It's All About the Confidence: An Unsupervised Approach for Multilingual Historical Entity Linking using Large Language Models

Anonymous ACL submission

Abstract

Despite the recent advancements in NLP with the advent of Large Language Models (LLMs), Entity Linking (EL) for historical texts remains challenging due to linguistic variation, noisy inputs, and evolving semantic conventions. Existing solutions either require substantial training data or rely on domain-specific rules that limit scalability. In this paper, we present MHEL-LLaMo (*Multilingual Historical Entity Linking with Large Language MOdels*[1]), an unsupervised ensemble approach combining a Small Language Model (SLM) and an LLM. MHEL-LLaMo leverages a multilingual bi-encoder (BELA) for candidate retrieval and an instruction-tuned LLM for NIL prediction and candidate selection via prompt chaining. Our system uses SLM's confidence scores to discriminate between easy and hard samples, applying an LLM only for hard cases. This strategy reduces computational costs while preventing hallucinations on straightforward cases. We evaluate MHEL-LLaMo on four established benchmarks in six European languages (English, Finnish, French, German, Italian and Swedish) from the 19th and 20th centuries. Results demonstrate that MHEL-LLaMo outperforms state-of-the-art models without requiring fine-tuning, offering a scalable solution for low-resource historical EL. Our error analysis reveals that 41% of false predictions exhibit semantic proximity to ground truth entities, highlighting the LLM's accurate disambiguation of historical references.

1 Introduction

Entity Linking (EL) is a crucial Information Extraction (IE) task, allowing researchers to automatically trace and link references to named entities to external Knowledge Bases (KBs) for semantic enrichment, document indexing, and linking to external sources. EL has been used in a variety of use cases, such as social network analysis (Groth and Dietze, 2024; Santini et al., 2024), media monitoring (Chen et al., 2022), and artwork and motif recognition (Santini et al., 2022; Sierra-Múnera et al., 2024).

Despite the remarkable performance of EL systems on contemporary benchmarks, this task remains challenging with historical texts due to complex document varieties, noisy inputs, multilinguality, and evolving syntactic and semantic conventions[2]. Some approaches have been proposed to adapt EL systems to deal with historical documents, relying either on supervised architectures (Linhares Pontes et al., 2022) that need training data and careful parameter tuning or rule-based constraints which do not scale well across different domains (Labusch and Neudecker, 2020). More recently, Large Language Models (LLMs) have been employed to perform EL on historical texts with zero-shot prompting (Boscariol et al., 2024; Graciotti et al., 2025b); however, the results obtained are still inferior to those of supervised domain-specific approaches (Labusch and Neudecker, 2020; Boros et al., 2020; Linhares Pontes et al., 2022).

The goal of this paper is to perform Multilingual Historical Entity Linking (MHEL) in an unsupervised manner by employing LLMs in combination with a Small Language Model (SLM). In specific, we use BELA (Plekhanov et al., 2023), a multilingual XLM-R-based bi-encoder (Conneau et al., 2020) for retrieving a set of candidates from Wikidata similar to a given mention. Then, we leverage *prompt chaining* to guide an LLM in either selecting the most plausible candidate from the set of Wikidata entities or predicting a NIL entity when

[1]This acronym was chosen because its pronunciation resembles the Spanish phrase *"Me llamo"*, which means *"My name is"*, a common way to declare someone's identity.

[2]For more details and examples, a comprehensive survey of IE tasks in historical texts has been presented in (Ehrmann et al., 2023).

no candidate corresponds to the entity referenced in the text. Based on the findings of (Ma et al., 2023), we investigate how confidence scores returned by the bi-encoder may improve the performance of LLMs by allowing them to discriminate between easy and hard samples. The main hypothesis of this work is that employing LLMs only for hard samples to refine the bi-encoder's results provides a threefold advantage:

1. the background knowledge of LLMs is leveraged to deal with complex named entity mentions;

2. by keeping the bi-encoder's results for easy samples, LLMs are prevented from hallucinating on simple cases;

3. by using LLMs only for hard samples, energy costs for model inference are minimized.

Therefore, the main contribution of this work is to propose an unsupervised ensemble architecture for MHEL called MHEL-LLAMO (*Multilingual Historical Entity Linking with Large Language MOdels*), where an LLM performs sequential NIL prediction and candidate selection over a block of entities returned by a bi-encoder only for hard samples, i.e., when confidence scores of the bi-encoder do not reach a minimum threshold. To prove the effectiveness of our approach, we test both a vanilla combination of BELA with LLMs and our ensemble solution MHEL-LLAMO on several datasets with texts spanning from the 19th to the 20th century written in six languages (English, Finnish, French, German, Italian and Swedish).

Our empirical findings show that MHEL-LLAMO outperforms most of the specialized models for MHEL and suggest that adaptively employing LLMs via prompt chaining to refine the bi-encoder's results leads to robust performance on multilingual historical texts belonging to different genres and represents a novel solution to democratize access to advanced IE techniques for low-resource historical languages. To the best of our knowledge, this is the first successful attempt to employ LLMs to adapt general-purpose EL systems to historical texts. The source code and data used for this study are publicly available[3].

[3] https://anonymous.4open.science/r/MHEL-LLAMO-B34F.

2 Related Work

In this section, we discuss related work on historical IE and EL.

2.1 Neural Approaches for MHEL

Historical texts present important challenges for EL systems due to the complexity of historical texts, the presence of OCR errors, and the high frequency of NIL entities, i.e., entities that do not appear in a KB. To solve these problems, several specialized MHEL systems have been proposed in recent years. *SBB* (Labusch and Neudecker, 2020) is a multilingual framework supporting English, German, and French languages, structured around three core modules: (1) candidate retrieval using word embeddings and k-NN search on Wikipedia titles, with temporal filtering based on document publication dates and Wikidata information; (2) candidate evaluation via BERT-based similarity scoring between mention contexts and Wikipedia entity sentences; and (3) candidate re-ranking using Random Forest classification on combined similarity scores. Despite the originality of this work, the proposed approach has potential disadvantages due to the use of Wikidata information for filtering candidates before retrieval, which may lead to a higher false negative rate.

The *L3i* system (Boros et al., 2020) also presents a neural approach for MHEL on English, French, and German. This method builds upon the end-to-end EL framework of (Kolitsas et al., 2018) by adding three key enhancements: an OCR error correction module, a cross-lingual vector index derived from multilingual Wikipedia entity representations, and post-processing heuristics that rerank candidates and identify NIL entities using DBpedia information. Building on this foundation, (Linhares Pontes et al., 2022) introduced MELHISSA, an enhanced version of the L3i system also applied to Swedish and Finnish. MELHISSA extends L3i by incorporating 18 specialized post-processing filters for candidate reranking and NIL entity prediction, utilizing string distance metrics such as Levenshtein distance alongside information from both Wikidata and DBpedia. Despite MELHISSA and L3i's robust performance across the HIPE-2020 (Ehrmann et al., 2020) and NewsEye (Hamdi et al., 2021) datasets, their extensive reliance on rule-based parameters and training data creates reproducibility challenges in the absence of large annotated datasets and extensive empirical tests.

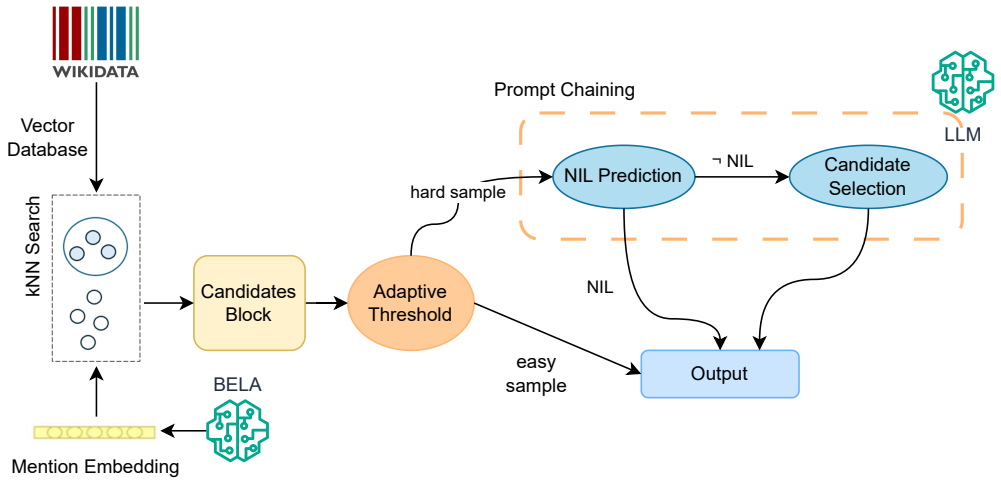

Figure 1: Overview of the MHEL-LLaMo architecture. The system combines BELA's bi-encoder for candidate retrieval, a KB consisting of a Faiss Index, and a lookup table for returning similar candidates from Wikidata with metadata, an adaptive threshold to filter easy samples and an instruction-tuned LLM for NIL prediction and candidate selection for hard samples.

To overcome the limitations of the approaches herein described, this work proposes to leverage the background knowledge of LLMs to perform NIL prediction and candidate selection in MHEL in an unsupervised way. While these models notably suffer from hallucinations, which may limit their efficiency in retrieving the correct candidate from Wikidata with a zero-shot approach, our idea is to adopt a pre-trained multilingual bi-encoder model for candidate retrieval. Further, we frame the NIL prediction and candidate selection tasks as multiple-choice question-answering tasks for LLMs, an activity for which instruction-tuned agents are thoroughly trained.

2.2 Historical IE with LLMs

Since the advent of LLMs, numerous approaches of these models for IE tasks on historical texts have been investigated. In (González-Gallardo et al., 2023, 2024), ChatGPT (Brown et al., 2020), and open-source LLMs (including LLaMa2 & 3, Mistral and Zephyr) were used in zero-shot prompting for Named Entity Recognition (NER) on the historical corpus HIPE-2022 (Ehrmann et al., 2022). While the experiments show that LLMs are generally capable of understanding IE instructions and recognizing simple classes of entities (e.g., person and location), they present notable issues due to the necessity of adopting post-processing strategies to control output formatting and remove instances of new classes of entities not present in the benchmark.

Recently, (Santini et al., 2024) compared open-source LLMs with models based on the GliNER (Zaratiana et al., 2024) architecture for NER on a corpus of 19th-century Italian philological texts. The main insight of this research was to show that LLMs exhibit important biases in the recognition of named entities, underperforming on domain-specific classes such as literary works. Moreover, (Graciotti et al., 2025b) presented a novel benchmark study on multilingual historical music periodicals where they tested GPT-4o mini and LLaMa3-70B in NER and EL with zero-shot prompting. While the results show how large instruction-tuned models surpass 50% of accuracy in both tasks using zero-shot prompting, the study also emphasize that these models perform poorly on long-tail entities, i.e., entities which occur less frequently in web-crawled corpora. To the best of our knowledge, MHEL-LLaMo is the first successful attempt to apply LLMs for MHEL.

3 Methodology

MHEL-LLaMo is based on three main components:

1. a candidate retrieval module based on BELA (Plekhanov et al., 2023);

2. a lookup table containing structured metadata about Wikidata entities, stored in an SQLite database;

3. an instruction-tuned LLM which performs NIL prediction and candidate selection based on a textual mention of an entity and a set of similar Wikidata candidates returned by the retrieval module.

A high-level representation of the approach is presented in Figure 1. Initially, BELA's XLM-R bi-encoder is used to encode an entity mention and its surrounding context in a vector (Conneau et al., 2020). This embedding is used to perform k-NN search on a dense vector index (Douze et al., 2024) of the same dimensionality, containing latent representations of Wikidata entities extracted from Wikipedia in 97 languages. The k-NN search returns the top-k most similar candidates to a mention embedding, ranked by cosine similarity.

The lookup table component is used to enrich the information about every candidate with structured information from Wikidata. For every candidate retrieved by the bi-encoder, the lookup table returns its label and description in the language of the text, along with the earliest date found in Wikidata and its entity type. This information allows the LLM to better discriminate the correct candidate based on KB information.

We then use an instruction-tuned LLM via prompt chaining to sequentially perform NIL prediction and candidate selection by evaluating the similarity between a text mention and each Wikidata candidate returned by the bi-encoder. As the pre-trained instruction-tuned LLM, we employed Mistral-24B[4] for texts in English, French, German, and Italian, Gemma-3-27B[5] for texts in Swedish, and Poro-2-8B[6] for texts in Finnish. These models were chosen for their strong performance across multiple benchmarks as well as their being open-source and versioned, thus supporting reproducibility.

Inspired by the findings of (Ma et al., 2023), we exploit the confidence scores of the retrieval module (i.e., the cosine similarities) to discriminate between easy and hard samples. Following their approach, we perform NIL prediction and candidate selection only if the scores of the candidates are below a certain threshold, thus considering a mention as a hard sample. If at least one candidate has a score higher than or equal to the threshold, we link the mention to the candidate with the highest similarity.

Our prompt chaining strategy for NIL prediction and candidate selection follows a two-step approach. First, we ask the instruction-tuned LLM to predict whether an entity is NIL by determining whether there is at least one exact match between a text mention and one of the candidates returned by the bi-encoder. If the answer is *"no"*, we classify the entity as NIL; otherwise, we run a separate prompt to ask the LLM to select the most similar entity among the candidates with respect to the mention. It is important to note that in this selection step the LLM can still return an empty set if it is not able to confidently determine which is the most plausible entity. If the LLM returns one entity from the candidate set, we consider it as a correct link; otherwise, we classify the entity as NIL. A detailed explanation of the prompts used is available in Appendix A.

This approach is completely unsupervised since it relies on the frozen weights of BELA and the instruction-tuned LLMs. The only two hyperparameters that must be specified are the number of candidates for each mention and the threshold for filtering hard samples, which we selected empirically by finding the best configuration on the development set of each dataset. The values used in our experiments for k-NN search and the thresholds applied to BELA confidence scores are reported in Appendix B.

4 Experimental Setup

4.1 Evaluation Datasets

We evaluated our approach using four established benchmarks: HIPE-2020 (Ehrmann et al., 2020), NewsEye (Hamdi et al., 2021), AJMC[7] (Romanello and Najem-Meyer, 2024), and MHERCL[8] (Graciotti et al., 2025a,b). Overall, these benchmarks include texts in 6 languages: English, Finnish, French, German, Italian and Swedish. They belong to various genres: newspapers, music periodicals and classical commentaries (i.e. scholarly notes to Greek classical works) and span from the nineteenth century (19C) to the twentieth century (20C). Statistics regarding number of documents and number of mentions in each test set and other characteristics are reported in Table 1. As shown in

[4] mistralai/Mistral-Small-24B-Instruct-2501
[5] google/gemma-3-27b-it
[6] LumiOpen/Llama-Poro-2-8B-Instruct

[7] HIPE-2020, NewsEye and AJMC are available on the Github repository of the HIPE-2022 Shared Task: github.com/hipe-eval/HIPE-2022-data
[8] github.com/polifonia-project/KE-MHISTO

the table, almost all datasets (except AJMC) have a percentage of NIL mentions higher than 25%.

4.2 Architecture Configuration

We evaluated MHEL-LLAMO using two distinct configurations:

- **MHEL-LLAMO**$_{van}$ (Vanilla Approach): This configuration uses the LLM for NIL prediction and candidate selection for every block of entities returned by BELA;
- **MHEL-LLAMO**$_{\theta}$ (Ensemble Architecture): This approach employs an adaptive threshold based on BELA's confidence scores to distinguish between easy and hard samples. Easy samples are those with high-confidence candidates, while hard samples require LLM intervention for NIL prediction and candidate selection.

Each configuration was tested under two settings: (a) *chain*: the LLM uses in sequence the NIL prediction prompt followed by the candidate selection prompt (see Section 3); (b) *single*: the LLM performs candidate selection and eventually predicts NIL using only the candidate selection prompt. This experimental design yields four distinct variants of MHEL-LLaMo, allowing us to analyze the individual and combined effects of the adaptive threshold mechanism and the prompt chaining strategy.

We determined the optimal hyper-parameters for candidate block size and threshold values by selecting configurations that achieved the best performance on each dataset's development set. These hyper-parameters are detailed in Appendix B, Table 4. Experiments were run on two NVIDIA L40S GPUs (2×46GB) with a total computational budget of approximately 51 GPU hours across all datasets and languages. To reduce computational costs, we reported results obtained with a single run.

4.3 Baseline systems

The results obtained by MHEL-LLaMo are measured using the F_1 metric, following (Ehrmann et al., 2022), and compared with those of SoTA models on the same datasets reported in (Linhares Pontes et al., 2022), (Ehrmann et al., 2022) and (Graciotti et al., 2025b). For MELHISSA (Linhares Pontes et al., 2022) we consider the results obtained by the best model without post-processing filters for a fair comparison. These baselines are described in Section 2. Results reported for mGENRE (De Cao et al., 2022), GPT-4o mini and LLaMa3.3-70B on MHERCL are taken from (Graciotti et al., 2025b). All models are tested in the entity disambiguation task, where the correct span of a mention and its surrounding context is given to the model to perform EL or NIL prediction.

5 Results

This section presents the experimental results for MHEL-LLaMo across four MHEL datasets: HIPE-2020, NewsEye, AJMC, and MHERCL. Table 2 shows the F_1 scores achieved by our approach compared to existing SoTA methods. Best results are highlighted in bold, while second-best results are underlined.

HIPE-2020 Performance: Our approach demonstrates strong performance across all three languages (English, German, French). MHEL-LLaMo$_{van}$ (chain) achieves the best performance on English (0.723 F_1), while MHEL-LLaMo$_{\theta}$ (chain) performs best on German (0.620 F_1) and achieves the best overall results. The vanilla single configuration excels on French (0.692 F_1). Notably, all MHEL-LLaMo variants substantially outperform the baseline systems by a large margin across different languages.

NewsEye Analysis: The results on NewsEye present a more challenging scenario, where MHEL-LLaMo outperforms all models in French and German, while MELHISSA maintains superior performance on Swedish and Finnish, probably due to better performance with NIL entities. To better understand this phenomenon, we conducted a thorough error analysis reported in Section 7. Nonetheless, MHEL-LLaMo$_{\theta}$ (chain) achieves the best performance on French (0.662 F_1) and German (0.556 F_1) and outperforms by a large margin the supervised system SBB.

AJMC Evaluation: On the AJMC dataset, MHEL-LLaMo demonstrates clear advantages, with the vanilla single configuration achieving the best results across all three languages: English (0.496), German (0.521), and French (0.635). This represents substantial improvements over the SBB and BELA baselines.

MHERCL Results: The MHERCL dataset showcases the strongest performance gains for our approach. MHEL-LLaMo$_{van}$ variants achieve

Dataset	Documents	Mentions	Languages	Period	Genre	% NIL
HIPE-2020	138	3,196	de, en, fr	19C-20C	newspapers	25.97
NewsEye	93	6,226	de, fi, fr, sv	19C-20C	newspapers	43.4
AJMC	44	1,090	de, en, fr	19C	classical commentaries	2.41
MHERCL	1,408 (sent.)	4,801	en, it	19C-20C	music periodicals	29

Table 1: Statistics of evaluation datasets including number of documents, mentions, languages, time period, genre, and percentage of NIL mentions.

the highest F_1 scores on both English (0.700) and Italian (0.698), significantly outperforming recent strong baselines including GPT-4o mini and LLAMA 3.3 70B, with an average improvement of 27% over these larger models used in zero-shot setting.

Configuration Analysis: Comparing the different MHEL-LLaMo variants reveals recurring patterns. MHEL-LLaMo$_\theta$ generally outperforms MHEL-LLaMo$_{van}$, particularly for German, French and Swedish. However, the adaptive threshold mechanism (θ) demonstrates its effectiveness primarily on texts belonging to the genre of press articles in HIPE-2020 and NewsEye, probably due to the higher percentage of high-frequency entities contained in these texts (popular places and famous personalities). The choice between chain and single prompt strategies appears dataset-dependent, with chaining showing advantages on HIPE-2020, NewsEye, and MHERCL, while the single prompt excels on AJMC where there is a lower ratio of NIL entities.

6 Discussion

The experimental results demonstrate that MHEL-LLaMo represents an important advancement in MHEL, achieving substantial improvements over SoTA approaches across multiple datasets. This section analyzes the key findings and their implications.

Performance Advantages: Our results reveal that MHEL-LLaMo variants consistently outperform existing SoTA models by substantial margins across three of four evaluated datasets. On HIPE-2020, our best configurations achieve improvements of up to 21% in F_1 score for English, 8% for German, and 10% for French over MELHISSA. Similarly, we observe substantial performance gains on AJMC, where MHEL-LLaMo achieves relative improvements of 30% on English, 4% on German and 35% on French over the SBB baseline. Most notably, on MHERCL, our approach demonstrates a remarkable average improvement of 27% over larger models (GPT-4o mini and LLAMA 3.3 70B), achieving F_1 scores of 0.700 for English and 0.698 for Italian. These results underscore the effectiveness of combining BELA's efficient candidate generation with the LLM's contextual understanding for refined NIL prediction and candidate selection. The approach proves particularly valuable for low-resource historical languages such as historical Italian, suggesting that the bi-encoder component effectively compensates for the LLM's limited exposure to historical language variants during pre-training.

Effectiveness of Ensemble Strategy: The empirical results validate the hypothesis proposed by (Ma et al., 2023), regarding ensemble strategies for IE tasks, though with important domain-specific limitations. The adaptive threshold mechanism (MHEL-LLaMo$_\theta$) demonstrates clear advantages on newspaper datasets (HIPE-2020 and NewsEye), where it achieves optimal performance for German and French on HIPE-2020, and for French and German on NewsEye. This effectiveness likely stems from the higher proportion of popular entities (prominent locations and well-known personalities) in newspaper texts, which BELA can confidently identify without LLM intervention. Conversely, the vanilla approach proves more effective on classical commentaries (AJMC) and music periodicals (MHERCL), where MHEL-LLaMo$_{van}$ achieves the best results. This divergence suggests that specialized domains with fewer popular entities benefit from consistent LLM processing rather than selective intervention, as the threshold-based filtering may incorrectly classify challenging domain-specific entities as easy samples.

Impact of Prompt Chaining Strategy: An important finding emerges regarding the relationship between prompt strategy and NIL entity prevalence. The chain configuration consistently outperforms

Dataset	Approach	F_1					
		de	**en**	**fi**	**fr**	**it**	**sv**
HIPE-2020	MELHISSA	0.573	0.597	-	0.630	-	-
	SBB	0.506	0.393	-	0.596	-	-
	L3i	0.481	0.546	-	0.602	-	-
	BELA	0.55	0.418	-	0.58	-	-
	MHEL-LLaMo$_{van}$ (single)	0.615	0.672	-	**0.692**	-	-
	MHEL-LLaMo$_{van}$ (chain)	0.587	**0.723**	-	0.679	-	-
	MHEL-LLaMo$_{\theta}$ (single)	0.618	0.615	-	0.672	-	-
	MHEL-LLaMo$_{\theta}$ (chain)	**0.62**	0.686	-	0.687	-	-
NewsEye	MELHISSA	0.547	-	**0.652**	0.542	-	**0.599**
	SBB	0.431	-	-	0.444	-	-
	BELA	0.3	-	0.3	0.422	-	0.41
	MHEL-LLaMo$_{van}$ (single)	0.488	-	0.47	0.593	-	0.504
	MHEL-LLaMo$_{van}$ (chain)	**0.556**	-	0.509	0.647	-	0.518
	MHEL-LLaMo$_{\theta}$ (single)	0.444	-	0.453	0.584	-	0.504
	MHEL-LLaMo$_{\theta}$ (chain)	**0.556**	-	0.479	**0.662**	-	0.521
AJMC	SBB	0.503	0.381	-	0.47	-	-
	BELA	0.443	0.338	-	0.437	-	-
	MHEL-LLaMo$_{van}$ (single)	**0.521**	**0.496**	-	**0.635**	-	-
	MHEL-LLaMo$_{van}$ (chain)	0.485	0.47	-	0.57	-	-
	MHEL-LLaMo$_{\theta}$ (single)	**0.521**	0.476	-	0.6	-	-
	MHEL-LLaMo$_{\theta}$ (chain)	0.479	0.463	-	0.57	-	-
MHERCL	mGENRE	-	0.47	-	-	0.37	-
	GPT-4o mini	-	0.6	-	-	0.51	-
	LLAMA 3.3 70B	-	0.61	-	-	0.48	-
	BELA	-	0.47	-	-	0.49	-
	MHEL-LLaMo$_{van}$ (single)	-	**0.7**	-	-	0.685	-
	MHEL-LLaMo$_{van}$ (chain)	-	**0.7**	-	-	**0.698**	-
	MHEL-LLaMo$_{\theta}$ (single)	-	0.65	-	-	0.627	-
	MHEL-LLaMo$_{\theta}$ (chain)	-	0.67	-	-	0.629	-

Table 2: F_1 performance comparison across four multilingual entity linking datasets. Best results are shown in **bold**, second-best results are underlined. Our MHEL-LLaMo variants consistently outperform baseline approaches, with particularly strong improvements on HIPE-2020, AJMC and MHERCL datasets.

Relation	HIPE-2020	NewsEye	AJMC	MHERCL	Tot.
False	15	8	16	20	59
Exact	4	8	4	2	18
Close	2	2	1	0	5
Related	2	6	3	2	13
Broader	0	1	0	0	1
Narrower	2	0	1	1	4

Table 3: Distribution of semantic relations between predicted and ground truth entities in a sample of 100 false predictions from MHEL-LLaMo.

the single prompt approach on datasets with high NIL percentages: HIPE-2020 (25.97% NIL), NewsEye (43.4% NIL), and MHERCL (29% NIL). This pattern strongly suggests that LLMs require explicit, separate prompting to accurately identify NIL entities. When using a single prompt that combines candidate selection and NIL prediction, the model exhibits a bias toward selecting existing candidates rather than predicting NIL, leading to degraded performance on datasets with substantial NIL proportions.

7 Error Analysis

To gain a better understanding of the accuracy of the proposed approach, we conducted an error analysis by sampling 100 false predictions of the best MHEL-LLaMo variants proportionally across all four datasets. The goal of this analysis was to understand if any pattern of semantic relation occurs between the ground truth annotations and the predicted entities. To estimate this, we classified each error in the sample according to 6 classes: false, exact, close, related, broader and narrower. Statistics related to these error classes for the four datasets are reported in Table 3. A detailed explanation of these error classes, with examples, is reported in Appendix C.

Our error analysis revealed that a consistent number of predictions (41/100) were either exactly referring to the entity mentioned or had a degree of semantic relation with the ground truth. Specifically, we found that 18 predictions were indeed correct (exact). This is primarily because many mentions, especially in NewsEye, are wrongly annotated as NIL and the entity they refer to was correctly retrieved by the model. This phenomenon can be attributed to KB updates, an issue which has already been investigated for EL datasets (Weichselbraun et al., 2019) and requires data curators to adopt versioning strategies to keep up with KB evolution. Another prominent category of errors consists of predictions that are somehow related to the ground truth. For instance, a reference to the Greek hero Odysseus can be linked by the system to the epic poem Odyssey. Moreover, close semantic relations can also appear when an entity may have multiple aliases in Wikidata and the system predicts a different identifier than the one in the ground truth. This is for example the case of the Greek god "Apollo", which in Wikidata has two distinct entries, "Apollo" (Q37340) and "Phoebus" (Q3382132). For more details on the cases found in our error analysis we refer to Appendix C. Finally, our error analysis revealed that a prominent source of errors is false NIL predictions, along with entities wrongly linked due to OCR errors in the surface form.

8 Conclusion

In this paper, we present MHEL-LLaMo, a novel unsupervised approach for MHEL that is inspired by ensemble techniques to combine a multilingual bi-encoder for candidate retrieval with instruction-tuned LLMs for candidate refining. The results obtained by MHEL-LLaMo on six European languages demonstrate not only better accuracy than SoTA specialized models but also robustness toward NIL entities (frequent in historical documents). The improvements achieved on historical benchmarks without the need for fine-tuning prove that this approach can offer a novel scalable solution for MHEL in low-resource scenarios, contributing to the democratization of NLP applications for various case studies in the humanities. Future work will investigate ways to take semantic proximity and perspectivism into account when evaluating MHEL systems (Aroyo and Welty, 2015; Valette, 2024) and ways to extend MHEL-LLaMo to other historical domains and time periods, focusing on the challenges of linking complex entities such as literary works, artworks and other cultural heritage objects.

9 Limitations

Our experiments revealed two important limitations. First, MHEL-LLaMo exhibits reduced performance on Nordic languages, particularly Finnish (best F_1: 0.509) and Swedish (best F_1: 0.521). This performance gap likely reflects the inherent linguistic complexities of these languages combined with potentially limited multilingual representation in both the bi-encoder and LLM components. The morphological richness and relatively

smaller training data availability for Nordic languages present challenges that our current architecture does not fully address.

Second, performance on classical commentaries (AJMC) remains lower compared to other genres, with F_1 scores ranging from 0.463 to 0.635. Detailed error analysis reveals that the majority of errors stem from linking mentions to literary works and cultural objects. This finding highlights a critical gap in current EL approaches, which are primarily optimized for person, location, and organization entities, and suggests the need for specialized techniques to handle cultural and literary entities that exhibit different referential patterns and contextual cues.

Furthermore, the BELA embeddings used are trained on a multilingual Wikipedia dump of 2023 (Plekhanov et al., 2023). While this issue may have an impact on recent texts, it does not significantly affect MHEL. Additionally, our error analysis revealed that gold-standard annotations in EL datasets, especially for NIL entities, may lose validity as Wikidata evolves. Although this issue does not originate from our approach, it may affect the validity of benchmarks for historical IE. Nevertheless, this issue can be tackled by adopting versioning techniques in data curation to explicitly state the specific KB release used for data annotation.

Finally, this paper has not explored the use of parameter efficient fine-tuning techniques, such as LoRA (Hu et al., 2022), to train smaller LLMs for NIL prediction and candidate selection. We consider this a promising research direction to reduce the computational cost of our approach and further increase its accessibility.

10 Ethical Considerations

While MHEL-LLaMo demonstrates strong performance, several potential risks warrant consideration. First, the reliance on LLMs for hard samples incurs environmental costs through GPU inference, though our adaptive approach mitigates this by processing only difficult cases. Second, both BELA and the instruction-tuned LLMs inherit biases from their training data, which may result in unequal performance across languages and cultures. Our experiments confirm this, showing reduced accuracy for underrepresented languages like Finnish and Swedish. Third, linking entities to Wikidata may perpetuate existing biases or gaps in the KB, particularly for non-Western historical figures and cultural artifacts. Finally, while our system aims to democratize access to advanced IE tools for humanities scholars, it may inadvertently privilege entities and narratives that are well-represented in web-crawled corpora, potentially marginalizing lesser-known historical references from underrepresented communities.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, pages 1–17. CEUR-WS Working Notes.

Marta Boscariol, Luana Bulla, Lia Draetta, Beatrice Fiumanò, Emanuele Lenzi, and Leonardo Piano. 2024. Evaluation of LLMs on Long-tail Entity Linking in Historical Documents. In *Posters and Demos, Workshops, and Tutorials of EKAW 2024*, volume 3967. CEUR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiyi Chen, Harald Sack, and Mehwish Alam. 2022. Analyzing social media for measuring public attitudes toward controversies and their driving factors: a case study of migration. *Social network analysis and mining*, 12(1):135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020. Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *European Conference on Information Retrieval*, pages 524–532. Springer.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: named entity recognition and linking in multilingual historical documents. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 423–446. Springer.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. Can ChatGPT Identify Entities in Historical Documents? In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 184–189. ISSN: 2575-8152.

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. 2024. Leveraging Open Large Language Models for Historical Named Entity Recognition. In *Linking Theory and Practice of Digital Libraries*, pages 379–395, Cham. Springer Nature Switzerland.

Arianna Graciotti, Nicolas Lazzari, Valentina Presutti, and Rocco Tripodi. 2025a. Musical heritage historical entity linking. *Artificial Intelligence Review*, 58(5):140.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025b. KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339, Vienna, Austria. Association for Computational Linguistics.

Paul Groth and Stefan Dietze. 2024. Understanding the impact of entity linking on the topology of entity co-occurrence networks for social media analysis. In *Knowledge Engineering and Knowledge Management: 24th International Conference, EKAW 2024, Amsterdam, The Netherlands, November 26–28, 2024, Proceedings*, volume 15370, page 69. Springer Nature.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G Moreno, and Antoine Doucet. 2021. A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking.

In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Kai Labusch and Clemens Neudecker. 2020. Named entity disambiguation and linking historic newspaper ocr with bert. *CLEF (Working Notes)*, 2696.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G Moreno, Emanuela Boros, Ahmed Hamdi, Antoine Doucet, Nicolas Sidere, and Mickaël Coustaty. 2022. Melhissa: a multilingual entity linking architecture for historical press articles. *International journal on digital libraries*, 23(2):133–160.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking. *arXiv preprint arXiv:2306.08896*.

Matteo Romanello and Sven Najem-Meyer. 2024. A named entity-annotated corpus of 19th century classical commentaries. *Journal of Open Humanities Data*, 10(1).

Cristian Santini, Laura Melosi, and Emanuele Frontoni. 2024. Named Entity Recognition in Historical Italian: The Case of Giacomo Leopardi's Zibaldone. In *Posters and Demos, Workshops, and Tutorials of EKAW 2024*, volume 3967, Amsterdam, Netherlands. CEUR.

Cristian Santini, Mary Ann Tan, Oleksandra Bruns, Tabea Tietz, Etienne Posthumus, Harald Sack, and 1 others. 2022. Knowledge extraction for art history: the case of vasari's the lives of the artists (1568). In *CEUR Workshop Proceedings*, volume 3234. CEUR-WS.

Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel. 2024. Melart: A multimodal entity linking dataset for art. *Transactions on Graph Data and Knowledge*, 2(2):8–1.

Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 111–115.

Albert Weichselbraun, Adrian M.P. Brasoveanu, Philipp Kuntschik, and Lyndon J.B. Nixon. 2019. Improving named entity linking corpora quality. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1328–1337, Varna, Bulgaria. INCOMA Ltd.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A NIL Prediction and Candidate Selection Prompts

This appendix presents the two prompts used in MHEL-LLaMo to perform entity disambiguation. The two prompts are used in a prompt chaining approach: first, the NIL prediction prompt is used to verify if the LLM finds a plausible match in the list of candidates; then the candidate selection prompt asks the LLM to pick the likeliest entity. Both prompts includes a task instruction, the input text with the target mention marked by `[ENT]` tags, contextual information related to the text (date, language, genre) and a JSON block listing candidate entities with associated labels, descriptions, types and temporal information.

A.1 NIL Prediction Prompt

System Prompt: You are a highly precise multilingual information extraction system specialized in disambiguating entities within noisy historical texts. Your task is to analyse the text provided by the user and determine if the reference marked by [ENT] tags can be associated or not to one of the candidate Wikidata entities provided in the JSON list. Always respond by saying either "yes" or "no". Do not generate Python code.

User Prompt: Read the input text written in `{language}`, published in `{document_date}` and belonging to the genre of `{genre}`.

Answer if the entity mentioned between the [ENT] tags in the input text corresponds to one of the candidate Wikidata entity provided in the json. Give a simple binary answer.

Input Text: `{annotated_text}`

Candidates: `{candidates_in_json}`

A.2 Candidate Selection Prompt

System Prompt: You are an effective multilingual information extraction system specialized in disambiguating entities within noisy historical texts. Your task is to analyse the text provided by the user and disambiguate the reference marked by [ENT] tags by selecting a Wikidata entity from a given list of candidates. Always respond by returning a JSON-formatted answer; do not generate Python code.

User Prompt: Read the input text written in `{language}`, published in `{document_date}` and belonging to the genre of `{genre}`.

Disambiguate the entity mentioned between the [ENT] tags by selecting the most appropriate Wikidata entity from the list of candidates.

Return the corresponding Wikipedia title and Wikidata ID of the selected entity in a JSON object formatted as follows: `{ "wikipedia_title": "", "wikidata_id": "" }`

Make sure to select both the Wikipedia title and the Wikidata ID from the provided list of candidates. Pay attention that the list of candidates may not include the entity mentioned. If none of the candidates match with high confidence the entity tagged with [ENT], return an empty json.

Input Text: `{annotated_text}`

Candidates: `{candidates_in_json}`

B Architecture Hyper-Parameters

This section reports the hyper-parameters used in the experiments. Threshold to filter easy samples was selected by computing the median of the scores of BELA for correct predictions on the development set. Instead, block size was estimated by measuring $Recall@K$ at 5 steps with $K = \{10, 20, 30, 40, 50\}$ and by finding the best step when there was no significant increment in $Recall@K$ on the development set, i.e. when $Recall@K_{i+10} - Recall@K_i < 0.01$.

Dataset	Lang	Block Size	Threshold
HIPE-2020	de	30	21.4
	en	20	21.24
	fr	20	21.24
NewsEye	de	30	21.5
	fi	20	19.57
	fr	20	21.35
	sv	20	20.43
AJMC	en	50	20.7
	de	50	21.5
	fr	20	17.92
MHERCL	en	20	21.5
	it	20	21.5

Table 4: Architecture hyper-parameters for MHEL-LLaMo experiments across datasets and languages.

C Semantic Relations in Error Analysis

This appendix provides detailed definitions and examples of the semantic relation categories used in our error analysis (Section 6). To better understand the nature of false predictions made by the MHEL-LLaMo model, we manually analyzed 100 incorrectly predicted entity links sampled proportionally across the four datasets. Each error was classified according to the semantic relationship between the predicted entity and the ground truth annotation, allowing us to distinguish between completely incorrect predictions and those that exhibit varying degrees of semantic proximity to the correct entity.

The classification scheme consists of six categories that capture different types of relationships between predictions and ground truth entities. Concise definitions of each category are reported below, which are further illustrated through concrete examples in Table 5.

False The predicted entity has no semantic relationship to the ground truth entity; the prediction is completely incorrect.

Exact The predicted entity refers to the same real-world entity as the ground truth, but uses a different Wikidata identifier.

Close The predicted entity is very similar to the ground truth entity, typically representing the same concept with minor differences (e.g., different time periods or alternative names).

Related The predicted entity has a clear semantic connection to the ground truth entity, such as

through thematic, conceptual, or associative relationships.

Broader The predicted entity represents a more general or superordinate concept that encompasses the ground truth entity.

Narrower The predicted entity represents a more specific or subordinate concept that is contained within or is a particular instance of the ground truth entity.

Semantic Relation	**Example**		
	Surface	Ground Truth	Prediction
False	Électre	Electra (Q733444)	Électre (Q3587592)
Exact	Sophocles	Sophocles (Q11950683)	Sophocles (Q7235)
Close	Epoque	L'Époque (1865-1869) (Q63171281)	L'Époque (1937–1949) (Q112206500)
Related	OEdipe	Oedipus at Colonus (Q294001)	Oedipus (Q130890)
Broader	Salford	Salford (Q47952)	Salford (Q1142118)
Narrower	11.	Iliad (Q8275)	Ilias, 11 Book (Q1658482)

Table 5: Examples of semantic relations between ground truth and predicted entities.