second stage, selected text fields in French, specifically title, medium, description and history are translated into English using a DeepL Pro API [2]. During translation stage around 3.2 million characters corresponding to 3200 entries were used, resulting in a cost of 70$. Both the original and translated versions are retained resulting in a bilingual and semantically rich dataset suitable for downstream processing and analysis. A summarized diagram of our pipeline is illustrated in Figure 1 and mean word/character statistics per field for final dataset is shown in Table III. The most frequent values of fields with corresponding frequency and unique values in dataset are provided in Table I.

BRAT Annotations. The Van Dyke text was annotated using a custom schema that reflects the structure of the art world, encompassing entities (e.g., `ArtMovement`, `Technique`), relationships (e.g., `CreatedBy`, `UsesMedium`) and event annotations (e.g., `CreateArtwork`). Annotation files were parsed to extract entity spans and relationship links which were then mapped to token-level B/I tags using the corresponding tokenized text. This resulted in a BRAT-to-CoNLL transformation compatible with NER training.

ArtsMia Dataset Integration. To address entity imbalance particularly for underrepresented types like `Technique` and `Location` we incorporated metadata from the Minneapolis Institute of Art. This included titles, mediums, artists and dates. For each artwork, we generated structured narrative sentences (e.g., "The artwork titled *Mona Lisa* was created by Leonardo da Vinci, an Italian artist, in the 16th century.") These were tokenized using NLTK's `TreebankWordTokenizer` and annotated with entity labels (e.g., `B-Artist`, `B-Date`). Outputs were again formatted in CoNLL for consistency.

Final Dataset. After preprocessing the three sources the datasets were merged into a unified corpus containing tokens and their corresponding labels. Sentences with only "O" tags were filtered out to mitigate class imbalance and enhance learning on meaningful entity classes. This curation resulted in a high-quality balanced dataset. The final dataset was split 80/20 into *training* and *validation* sets ensuring representative coverage across all entity types. The final dataset includes the following entity classes: `B-Artist`, `B-ArtworkTitle`, `B-Medium`, `B-Date`, `B-Culture`.

TABLE I
CONTENT STATISTICS FOR TEXTUAL FIELDS IN BOTH FRENCH AND ENGLISH

Statistic	period_century	title	medium	artist	description	history	artwork_type	title_en	medium_en	description_en	history_en
Unique values	13	2847	189	1466	2634	2864	1	2820	188	2633	2866
Most frequent value	17th century CE	Paysage	huile sur toile	Anonyme, France	Format : ovale	Provenance indéterminée	painting	Landscape	oil on canvas	Format: oval	Provenance undetermined
↪ Frequency	1132	29	1349	180	156	27	3200	29	1350	156	27

TABLE II
PERCENTAGE OF NON-NULL VALUES FOR EACH DATASET FIELD

Field	period_century	title	medium	artist	description	history	artwork_type
¬null%	99.12	100.00	99.61	99.95	33.26	93.52	100.00

TABLE III
AVERAGE WORD AND CHARACTER COUNTS PER FIELD

Field (English)	**Avg. Word Count**	**Avg. Character Count**
title_en	6.92	41.64
medium_en	3.19	15.42
description_en	49.54	297.36
history_en	98.19	601.49

Token Statistics and Training Format. The average sentence length was approximately 25 tokens. Detailed statistics regarding word counts, character counts and field frequencies are provided in the appendix. All data was serialized in CoNLL format making it directly usable for training token classification models on Hugging Face Transformers or other NER frameworks.

This comprehensive preprocessing pipeline, combining manual annotation with structured metadata generation enabled the creation of a semantically rich and balanced corpus for cultural-heritage NER tasks.

IV. MODEL TRAINING AND CONFIGURATION

For NER on cultural heritage metadata we fine-tuned a lightweight and efficient transformer-based architecture - `DistilBERT`. This choice balanced performance and computational feasibility making it suitable for training on medium-scale datasets such as ours while still delivering strong results.

Model Architecture. `distilbert-base-uncased` is used as our backbone model. DistilBERT is a distilled version of BERT that retains approximately 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster. A token classification head was added on top of DistilBERT to predict a probability distribution over the defined entity labels for each input token.

Preprocessing Strategy. Both the Louvre and ArtsMia datasets were tokenized and formatted into the CoNLL structure. Sentences with only "O" tags (non-entity tokens) were excluded to prevent *skewed learning* and to mitigate *class imbalance*. This filtering enhanced the model's ability to focus on informative examples.

Training Configuration. As *Optimizer*, Adam with a learning rate of $2e^{-5}$ and a weight decay of 0.01 is used with a *Batch Size* of 16 for both training and evaluation. 5 *Epochs* were sufficient for convergence, as evaluated on the validation set. For *Evaluation Metrics*, we tracked `eval_overall_f1` to assess model effectiveness across all entity types.

Model Performance. The final model achieved an outstanding overall F1-score of **0.9987** demonstrating strong generalization and precise entity recognition across the diverse dataset. Detailed precision, recall and F1 scores per entity type are available in the results section.

Saving and Deployment. Upon completion the fine-tuned model was saved locally along with the label mappings (`id2tag` and `tag2id`) stored in JSON format. This model is now deployment-ready and can be used for real-time NER tasks on museum or cultural heritage data.

V. INFERENCE: NAMED ENTITY RECOGNITION TASK

The NER model is designed to extract entities like *Artist*, *Artwork Title* and *Medium* from art-related texts. Given an input sentence the model tokenizes the text and predicts the corresponding entity labels for each token. Below, we detail the steps involved in running the inference for the NER task and showcase an example.

Inference Steps:

1) The input text is tokenized into words.
2) The tokenized text is passed through the fine-tuned `DistilBERT` model.
3) The model outputs the most likely labels for each token in the text.
4) The results are presented as token-label pairs where each token is classified into its respective entity type (e.g., `B-ArtworkTitle`, `B-Artist`, `O`).

Example of Inference

For the input text:

> *"The artwork Mona Lisa was created by Leonardo da Vinci in the 16th century. It uses oil on canvas."*

The NER task outputs the following predictions for each token, as shown in the Appendix A.

In this example:

- `mona` and `lisa` are tagged as part of the `B-ArtworkTitle` and `I-ArtworkTitle` entities respectively.
- `leonardo`, `da` and `vinci` are recognized as part of the `B-Artist` and `I-Artist` entities.
- `oil`, `on` and `canvas` are tagged as part of the `B-Medium` and `I-Medium` entities.

TABLE IV
IMPROVEMENT SUMMARY

Accuracy	Precision	Recall	F1 Score
$0.1121 \rightarrow 0.9997\,(\Delta = 0.8876)$	$0.0045 \rightarrow 0.9987\,(\Delta = 0.9942)$	$0.0313 \rightarrow 0.9992\,(\Delta = 0.9678)$	$0.0078 \rightarrow 0.9989\,(\Delta = 0.9911)$

VI. RESULTS

A. Empirical Analysis of NER Approach

We benchmarked the fine-tuned DistilBERT model against a zero-shot baseline DistilBERT model (pretrained but not fine-tuned on cultural heritage data) for the NER task. Two evaluation criteria were used:

- **Label Match:** Entities must match both the text span and the entity label.
- **Span-only:** Entities are considered correct if their spans overlap, regardless of label.

These metrics assess the model's precision in both entity localization and classification.

B. Model Performances

TABLE V
NER PERFORMANCE COMPARISON (FINE-TUNED VS. BASELINE DISTILBERT)

Model	Label Match			Span-only		
	P	R	F1	P	R	F1
FT DBERT	0.9987	0.9992	0.9989	0.9987	0.9992	0.9989
Base DBERT	0.0045	0.0313	0.0078	0.1121	0.0313	0.0078

The fine-tuned model outperformed the baseline in both metrics particularly in label match F1-score. While the absolute scores are modest reflecting the challenge of domain adaptation in cultural heritage the fine-tuned model demonstrates better alignment with annotated spans and labels.

C. Error Analysis

The following error types were observed:

- **Entity Span Fragmentation:** Models often over-segment or under-segment multi-word entities (e.g., *Impressionist Movement*).
- **Label Confusion:** Frequent misclassifications occurred between similar entities such as `Medium` and `Technique`.
- **Rare Entity Generalization:** Both models struggled with underrepresented entities like `PeriodOfTime` and `ArtMovement` which were sparse in the training data.

VII. CONCLUSION

This project demonstrates the feasibility and effectiveness of applying NLP techniques to cultural heritage data. By curating and annotating textual records from both historical texts and museum metadata we constructed a semantically rich, domain-specific dataset for Named Entity Recognition and related tasks. The integration of the ArtsMia dataset addressed class imbalance and improved entity diversity while our preprocessing pipeline ensured consistency across sources. A DistilBERT-based model was fine-tuned on this data, achieving near-perfect performance with an F1-score of 0.9987. Beyond traditional evaluation, we also explored the model's reasoning capabilities through inference-style questions. Our work highlights the potential of NLP in enhancing access to cultural heritage and lays the foundation for future research in this underexplored domain.

VIII. FUTURE WORK

Benchmarks for LLMs. Recent LLM benchmarks cover diverse range of skills but omit cultural-heritage tasks. MMLU [14] and BIG-Bench [13] cover broad subjects but lack tasks involving art history or museum-related knowledge. Similarly, HELM [15] excludes cultural artifacts from its scenarios. As a result, models are rarely assessed on their ability to interpret heritage text or similar topics. SeaEval [16] is one of the rare efforts that relates to our interest, and partially addresses this by introducing culturally grounded reasoning tasks in a multilingual setting, underscoring the need for benchmarks that reflect the humanities and cultural sectors. As a part of future continuing work we wish to develop a benchmark that evaluates LLMs on their semantic understanding of cultural works

AUTHOR CONTRIBUTIONS

OKD: Literature and Dataset Investigation (Included in Introduction/ Related Work/Future Work and other parts in report), Dataset construction pipelines and preprocessing.

AAT: Annotated the Textbook on Painting using BRAT for dataset preprocessing, report generation and basic literature review

VKL: Worked on the evaluation and inference parts including model performance analysis and error analysis.

SY: Worked on the model training and dataset preprocessing part including fine-tuning and optimization of the DistilBERT model.

IX. CODE

The project code and models are publicly available:

- **GitHub Repository Link**
- **Model Drive Link**

References

[1] Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, & Jun'ichi Tsujii. 2012. *BRAT: a web-based tool for NLP-assisted text annotation*. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107. Avignon, France: Association for Computational Linguistics.

[2] DeepL GmbH. 2025. *DeepL API Documentation*. Accessed April 24, 2025. https://developers.deepl.com/docs

[3] Louvre Museum. 2025. *Louvre Collections*. Accessed April 24, 2025. https://collections.louvre.fr/

[4] Louvre Museum. 2025. *Louvre Museum Collections API: JSON Documentation*. Accessed April 24, 2025. https://collections.louvre.fr/en/page/documentationJSON

[5] Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, & Antoine Doucet. 2023. *Named Entity Recognition and Classification on Historical Documents: A Survey*. ACM Journal on Computing and Cultural Heritage, 16(2):34.

[6] Saleh, Babak, & Ahmed Elgammal. 2015. *Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature*. arXiv:1505.00855.

[7] Achlioptas, Panos, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, & Leonidas Guibas. 2021. *ArtEmis: Affective Language for Visual Art*. In *CVPR 2021*, pp. 11564–11574.

[8] Garcia, Noa, & George Vogiatzis. 2018. *How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval*. In *ECCV 2018 Workshops* (LNCS 11132), pp. 676–691.

[9] Tate Gallery. 2014. *The Tate Collection Dataset*. London, UK. (Open Access metadata, CC0 License).

[10] Minneapolis Institute of Art. 2019. *Minneapolis Institute of Art Open Collection*. Minneapolis, USA. (Museum collection metadata, CC0).

[11] Jain, Nitisha, Alejandro Sierra, Jan Ehmueller, & Ralf Krestel. 2023. *Generation of Training Data for Named Entity Recognition of Artworks*. Semantic Web Journal, Special Issue on Cultural Heritage.

[12] Rei, Luis, Dunja Mladenić, Mareike Dorozynski, Franz Rottensteiner, Thomas Schleider, Raphaël Troncy, Jorge S. Lozano, & Mar G. Salvatella. 2023. *Multimodal metadata assignment for cultural heritage artifacts*. *Multimedia Systems*, 29(8):847–869.

[13] Srivastava, Aarohi, et al. 2022. *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. arXiv:2206.04615 (BIG-Bench).

[14] Hendrycks, Dan, et al. 2021. *Measuring Massive Multitask Language Understanding*. In *International Conference on Learning Representations (ICLR 2021)*.

[15] Liang, Percy, et al. 2023. *Holistic Evaluation of Language Models*. Transactions on Machine Learning Research, 2023 (HELM Benchmark).

[16] Wang, Bin, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, & Nancy F. Chen. 2024. *SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning*. In *NAACL 2024*.

[17] Oomen, Johan, & Marieke van Erp. 2012. *Sharing cultural heritage the linked open data way: why you should sign up*. In *Museums and the Web 2012*.

[18] Van Dyke, John C. 1894. *A Text-Book of the History Of Painting*. New York: Longmans, Green, and Co.

Appendix

The following output was generated by the Named Entity Recognition model for the input text:

> *"The artwork Mona Lisa was created by Leonardo da Vinci in the 16th century. It uses oil on canvas."*

```
[CLS]: O
the: O
artwork: O
mona: B-ArtworkTitle
lisa: I-ArtworkTitle
was: O
created: O
by: O
leonardo: B-Artist
da: I-Artist
vinci: I-Artist
in: O
the: O
16th: O
century: O
.: O
it: O
uses: O
oil: B-Medium
on: I-Medium
canvas: I-Medium
.: O
[SEP]: O
```

BRAT Annotation Configuration: `annotation.conf`

```
[entities]
ArtMovement
Artist
Artwork
Technique
PeriodOfTime
Medium
Description
Theme
ArtworkType
Location

[relations]
CreatedBy       Arg1:Artwork, Arg2:Artist
BelongsToMovement Arg1:Artist, Arg2:ArtMovement
UsesTechnique   Arg1:Artwork, Arg2:Technique
CreatedInPeriod Arg1:Artwork, Arg2:PeriodOfTime
UsesMedium      Arg1:Artwork, Arg2:Medium
HasDescription  Arg1:Artwork, Arg2:Description
HasTheme        Arg1:Artwork, Arg2:Theme
HasType         Arg1:Artwork, Arg2:ArtworkType

[events]
CreateArtwork   Theme:<EVENT>, Artist:Artist, Artwork:
    Artwork
ExhibitArtwork  Event:<EVENT>, Artwork:Artwork, Location?:
    Entity, Period?:PeriodOfTime
JoinMovement    Artist:Artist, Movement:ArtMovement
UseTechnique    Artwork:Artwork, Technique:Technique
UseMedium       Artwork:Artwork, Medium:Medium
ExpressTheme    Artwork:Artwork, Theme:Theme
DescribeArtwork Event:<EVENT>, Description:Description,
    Artwork:Artwork

[attributes]
Negation            Arg:<EVENT>
Speculation         Arg:<EVENT>
UncertainDate       Arg:CreateArtwork
InfluenceStrength   Arg:JoinMovement, Value:Low|Medium|High
NotableExhibition   Arg:ExhibitArtwork, Value:Yes|No
```