



Repurposing and Adapting Language Resources for a Historical Italian Large Language Model Benchmark

SHIBINGFENG ZHANG

MARIIA LEVCHENKO

CHIARA MANCA

FIAMMETTA SABBA

PAOLA MARIA CARMELA ITALIA

GIOVANNI COLAVIZZA

**Author affiliations can be found in the back matter of this article*

COLLECTION:
LANGUAGE
DATASETS REUSE:
OPPORTUNITIES,
CHALLENGES, AND
BEST PRACTICES

DISCUSSION PAPER

[ubiquity press

ABSTRACT

We introduce the first benchmark specifically designed to evaluate the capabilities of large language models on historical Italian. While benchmarks for modern Italian are available, historical varieties of the language remain underexplored in this context. To address this gap, we repurpose and adapt several high-quality digital resources to construct a novel benchmark. The result spans a range of tasks, including word sense disambiguation, syntactic structure prediction, intertextuality detection, named entity recognition, authorship verification, and chronological detection, with source texts ranging from the High Middle Ages to the nineteenth century.

We evaluate several state-of-the-art large language models in a zero-shot setting. Our findings show that while these models exhibit promising performance, the benchmark remains far from saturated, suggesting that the capabilities of large language models on historical languages still have considerable margin for improvement. Our work offers a standardized evaluation framework to support the selection of models for annotating digital archives and to guide the development of linguistically inclusive language models.

CORRESPONDING AUTHOR:

Shibingfeng Zhang

Department of Classical
Philology and Italian Studies,
University of Bologna,
Bologna, Italy

shibingfeng.zhang@unibo.it

KEYWORDS:

Large Language Models; Italian
studies; Digital Humanities;
Digital Archives; Information
Extraction; Language
Resources Reuse

TO CITE THIS ARTICLE:

Zhang, S., Levchenko, M.,
Manca, C., Sabba, F., Italia, P.
M. C., & Colavizza, G. (2026).
Repurposing and Adapting
Language Resources for a
Historical Italian Large
Language Model Benchmark.
*Journal of Open Humanities
Data*, 12: 119. DOI: <https://doi.org/10.5334/johd.538>

1 CONTEXT AND MOTIVATION

In this paper, we present our work of reusing language resources in Italian to construct a benchmark that serves to assess the ability of general purpose large language models (LLMs) in understanding historical Italian. Our benchmark is constructed by adapting several existing Italian linguistic resources, adhering to the standards of similar benchmarks in the field. To the best of our knowledge, our work represents the first benchmark with such scope. The benchmark, together with its evaluation code, are publicly available.¹

Currently, no standardized framework exists for evaluating LLMs on historical Italian, despite growing interest in applying these models to cultural heritage materials. To address this research gap, we introduce a benchmark with two primary aims. Firstly, in the development of digital archives, such as historical documents, manuscripts, and correspondence, it is a common practice to enrich these collections with annotations such as syntactic information and named entities with the purpose to support subsequent scholarly analysis (Corbetta et al., 2025; Cristofaro et al., 2023; Manca et al., 2025). Traditionally, this annotation process has relied heavily on the manual efforts of domain experts, making it time-consuming and resource-intensive. LLMs, on the other hand, offer a cost-effective alternative that reduces the need for manual labor. However, they remain prone to errors and typically require post-editing. In this context, a benchmark tailored to historical Italian can guide model selection for large-scale digital archive projects, helping to identify the most effective LLM for the task and thereby minimizing the manual effort required. Secondly, establishing a standardized means of evaluating LLMs on historical Italian facilitates the development of language models that are more capable in handling diachronic linguistic variation.

In this section, we provide the research background by reviewing the synergistic development of benchmarks and language models (Section 1.1) in recent years, and the current state of existing benchmarks for natural language understanding (Section 1.2). Finally, we summarize our motivations and contributions.

1.1 THE DEVELOPMENT OF BENCHMARKS

In the field of Natural Language Processing (NLP), benchmarks serve as the foundational instruments for standardized evaluation, providing the metrics necessary to quantitatively assess a language model's capabilities. The architecture and methodology of these benchmarks have undergone a profound evolution together with the rapid advancements in language model architectures over the last several decades (Ni et al., 2025).

This progression originated with statistical language models, which relied on probabilistic N-gram distributions, and transitioned into neural network-based models such as LSTM (Hochreiter & Schmidhuber, 1997) that utilized dense vector representations to capture deeper semantic relationships. Vaswani et al. (2017) proposed the attention-based Transformer architecture, which established a new paradigm in the design of language models. Initially, the most prominent implementations of this architecture were discriminative models. Discriminative models, such as BERT (Devlin et al., 2019) and RoBERTa (Y. Liu et al., 2019), utilize the encoder stack and are typically trained on a Masked Language Modeling and Next Sentence Predicting objectives. They serve primarily as feature extractors of the input text, generating dense vector embeddings that capture bidirectional context. These embeddings are highly effective for downstream tasks like sentiment analysis, named entity recognition, and document classification, where a good understanding of the input text is required. Generally, these discriminative models are pre-trained on vast quantities of unlabeled text in an unsupervised manner, then adapted for specific applications through fine-tuning on task-specific datasets (Naveed et al., 2025). Consequently, the benchmarks traditionally used to evaluate these models consist of relatively small, single-task datasets partitioned into training, development, and test sets. The language models are fine-tuned or trained on the training and development sets, then tested on the test set.

The recent rapid development of generative LLMs and their emergent capabilities, such as few-shot and zero-shot learning, has fundamentally transformed the different NLP tasks into a unified generative format. The popular LLMs include the Llama series (Touvron et al.,

¹ Github: <https://github.com/Zhangshibf/Benchmarking-Large-Language-Model-on-Historical-Italian>
Zenodo: <https://zenodo.org/records/20704200>.

2023), DeepSeek (A. Liu et al., 2025), GPT series (Brown et al., 2020), etc. Unlike the discriminative models, these LLMs are general purpose with versatile capabilities. Consequently, the benchmarks evolved alongside LLMs to assess a wider range of abilities such as reasoning, knowledge, and cultural alignment. Many of these benchmarks are designed for evaluation in zero-shot or few-shot settings. These methods require very little or no training data, which aligns more closely with how these models are used in practice (Ni et al., 2025). The multiple-choice question-answering format adopted by many of these benchmarks has become a popular paradigm, as it permits automatic and unambiguous scoring. Other benchmarks (Bai et al., 2024; Zheng et al., 2023) instead adopt an LLM-as-judge approach to evaluate open-ended questions. To thoroughly assess the abilities of LLMs in a given language, it is common practice to report their performance on a multi-task benchmark such as BIG-Bench (Srivastava et al., 2023), which aggregates a diverse collection of tasks spanning reasoning, knowledge, and language understanding, thereby providing a more comprehensive picture of model capabilities than any single task could offer.

1.2 BENCHMARKING HISTORICAL LANGUAGES

Various benchmarks exist to measure language models' performance on historical languages. This section reviews the benchmark construction works that served as primary references for our study and guided our benchmark design and provide a comprehensive review of the research background. These benchmarks can be categorized based on the type of language models they aim to evaluate. For discriminative language models, tasks are often formulated in diverse formats, such as passage retrieval and classification, depending on the task content. In contrast, benchmarks for generative language models typically adopt a different paradigm by converting different tasks into a generative text-to-text format.

Caffagni et al. (2025) proposes to test discriminative language models' abilities in Latin using the task of identifying intertextual biblical references within patristic commentaries, specifically Augustine's *De Genesi ad litteram*. The benchmark frames the problem as an information retrieval task: given a passage from the commentary as the query, the model must retrieve the most relevant verse from a collection of biblical passages from either the Vulgate or the Vetus Latina versions. A reference can manifest in three primary forms: 1. direct citation: verbatim or near-verbatim reproduction, 2. paraphrase: rephrasing while preserving core meaning, and 3. allusion: indirect evocation with minimal lexical overlap. While fine-tuning helped, the models they tested still showed low recall, struggling to identify these subtle references. Tian et al. (2021) and Stopponi et al. (2024) created benchmarks to test discriminative language models on historical Portuguese and ancient Greek, respectively. These benchmarks are created utilizing word relatedness scores provided by human evaluators and various corpora resources. L. Liu et al. (2025) proposed a shared NER evaluation corpus for 18th-century historical texts, built by normalizing annotations across four heterogeneous datasets, and used it to benchmark language models across different evaluation modes.

WenMind (Cao et al., 2024) is a benchmark designed to evaluate generative LLMs capabilities in Chinese Classical Literature and Language Arts. This benchmark comprises 4,875 question-answer pairs for 42 fine-grained tasks. Tasks include topic classification of ancient Chinese prose, reading comprehension, named entity recognition, punctuation, word explanation, function words, homophones, and polysemy disambiguation, etc. Data in this benchmark are collected from the internet, open-source datasets, or generated by LLMs and are then manually examined. They evaluated 31 LLMs on this benchmark, including general-purpose models and LLMs specialized in ancient Chinese. The results show that models pre-trained primarily on Chinese corpora generally outperform models first pre-trained on English data and subsequently fine-tuned on Chinese data. Surprisingly, models specifically tailored for ancient Chinese perform significantly worse than general-purpose Chinese LLMs of comparable size. This counterintuitive finding highlights the importance of domain-specific benchmarks for effective LLM selection. Similarly, CHENGYU-Bench (Fu et al., 2025) is a benchmark which aims to evaluate LLMs capabilities for Chinese idioms understanding via different tasks.

With regard to the Italian language, benchmarks such as CALAMITA (Attanasio et al., 2024) have been introduced to evaluate LLM performance on diverse text genres and tasks. In addition, datasets such as KIND (Paccosi & Aprosio, 2022), developed for named entity recognition, can be used to assess specific capabilities of LLMs. However, the scope of these resources remains

confined to modern Italian. Several studies have investigated the capabilities of language models on historical Italian. For example, Corbetta et al. (2024) explored the feasibility of improving dependency parsing annotation generation by training language models on annotated modern Italian data. Santini et al. (2026) introduced a named entity recognition dataset based on Italian texts spanning from the early 18th century to the late 20th century, and evaluated the performance of instruction-tuned models as well as the fine-tuned GLiNER (Zaratiana et al., 2024) model on this dataset. Nevertheless, these works always have a specific focus on a certain task or specific text genre and, to the best of our knowledge, there is currently no benchmark designed with the aim to assess comprehensively LLMs' natural language understanding capabilities in historical Italian. This identified gap provides the central motivation for our work.

Following examples of similar works, we formatted our benchmark into a question-answering framework. We selected a diverse range of historical Italian language resources, which were then repurposed to construct a benchmark for assessing the capabilities of LLMs in understanding historical Italian. The principles guiding our resource collection, along with the selected datasets, are presented in Section 2.

The contributions of our work are:

- We introduce the first benchmark specifically designed to evaluate the performance of LLMs on historical Italian.
- We conduct an evaluation of several frontier LLMs using the proposed benchmark in order to provide reference for future works and highlight current model capabilities and limitations.

2 DESCRIPTION OF THE DATASETS

During the dataset selection process for the historical Italian benchmark, we adhered to several principles:

1. **Comprehensiveness:** The datasets should be as diverse as possible, encompassing various genres such as poetry, correspondence, commentaries, and novels.
2. **Temporal Breadth:** The resources should cover different periods of the Italian language, from the emergence of its earliest form in the High Middle Ages to the 19th century.
3. **Author Diversity:** To ensure the benchmark remains unbiased toward a single figure, each author is represented only once. For instance, despite the abundance of resources related to Dante Alighieri, we selected only the Dante Source dataset. This restriction ensures the benchmark evaluates proficiency in historical Italian as a whole, rather than specializing in a single historical figure's work.
4. **High Quality:** The annotations of the selected dataset have to be manually created or manually reviewed.

Table 1 provides a general overview of the datasets leveraged for the construction of the benchmark. As can be seen, beyond genre and chronological diversity, the datasets also span diatopic and geographic variation. *LeggoManzoni* is the clearest case: for the creation of *Quarantana*, Manzoni revised the novel's language toward living Florentine usage, though Lombard traces remain. The Classense letters carry a Romagnol provenance from Ravenna and the Bellini letters a Sicilian one, from Catania, both in a cultivated epistolary register. The two poetic resources are predominantly Tuscan: the Cavalcanti sonnets and Dante's works belong to the Florentine Duecento–Trecento tradition, although the Duecento sources mapped in DanteSource reach further vernacular schools, including the Sicilian. Biblioteca Italiana is diatopically heterogeneous by construction, involving 86 authors spanning from the 13th century to modern era. The details of these datasets are presented below.

2.1 LEGGOMANZONI

*LeggoManzoni*² is a digital scholarly platform developed at the University of Bologna that makes available the full text of Alessandro Manzoni's *I Promessi Sposi* in its canonical 1840–42

² <https://projects.dharc.unibo.it/leggomanzoni/>.

DATASET	GENRE	PERIOD	MAIN AUTHOR(S)	DESCRIPTION
LeggoManzoni	Novel & commentaries	19th c. (comm. 1893–2021)	A. Manzoni	Full text of <i>I Promessi Sposi</i> with 40 critical commentaries
Classense (Canneti–Fiacchi)	Correspondence	Early 18th c. (1711–1720)	P. Canneti, M. Fiacchi	129 letters with named-entity annotations
Bellini Digital Correspondence	Correspondence	Early 19th c.	V. Bellini	29 letters with named-entity annotations
Biblioteca Italiana	Various genres	From 13th c. to modern era	Multiple authors	1,343 books from a broad digital library
Cavalcanti <i>Rhymes</i>	Poetry	13th c.	G. Cavalcanti	22 sonnets annotated with dependency syntax
DanteSource	Poetry	14th c.	D. Alighieri	Ontology of primary sources referenced by Dante in his works

Table 1 Overview of the datasets used to construct the historical Italian benchmark.

edition (the *Quarantana*), accompanied by forty critical commentaries spanning over a century of Manzoni scholarship, from Petrocchi (1893–1902) to Raimondi and Bottoni (2021). The commentaries are encoded in TEI-XML (Levchenko et al., 2025) and annotate semantically significant passages of the novel with linguistic glosses, etymological notes, and contextual explanations. Together, they represent the full collection of Manzoni’s critical reception, from late 19th-century philology to contemporary annotated editions.

2.2 CLASSENSE’S CANNETI-FIACCHI CORRESPONDENCE

The Classense dataset contains the correspondence between Pietro Canneti and Mariangelo Fiacchi, two Camaldolese monks active at the Biblioteca Classense in Ravenna, Italy during the early 18th century. The whole correspondence consists of over 600 letters spanning the period 1700–1730. The letters of *Busta 10* (1711–1720) have been encoded in TEI-XML by the LUDI laboratory at the University of Bologna (Sabba, 2025).³ The encoding includes named entity markup for people, places, organizations, and bibliographic references. The 129-letter corpus contains a total of 5,546 annotated entities, with an average of 43 entities per document and an average document length of approximately 3,700 characters (Sabba et al., 2025).

2.3 THE BELLINI DIGITAL CORRESPONDENCE

The Bellini Digital Correspondence (Cristofaro et al., 2023, 2025)⁴ is a collection of letters written by the renowned composer Vincenzo Bellini from Catania, Italy during the early 19th century. The corpus is encoded following the TEI guidelines. The encoding includes named entity markup for person, group of persons, location, organization, work, and music terms. The Bellini Digital Correspondence consists of 31 letters, with an average length of 1162 characters per letter. These letters contain around 16 entities each.

2.4 BIBLIOTECA ITALIANA

Biblioteca Italiana,⁵ meaning “Italian Library” in Italian, is a digital library of books representing Italian cultural and literary tradition from the Middle Ages to the 20th century, with more than 3,500 books included in its catalog. Around 1,600 books included in the Biblioteca Italiana are available in the text format of TEI-XML. Of these digital editions, 1,343 books are written in Italian.

³ <https://digilettclass.unibo.it/methodology>.

⁴ <https://bellinicornespondence.cnr.it/documentazione/>.

⁵ <https://www.bibliotecaitaliana.it/>. The full list of authors is available at <https://www.bibliotecaitaliana.it/catalogo>.

2.5 CAVALCANTI RHYMES

Corbetta et al. (2025) developed a manually annotated corpus of 22 sonnets from Ercole Rivalta's edition of Guido Cavalcanti's *Rhymes* (Cavalcanti, 1902). The collection includes tokenization, lemmatization, morphological tagging, and dependency syntax.

2.6 DANTESOURCE

The DanteSource platform (Bartalesi & Meghini, 2016; Bartalesi et al., 2018) serves as an open-access web application dedicated to mapping the primary sources utilized by Dante Alighieri. In a literary context, primary sources are defined as the external texts and authors that a writer cites or references within their own compositions (Hellqvist, 2010). Typically, these intertextual connections are documented within scholarly commentaries.

By synthesizing data from various Dantean commentaries, the original developers created a specialized ontology. This framework uses the Resource Description Framework Schema to translate complex literary relationships into a structured, machine-readable format.

3 REUSE PROCESS AND METHODOLOGY

Following the selection phase, we process and review each resource to develop the tasks for the historical Italian benchmark. This section presents the processing methodology for each dataset, the tasks derived from them, the specific LLM capabilities these tasks are designed to evaluate, and the evaluation metrics. To address position bias, where models favor options according to their position within the context rather than their content (Wang et al., 2025), the candidate options in all multiple-choice and ranking tasks are shuffled, so that the correct answer appears in varying positions or orderings across questions.

3.1 LEGGOMANZONI: WORD SENSE DISAMBIGUATION

I Promessi Sposi, together with its commentaries, are repurposed to create a word sense disambiguation (WSD) task in the benchmark.

In the resources of *LeggoManzoni* project, there are 15,172 phrases of *I Promessi Sposi* annotated by commentary. To ensure the quality and significance of the dataset, we retained only those phrases annotated by more than one commentator, as these are considered more significant and worthy of interpretive attention. This filtering process resulted in 1,492 entries. Initial manual examination revealed that many comments were unsuitable for a WSD task, as they focused on historical context or literary appreciation rather than semantic explanation. Consequently, we conducted a screening and cleansing phase using the ChatGPT 5.2 API, prompting the system to identify entries suitable for WSD. This yielded 817 entries. Each entry consists of:

- **Target phrase:** The specific phrase whose meaning is explained by the answer.
- **Context:** The excerpt from *I Promessi Sposi* containing the target phrase.
- **Answer:** A synthesized explanation of the target phrase's meaning, created by aggregating comments from multiple commentaries.

The subsequent stage of the processing pipeline involves generating distractors for the WSD task in order to formulate the task into multiple-choice questions. First, the 817 entries were filtered by calculating the semantic similarity between the target phrase and the answer using the multilingual sentence transformer model *sentence-transformers/paraphrase-multilingual-mpnet-base-v2* (Reimers & Gurevych, 2020). Only entries with a similarity score below 0.7 were retained. These expressions are considered more challenging, as the answers are less obvious.

For negative answer generation, single-word expressions were handled by identifying synonyms from the Italian dictionary developed by Ciaccio et al. (2025), derived from an XML dump of the Italian Wiktionary.⁶ These synonyms were then used to generate distractors. For multi-word expressions, the negative options were generated by direct prompting. Finally, all resulting entries were manually reviewed by native Italian speakers to ensure that the distractors do not semantically overlap with the correct answer. This process resulted in a final set of 200

⁶ <https://it.wiktionary.org>.

multiple-choice WSD questions, each composed of a context, a target phrase, and six candidate answers that explain the target phrase's meaning in the context, of which only one is correct. **Box 1** shows an entry in the WSD task.

The prompts utilized throughout this process are available in **Appendix A**.

```
"target phrase": "somigliare",
"context": "La costiera, formata dal deposito di tre grossi
torrenti, scende appoggiata a due monti contigui, l'uno detto
di san Martino, l'altro, con voce lombarda, il Resegone , dai
molti suoi cocuzzoli in fila, che in vero lo fanno somigliare
a una sega: talché non è chi, al primo vederlo, purché sia di
fronte, come per esempio di su le mura di Milano che guardano a
settentrione, non lo discerna tosto, a un tal contrassegno, in
quella lunga e vasta giogaia, dagli altri monti di nome più oscuro
e di forma più comune.",
"candidates": ["Verbo che significa «imitare, ritrarre» le qualità
o i modi altrui",
"Verbo transitivo che significa «paragonare, confrontare» due cose
tra loro",
"Verbo riflessivo con valore di «mettersi insieme, associarsi» a
qualcuno",
"Verbo che indica «rassomigliare, essere simile» e si costruisce
con di",
"Verbo che può reggere sia l'oggetto diretto sia un complemento
introdotta da a",
"Verbo intransitivo che vale «apparire, mostrarsi» in un certo
modo o aspetto"],
"gold": "Verbo che può reggere sia l'oggetto diretto sia un
complemento introdotto da a"
```

Box 1 An example entry of the word sense disambiguation task derived from LeggoManzoni.

3.2 CLASSENSE'S CANNETI-FIACCHI CORRESPONDENCE & THE BELLINI DIGITAL CORRESPONDENCE: NAMED ENTITIES RECOGNITION

Both the Classense and Bellini datasets were derived from TEI-XML scholarly editions, and in both cases the original XML files were parsed to extract the letter text, entity annotations, and correspondence metadata (sender, recipient, date, place). The extracted data were serialized into a flat JSON format.

For the Classense corpus, indirect person references (e.g., *Vostra Paternità, il Signore*) were excluded, as these are referential expressions rather than named entities. A representative subset of 25 letters was selected from the full corpus of 129 letters, with the aim of preserving the overall distribution of entity types. A chi-square goodness-of-fit test confirms that the subset does not significantly differ from the full corpus in this regard ($\chi^2 = 5.12, p = 0.163$). For the Bellini corpus, two letters written in French were excluded, while the remaining 29 letters were included in the benchmark.

Classense's Canneti-Fiacchi Correspondence and Bellini Digital Correspondence share a unified evaluation pipeline. Each document is processed individually: the model receives the full text of the letter, along with a system prompt specifying the entity types and instructions to preserve the original spelling.

The structured output from the LLM consists of predicted entities represented as key-value pairs of the entity text and its type (Lepagnol et al., 2026). To enable evaluation, these predictions are mapped back onto the source text of each letter. This is achieved by sequentially scanning the text and, for each predicted entity, matching its text to the first occurrence that has not already been assigned to a previous prediction. Evaluation is then performed in two modes: *strict* matching requires identical character spans and matching type; *fuzzy* identifies a prediction as true positive if the predicted entity type is correct and the overlap ratio between predicted

entity and ground truth exceeds 0.5, computed as $\frac{|LCS|}{|predicted| + |gt| - |LCS|}$ where $|LCS|$ is the length of the longest common substring. The evaluation metrics are macro-averaged precision, recall, and F1. [Box 2](#) shows an example of the NER task from the Classense Correspondence.

```
"text": "Molto Reverendo Padre mio Osservandissimo Con la
gratissima lettera di Vostra Paternitàricevo la nota de' libri
di Faenzae insieme la certa notizia dell'offerta già fatta dal mio
Marozzi. Con questo lume ora potrò regolarli nel rispondere alla
richiesta di dieci scudi. Mi spiace che ancora non siano giunti i
due Manoscritti del Quadriregio;",
"entities": [{"text": "Faenza", "start": 115, "end": 121, "type":
"LOC"},
{"text": "Marozzi", "start": 179, "end": 186, "type": "PER"},
{"text": "due Manoscritti del Quadriregio", "start": 310, "end":
341, "type": "WORK"}]
```

Box 2 An example entry of the named entity recognition task from the Classense Correspondence.

3.3 BIBLIOTECA ITALIANA: CHRONOLOGICAL RANKING & AUTHORSHIP VERIFICATION

Two tasks were developed using resources from the Biblioteca Italiana: chronological ranking and authorship verification. In the chronological ranking task, the LLM is presented with four text snippets and is required to arrange them in order from most recent to most ancient. This setup aims to evaluate the model's ability to identify diachronic linguistic shifts across different historical periods. The authorship verification task provides the LLM with four text snippets, where three are written by the same author and one belongs to a different author of the same historical period. The model is then required to identify the outlier. This task aims to test the model's ability to detect idiosyncratic writing styles.

The Biblioteca Italiana covers a wide range of text genres, including poetry, documents, and oratory. For the construction of the benchmark, the poetic and narrative genres were selected because they contain the highest number of unique authors within the collection. This provides sufficient resources for the development of both tasks. In order to avoid bias toward a specific author, books from authors present in other tasks were excluded.

The first step involves chunking the content of books into snippets with lengths between 200 and 600 characters. The text is split at natural sentence endings to ensure the context remains complete. To calculate the semantic similarity between snippets, the *sentence-transformers/paraphrase-multilingual-mpnet-base-v2* model from the SentenceTransformer ([Reimers & Gurevych, 2020](#)) library is utilized to encode all text snippets.

To construct a question for the chronological ranking task, a random anchor snippet is selected, and its cosine similarity is calculated against all other snippets from books of the same genre written by different authors. The three snippets most similar to the random snippet are then selected, while ensuring that the birth dates of the four authors are at least 100 years apart. This method aims to create a challenging task where models must rely on subtle diachronic linguistic cues rather than superficial semantic similarity to determine the correct chronological order.

To construct a question for the authorship verification task, a random snippet is selected, and a search is performed for two other snippets that belong to the same author but are most dissimilar from the initial selection. Finally, an additional snippet is selected from a different author that is most similar to the original snippet, ensuring the birth date distance is less than 50 years. This setup ensures that the task cannot be solved by simply identifying topical overlaps or diachronic linguistic characteristics, forcing the model to capture deeper, inherent stylistic features.

In total, 100 questions are created for each task, consisting of 50 questions using snippets from the poetic genre and 50 from the narrative genre. The evaluation metric for chronological ranking task is the mean of Spearman correlation coefficient. The evaluation metric for authorship verification is accuracy. [Box 3](#) shows an example of the chronological ranking task. [Box 4](#) shows an example of the authorship verification task.

```
{"year": 1855,  
"text": "del cielo azzurro. - È l'usignolo, o Rosa! -\\\\\\Il mare,  
al buio, fu cattivo. Urlava\\\\sotto gli schiocchi della folgore!  
Ora\\\\qua e là brilla in rosa la sua bava.\\\\\\Intorno a mucchi  
d'alga ora si dora\\\\la bava sua lungi da lui. S'effonde\\\\l'alito  
salso alla novella aurora."  
},  
{"year": 1754,  
"text": "la bellicosa degli augei reina.\\\\Tremava tutta e si  
battea la guancia,\\\\del contumace suo furor pentita,\\\\\\la  
superba Lamagna; e del suo sangue\\\\tinto e satollo alfin sorgea  
l'olivo.\\\\All'apparir che fea sulle gelate"  
},  
{"year": 1474,  
"text": "S'accosta alla spelunca spaventosa,\\\\e percuote a gran  
colpo con un'asta\\\\quella ferrata porta, mezzo rósa\\\\da' tarli e  
da la ruggine più guasta.\\\\L'Invidia, che di carne venenosa\\\\allora  
si pascea d'una cerasta,\\\\levò la bocca alla percossa grande\\\\da  
le amare e pestifere vivande."  
},  
{"year": 1313,  
"text": "E sopra l'acqua del fossato gito,\\\\l'aguto dardo si  
recava in mano,\\\\e al petto si ponea 'l ferro pulito,\\\\e 'n  
terra l'asta, dicendo: «O villano\\\\Amor, che m'hai condotto a tal  
partito,\\\\ch'i' moro in questo modo tanto strano!\\\\e pure, innanzi  
ch'i' voglia più stare\\\\in cotal vita, mi vo' disperare."}
```

Box 3 An example of the
chronological ranking task.
\\\\ represents a line break
present in the poem.

```
"texts": [  
"Non di risposta mi par ch'abbia bisogno, ma d'alcuna dimanda.  
Ditemi, dunque, perché nel giudizio s'allontana dal parlar  
d'Aristotele: il qual, se ben mi ricordo, vuol che così le voci  
come i nomi peregrini siano riguardati con maggior meraviglia?  
Ma, di grazia, non ci fermiamo su' testi, se non c'è conceduta  
commodità di rivolger le carte con lungo studio: perché altrimenti  
converrebbe ch'io invocassi più volte la memoria che non fanno  
tutt'i poeti le Muse.",  
"Fu accettata dalla Reina questa risposta, ma con tutto ciò tornò  
il signor Giovanni a dire: Può essere che questa ragione abbia  
luogo fra noi, ma non sarebbe già stata accettata in quel paese  
dove si sogliono misurare i corpi con una cintola di certa misura,  
nella quale se per avventura alcuno non poteva capire, era tenuto  
per disonorato e per uno della greggia d'Epicuro. E parimente  
i Lacedemoni con aspre leggi castigavano gli uomini grassi,  
conoscendo che così fatti corpi erano inutili a se stessi e agli  
altri.",  
"La confuse, se c'era, e io poteva farlo con tale esempio;  
ma s'ella c'era, doveva esserci per rispetto della quantità  
significata o della relazione, o pur per rispetto delle diverse  
misure della quantità.",  
"Non sarà dunque sconvenevole a l'epico, che somiglia l'uomo  
d'arme, usare alcuna saetta tolta da la faretra d'Ovidio; la  
qual vada a ferire in modo che la piaga porti seco il diletto  
accompagnato con la meraviglia."  
],  
"solution": 1
```

Box 4 An example of the
authorship verification task.
The second text is produced by
Stefano Guazzo, while the rest
three belong to Torquato
Tasso.

3.4 CAVALCANTI *RHYMES*: HEAD PREDICTION & DEPENDENCY RELATIONSHIP PREDICTION

The 22 manually annotated sonnets in *Rhymes* contain a total of 65 sentences, encoded in the CoNLL-U format⁷ with information on lemmatization, part-of-speech tagging, and dependency parsing. Two tasks are derived from this dataset to test the syntactic parsing capabilities of LLMs: head prediction and dependency relation prediction.

In the head prediction task, given a sentence, a selected token, and four potential head tokens, the LLM is prompted to identify which one is the true head of the selected token. Distractors are selected by prioritizing tokens that share the same case, gender, number, and part-of-speech features as the real head. One instance is created from each sentence. In the dependency relation prediction task, given a sentence, a selected token, its head, and four possible dependency relations, the LLM is prompted to identify the correct relationship. The distractor relations are selected by querying the most frequent dependency relations of tokens sharing the same part-of-speech feature as the selected token.

One head prediction question is created for each sentence. Two dependency relationship questions are created for each dependency relationship type. In total, 65 questions are created for the head prediction task, 70 questions are created for the dependency relation prediction task. The evaluation metric for both tasks is accuracy. Box 5 shows one example of each task.

Box 5 Examples of head prediction task (above) and dependency relation prediction task (below) derived from the Cavalcanti's sonnets.

```
"selected_token": "un",
"true_head": "pietà",
"distractors": ["amore","noia","gioia"],
"text": "Ma quando sento che sì dolce sguardo dentro da li occhi
mi passò a lo core e posevi uno spirito di gioia, di farne a lei
merzè già mai non tardo: così pregata foss'ella d'amore ch'un poco
di pietà no i fosse a noia.",

"selected_token": "una",
"deprel": "determiner",
"distractors": ["possessive determiner","nominal
subject","object"],
"text": "Allora par che ne la mente piova una figura di donna
pensosa, che vegna per veder morir lo core."
```

3.5 DANTESOURCE: INTERTEXTUALITY DETECTION

The DanteSource dataset serves as the foundation for an intertextuality detection task designed to evaluate the ability of LLMs to identify textual correspondences between Dante Alighieri's Italian works and their vernacular sources in Duecento poetry.

Expert-annotated citations from DanteSources were cross-referenced against full-text poems from the Biblioteca Italiana, yielding 416 verified positive citation pairs drawn from 60 source poems by 12 poets. Only citations classified as *concordanza stringente*, which means direct verbal or lexical correspondences, were retained, as looser thematic similarities are insufficient to constitute a well-defined ground truth. Each source poem was segmented into four-line chunks, with the one containing the cited lines labeled as the positive chunk and all remaining chunks from the same poem serving as hard negatives. This design ensures that the task cannot be solved by surface topical similarity or period vocabulary alone, since both the positive and negative candidates share the same author, style, and lexical register.

The task is structured as a binary choice problem: given a passage from Dante and two candidate chunks from the same source poem, the model must identify which chunk holds an intertextual relationship with the passage. The evaluation set comprises 219 questions, with accuracy as the evaluation metric. Box 6 shows an example of the intertextuality detection task.

⁷ <https://universaldependencies.org/format.html>.

```
"dante_text": "Già eran quasi che aterzate l'ore del tempo che
onne stella n'è lucente, quando m'apparve Amor subitamente, cui
essenza membrar mi dà orrore.",
"text_a": "ch'adesso con' fu 'l sole, sì tosto lo splendore fu
lucente, né fu davanti 'l sole; e prende amore in gentilezza
loco",
"text_b": "così propriamente come calore in clarità di foco. Foco
d'amore in gentil cor s'apprende come vertute in petra preziosa,"
```

Box 6 An entry of the Dante intertextuality detection task. In the shown example, *text_a* has an intertextual relation with *dante_text*.

4 OUTCOMES AND EXPERIENCE

To provide insight into LLMs' capabilities in historical Italian and offer a reference for future work, we evaluated the historical Italian benchmark using five LLMs covering a range of scales: the state-of-the-art DeepSeek-V3.2, Claude-Opus-4.6, and GPT-5.2, selected as popular, widely applied models that were the most recent releases at the time of writing, alongside the relatively smaller-scale Minerva-7B-instruct-v1.0 (Orlando et al., 2024), the first family of LLMs trained from scratch on Italian data, and Llama 3.1 8B Instruct (Grattafiori et al., 2024).

We evaluate the five forced-choice tasks (word sense disambiguation, head prediction, dependency relation prediction, intertextuality detection, and authorship verification) with accuracy. Each presents a closed set of candidate answers containing exactly one correct option, which makes them single-label classification problems. The candidate sets are balanced by construction, as every question offers the same number of options within the same task, and the position of the correct answer is randomized across questions. The remaining two tasks are scored differently: named entity recognition with precision, recall, and F1, and chronological ranking with the Spearman correlation coefficient.

Evaluation results, except for the NER tasks, are presented in Table 2. Table 3 presents the results on the two NER tasks. As can be seen, the performance varies greatly across models. Claude demonstrates to be the best performance model overall except on the Dante intertextuality task. Smaller LLMs lag substantially behind larger ones across most tasks, with Authorship Verification as the notable exception: both Minerva and Llama achieve results close to those of DeepSeek and GPT. Tasks developed on Cavalcanti dataset and Biblioteca Italiana dataset demonstrated to be harder. Almost all models, except for Claude, struggle on the authorship verification task and achieve accuracy close to the random guessing level of 0.25.

As for the NER tasks, Claude achieves better results than DeepSeek and GPT across both datasets under all settings. Despite its relatively small size, Llama achieves results close to those of larger models on the Bellini dataset, though the performance gap widens on Classense. By contrast, Minerva's evaluation results on NER are close to zero under all settings. A close examination on the results reveals that Minerva produces outputs dominated by hallucinations: it repeatedly identifies entities that never appear in the input. In general, the Bellini dataset proves to be more challenging than the Classense one, with LLMs performance consistently lower in all settings. Another pattern that can be observed is that there still is much space for improvement for both datasets, even under the fuzzy settings. To further investigate the models' performance, Table 4 and Table 5 report the per-class results on the NER tasks. For both datasets, the *LOC* class is the simplest, consistently achieving the highest F1 score among all classes, while classes

Table 2 Performance results on the historical Italian benchmark with 95% confidence intervals across rounds. The evaluation metric is Spearman correlation for Chronological Ranking and accuracy for the other tasks. *Random* indicates the expected value of each metric under completely random guessing.

MODEL	MANZONI	CAVALCANTI		DANTE	BIBLIOTECA ITALIANA	
	WORD SENSE DISAMBIGUATION	HEAD PREDICTION	DEPREL PREDICTION	INTERTEXTUALITY DETECTION	CHRONOLOGICAL RANKING	AUTHORSHIP VERIFICATION
DeepSeek	0.890 ± 0.022	0.533 ± 0.154	0.467 ± 0.020	0.848 ± 0.013	0.419 ± 0.123	0.313 ± 0.014
Claude	0.933 ± 0.038	0.733 ± 0.080	0.671 ± 0.035	0.886 ± 0.000	0.693 ± 0.126	0.747 ± 0.014
GPT	0.776 ± 0.293	0.735 ± 0.282	0.543 ± 0.061	0.910 ± 0.096	0.502 ± 0.062	0.310 ± 0.388
Minerva-7B	0.159 ± 0.013	0.274 ± 0.067	0.269 ± 0.128	0.470 ± 0.000	-0.088 ± 0.121	0.300 ± 0.000
Llama-3.1-8B	0.637 ± 0.031	0.308 ± 0.153	0.386 ± 0.061	0.553 ± 0.049	0.117 ± 0.132	0.250 ± 0.050
Random	0.167	0.25	0.25	0.5	0	0.25

Table 3 Performance on the NER tasks with 95% confidence intervals across rounds. The scores are macro-averaged.

MODEL	BELLINI									
	STRICT					FUZZY				
	P	R	F1	R	F1	P	R	F1	R	F1
DeepSeek	0.455 ± 0.013	0.388 ± 0.003	0.401 ± 0.004	0.631 ± 0.014	0.512 ± 0.003	0.545 ± 0.005	0.630 ± 0.050	0.594 ± 0.033	0.604 ± 0.039	0.680 ± 0.039
Claude	0.462 ± 0.015	0.531 ± 0.009	0.487 ± 0.011	0.641 ± 0.006	0.725 ± 0.013	0.672 ± 0.008	0.731 ± 0.010	0.696 ± 0.012	0.710 ± 0.008	0.789 ± 0.006
GPT	0.422 ± 0.111	0.390 ± 0.172	0.397 ± 0.145	0.593 ± 0.086	0.542 ± 0.175	0.556 ± 0.136	0.643 ± 0.034	0.668 ± 0.052	0.652 ± 0.044	0.700 ± 0.024
Minerva-7B	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.080 ± 0.000	0.003 ± 0.000	0.006 ± 0.000	0.080 ± 0.000
Llama-3.1-8B	0.359 ± 0.030	0.366 ± 0.065	0.333 ± 0.048	0.522 ± 0.008	0.496 ± 0.052	0.472 ± 0.030	0.495 ± 0.062	0.458 ± 0.035	0.467 ± 0.040	0.589 ± 0.043

Table 4 Per-class F1 on the Bellini NER dataset, strict (S) and fuzzy (F) matching, with 95% confidence intervals across rounds.

CLASS	SUPPORT	DEEPSEEK		CLAUDE		GPT		LLAMA-3.1-8B	
		S	F	S	F	S	F	S	F
LOC	124	0.501 ± 0.052	0.736 ± 0.031	0.625 ± 0.018	0.896 ± 0.009	0.562 ± 0.014	0.796 ± 0.046	0.468 ± 0.078	0.629 ± 0.065
MISC	10	0.000 ± 0.000	0.000± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
MUSIC_TERM	56	0.098 ± 0.040	0.223 ± 0.094	0.438 ± 0.026	0.673 ± 0.018	0.184 ± 0.029	0.332 ± 0.105	0.044 ± 0.020	0.096 ± 0.051
ORG	13	0.202 ± 0.019	0.270 ± 0.026	0.415 ± 0.014	0.646 ± 0.021	0.342 ± 0.114	0.513 ± 0.036	0.320 ± 0.081	0.529 ± 0.222
PER	194	0.411 ± 0.016	0.618 ± 0.005	0.490 ± 0.029	0.757 ± 0.016	0.430 ± 0.006	0.658 ± 0.022	0.332 ± 0.037	0.512 ± 0.007
PER_GROUP	9	0.199 ± 0.098	0.274 ± 0.257	0.145 ± 0.007	0.291 ± 0.013	0.205 ± 0.062	0.318 ± 0.094	0.000 ± 0.000	0.108 ± 0.055
WORK	22	0.403 ± 0.138	0.677 ± 0.091	0.520 ± 0.000	0.880 ± 0.000	0.441 ± 0.112	0.808 ± 0.077	0.341 ± 0.048	0.547 ± 0.102

Table 5 Per-class F1 on the Classense NER dataset, strict (S) and fuzzy (F) matching, with 95% confidence intervals across rounds.

CLASS	SUPPORT	DEEPSEEK		CLAUDE		GPT		LLAMA-3.1-8B	
		S	F	S	F	S	F	S	F
LOC	128	0.744 ± 0.026	0.754 ± 0.027	0.876 ± 0.014	0.885 ± 0.011	0.741 ± 0.039	0.748 ± 0.039	0.603 ± 0.041	0.651 ± 0.075
ORG	61	0.372 ± 0.083	0.420 ± 0.057	0.347 ± 0.041	0.347 ± 0.041	0.338 ± 0.099	0.364 ± 0.115	0.245 ± 0.107	0.292 ± 0.104
PER	274	0.693 ± 0.065	0.735 ± 0.054	0.771 ± 0.013	0.843 ± 0.002	0.727 ± 0.024	0.793 ± 0.009	0.519 ± 0.065	0.628 ± 0.030
WORK	169	0.419 ± 0.047	0.514 ± 0.024	0.530 ± 0.015	0.603 ± 0.005	0.499 ± 0.037	0.590 ± 0.043	0.292 ± 0.011	0.389 ± 0.111

such as *ORG*, *PER*, and *WORK* pose a greater challenge. It is worth noting that the special class *MUSIC_TERM* in *Bellini* yields markedly different results across models. While DeepSeek, Llama, and GPT struggle, with F1 scores below 0.2 in the strict setting and below 0.33 in the setting, Claude achieves relatively strong scores of 0.438 and 0.673, respectively. This suggests that recognizing specialized, domain-specific entities such as musical terminology could depend strongly on the individual model, making model choice far more consequential for niche entity types than for common ones like *LOC* or *PER*.

It is worth highlighting that this benchmark has several limitations, which stem from two sources. The first consists of limitations introduced by humans during construction, such as biases that the annotators of the original datasets may have introduced during annotation. We seek to address this issue by selecting only high quality data that are manually annotated by multiple annotators. The second consists of limitations inherent to LLMs themselves, such as data contamination. Data contamination refers to the case in which an evaluation dataset is already included in the training phase of an LLM, making evaluation results on that dataset unreliable (Ravaut et al., 2024). This issue is difficult, if not impossible, to resolve completely during benchmark construction, since many LLMs are closed-source and information about their training data is unavailable. Several measures are nonetheless taken to mitigate it. The texts involved in this benchmark, such as *I Promessi Sposi* and *Vita Nova*, are almost certain to be present in the training data of LLMs given their popularity and literary significance. The labels in the benchmark, however, are not the source texts themselves but are instead derived from other resources. For example, these include senses synthesized by aggregating multiple commentaries and dictionaries, syntactic dependencies paired with the most convincing distractors, and scholarly judgments of intertextuality. Consequently, prior exposure to a source text does not directly reveal the correct answer, as solving each task requires the derived annotation rather than verbatim recall of the original. This argument is admittedly weaker for the chronological ranking and authorship verification tasks, where the publication date and authorship effectively constitute the answer and could in principle be recovered from memorized metadata. To limit this risk, the snippets are deliberately selected by filtering based on semantic similarity, so that the tasks cannot be solved through surface-level similarity alone. While the complete absence of contamination cannot be guaranteed, these design choices ensure that memorization of the source material confers only limited advantage on the evaluation tasks.

In general, the historical Italian benchmark demonstrates to be a non-saturated benchmark, with none of the models achieving near-perfect performance. This indicates that the linguistic nuances, syntactical structures, and other specific factors inherent in pre-modern Italian texts still pose a significant challenge for LLMs.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the first benchmark specifically designed to evaluate the capabilities of large language models on historical Italian. By repurposing and adapting a diverse set of high-quality, manually-annotated language resources of various text genres, we constructed a multi-task evaluation framework that assesses LLMs on word sense disambiguation, named entity recognition, syntactic parsing, intertextuality detection, diachronic detection, and stylistic analysis.

Our evaluation of several state-of-the-art LLMs on this benchmark yields several findings. First, performance varies greatly across tasks and models. Large-scale LLMs achieve results superior to that of the smaller models, with Claude achieving the best performance in general. Second, the benchmark remains far from saturated; even the best-performing model's performance is still far from perfection. This indicates that the linguistic nuances and diachronic variations inherent in pre-modern Italian pose substantial challenges for contemporary LLMs.

The primary contribution of this work is a standardized evaluation framework that serves two critical functions: (1) guiding digital humanists in selecting appropriate models for automating or assisting annotation workflows in digital archives, and (2) providing the NLP community with a helpful tool to the development of more linguistically inclusive and historically-aware language models.

Future work will proceed in two directions. First, we invite the integration of additional historical Italian resources to expand the benchmark's coverage. Second, we aim to explore whether fine-tuning or in-context learning strategies can improve performance on the most challenging tasks.

APPENDIX A

Given a Context and a collection of explanations of a short phrase in the context, you need to process it for a historical language understanding task. Perform the following:

1. Determine Suitability for language understanding task: An entry is suitable ONLY if the explanations focus ****primarily on the linguistic meaning, etymology, or semantic interpretation**** of a phrase within the Context

An entry is ****NOT suitable if**** the explanations focus on:

- Historical background or events.
- Narrative, plot, or character analysis.
- Literary or stylistic appreciation.
- General commentary without clear linguistic explanation.

If you are uncertain or if the explanations are ambiguous, mark the entry as ****NOT suitable****.

2. Generate output:

- If "suitable entry" is true:
 - "target phrase" must be the phrase in the Context that the explanations refer to.
 - "answer" must be a concise explanation of the target phrase created based only on the provided Explanations. Do NOT include parentheses, quotes, slashes, or the target phrase itself.
- If "suitable entry" is false:
 - "target phrase" must be null.
 - "answer" must be null.

If there is any uncertainty about suitability, mark the entry as not suitable.

Return a JSON object strictly in this format: "suitable entry": < Boolean > , "answer": < String or null > , "target phrase": < String or null >

—

EXAMPLE

Context: "Ma gl'illustri Campioni che in tal Arringo fanno messe di Palme e d'Allori, rapiscono solo che le sole spoglie più sfarzose e brillanti"

Explanations:

1. battaglia.
2. arengo, recinto per la battaglia, per la giostra militare, quindi gara.
3. arengo, letteralmente indica il campo entro il quale si svolgevano i tornei, le gare, e quindi, per metonimia, il combattimento stesso.

Expected Output: "suitable entry": true, "answer": "Campo o recinto destinato a battaglie o gare, e per estensione il combattimento o la competizione stessa.", "target phrase": "arringo"

—

NOW PROCESS THE FOLLOWING ENTRY

Context: context

Explanations: explanations

Box 7 Prompt used for selecting entries suitable for word sense disambiguation task in the commentaries of project *LeggoManzoni*.

You are creating distractors for a multiple-choice question in a historical Italian language understanding task.

Task: Generate plausible incorrect options (distractors) for a question asking for the meaning of a target word in historical Italian.

Input Details:

- Target word to be explained: {target}
- One correct option (already provided): {answer}
- **Possible alternative meanings/senses (for reference): {sins}**

Requirements:

1. Generate 5 distractors that are plausible but have meanings different from the provided correct option.
2. Tone & Style: Match the linguistic register and explanatory style of the provided correct option.
3. Content Guidelines:
 - Distractors be semantically different to the correct option
 - Ensure distractors are plausible for Italian language
 - **Consider the provided alternative meanings ({sins}) as inspiration when suitable**
 - **If the provided alternatives aren't suitable, generate based on your knowledge of historical Italian**
4. Output Format: Return ONLY a Python list of strings containing exactly 5 distractors.

Example Structure: ["distractor 1", "distractor 2", "distractor 3", "distractor 4", "distractor 5"]

Important: Do not include the correct answer in your list. Generate only the 5 distractors.

Box 8 Prompt used for generating distractors for the word sense disambiguation task. The bold text is present only when generating distractors for single-word phrases.

ACKNOWLEDGEMENTS

We would like to thank the creators and curators of the datasets used in this work to generate the benchmarks.

FUNDING STATEMENT

The research leading to these Results is funded by:

The European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement FutureData4EU 101126733.

The central budget of the University of Bologna under the 'Departments of Excellence' initiative of the Italian Ministry of University and Research – Department of Classical Philology and Italian Studies.

The European Union – NextGenerationEU, PNRR Mission 4 Component 2 Investment 1.3, with Unique Project Code J33C22002850006.

The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Shibingfeng Zhang: Conceptualization, Data curation, Software, Project administration, Writing – original draft, Writing – review & editing.

Mariia Levchenko: Data curation, Software, Methodology, Writing – original draft, Writing – review & editing.

Chiara Manca: Data curation, Software, Methodology, Writing – original draft, Writing – review & editing.

Fiammetta Sabba: Conceptualization, Supervision.

Paola Maria Carmela Italia: Conceptualization, Supervision.

Giovanni Colavizza: Conceptualization, Supervision, Writing – review & editing.

Zhang et al.
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.538

16

AUTHOR AFFILIATIONS

Shibingfeng Zhang  orcid.org/0009-0005-7320-9088

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Mariia Levchenko  orcid.org/0009-0008-2649-0729

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Chiara Manca  orcid.org/0000-0003-4037-1175

Department of Cultural Heritage, University of Bologna, Ravenna, Italy

Fiammetta Sabba  orcid.org/0000-0002-9786-6825

Department of Cultural Heritage, University of Bologna, Ravenna, Italy

Paola Maria Carmela Italia  orcid.org/0000-0002-6281-7708

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Giovanni Colavizza  orcid.org/0000-0002-9806-084X

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Department of Communication, University of Copenhagen, Copenhagen, Denmark

REFERENCES

- Attanasio, G., Basile, P., Borazio, F., Croce, D., Francis, M., Gili, J., Musacchio, E., Nissim, M., Patti, V., Rinaldi, M., et al. (2024). CALAMITA: Challenge the abilities of LAngeuage models in ITAlian. *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 1054–1063).
- Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., Lin, Z., Su, W., Ge, T., Zheng, B., et al. (2024). MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7421–7454). <https://doi.org/10.18653/v1/2024.acl-long.401>
- Bartalesi, V., & Meghini, C. (2016). Using an ontology for representing the knowledge on literary texts: The Dante Alighieri case study. *Semantic Web*, 8(3), 385–394. <https://doi.org/10.3233/SW-150198>
- Bartalesi, V., Meghini, C., Metilli, D., Tavoni, M., & Andriani, P. (2018). A web application for exploring primary sources: The DanteSources case study. *Digital Scholarship in the Humanities*, 33(4), 705–723. <https://doi.org/10.1093/llc/fqy002>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Caffagni, D., Cocchi, F., Mambelli, A., Tutrone, F., Zanella, M., Cornia, M., Cucchiara, R., et al. (2025). Benchmarking BERT-based Models for Latin: A Case Study on Biblical References in Ancient Christian Literature. *Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science*.
- Cao, J., Liu, Y., Shi, Y., Ding, K., & Jin, L. (2024). WenMind: A Comprehensive Benchmark for Evaluating Large Language Models in Chinese Classical Literature and Language Arts. *Advances in Neural Information Processing Systems*, 37, 51358–51410. <https://doi.org/10.52202/079017-1626>
- Cavalcanti, G. (1902). *Le rime di Guido Cavalcanti*. N. Zanichelli.
- Ciaccio, C., Miaschi, A., & Dell’Orletta, F. (2025). Evaluating lexical proficiency in neural language models. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1267–1286). Association for Computational Linguistics. <https://aclanthology.org/2025.acl-long.64/>
- Corbetta, C., Colombi, A. E., Moretti, G., & Passarotti, M. (2025). What Is Better for Syntactic Parsing? A Comparison Between Supervised and Unsupervised Models on Dante and Cavalcanti. *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)* (pp. 292–303).
- Corbetta, C., Moretti, G., & Passarotti, M. (2024). Join together? combining data to parse Italian texts. *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 251–257).
- Cristofaro, S., Del Grosso, A. M., Mazzagufo, L., Sichera, P., & Spampinato, D. (2023). Bellini Digital Correspondence: A Model for Making Collaborative Digital Scholarly Editions. *2023 7th IEEE Congress on Information Science and Technology (CiSt)* (pp. 615–620). <https://doi.org/10.1109/CiSt56084.2023.10409920>
- Cristofaro, S., Del Grosso, A. M., Mazzagufo, L., Sichera, P., & Spampinato, D. (2025). Implementing collaborative digital scholarly editions: insights from Bellini Digital Correspondence. *International Journal of Information Science and Technology*, 9(1), 42–53.

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *NAACL-HTL* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Fu, Y., Huang, Z., Yang, L., Lu, Y., & Dai, Z. (2025). Chengyu-bench: Benchmarking large language models for chinese idiom understanding and use. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2355–2366). <https://doi.org/10.18653/v1/2025.emnlp-main.119>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv*. <https://arxiv.org/abs/2407.21783>
- Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2), 310–318. <https://doi.org/10.1002/asi.21256>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Lepagnol, P., Ghannay, S., Gerald, T., Servan, C., & Rosset, S. (2026). Format Matters: A Critical Evaluation of Output Formats for Prompting LLMs in SLU and NER. In S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, & A. Toral (Eds.), *Proceedings of the fifteenth language resources and evaluation conference (lrec 2026)* (pp. 7485–7497). European Language Resources Association (ELRA). <https://doi.org/10.63317/3osjjdr778fh>
- Levchenko, M., Nava, B., & Russo, E. (2025). TEI encoding as a unified structure for multilingual digital editions: The LeggoManzoni case study. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), *Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale. proceedings del XIV convegno annuale AIUCD2025* (p. 663). AIUCD. <https://doi.org/10.6092/unibo/amsacta/8380>
- Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al. (2025). DeepSeek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*. <https://api-docs.deepseek.com/>
- Liu, L., Vlachidis, A., Crymble, A., Lee, D., & Humbel, M. (2025). Towards Comparable Historical NER: Building a Shared Evaluation Corpus for 18th-Century Historical Texts. *Anthology of Computers and the Humanities*, 3, 968–982. <https://doi.org/10.63744/dwCJ80qvwAtr>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692. <https://arxiv.org/abs/1907.11692>
- Manca, C., Sabba, F., Sorbara, B., Tripodi, S., et al. (2025). Verso l'edizione digitale del carteggio Canneti-Fiacchi. In *Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale, proceedings del xiv convegno annuale aiucd 2025* (pp. 184–189). AIUCD.
- Manzoni, A. (1893–1902). I promessi sposi, raffrontati sulle due edizioni del 1825 e 1840: Con un commento storico, estetico e filologico di policarpo petrocchi. In P. Petrocchi (Ed.) Vols. 4. G. C. Sansoni.
- Manzoni, A. (2021). I promessi sposi. In E. Raimondi & L. Bottoni, (Eds.). Carocci.
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1–72. <https://doi.org/10.1145/3744746>
- Ni, S., Chen, G., Li, S., Chen, X., Li, S., Wang, B., Wang, Q., Wang, X., Zhang, Y., Fan, L., et al. (2025). A survey on large language model benchmarks. *arXiv*. <https://arxiv.org/abs/2508.15361>
- Orlando, R., Moroni, L., Huguet Cabot, P.-L., Conia, S., Barba, E., Orlandini, S., Fiameni, G., & Navigli, R. (2024). Minerva LLMs: The first family of large language models trained from scratch on Italian data. In F. Dell'Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), *Proceedings of the tenth italian conference on computational linguistics (clit-it 2024)* (pp. 707–719). CEUR Workshop Proceedings. <https://aclanthology.org/2024.clcit-1.77/>
- Paccosi, T., & Aprosio, A. P. (2022). KIND: an Italian multi-domain dataset for named entity recognition. *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 501–507). <https://doi.org/10.63317/4364mopok745>
- Ravaut, M., Ding, B., Jiao, F., Chen, H., Li, X., Zhao, R., Qin, C., Xiong, C., & Joty, S. (2024). How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*, 121.
- Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- Sabba, F. (2025). The PNRR project 'DigiLetClass – Digitizing letters of Classense Library'. *Bibliothecae.it*, 14(1), 224–231. <https://doi.org/10.60923/issn.2283-9364/23384>

- Sabba, F., Manca, C., Sorbara, B., & Tripodi, S. (2025). Carteggio Canneti-Fiacchi (v1.0.2) [Epistolari classensi tra edizione e digitalizzazione, Ravenna]. <https://doi.org/10.5281/zenodo.15683491>
- Santini, C., Barzaghi, S., Sernani, P., Frontoni, E., Melosi, L., & Alam, M. (2026). ENEIDE: A High-Quality Silver Standard Dataset for Named Entity Recognition and Linking in Historical Italian. *arXiv preprint arXiv:2603.29801*. <https://doi.org/10.63317/3hwbiiiv85eub>
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.
- Stopponi, S., Peels-Matthey, S., & Nissim, M. (2024). AGREE: a new benchmark for the evaluation of distributional semantic models of ancient Greek. *Digital Scholarship in the Humanities*, 39(1), 373–392. <https://doi.org/10.1093/lhc/fqad087>
- Tian, Z., Jarrett, D., Torres, J. E., & Amaral, P. (2021). BAHF: Benchmark of assessing word embeddings in historical Portuguese. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 113–119). <https://doi.org/10.18653/v1/2021.latechclfl-1.13>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv*. <https://arxiv.org/abs/2302.13971>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Z., Zhang, H., Li, X., Huang, K.-H., Han, C., Ji, S., Kakade, S., Peng, H., & Ji, H. (2025). Eliminating position bias of language models: A mechanistic approach. *International Conference on Learning Representations, 2025*, 91212–91239.
- Zaratiāna, U., Tomeh, N., Holat, P., & Charnois, T. (2024). GLiNER: Generalist model for named entity recognition using bidirectional transformer. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5364–5376). <https://doi.org/10.18653/v1/2024.naacl-long.300>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36, 46595–46623. <https://doi.org/10.52202/075280-2020>

TO CITE THIS ARTICLE:

Zhang, S., Levchenko, M., Manca, C., Sabba, F., Italia, P. M. C., & Colavizza, G. (2026). Repurposing and Adapting Language Resources for a Historical Italian Large Language Model Benchmark. *Journal of Open Humanities Data*, 12: 119. DOI: <https://doi.org/10.5334/johd.538>

Submitted: 15 March 2026

Accepted: 24 June 2026

Published: 11 September 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.