

Predicting the Fictional Time and Space of French Theatre Plays by Using Large Language Models

Matteo Romanello, Simon Gabay
Unité d'Humanités Numériques
Université de Genève
Geneva, Switzerland
{name.surname}@unige.ch

Nicola Carboni
School of Information Sciences
University of Illinois Urbana-Champaign
Champaign, USA
carboni@illinois.edu

Abstract—Research in literary geography has traditionally focused on cartography overlooking the diachronic dimension of literary spaces. To analyse both the temporal and spatial dimensions of a text, we collect 600 pre-editions automatically produced from digital facsimiles, and use Large Language Models (LLMs) to extract their spatiotemporal features. To determine their accuracy and adequacy, we evaluate and compare a representative sample of LLMs using three types of prompts: metadata, excerpts from the text, and a combination of information extracted from the text. Thanks to an *ad hoc* evaluation grid established for this experiment, the results are compared to a manually annotated corpus in order to evaluate the performance of various LLMs and the impact of different types of information provided. CLAUDE 3.7 SONNET appears to be the best model for this task, followed closely by the open-weight DEEPSEEK R1 (671B). Among smaller models, MISTRAL SMALL constitutes a valid cost-efficient alternative to more expensive LLMs.

Index Terms—Computational humanities, Large Language Models, Literary Geographies, Data Curation, Historical Data.

I. SPATIO-TEMPORAL HORIZON IN LITERATURE

Toponyms define a “geographic horizon” of literature in constant movement. Depending on maritime discoveries, literary fashions, or deeper artistic movements, authors tend to write about different places in their work. However, these places do not only have geographic coordinates but also diachronic ones: Ancient Rome is not Borgia’s Rome, and Biblical Jerusalem is not the one of the Crusades. For this reason, simultaneously understanding the geographic and temporal contexts of literary works is vital to exploring cultural trends. Yet, any analysis is hindered by the fact that spatio-temporal information is usually under-documented and/or unstructured and therefore difficult to exploit, especially at scale.

For several years now, geography has become a popular reading tool for literary scholars [1], using traditional [2] or digital [3] methods. By mentioning places, authors create imaginaries or landscapes, the study of which has become essential to a new critical reading of literature, especially via maps [4]. However, in this geographical turn of literary criticism, not enough attention has yet been paid to the fact that places are anchored not only in space but also in time.

M. Collot [5] has recently distinguished two main modes of this geographic reading of texts: geocriticism, which studies

the representations of a place in texts (e.g., the “literary memory” of a town), and geopoetics, which studies the presence of places in a work (e.g., the relationship between space and a literary genre). In this paper, we focus on the latter approach and evaluate the contribution of computational methods to geopoetics for a specific literary genre over a long period of time: 17th century French theatre.

II. RELATED WORK

In this section we briefly review related work in three distinct areas: the automatic extraction of time-related (a) and space-related information (b) from literary texts, as well as the usage of Large Language Models (LLMs) in computational literary studies (CLS) (c).

a) Time: Structured information about the temporal dimension of plays is rarely documented in existing knowledge bases. For example, at the time of this writing, only 30 of the 26,000 plays documented in Wikidata are linked with their temporal context. As structured data about the fictional time of plays is currently lacking, it is crucial to devise automatic approaches to gauge this information from the text itself. This task is still largely unexplored in computational literary studies [6], with the exception of a recent study on distinguishing historical from contemporary novels in Danish and Norwegian fiction [7].

b) Space: Although the extraction and resolution of geographical place names contained in literary texts is not the main focus of this article, we should point out that there exists abundant literature on this topic. The extraction of place names from historical digitised texts shares the same issues as other entity types (e.g., people), such as noisy OCR and historical variations in their naming [8]. Furthermore, the resolution of historical place names (often referred to as geo-coding) suffers from the limited coverage of geographical knowledge bases, which often lack historical forms of place names [9].

c) LLMs: Scholars in the field of CLS have just begun to explore the potential and implications of using LLMs for the analysis, annotation, and interpretation of literary texts. While a survey published in 2023 about machine learning approaches in CLS did not feature any generative approaches [6], recent works have employed such models for the analysis of various textual and narratological aspects, such as measuring the

Research has been funded by the FNS project N°220833.

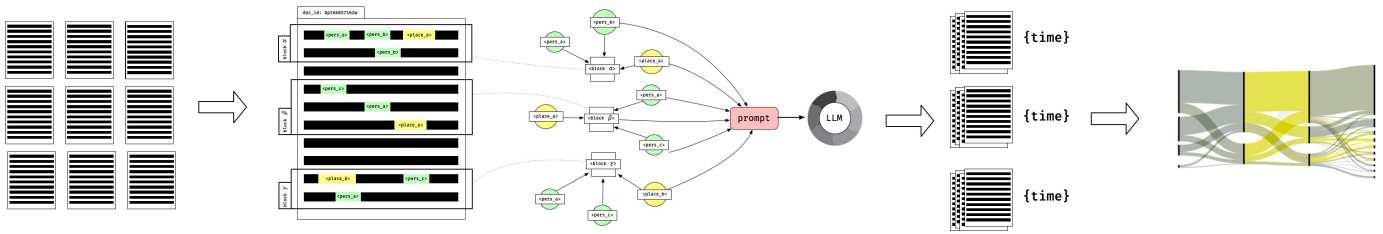


Fig. 1: Schematic representation of our hybrid approach to spatio-temporal enrichment.

passage of time in fiction [10], predicting narrative twists [11], detecting focalisation (i.e., identifying the perspective from which a narrative is presented) [12], performing authorship attribution [13], and capturing the transfer of knowledge about family relations between characters of a play [14]. However, as it was rightly noted [15], these initial explorations need to mature into a more systematic evaluation of LLMs’ capabilities, strengths, and limitations, an endeavour to which the present paper aims to contribute.

III. DATA

The automatic analysis of time and space in literary works is not a trivial endeavour. Notably intricate cases are travel stories where the characters move around the world, like Verne’s *Le Tour du monde en quatre-vingts jours*, or novels whose action spans several decades, such as García Márquez’s *Cien años de soledad*. To ensure consistency and analytical relevance, we decided to focus on theatre plays, as they are generally characterised by a unity of fictional time and space, especially in the French tradition [16].

We use an initial corpus of approximately 600 French theatre plays from the 17th century, pre-edited in XML-TEI format [17], [18]. The corpus (statistics in Table I) was compiled using a dedicated pipeline developed in a previous study [19] and contains plays sourced from *Gallica*¹. The textual data has been extracted from digital facsimiles via a combination of layout analysis, optical character recognition, linguistic normalisation, and named entity recognition (NER) through a model trained on the FrEM corpus [20]. It is therefore a noisy corpus, unlike that of *Théâtre Classique*². We chose to create our own corpus, as we plan to go beyond *Théâtre Classique*’s restrictive licences (CC BY-NC-ND) that hinder optimal data utilisation.

Aspect	Number
Documents	594
Tokens	12,885,306
Extracted entities	287,389

TABLE I: Basic corpus statistics.

IV. METHODS

In order to identify the fictional time and space of theatre plays, we devise a hybrid method that combines traditional NER and entity linking (EL) with the usage of LLMs. We benchmark three different prompt types against an *ad hoc* annotated dataset by using an LLM-judge approach.

¹<https://gallica.bnf.fr>

²<https://www.theatre-classique.fr>

A. Hybrid Approach to Spatio-temporal Processing

We identify, for each document in the corpus, the place and time in which the action is set, and we extrapolate the time dimension based on the recognised entities and the text. The rationale for this approach is simple: the named entities mentioned in a document provide useful contextual cues for situating a play on the temporal and geographical axes. In particular, we use a hybrid approach that combines traditional NER and EL with the usage of LLMs. For NER we employ the model described in [20], and for EL the off-the-shelf *entity-fishing* tool [21] via its spaCy wrapper³.

We leverage NER and EL to build an entity-graph for each document in our corpus (see Fig. 1); this graph is then used to identify central entities (i.e., the most frequently mentioned ones) and to select from the document salient sentences that mention them (extractive selection). We use this summary as input to an LLM to predict its spatio-temporal dimension.

This hybrid approach seeks to address the challenges inherent in a solely LLM-based method while relying on LLMs for tasks at which they tend to excel (e.g., reasoning over text). For example, we purposefully decided not to use LLMs for named entity processing given their limitations on historical texts [22]. Furthermore, working with a summary rather than the entire document mitigates the issue that the performance of LLMs typically degrades when working with longer contexts [23], which is often the case with literary texts. The choice of using a summary based on salient entities, as opposed to abstractive summarisation, aims to reduce the risk of factual inconsistencies that are often introduced by abstractive summarisation [24].

B. Large Language models (LLMs) and Prompting

As the LLM landscape is in rapid and constant evolution, in choosing the models to benchmark, we did not strive for exhaustivity with regard to the spectrum of LLMs we cover. We rather aimed at identifying a significant sample of models that a) includes the more recent reasoning LLMs and b) represents the variety of situations in which researchers might find themselves. To this end, we group LLMs into three tiers: *Tier 1* models can run on consumer hardware without the use of GPUs; *Tier 2* models usually require high-end hardware with at least one GPU; lastly, *Tier 3* models require several GPUs available simultaneously or are only available via paid APIs.

³<https://github.com/Lucaterre/spacyfishing>

Model	Provider	Tier	Params	Context
PHI 4 MINI	Ollama	1	3.8B	128k
GEMMA 3	Ollama	1	12B	128k
MISTRAL SMALL	Ollama	2	24B	32k
DEEPSEEK R1 (14B)	Ollama	2	14B	128k
DEEPSEEK R1 (32B)	Ollama	2	32B	128k
GPT 4O	OpenAI API	3	200B	128k
O1 MINI	OpenAI API	3	100B	128k
DEEPSEEK R1 (671B)	DeepSeek API	3	671B	128k
CLAUDE 3.7 SONNET	Anthropic API	3	>100B	200k

TABLE II: Basic information about the evaluated LLMs.

The LLM prompt devised for this task includes a JSON representation of the input document, as well as a list of key rules to follow; additionally, it specifies that the response should consist exclusively of a JSON object with a pre-defined structure. In this study, we experiment with three types of prompting:

- **Metadata** This prompt contains only the document metadata, namely author(s), title, and publication date.
- **Excerpt** This prompt provides the document’s metadata as well as a text excerpt of 400 tokens. As documents’ *incipits* often contain less informative text for the task at hand, such as dedications and prefaces, we extract these excerpts from the middle of the document.
- **Summary** This prompt combines the document’s metadata with an automatically generated summary, based on extracted person and place entities (see Fig. 2 for an example). In this summary, the top-5 person/place entities are listed, along with their frequency, their containing sentences and names.

C. Evaluation

1) *Benchmark data:* With the help of a domain expert, we manually annotated a random sample of 60 documents (corresponding approximately to 10% of the whole corpus).⁴ For each document, the expert was asked to provide details following the same template and guidelines that are included in the LLMs’ prompts (see Table III). Successively, annotations were compared with sample outputs generated by LLMs on the validation set in order to ensure adequate consistency, especially in the granularity of descriptions.

The spatio-temporal annotation of the data is particularly complex. The *pastorale* plays have a particularly vague spatio-temporal anchoring, as they can be set as much in ancient Greece as in modern France without any notable difference (e.g., character names, theme). We strive to provide annotations whenever possible, rather than leaving fields blank. Although impossible to date precisely, Greek myths are chronologically situated before the ancient period. The *pastorale* plays, while not always explicitly set in ancient Greece, remain extremely influenced by this period.

⁴Code and data used for this study can be found at <https://github.com/textEnt/chrono-spatial-processing/>.

```

{
  "metadata": {
    "author": "Genest, Charles-Claude",
    "title": "Zélonide, princesse de Sparte . Tragedie",
    "publication_date": "1682",
    "document_id": "bpt6k9807756q"
  },
  "context": {
    "people": {
      "top_1_person": {
        "entity": {
          "label": "Pyrrhus",
          "frequency": 52
        },
        "related_sentences": [
          "Mais cette jeune ardeur qui vous porte aux combats, Seigneur, aurait
          ↪ besoin d'armes et de Soldats Malgré ces hauts désirs notre Ville
          ↪ déserte Sans pouvoir se défendre à Pyrrhus est ouverte.",
          "..."]
        ]
      },
      "top_5_persons": [
        "Pyrrhus",
        "Pyrrhus",
        "Pyrrhus",
        "Phyllus",
        "Spartiates"
      ]
    },
    "places": {
      "top_1_place": {
        "entity": {
          "label": "Sparte",
          "frequency": 52
        },
        "related_sentences": [
          "D'autres ont prétendu disputer à Zélonide le titre de parfaite Héroïne
          ↪ Outre que la perfection absolue n'est pas toujours nécessaire aux
          ↪ Héros de la Tragédie, j'ai à répondre encore qu'on ne sait pas bien
          ↪ toutes les circonstances de la rupture de Zélonide avec Cléonime, et
          ↪ de son engagement avec Acorate mais que toutes les langes qu'on lui
          ↪ donne à Sparte, et les acclamations que font pour elle tant de Sages
          ↪ Vieillards, montrent assez qu'ils la regardaient comme une Princesse
          ↪ Héroïque:",
          "..."]
        ]
      },
      "top_5_places": [
        "Sparte",
        "Grèce",
        "France",
        "Paris",
        "Rome"
      ]
    }
  }
}

```

Fig. 2: JSON summary of document bpt6k9807756q. For the sake of conciseness and readability, we reproduce only the first related sentence (out of five) for each top person/place entity.

2) *Evaluation method:* For the task evaluation, we used the LLM-as-a-Judge approach [25] which consists in using an LLM to perform the assessment of the models’ predictions. This approach comes with two main advantages: its scalability, and a certain degree of flexibility in the model’s judgement (for example, an LLM-Judge is able to determine that “Hungary” and “Hongrie” are equivalent as they refer to the same place, and should thus be considered equivalent).

We perform a zero-shot reference-based evaluation as we ask the LLM-Judge to compare the model’s predictions against the ground truth, based on the evaluation criteria contained in the LLM-Judge’s prompt. In order to improve the quality and robustness of the assessment, we decompose the evaluation criteria into finer-grained scores: two for the fictional period (string, interval) and two for the fictional location (string, Wikidata QID). Our model assigns scores to predictions based on each criterion: 1 for identical or equivalent matches, 0.5 for partial overlap, and 0 for non-matches. Taking the case of the example in Table III, if the model’s response for the field `period_timeframe` is `-500, -300` (where the ground

Field	Value	Description
document_id	bpt6k9807756q	The document identifier
author	Genest, Charles-Claude	Metadata: author name(s)
title	Zelonide, princesse de Sparte. Tragedie	Metadata: title
publication_date	1682	Metadata: publication date
period	Ancient Greece	The historical period in which the action of the play is set
period_timeframe	-272, -272	The start and end of the historical period in which the action of the play is set, formatted as [±Y]YYYY
period_reason	Siege of Sparta by Pyrrhus of Epirus	A detailed explanation of how it was possible to identify the historical period (textual cues, reasoning steps, etc.)
preferred_location	Sparta	The geographic location where the action of the play takes place
acceptable_locations	Sparta Greece	All geographic locations that can be considered as acceptable correct answers
location_QID	Q5690	The Wikidata QID of the location where the action of the play takes place
location_acceptable_QIDs	Q5690 Q41	The Wikidata QIDs of all locations that can be considered as acceptable correct answers
location_reason	Sparta is mentioned	A detailed explanation of how it was possible to identify the location (textual cues, reasoning steps, etc.)

TABLE III: Example of an annotated document used for benchmarking.

truth annotation has -272 , -272), the assigned score will be 0.5 , which corresponds to partial overlap.

V. RESULTS AND DISCUSSION

A. Validation and selection of an LLM-Judge

When using an LLM-Judge approach for evaluation, it becomes essential to determine the extent to which the assessment of the LLM-Judge is aligned with that of human annotators.

To this end, we let two human annotators and a representative sample of LLM-Judge models assign scores to the models’ predictions on the validation set, for a total of 135 predictions (5 documents \times 3 prompts \times 9 models). We then computed the inter-annotator agreements (IAA) among human annotators, as well as between human annotators and LLM-Judges (see Table IV). The LLM-Judge whose evaluation is most similar to human evaluation is O1 MINI with an average IAA score of 0.90 (Krippendorff’s α), followed closely by DEEPSEEK R1 (671B) with 0.89. Based on this evidence, we used O1 MINI for the evaluation results discussed in the remainder of the paper.

TABLE IV: Inter-annotator agreement (IAA) scores for the human annotators and the LLM-judge candidates.

	Loc. QID	Loc. Str.	Per. Str.	Per. Interv.	Avg. IAA
Human annotators	0.94	0.97	0.74	0.88	0.89
DEEPSEEK R1 (670B)	0.96	0.94	0.74	0.90	0.89
CLAUDE 3.7 SONNET	0.89	0.97	0.81	0.88	0.89
MISTRAL SMALL	0.79	0.97	0.71	0.87	0.83
O1 MINI	0.96	0.98	0.76	0.91	0.90
GEMMA 3	0.55	0.96	0.69	0.88	0.77

B. Accuracy of LLM predictions

To evaluate the accuracy of LLM predictions on this task, we compute a **strict accuracy** (SACC) as well as a **lenient accuracy** (LACC). In the strict regime, we consider an answer to be correct only when it was assigned a score of 1.0 , thus indicating that the model’s response coincides with the ground truth. However, given the degree of fuzziness of this task, such strict accuracy does not help in capturing answers that, despite not coinciding with the ground truth, can nevertheless be considered *acceptable* and thus be used reliably for further analysis. Thus, in the lenient regime, we consider answers with a score of 0.5 and higher as correct.

The results in Table V indicate that the examined models struggle to match the **degree of precision** of domain experts in making predictions, particularly for fictional time compared to fictional space. In the SACC regime, fictional time prediction accuracy is lower than fictional space prediction, but this pattern reverses in the LACC regime. This result seems to confirm the observation that predicting fictional time, by nature, entails a higher degree of fuzziness. Generally, Tier 3 models significantly outperform those in Tier 1 and 2, with the notable exception of MISTRAL SMALL, a 24B multimodal model whose LACC accuracy closely resembles that of models several orders of magnitude larger.

An interesting insight we gained from this evaluation is that the evaluated LLMs cannot *reliably* predict the Wikidata identifier (QID) of the fictional location of a play, even when they do predict the correct location string. This tendency to hallucinate about QIDs can be seen in the existing gap between location string accuracy and location QID accuracy. MISTRAL SMALL’s prediction for the tragedy *Alcide* by Jean Galbert de Campistron gives a representative example of such hallucinations. The model predicts “Greece” (instead of Oechalia) as the fictional location but assigns to it the QID corresponding

TABLE V: Accuracy of the models’ predictions on the test set, with strict and lenient accuracy, on four fine-grained scores: fictional location QID (loc. QID), fictional location string (loc. str.), fictional period string (per. str.), and fictional period interval (per. interv.). For each model, we evaluated three different prompts: metadata, excerpt and summary. Best scores are marked in **bold**, while second-best scores are underlined.

Model	Prompt	STRICT ACCURACY				LENIENT ACCURACY			
		Loc. QID	Loc. Str.	Per. Str.	Per. Interv.	Loc. QID	Loc. Str.	Per. Str.	Per. Interv.
PHI 4 MINI	metadata	1.72	0.00	0.00	0.00	1.72	0.00	1.72	1.72
	excerpt	1.72	0.00	0.00	0.00	1.72	0.00	3.45	0.00
	summary	1.72	1.72	0.00	1.72	1.72	3.45	1.72	3.45
GEMMA 3	metadata	3.45	8.62	1.72	6.90	3.45	55.17	79.31	74.14
	excerpt	1.72	8.62	1.72	13.79	3.45	62.07	77.59	79.31
	summary	22.41	<u>37.93</u>	0.00	6.90	22.41	60.34	74.14	75.86
MISTRAL SMALL	metadata	8.62	18.97	0.00	<u>17.24</u>	29.31	53.45	81.03	81.03
	excerpt	5.17	12.07	0.00	10.34	20.69	39.66	72.41	68.97
	summary	13.79	<u>37.93</u>	1.72	24.14	32.76	67.24	81.03	84.48
DEEPSEEK R1 (14B)	metadata	1.72	17.24	0.00	6.90	5.17	31.03	55.17	51.72
	excerpt	0.00	20.69	0.00	10.34	1.72	48.28	53.45	62.07
	summary	1.72	25.86	0.00	6.90	3.45	43.10	51.72	51.72
DEEPSEEK R1 (32B)	metadata	8.62	27.59	3.45	1.72	24.14	55.17	62.07	68.97
	excerpt	5.17	17.24	3.45	10.34	22.41	46.55	72.41	72.41
	summary	5.17	41.38	0.00	5.17	10.34	65.52	68.97	74.14
GPT 4O	metadata	10.34	12.07	1.72	10.34	34.48	43.10	53.45	53.45
	excerpt	10.34	8.62	1.72	<u>17.24</u>	41.38	56.90	77.59	79.31
	summary	29.31	36.21	1.72	8.62	<u>44.83</u>	<u>68.97</u>	<u>82.76</u>	79.31
O1 MINI	metadata	13.79	13.79	3.45	15.52	31.03	44.83	43.10	50.00
	excerpt	18.97	20.69	1.72	15.52	43.10	58.62	56.90	60.34
	summary	32.76	46.55	0.00	12.07	41.38	63.79	62.07	67.24
DEEPSEEK R1 (671B)	metadata	20.69	24.14	3.45	13.79	41.38	56.90	68.97	72.41
	excerpt	17.24	18.97	1.72	6.90	43.10	58.62	70.69	79.31
	summary	<u>37.93</u>	46.55	1.72	8.62	53.45	67.24	79.31	77.59
CLAUDE 3.7 SONNET	metadata	36.21	<u>37.93</u>	20.69	10.34	<u>44.83</u>	<u>68.97</u>	77.59	81.03
	excerpt	20.69	24.14	3.45	13.79	50.00	67.24	89.66	<u>82.76</u>
	summary	39.66	46.55	<u>6.90</u>	12.07	53.45	79.31	<u>82.76</u>	74.14

to Isaac Newton (Q935) instead of Oechalia’s (Q3594069) or Greece’s (Q41). These hallucinations, however, can easily be corrected by querying Wikidata at post-processing time, provided that the location string has been correctly predicted.

Among the tested prompts — metadata, excerpt, and summary (see Section IV-B) — the summary prompt tends to perform best. Interestingly, reasoning models such as DEEPSEEK R1 (671B) and CLAUDE 3.7 SONNET seem more sensible than the others to the noisy information present in the automatically generated summaries, which at times act as a confounding factor for the model. A telling example of this phenomenon comes from DEEPSEEK R1 (671B)’s predictions on the comedy *Les Ménechmes* by Jean de Rotrou. When prompted with the document excerpt, the model correctly predicts the fictional location as “Epidamnus”; the model’s reasoning shows that here the model is using prior knowledge about the fictional location of Plautus’ comedy, of which Rotrou’s is an adaptation. Instead, when prompted with the document summary, the model erroneously predicts the location string as “Sicily”, as Sicily is a frequently referenced location in the play.

C. Scalability of the approach

As mentioned above, the corpus we used for this study is a tiny fraction of a much bigger corpus which is currently being produced and whose expected size is ca. 10k documents (i.e., theatre plays).

Thus, the cost and runtime of the evaluated LLMs are important factors for selecting the model to be deployed for inference on the final corpus.

In Table VI, we estimate costs and runtime for all evaluated models. These figures are based on model predictions on 59 evaluation documents. We computed average processing time and token consumption (only for API-queried models). API models are

TABLE VI: Average-based estimates of cost and computing time for processing 1k documents.

Model	Cost (USD)	Time (hrs)
PHI 4 MINI	0.35	3.21
GEMMA 3	0.46	4.15
MISTRAL SMALL	0.97	8.83
DEEPSEEK R1 (14B)	2.11	19.20
DEEPSEEK R1 (32B)	4.79	43.57
GPT 4O	3.52	1.08
O1 MINI	4.78	1.94
DEEPSEEK R1 (671B)	2.87	17.61
CLAUDE 3.7 SONNET	6.55	1.45

priced based on the token cost announced by the API provider, while local models are priced based on the hourly cost of operating on the computing infrastructure⁵.

This benchmark provides some useful guidance for choosing an LLM based on available financial and computing resources. Processing a 10k document corpus with the most accurate model, CLAUDE 3.7 SONNET, would cost slightly less than 70 USD. This cost can be halved by trading some accuracy and choosing DEEPSEEK R1 (671B). MISTRAL SMALL offers a cost-efficient alternative, as this smaller model (24B parameters) runs effectively on high-end consumer hardware. It is suitable for multiple corpus processing or tight budgets, though post-processing is needed to eliminate hallucinated Wikidata QIDs for fictional locations.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a hybrid approach to predicting the fictional time and location of theatre plays that combines named entity recognition and linking with the usage of LLMs. This approach, based on automatically generated document summaries, proves beneficial as it provides LLMs with useful contextual information and performs better than a random document excerpt for most of the tested models, with the exception of CLAUDE 3.7 SONNET and, partly, DEEPSEEK R1 (671B). While we evaluated LLMs in a zero-shot setting, it remains to be assessed to what extent they can benefit from in-context learning or fine-tuning for this specific task. Moreover, as far as named entity extraction is concerned, a systematic evaluation of existing tools on French literary texts is still missing and will be highly useful in order to better understand its impact on the prediction of fictional time and space.

To sum up, the contribution of our paper is twofold. Firstly, our work contributes to a better understanding of the capabilities of LLMs when used for the analysis and interpretation of literary texts, an area where systematic evaluations are currently lacking. Secondly, we created and published a manually curated benchmark dataset that can be used to evaluate new models as they become available.

REFERENCES

- [1] N. Alexander and D. Cooper, Eds., *The Routledge Handbook of Literary Geographies*. London: Routledge, 2024. [Online]. Available: <https://doi.org/10.4324/9781003097761>
- [2] M. Collot, *Pour une géographie littéraire*. Paris: José Corti, 2014.
- [3] F. Moretti, *Atlas of the European Novel 1800-1900*. London and New York: Verso, 1998.
- [4] B. Piatti, "Mit Karten lesen. Plädoyer für eine visualisierte Geographie der Literatur," *Textwelt-Lebenswelt. Interpretation interdisziplinär*, vol. 10, pp. 261–288, 2012. [Online]. Available: https://www.literaturatlas.eu/files/2012/03/Piatti_TextweltLebenswelt.pdf
- [5] M. Collot, "Tendances actuelles de la géographie littéraire," *Histoire de la recherche contemporaine*, vol. 10, pp. 37–43, 2021. [Online]. Available: <https://doi.org/10.4000/hrc.5514>
- [6] H. O. Hatzel, H. Stiemer, C. Biemann, and E. Gius, "Machine Learning in Computational Literary Studies," *it - Information Technology*, vol. 65, no. 4, pp. 200–217, 2023. [Online]. Available: <https://doi.org/10.1515/itit-2023-0041>

⁵The benchmark was run on a single cluster node of the high-performance computing cluster of the University of Geneva, equipped with two NVIDIA TITAN XP GPUs (12GB vMem, computing capability 6.1)

- [7] J. Bjerring-Hansen, A. Al-Laith, D. Hershovich, A. Conroy, and S. Ørtoft Rasmussen, "Literary Time Travel: Distinguishing Past and Contemporary Worlds in Danish and Norwegian Fiction," in *Proceedings of the Computational Humanities Research Conference 2024*, vol. 3834. Aarhus, Denmark: CEUR-WS, 2024, pp. 772–787. [Online]. Available: <https://ceur-ws.org/Vol-3834/paper19.pdf>
- [8] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, "Named Entity Recognition and Classification in Historical Documents: A Survey," *ACM Computing Surveys*, Jun. 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3604931>
- [9] H. Kiiskinen, A. Nivala, J. Westerlund, and J. Saarelainen, "Extracting Geographical References from Finnish Literature. Fully Automated Processing of Plain-Text Corpora," *Journal of Computational Literary Studies*, vol. 2, no. 1, Jan. 2024. [Online]. Available: <https://jcls.io/article/id/3584/>
- [10] T. Underwood. (2023) Using GPT-4 to measure the passage of time in fiction. [Online]. Available: <https://tedunderwood.com/2023/03/19/using-gpt-4-to-measure-the-passage-of-time-in-fiction>
- [11] —, "Can language models predict the next twist in a story?" Jan. 2024. [Online]. Available: <https://tedunderwood.com/2024/01/05/can-language-models-predict-the-next-twist-in-a-story>
- [12] R. M. M. Hicke, Y. Bizzoni, P. Feldkamp, and R. D. Kristensen-McLachlan, "Says Who? Effective Zero-Shot Annotation of Focalization," 2024. [Online]. Available: <https://arxiv.org/abs/2409.11390>
- [13] R. M. M. Hicke and D. Mimno, "T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models," in *Proceedings of the Computational Humanities Research Conference 2023*, ser. CEUR Workshop Proceedings, A. Šeja, F. Jannidis, and I. Romanowska, Eds., vol. 3558. Paris, France: CEUR, Dec. 2023, pp. 274–302. [Online]. Available: <https://ceur-ws.org/Vol-3558/#paper2757>
- [14] J. Pagel, A. Pichler, and N. Reiter, "Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama," in *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, and S. Szpakowicz, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1–10. [Online]. Available: <https://aclanthology.org/2024.latechclfl-1.1>
- [15] F. Ciotti, "Gli LLM come lettori modello artificiali." AIUCD, 2024. [Online]. Available: <https://art.torvergata.it/handle/2108/386163>
- [16] P. Corneille, *Le Théâtre de P. Corneille*. Rouen: A. Courbé et G. de Luyne, 1660, ch. Discours des trois unités, d'action, de jour, et de lieu, pp. v–xxxvii. [Online]. Available: <https://gallica.bnf.fr/ark:/12148/bpt6k12802528/f13.item>
- [17] A. Pinche, K. Christensen, and S. Gabay, "Between Automatic and Manual Encoding," in *TEI 2022 conference : Text as data*, 2022. [Online]. Available: <https://hal.science/hal-03780302>
- [18] S. Gabay, A. Pinche, K. Christensen, and J.-B. Camps, "SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles," *Journal of Data Mining and Digital Humanities*, Dec. 2024. [Online]. Available: <https://hal.science/hal-04343404>
- [19] S. Gabay and T. Clérice, "The Birth of French Orthography. A Computational Analysis of French Spelling Systems in Diachrony," in *Proceedings of the Computational Humanities Research Conference 2024*. Aarhus, Denmark: CEUR Workshop Proceedings, 2024, pp. 246–264. [Online]. Available: <https://ceur-ws.org/Vol-3834/paper21.pdf>
- [20] P. Ortiz Suarez and S. Gabay, "A Data-driven Approach to Named Entity Recognition for Early Modern French," in *Proceedings of the 29th International Conference on Computational Linguistics*, no. 2022.coling-1.327. Gyeongju, South Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3722–3730. [Online]. Available: <https://hal.science/hal-04110765>
- [21] L. Foppiano and L. Romary, "Entity-fishing: A DARIAH Entity Recognition and Disambiguation Service," *Journal of the Japanese Association for Digital Humanities*, vol. 5, no. 1, pp. 22–60, 2020. [Online]. Available: https://doi.org/10.17928/jjadh.5.1_22
- [22] C.-E. González-Gallardo, E. Boros, N. Girdhar, A. Hamdi, J. G. Moreno, and A. Doucet, "Yes but.. Can ChatGPT Identify Entities in Historical Documents?" in *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2023, pp. 184–189. [Online]. Available: <https://doi.org/10.1109/JCDL57899.2023.00034>
- [23] M. Levy, A. Jacoby, and Y. Goldberg, "Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large

- Language Models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024, pp. 15 339–15 353. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.818>
- [24] F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, and B. Xiang, “Entity-level Factual Consistency of Abstractive Text Summarization,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 2727–2733. [Online]. Available: <https://doi.org/10.18653/v1/2021.eacl-main.235>
- [25] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo, “A Survey on LLM-as-a-Judge,” Mar. 2025. [Online]. Available: <https://arxiv.org/abs/2411.15594>