KE-MHISTO: Towards a Multilingual Historical Knowledge Extraction Benchmark for Addressing the Long-Tail Problem

Arianna Graciotti[*,1], **Leonardo Piano**[*,2], **Nicolas Lazzari**[*,3], **Enrico Daga**[4], **Rocco Tripodi**[5], **Valentina Presutti**[1], **Livio Pompianu**[2]

[1]University of Bologna, [2]University of Cagliari, [3]University of Pisa,
[4]Open University, [5]Ca' Foscari University of Venice

[*]Equal contribution. **Correspondence:** arianna.graciotti@unibo.it

Abstract

Large Language Models (LLMs) face significant challenges when queried about long-tail knowledge, i.e., information that is rarely encountered during their training process. These difficulties arise due to the inherent sparsity of such data. Furthermore, LLMs often lack the ability to verify or ground their responses in authoritative sources, which can lead to plausible yet inaccurate outputs when addressing infrequent subject matter. Our work aims to investigate these phenomena by introducing KE-MHISTO, a multilingual benchmark for Entity Linking and Question Answering in the domain of historical music knowledge, available in both Italian and English. We demonstrate that KE-MHISTO provides significantly broader coverage of long-tail knowledge compared to existing alternatives. Moreover, it poses substantial challenges for state-of-the-art models. Our experiments reveal that smaller, multilingual models can achieve performance comparable to significantly larger counterparts, highlighting the potential of efficient, language-aware approaches for long-tail knowledge extraction. KE-MHISTO is available at: https://github.com/polifonia-project/KE-MHISTO.

1 Introduction

Modern Knowledge Extraction (KE) methods rely on the *pre-train-then-finetune* paradigm, leveraging large language models (LLMs) as foundational components. These can be viewed as *soft knowledge bases* (Youssef et al., 2023), making them a cornerstone for KE tasks, including Named Entity Recognition (NER), Entity Linking (EL), Relation Extraction (RE), and Question Answering (QA) (Wu et al., 2020; De Cao et al., 2022; De Cao et al., 2021; Mallen et al., 2023; Sun et al., 2024).

However, LLMs' knowledge retention is closely linked to the frequency of information in the pre-training data. LLMs struggle when probed for so-called *long-tail knowledge* (Hogan et al., 2025; Li et al., 2024; Jiang and Joshi, 2024; Sun et al., 2024). The long-tail problem involves difficulties in handling queries about lesser-known topics due to limited supporting evidence in pre-training data. According to Hogan et al. (2025), these long-tail queries seek domain-specific factual information, making them particularly challenging for LLMs. While larger models improve retention and accuracy in performing the task (e.g. in question-answering), they provide only modest benefits for infrequent knowledge (Kandpal et al., 2023; Mallen et al., 2023).

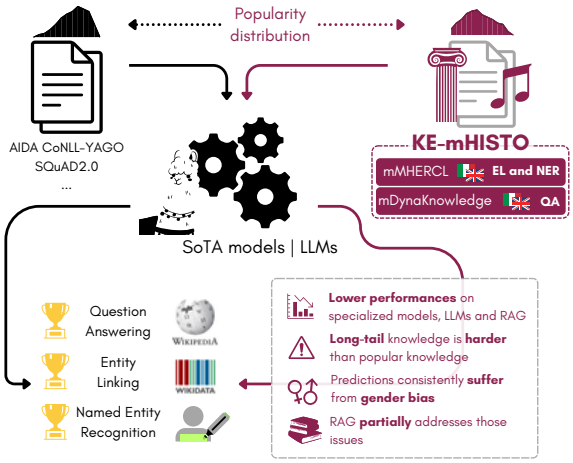


Figure 1: SoTA models are evaluated on benchmarks that focus on popular entities. KE-MHISTO evaluates LLM performance on long-tail knowledge.

Benchmarks focused on historical documents provide an opportunity for evaluating SotA KE methods on real-world data inherently characterised by long-tail knowledge (Arora et al., 2024a). Historical documents exemplify the challenges of long-tail KE, as information about historical entities and events is sparse in the pre-training datasets used to develop LLMs, which are typically based on large-scale general-purpose knowl-

[0*]Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 20316–20339

edge bases (KB), such as Wikipedia. Consequently, models relying on LLMs often fail to perform adequately when performing tasks like NER or EL (Ehrmann et al., 2020c,b, 2023) on such input data, or answering questions about historical entities (Graciotti et al., 2024). These challenges are even more pronounced in non-English languages (Holtermann et al., 2024), which are typically underrepresented in both pre-training corpora and downstream NLP resources. However, the insufficient availability of multilingual benchmarks specifically tailored to the long-tail problem hinders progress. In this paper, we present **KE-MHISTO**: a multilingual Historical KE benchmark in English and Italian comprising high-quality datasets for evaluating EL and QA in the long tail. KE-MHISTO includes: 1. **MMHERCL**, the **m**ultilingual **M**usical **H**eritage named entities **R**ecognition, **C**lassification and **L**inking benchmark, which extends the existing English-only MHERCL-EN dataset (Graciotti et al., 2025) by introducing **MHERCL-ITA**, a novel Italian-language expansion. To the best of our knowledge, MHERCL-ITA is the first gold standard for historical NER and EL in Italian. 2. **MDYNAKNOWLEDGE**, which expands the free-form historical QA benchmark DynaKnowledge (Graciotti et al., 2024) by releasing **DynaKnowledge-EN** (DK-EN), which increases DynaKnowledge's number of English samples by an order of magnitude, and **DynaKnowledge-ITA** (DK-ITA), a novel Italian-language expansion. KE-MHISTO shares neither time period nor topical overlap with mainstream benchmarks. We provide an extensive evaluation of KE-MHISTO, testing both task-specialised SotA models and LLMs across all tasks. We find that KE-MHISTO presents significant challenges for all tested models, underscoring its effectiveness in evaluating LLM limitations on long-tail knowledge. As a manually annotated benchmark derived from non-web historical sources, it also exposes biases in widely used knowledge bases (KBs), such as Wikidata, offering an opportunity to identify and mitigate them. Our experiments further reveal that smaller models trained in a multilingual setting achieve performance comparable to that of significantly larger models, highlighting the potential of efficient, language-aware approaches for long-tail KE, in line with the findings of (Feith et al., 2024).

2 Methodology

The long tail problem presents issues that deserve a dedicated methodological effort towards a high-quality, fully supervised benchmark. Therefore, we identified four steps: 1. corpus identification, 2. annotators' selection and training, 3. benchmark construction, and 4. evaluation, that we illustrate.

Corpus identification To maximise long-tail knowledge, we look into resources that include historical entities and facts. However, to accurately measure popularity, the corpus must cover entities that exist on Wikipedia and Wikidata. Furthermore, the corpus should also include resources in languages other than English. We select the Polifonia Corpus (Polifonia Corpus), a diachronic, multilingual, and modular resource specialised in the musical heritage (MH) domain. The corpus spans six languages: English, Dutch, German, Spanish, Italian, and French. While resources differ across languages (authoritative 19C music magazines in Spanish are different from the same editorial endeavours in German), each module maintains equivalence. We select the *Periodicals Module*, consisting of digitised articles from 19th-century musical journals, reflecting contemporary discussions on musical practice and criticism. Due to its nature and OCR-derived text, this part of the corpus presents additional challenges related to transcription errors, not contemporary lexicon, and non-standard syntactic structures, making it an ideal benchmark for evaluating KE methods in long-tail, multilingual settings.

Annotators should be fluent in the languages and the domain to ensure they can select meaningful samples and questions. In our work, annotations are performed by eight undergraduate students of the Foreign Languages and Literature program.

Benchmark construction In this step, we aim to ensure that benchmark samples cover knowledge within a source known to have been seen by LLMs during pre-training, ensuring the feasibility of the task in a zero-shot setting. This step is articulated in four phases. 1. Each annotator is presented with a set of periodicals from the corpus. 2. Next, they are asked to identify entity mentions in text and link them to corresponding Wikidata entries when available or to mark them as `NIL`[1]. 3. Annotators

[1]As detailed annotations guidelines for this step, we refer to those included in Graciotti et al. (2025).

retrieve corresponding Wikipedia pages and extract a passage relevant to the mentioned entity. 4. A question is then formulated based on this passage. As a result, we obtained a collection of sentences associated with entities for EL, as well as a question and its corresponding answer for QA, all with integrated knowledge. Quality assurance was performed by Inter Annotator Agreement (IAA). The approach is designed for extensibility across resource types and languages. Here, we apply it to historical periodicals in English and Italian. The output is presented in Section 3.

Evaluation To demonstrate how KE-MHISTO is an essential tool for advancing research aiming at addressing the long-tail problem, we perform an analysis comparing entity popularity in competing benchmarks. Following Arora et al. (2024a) and Mallen et al. (2023), we treat page views as a proxy for entity prominence in online discourse and compute popularity using QRank[2], which ranks entities based on view statistics across Wikimedia projects, such as Wikipedia, Wikitravel, and Wikibooks. We use QRank as our primary popularity proxy while experimenting with two alternative measures: entity degree in the Wikidata KG (calculated by counting all connections each entity has within the KG, treating it as undirected) following Arora et al. (2021), and Wikipedia article quality scores (computed by extracting structural features from article wikitext and applying the language-agnostic weighting model by (Johnson, 2021)) following Arora et al. (2024b). Detailed results using these alternative popularity proxies are presented in Appendix D.2. In addition, we perform extensive experiments in two modalities:1. a zero-shot prompt (for both the EL and QA task) and 2. a retrieval augmented one, where we incorporate the first paragraph of the subject entity in the context (only for the QA task). The evaluation is discussed in Section 4.

3 The KE-MHISTO Benchmark

KE-MHISTO is divided into two parts: one for the EL and the other for the QA task. Furthermore, each part has a separate module for each language. Figure 1 provides an overview of the resource. In what follows, we illustrate each component in detail.

[2] https://github.com/brawer/wikidata-qrank/tree/main

MMHERCL is an EL benchmark comprising sentences from historical periodicals in the music domain. The resource was preliminarily introduced in Graciotti et al. (2025) based on British periodicals dated from 1823 to 1900. In this work, we add MHERCL-ITA, a resource constructed from equivalent periodicals in Italian: *L'arpa: giornale letterario, artistico, teatrale* (Periodico 8670) and *Rivista nazionale di musica* (Periodico 8654). MHERCL-ITA includes 533 sentences extrapolated from 20 periodicals issues, with a total of 2,431 manually annotated entities. We stress that a significant proportion of entities in both datasets (30% in MHERCL-EN and 28% in MHERCL-ITA) lack a corresponding Wikidata entry (NIL entities), highlighting gaps in existing KBs and demonstrating the nature of MMHERCL in covering long-tail knowledge. We report full statistics comparing MHERCL-EN and MHERCL-ITA in Table 6a in Appendix B.

The IAA for MHERCL-EN is reported in Graciotti et al. (2025), with a score of 0.82 Krippendorff's alpha for nominal-scale data (Hayes and Krippendorff, 2007). To assess annotation quality for MHERCL-ITA, we sample 200 sentences (12,775 tokens, 1,176 NE mentions) and conduct an annotation by two independent annotators of NER and EL information. The computed IAA, using the same score, is 0.82, consistent with MHERCL-EN, indicating high annotation reliability.

MDYNAKNOWLEDGE (QA) MDYNAKNOWLEDGE is an open-form QA benchmark derived from MMHERCL. Each question can be answered by taking one or more sentences from the entity's Wikipedia page. The first steps towards MDYNAKNOWLEDGE were introduced by Graciotti et al. (2024), where an initial concept of DynaKnowledge with 82 samples was presented. DK-EN significantly expands upon this by scaling the English dataset to 567 QA pairs and by introducing the novel Italian dataset DK-ITA. Notably, the average provenance length, representing the supporting passage length in Wikipedia, is longer in DK-ITA, indicating structural differences between Italian and English. Additionally, the average answer length is significantly higher in DK-ITA, which may be attributed to linguistic differences in expressing factual information across languages. Table 6b in Appendix B reports the dataset statistics. Appendix C provides an example from each dataset

in mDynaKnowledge, which includes DK-EN (English) and DK-ITA (Italian).

4 Evaluation

In this section, we evaluate KE-MHISTO in two ways. First, we perform a long-tail analysis focusing on the popularity distribution of entities, comparing with alternative solutions (Section 4.1). Next, we report on extensive experiments in zero-shot and retrieval augmented settings, with State-of-the-Art (SotA) pre-trained LLMs (Section 4.2). We discuss our findings in Section 4.3.

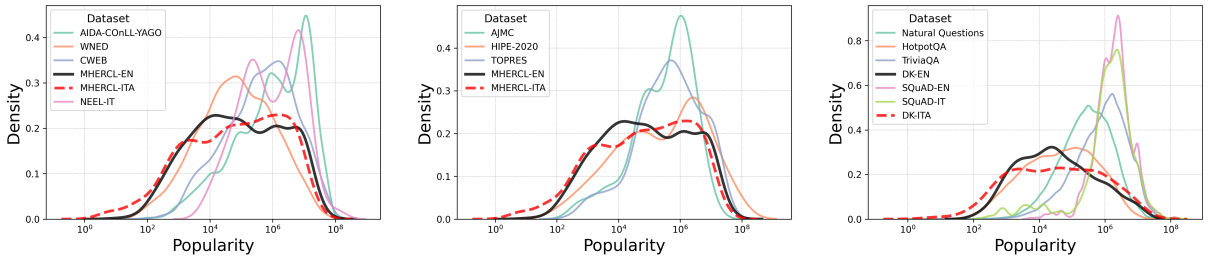
4.1 Long-tail Analysis

We analyse the popularity distribution of KE-MHISTO by comparing mMHERCL with existing EL benchmarks and mDynaKnowledge with existing QA benchmarks. In the context of the QA task, popularity refers to the QRank score of the Wikipedia page from which the entity or QA pair is derived. For EL and QA, we use the datasets included in (Petroni et al., 2021, KILT). Since KILT lacks Italian resources, we add (Basile et al., 2015, NEEL-IT-TWITTER) for EL and (Croce et al., 2018, SQuAD-it) for QA. Additionally, we incorporate (Rajpurkar et al., 2018, SQuAD2.0) for English QA. To extend the comparison to historical EL datasets, we use those from (Ehrmann et al., 2020c, HIPE-2020).

Figure 2a compares the popularity distribution of mMHERCL with the other EL datasets considered in this study. Entities in MHERCL-EN and MHERCL-ITA exhibit a higher concentration of long-tail entities compared to (Hoffart et al., 2011, AIDA CoNLL-YAGO), which primarily focuses on highly popular entities, and (Guo and Barbosa, 2018, WNED-WIKI) and (Gabrilovich et al., 2013, WNED-CWEB), which cover entities of intermediate popularity. This distinction highlights the focus of mMHERCL on underrepresented entities, differentiating it from widely-used EL benchmarks. Figure 2b further contrasts mMHERCL with historical EL datasets, which are often considered long-tail resources. While (Ehrmann et al., 2020c, HIPE-2020), (Romanello and Najem-Meyer, 2024, AjMC), and (Coll Ardanuy et al., 2022, TopRes19th) also target less popular entities, MHERCL-EN and MHERCL-ITA demonstrate enhanced coverage of them.

Figure 2c presents the popularity distribution of entities in DK-EN and DK-ITA in comparison with QA datasets from KILT, including (Kwiatkowski et al., 2019, Natural Questions) and (Yang et al., 2018, TriviaQA). Additionally, we consider SQuAD2.0 and (Croce et al., 2018, SQuAD-it), which are not included in KILT but provide further insights into QA dataset popularity distributions. DK-EN and DK-ITA exhibit a stronger focus on less popular entities, particularly in contrast to Natural Questions, SQuAD2.0, and SQuAD-it, which are dominated by highly popular entities. HotpotQA distribution resembles DK-EN, but the latter remains more skewed toward long-tail entities, with a mean shifted toward lower popularity.

Despite some datasets, such as WNED-WIKI and HotpotQA, appearing to target less popular entities, SotA models achieve high performance on the respective tasks. WNED-WIKI is based on web documents that are closer to PLM pre-training data, leading to a SotA F1 score of 0.90 (Ayoola et al., 2022a). Similarly, HotpotQA features a SotA Exact Match (EM) score of 0.72 (Zhang et al., 2024), despite having a popularity distribution similar to that of DK-EN. We report a full comparison in Appendix D. Finally, the domain specificity of the Polifonia Corpus has an influence on the distributions, making mMHERCL and mDynaKnowledge ideal resources to investigate novel solutions to the long-tail problem. Statistical analysis using Mann-Whitney U tests (detailed results in Appendix D.3) confirms these distributional characteristics across popularity indices. WNED, while a mainstream benchmark, is statistically similar to MHERCL-EN in terms of QRank but diverges from HIPE, another historical benchmark, suggesting that long-tail similarity alone does not capture domain-specific complexity. MHERCL-EN demonstrates greater similarity to HIPE than WNED across all popularity indices, indicating that MHERCL-EN better preserves the linguistic and structural challenges of historical data while covering a broader popularity spectrum. MHERCL-ITA differs significantly from NEEL-IT across all metrics. While Figure 2 may suggest similar popularity distributions, this apparent similarity reflects the tendency of English-centric datasets to include entities with higher overall popularity. When distributions are analyzed separately by language, the distinction becomes clear, with MHERCL-ITA being markedly different from NEEL-IT in terms of entity popularity patterns.



(a) Comparison of MHERCL-EN/ITA, KILT's EL datasets (AIDA-COnLL-YAGO, WNED, CWEB) and NEEL-it.

(b) Comparison of MHERCL-EN/ITA, and other historical EL benchmarks (AJMC, HIPE-2020, TopRes19th).

(c) Comparison of DK-EN, DK-ITA, Natural Questions, HotpotQA, TrivialQA, SQuAD2.0 and SQuAD-IT.

Figure 2: Popularity (QRank) distribution (computed using Kernel Density Estimation) of named entities across selected benchmarks in English and Italian. Only entities with a valid QID are considered. NIL entities are excluded.

4.2 Experimental evaluation

We investigate the impact of KE-MHISTO popularity distribution on the performances of NER, EL, and QA models. For NER, we experiment with GLiNER (Zaratiana et al., 2024), NuExtract[3], and LLAMA 3.3 70B (Dubey et al., 2024) and GPT-4 o-mini[4] LLMs[5] and evaluate models using a strict span-based offset Micro-F1 (Lu et al., 2022) on MHERCL-EN and MHERCL-ITA. For EL, we experiment with mGENRE (Cao et al., 2022), BELA (Plekhanov et al., 2023), and GPT-4o mini and LLAMA 3.3 70B (Dubey et al., 2024) LLMs. We rely on two similar zero-shot prompts[6], where one explicitly instructs the model to answer NIL when a suitable entity does not exist. We evaluate models using the F1 score defined by Ehrmann et al. (2020a) on MHERCL-ITA and MHERCL-EN. For QA we experiment with GPT-4o mini and LLAMA3.3 70B in zero-shot[7] and Retrieval Augmented Generation (RAG) fashion. For the RAG setting, we include in the prompt the first section of the Wikipedia page for the named entities predicted by LLAMA 3.3 70B and linked by GPT-4o mini. The prompt is divided into two main strategies based on in-context learning (ICL) and chain-of-thought (CoT) prompting[8]. We evaluate the models on DK-EN and DK-ITA using *LLM-as-a-judge* (Zheng et al., 2023). We instruct LLAMA 3.3 70B[9] to judge whether a response is correct, incorrect, or partially correct and compute F1 score and accuracy by considering partially correct answers as false positives (*strict*) and true positives (*soft*). We considered the use of an LLM-as-a-judge to be a reliable choice, as its task was limited to comparing predicted and gold answers, thus not affected by the model's potential lack of knowledge. To ensure judgment reliability, we evaluated responses across a full run involving both DK-EN and DK-ITA, totaling 1,545 evaluated samples. This assessment showed the model achieved a precision of 95%.

Model	MHERCL-ITA	MHERCL-EN
GLiNER-multi	0.63	0.59
NuExtract 1.5	0.44	0.43
LLAMA 3.3 70B	**0.67**	**0.63**
GPT-4o mini	0.57	0.58

Table 1: NER results on MHERCL-EN and MHERCL-ITA.

Named Entity Recognition Table 1 shows the NER results computed on MMHERCL. LLAMA 3.3 70B outperforms other models on both MHERCL-ITA and MHERCL-EN. Interestingly, it achieves better results in the Italian dataset despite being mostly trained on English corpora. GLiNER achieves competitive performances compared to LLAMA 3.3 and outperforms GPT-4 despite orders of magnitude fewer parameters.

Entity Linking Table 2 presents the EL results obtained using SotA multilingual models and LLMs. LLMs outperform specialised models when instructed to predict NIL. This is particularly evident in the Italian dataset, where entities are generally less popular than in the English dataset. We felt the need to further analyse popularity-stratified performance across EL models on benchmarks with

[3] https://huggingface.co/numind/NuExtract-1.5

[4] https://openai.com/index/learning-to-reason-with-llms/

[5] The prompt used for LLMs is reported in Table 14 of the Appendix

[6] Tables 15 and 16 in Appendix

[7] Tables 19 and 20 Appendix

[8] ICL in Tables 21, 22, CoT in Tables 23 and 24

[9] Prompt of Table 25 in Appendix

Model	MHERCL-ITA	MHERCL-EN
mGENRE	0.37	0.47
BELA	0.49	0.47
GPT-4o mini (+ NIL)	0.33 (**0.51**)	0.44 (0.60)
LLAMA 3.3 70B (+ NIL)	0.38 (0.48)	0.51 (**0.61**)

Table 2: EL results on MHERCL-ITA and MHERCL-EN.

Dataset	Model	10%	20%	50%	80%
MHERCL-EN	mGENRE	0.00	0.00	0.00	0.24
	BELA	0.00	0.00	0.13	0.49
	GPT-4o	0.00	0.00	0.00	0.09
	LLAMA	0.00	0.00	0.00	0.22
HIPE2020	mGENRE	0.00	0.00	0.00	0.11
	BELA	0.00	0.00	0.00	0.11
	GPT	0.00	0.00	0.00	0.11
	LLAMA	0.00	0.00	0.00	0.06
WNED-WIKI	mGENRE	**0.70**	**0.70**	**0.81**	**0.81**
	BELA	0.00	0.00	0.33	0.60
	GPT	0.60	0.60	0.26	0.53
	LLAMA	**0.70**	**0.70**	0.41	0.57

Table 3: Stratified F1 performance of the models tested on the EL task comparing MHERCL-EN to HIPE2020 and WNED, datasets with similar QRank popularity distributions. Results are reported at different popularity percentiles.

similar popularity distributions to MHERCL-EN. The results, reported in Table 3, show that the models tested perform better on low-popularity entities in WNED than in MHERCL-EN. This highlights that EL is more challenging in MHERCL-EN than in other mainstream benchmarks with similar popularity distributions, such as WNED-WIKI. While LLMs such as LLaMA-3 and GPT-4-o-mini perform well on average, their performance degrades on the lowest popularity strata, where specialized models like BELA and mGENRE can outperform LLMs. This degradation is particularly pronounced in MHERCL-EN, where the challenging nature of historical domain entities amplifies the difficulty for general-purpose models. Stratified results highlight that HIPE-2020 is also very challenging, further validating the added value of using historical documents for studying real-world long-tail problems. An additional complexity factor in historical documents is OCR quality. For a systematic analysis of the impact of OCR errors in source materials in MHERCL-EN and HIPE-2020, we refer to Table 8 in Graciotti et al. (2025). OCR errors are more common in incorrectly linked mentions than in correctly linked ones across all tested models. We repeat the same tests on MHERCL-ITA, and we obtain comparable results, with mentions containing OCR errors occurring more frequently among the incorrectly linked ones (see Appendix F for detailed breakdown).

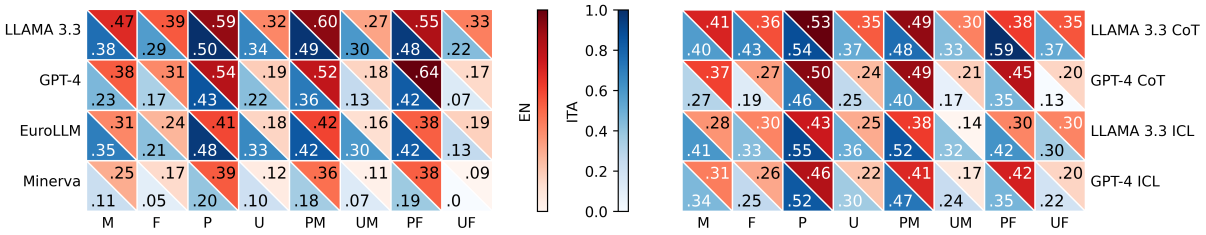
Question Answering Table 4 shows the QA results. LLAMA 3.3 outperforms all the other models in both Italian and English. Interestingly, RAG prompting only enhances GPT-4 performances, while LLAMA 3.3 works best in a zero-shot setting in all settings, having the lowest amount of empty answers and the highest amount of partially correct answers in both datasets. Multiple reasons explain this behaviour, such as the accuracy of the NER and EL phases, which might result in partial and possibly wrongly aligned contextual information. Despite its smaller size, EuroLLM (Martins et al., 2024) achieves competitive performances compared to LLAMA in Italian, surpassing GPT4 o-mini and always producing an answer for each question. Finally, our results suggest that a model specifically fine-tuned for the Italian language, Minerva-7b (Orlando et al., 2024), underperforms in the Italian language. This result supports the fact that training LLMs in multiple languages is beneficial for multi-lingual support.

4.3 Discussion

In this section we discuss our findings. Figure 3 shows the influence of gender, popularity, and their combination on EL and QA performances. Regardless of the task, popular entities lead to better performances than unpopular ones. Similarly, there is a consistent gender bias which is emphasized when only unpopular entities are considered. This can be seen mostly in EL (Figure 4) and zero-shot QA (Figure 3a). Interestingly, the general trend of male performance greater than female is reversed when considering popular female entities, suggesting a possible *Simpson's paradox* (Simpson, 1951). However, a closer analysis reveals that this effect is due to the underrepresentation of female entities: the few popular female entities present in the dataset are well-known figures, which reduces the likelihood of recognition and disambiguation errors. The same argument holds partially for RAG QA as well (Figure 3b). Differently from the others, however, relying on RAG allows a consistent enhancement over unpopular

Model	DK-ITA								DK-EN							
	Strict ↑		Soft ↑		✓ ↑	× ↓	≈↓	ε ↓	Strict ↑		Soft ↑		✓ ↑	× ↓	≈↓	ε ↓
	F1	Acc	F1	Acc	(%)	(%)	(%)	(%)	F1	Acc	F1	Acc	(%)	(%)	(%)	(%)
EuroLLM-9b	0.41	0.26	0.64	0.47	0.26	0.53	0.21	**0.0**	0.29	0.17	0.49	0.32	0.17	0.34	0.16	0.34
Minerva-7b	0.15	0.08	0.27	0.15	0.08	0.10	**0.07**	0.74	0.26	0.15	0.47	0.31	0.15	0.51	0.16	0.18
LLAMA 3.3 70B	**0.48**	**0.31**	**0.69**	**0.52**	**0.31**	0.42	0.21	0.06	**0.49**	**0.32**	**0.70**	**0.54**	**0.32**	0.32	0.22	**0.14**
w/ ICL RAG	0.46	0.30	0.64	0.47	0.30	0.22	0.17	0.30	0.37	0.22	0.55	0.38	0.22	0.15	0.16	0.47
w/ CoT RAG	0.46	0.30	0.66	0.49	0.30	0.25	0.19	0.27	0.44	0.29	0.68	0.51	0.29	0.21	0.23	0.28
GPT4 o-mini	0.33	0.20	0.50	0.33	0.20	0.12	0.13	0.55	0.34	0.21	0.53	0.36	0.21	0.28	0.15	0.36
w/ ICL RAG	0.42	0.27	0.58	0.41	0.27	0.13	0.14	0.46	0.36	0.21	0.49	0.32	0.22	**0.14**	**0.10**	0.54
w/ CoT RAG	0.37	0.23	0.51	0.34	0.23	**0.08**	0.11	0.58	0.38	0.23	0.55	0.38	0.23	0.17	0.14	0.46
In common					49	6	17	3					69	14	15	53

Table 4: F1 and accuracy scores on DK-ITAand DK-ENaveraged across three different runs. Percentage of correct (✓), incorrect (×), partially correct (≈), and not given answers (ε) is reported. The shaded row (*In common*) reports the number of equal predictions from all the models.



(a) Zero-shot QA

(b) RAG QA

Figure 3: QA performances on popular (P), unpopular (U), male (M), or female (F) entities. The top half refers to the English language, and the bottom half to the Italian language. Darker colours are used to indicate higher performances.

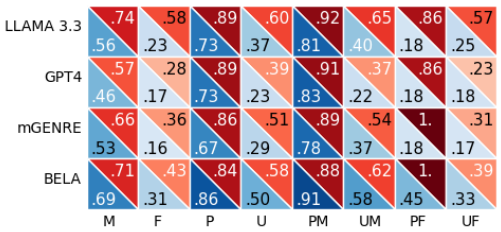


Figure 4: EL performances on same setting as Figure 3.

entities. Even though in Table 4 performances are degraded when using RAG, Figure 3b suggests that RAG enhances performances when it is possible to identify meaningful entities in the question and accurately align them to obtain the Wikipedia section, being especially beneficial when questions are about unpopular female person named entities.

The small number of common correct, partially correct, incorrect, and not given answers across all runs of all models, reported in the last row of Table 4, indicates that the models exhibit divergent behaviour, making different errors and correctly predicting different samples. By examining common errors—both common incorrect and common partially correct cases—we can identify dominant error patterns. For both DK-EN and DK-ITA, a significant portion of common incorrect cases stems from *temporal reasoning* errors. Temporal reasoning involves answering questions that require nuanced temporal understanding, such as in the question *In which conservatory was François-Joseph Fétis a teacher in 1821?*. This category accounts for 57% of the common incorrect cases in DK-EN and 50% in DK-ITA. Examples of this error pattern are reported in Appendix H.1. Similarly, a dominant pattern among common partially correct cases is the challenge of recalling complete *lists* as answers. In these cases, models retrieve partial lists instead of the full expected answer. For example, in *By whom was the Tragédie en musique genre invented?*, where the correct answer is *by Lully and his librettist Quinault*, all models cor-

	MHERCL-EN / MHERCL-ITA		
	Male (%)	**Female (%)**	**Tot.**
Person	85% / 75%	15% / 25%	100.0%
NIL	27% / 28%	55% / 50%	30.5% / 35%

Table 5: Proportion of persons in MHERCL-EN and MHERCL-ITA by genre and whether they are not in Wikidata (NIL).

rectly retrieve Lully, but none recall his librettist Quinault. This issue, also highlighted by Hogan et al. (2025), accounts for 67% of common partially correct cases in DK-EN and 35% in DK-ITA. Examples of this error pattern are reported in Appendix H.2. A detailed qualitative analysis of errors is challenging due to the free-form nature of our dataset and will be addressed in future work.

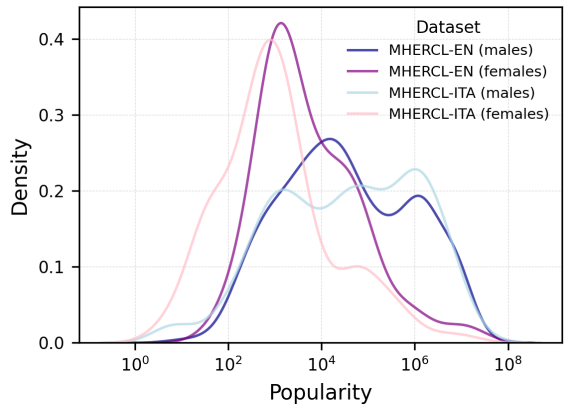


Figure 5: Popularity distribution (computed using KDE) of MMHERCL male and female named entities.

It is worth observing gender bias in KE-MHISTO.[10] The distribution of person NEs in MMHERCL and MHERCL-ITA reveals disparities in gender representation and KB coverage, as shown by Table 5. Male entities account for the majority of person NEs in both MHERCL-EN (85%) and MHERCL-ITA (75%). Female entities constitute only 15% of person NEs in MHERCL-EN and 25% in MHERCL-ITA.When examining the Wikidata coverage of male and female entities, the disparity becomes even more pronounced. Male entities are more frequently represented in it (73% for MHERCL-EN and 72% for MHERCL-ITA). In contrast, female entities are less likely to be present, being disproportionately classified as NIL (55% in MHERCL-EN and 50% in MHERCL-ITA), underscoring the gap in Wikidata coverage for representation of female historical figures. Figure 5 compares the popularity distribution of male and female NEs in MMHERCL and MHERCL-ITA. Male entities exhibit a broader distribution, including higher densities in the mid-to-high popularity ranges. Female entities are concentrated in the lower popularity ranges. Distribution is similar in both datasets. These findings reinforce the challenges posed by historical datasets like KE-MHISTO, where biases in gender representation and KB coverage are intertwined with the difficulty of long-tail EL.

[10] In our study, we limit our analysis to binary gender categories, which, while not capturing real-world diversity, provide a starting point for defining our method. We classify a person named entity as *female* if the pronoun *she* is used in the source document in which it occurs or on its Wikipedia page. Similarly, we classify an entity as *male* if referred to with the pronoun *he* in these contexts.

5 Related Work

The long-tail problem in KE refers to the difficulty of handling queries related to lesser-known topics with limited supporting evidence in pre-training corpora.

Long-tail queries, seeking domain-specific factual information, is challenging for LLMs (Hogan et al., 2025), since their knowledge retention is strongly correlated with the frequency of information in pre-training data, particularly on larger models (Kandpal et al., 2023). Mallen et al. (2023) introduce PopQA, a benchmark derived from Wikidata with controlled entity popularity. They find that larger models improve QA accuracy but provide limited gains for infrequent knowledge while RAG enhances long-tail performance, but introduces errors for popular entities. Similarly, Sun et al. (2024) introduces Head-to-Tail, a QA benchmark constructed from DBpedia, reporting a systematic decline in the performances for rare entities. Maekawa et al. (2024) develop WiTQA, a dataset incorporating entity and relation questions across different popularity levels using Wikidata and Wikipedia. They confirm that RAG does not consistently improve LLM accuracy and that an adaptive strategy —-selectively employing retrieval based on entity and relation frequency —- yields better results. Graciotti et al. (2024) introduces DynaKnowledge, a manually annotated QA benchmark constructed from multilingual historical documents, many digitized for the first time, composed of question-answer pairs about musicians from the 18th and 19th centuries.

Research on KE from historical texts has primarily focused on NER and EL. Ehrmann et al. (2023) review historical NER resources, highlighting the HIPE-2020 (Ehrmann et al., 2020c) and HIPE-2022 (Ehrmann et al., 2022a) evaluation campaigns, which include EL annotations. HIPE-2022 spans six datasets in English, Finnish, French, German, and Swedish, sourced from historical newspapers and classical commentaries (18C–20C), including NewsEye (Hamdi et al., 2021), SoNAR (Menzel et al., 2021), Le Temps (Ehrmann et al., 2016), Living with Machines, which annotates toponyms in 19th-century British newspapers in TopRes19th, and the annotated classical commentaries of AjMC[11]. Blouin et al. (2024) conduct NER and EL on historical Chinese newspapers (1872–1949) and bilingual Chinese-English biographies from the early 20th century. Arora et al. (2024a) address entity disambiguation challenges in historical documents, with a focus on individuals absent from modern KBs. Graciotti et al. (2025) focus on the music domain with MHERCL-EN, a manually annotated benchmark extracted from 19C English-language periodicals containing underrepresented or absent entities in major KBs.

Research on QA for historical documents is comparatively less developed than for NER and EL. ArchivalQA (Wang et al., 2022) introduces a large-scale QA dataset designed for temporal news archives that classify questions by difficulty and temporal expressions, enabling the evaluation of models on long-term news corpora. Similarly, ChroniclingAmericaQA (Piryani et al., 2024) builds QA pairs from a newspaper collection spanning 120 years, which also tests models on noisy OCR text, corrected transcriptions, and scanned images, addressing real-world challenges. Unlike web-based QA resources, both datasets focus on temporal knowledge retrieval and linguistic shifts, highlighting the limitations of existing LLMs in handling diachronic corpora. QUANDHO (Menini et al., 2016) provides a historical QA dataset in Italian covering Italy's early 20th-century history. It includes manually classified questions, question-answer pairs, and lexical answer type annotations, supporting domain adaptation for QA systems.

KE-MHISTO goes beyond existing work in three key aspects. Differently from previous datasets primarily derived from web-based or semi-automatically generated sources, KE-MHISTO introduces a novel manual annotation methodology designed to scale across multiple languages and historical domains. By systematically curating and digitizing multilingual historical documents our approach ensures high-quality entity annotations while accommodating their linguistic and structural variability. Secondly, it integrates historical NER and EL with historical QA, providing a unified benchmark for evaluating long-tail KE across multiple tasks rather than only focusing on NER and EL. Finally, KE-MHISTO explicitly connects entities to Wikipedia and Wikidata, differently from other historical QA datasets.

MMHERCL contains MHERCL-ITA, to the best of our knowledge the first gold-standard resource of historical texts annotated for NER and EL in Italian. MDYNAKNOWLEDGE, extends the benchmark introduced in Graciotti et al. (2024) by releasing DK-EN, which increases the English-language sample from 82 to 567 samples, improving coverage of long-tail entities and DK-ITA, the first Italian QA benchmark addressing long-tail factual knowledge about entities in diachronic corpora and fully linked to Wikipedia and Wikidata.

6 Conclusion

In this paper we presented a novel multilingual benchmark tailored to the long-tail problem. Going beyond existing alternative benchmarks, KE-MHISTO consists mostly of entities with low popularity, and it is fully linked to Wikipedia and Wikidata. Our findings indicate that SotA models struggle on KE-MHISTO, performing significantly worse than on alternative benchmarks. Furthermore, models exhibit systematic failures on specific segments of the dataset, revealing biased performance. In addition, smaller models trained in a multilingual setting achieve performance comparable to significantly larger models. This aspect highlights the potential of efficient, language-aware approaches for long-tail KE. In the future, we plan to keep nurturing this resource, adding equivalent datasets for the other languages covered by the Polifonia Corpus, such as Spanish, French, German, and Dutch. Our contribution is an added milestone towards supporting the systematic evaluation of EL and QA methods in a multilingual, long-tail setting.

[11] https://mromanello.github.io/ajax-multi-commentary/

7 Limitations

In what follows, we discuss the limitations of KE-MHISTO and of our experimental analysis.

OCR errors are preserved in MMHERCL since they require models to couple with issues generally found in long-tail documents, especially when knowledge originates from outside the web. The models we use in our experimental settings are not specifically trained to couple with these errors. While it is reasonable to assume that the large pretraining corpus used to train LLMs and their high number of parameters allows the model to understand the content of the text despite those errors, other models that rely on a smaller language model might suffer from it.

NILs in MMHERCL account for a consistent subset of the entities that must be linked by one of the methods of Table 2 (see Table 6a). While we account for this aspect in LLMs by explicitly instructing them to predict NILs, we always consider all the predictions of mGENRE and BELA, regardless of their confidence score. Hence, their performances are bounded by the number of NILs in the dataset. Nonetheless, the performances of Table 2 show that they still achieve competitive results, which could be enhanced by integrating NIL-aware techniques (Graciotti et al., 2025).

Multi-lingual prompts are only used in the QA task since the question-answer pairs are expressed in English in DK-EN and in Italian in DK-ITA. Even though the results in the NER and EL tasks obtain comparable results in both languages (Tables 1 and 2), language-specific prompts might enhance the performances of LLM-based approaches.

Entity names change over time such as in the case of locations. For example, the name of a theater can change throughout time. Hence, an historical document might refer to the same location as a contemporary one, but with a different name. This is not explicitly handled in KE-MHISTO and can lead to inconsistent results. In our setting, alternative names are not considered in the EL task and are opaquely handled by the LLM-as-a-judge in the QA task. Nonetheless, an evaluation setting that actively considers this aspect might provide results that better translate to real-world settings.

The Wikipedia abstract – i.e. the first paragraph in the Wikipedia article – used for the contextual information in the RAG prompts is not guaranteed to be informative for a specific question. While it is supposed to contain the most important information for an entity[12], questions on long-tail knowledge are designed to target specific content of an entity. A more sophisticated approach that retrieves relevant passages from the whole Wikipedia page can produce more accurate results and possibly overcome the issues described in Section 4.3.

The form of questions in DK-EN require answers that are not limited to closed-form answers, but might also require basic deductions, extensive background knowledge, or in general they must leverage a large amount of context to be answered. While this allows testing different aspects of models, we evaluate all the answers using the same experimental settings. A more fine-grained evaluation setting can help better understand which are the main limitations of those models in understanding long-tail knowledge.

Popularity of Wikipedia page is used as a proxy to classify long-tail entities and estimate the amount of content seen by an LLM on that entity, as done in Arora et al. (2024a) and Mallen et al. (2023). Nonetheless, some pages might have a relatively low amount of views on Wikipedia without being long-tail entities. For instance, the Wikipedia page of Google might be less popular than the Wikipedia page of Mozart, since the latter aggregates encyclopedic knowledge on an entity while other resources (such as the current official websites) might be more informative for the former and hence receive more visits.

Music periodicals of the 19th century are assumed to be a reasonable representative of long-tail knowledge. Nonetheless, they cover a niche domain and might hence be intrinsically biased. Resources that focus on other domains, or different periods, may frame long-tail knowledge differently.

8 Ethical considerations

The creation of KE-MHISTO was carried out by eight undergraduate students from the University of Bologna's Foreign Languages and Literature program as part of a curricular internship. These students were selected by assessing their fluency in the relevant languages and their familiarity with the domain, ensuring that they possessed the necessary

[12] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

linguistic and contextual expertise for the task. To maintain high-quality and consistent annotations, the students underwent task-specific training. This training was designed not only to maximize annotation consistency but also to introduce them to research methodologies in the construction of linguistic resources, particularly within the domain of natural language processing applied to the cultural heritage sector and digital humanities. Beyond technical training, the students participated in dedicated sessions held by the authors of the paper to reflect on the ethical implications of generative AI, particularly regarding its limitations in handling niche and less popular knowledge domains. Additionally, discussions focused on the gender biases present in large language models (LLMs), fostering awareness of the risks and responsibilities associated with AI-driven content generation. To provide flexibility, the students were allowed to work remotely. However, their work was closely supervised by the authors of this paper through bi-weekly meetings, ensuring continuous guidance and quality control. The authors also remained available via email to promptly address any concerns or questions.

9 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746. Nicolas Lazzari is supported by the FAIR – Future Artificial Intelligence Research Foundation as part of the grant agreement MUR n. 341, code PE00000013 CUP 53C22003630006. Leonardo Piano acknowledges financial support under the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP – funded by the European Union - NextGenerationEU - Mission 4 "Education and Research", Component 1 "Enhancement of the offer of educational services: from nurseries to universities" - Investment 4.1 -CUP F22B22000830007.

References

Anthropic. 2023. Model card and evaluations for claude models. Technical Report. Accessed: 2023-07-08.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. 2024a. Contrastive entity coreference and disambiguation for historical texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186, Miami, Florida, USA. Association for Computational Linguistics.

Akhil Arora, Alberto Garcia-Duran, and Robert West. 2021. Low-rank subspaces for unsupervised entity linking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8054, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akhil Arora, Robert West, and Martin Gerlach. 2024b. Orphan articles: The dark matter of wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):100–112.

Giuseppe Attardi, Daniele Sartiano, Maria Simi, and Irene Sucameli. 2016. Using embeddings for both entity recognition and linking in tweets. In *Proceedings of the Evalita 2016*, Pisa, Italy. Università di Pisa. Accessed: 2023-12-15.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. 2022a. Improving entity disambiguation by reasoning over a knowledge base. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912, Seattle, United States. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022b. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. neel-it-twitter-master 09122015.

Baptiste Blouin, Cécile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 385–394, Torino, Italia. ELRA and ICCL.

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. *Trans. Assoc. Comput. Linguistics*, 10:274–290.

Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel C. S. Wilson. 2022. A dataset for toponym resolution in nineteenth-century english newspapers. *Journal of Open Humanities Data*.

Danilo Croce, Giorgio Brandi, and Roberto Basili. 2019. Deep bidirectional transformers for italian question answering. In *Proceedings of a Conference (if applicable, otherwise remove this line)*, Via del Politecnico 1, 00133 Roma, Italy. University of Roma, Tor Vergata. Published under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations*, Online, Austria. OpenReview.net.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic Evaluation of NER Systems on Old Newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107, Bochum, Germany. Bochum, Germany, Bochumer Linguistische Arbeitsberichte.

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).

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020a. Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 524–532. Springer.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022a. Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Advances in Information Retrieval*, volume 13186, pages 347–354, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Alex Fluckiger, and Simon Clematide. 2020b. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, page 38, Thessaloniki, Greece.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020c. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 288–310, Berlin, Heidelberg. Springer-Verlag.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022b. Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF)*, volume 3180, pages 1038–1063, Aachen. CEUR-WS.

Tomás Feith, Akhil Arora, Martin Gerlach, Debjit Paul, and Robert West. 2024. Entity insertion in multilingual linked corpora: The case of wikipedia. *arXiv preprint arXiv:2410.04254*.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0). Web Download. Available at `http://lemurproject.org/clueweb09/` and `http://lemurproject.org/clueweb12/`.

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

Arianna Graciotti, Valentina Presutti, and Rocco Tripodi. 2024. Latent vs explicit knowledge representation: How ChatGPT answers questions about low-frequency entities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10172–10185, Torino, Italia. ELRA and ICCL.

Zhaochen Guo and Denilson Barbosa. 2018. Robust Named Entity Disambiguation with Random Walks. *Semantic Web Journal*, 9(4):459–479.

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*, SIGIR '21, page 2328–2334, New York, NY, USA. Association for Computing Machinery.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for

coding data. *Communication Methods and Measures*, 1(1):77–89.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Aidan Hogan, Xin Luna Dong, Denny Vrandečić, and Gerhard Weikum. 2025. Large language models, knowledge graphs and search engines: A crossroads for answering users' questions. *Preprint*, arXiv:2501.06699.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4476–4494.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Ming Jiang and Mansi Joshi. 2024. CPopQA: Ranking cultural concept popularity by LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 615–630, Mexico City, Mexico. Association for Computational Linguistics.

Isaac Johnson. 2021. Language-agnostic quality. Accessed: 2025-05-27.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Honolulu, Hawaii, USA. JMLR.org.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. 2024. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370, Miami, Florida, USA. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. 2024. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. 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 5506–5521, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Stefano Menini, Rachele Sprugnoli, and Antonio Uva. 2016. "who was pietro badoglio?" towards a QA system for Italian history. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 430–435, Portorož, Slovenia. European Language Resources Association (ELRA).

Sina Menzel, Hannes Schnaitter, Josefine Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner, and Georg Rehm. 2021. *Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten*, pages 229–258. De Gruyter Saur, Berlin, Boston.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, Roberto Navigli, et al. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. *Proc. of CLiC-it*.

Periodico 8654. Rivista Nazionale di Musica. Emeroteca Digitale Italiana. Periodico 8654. https://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8670. Accessed February 2025.

Periodico 8670. L'Arpa: Giornale Artistico, Teatrale. Emeroteca Digitale Italiana. Periodico 8670. https://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8670. Accessed February 2025.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2038–2048, New York, NY, USA. Association for Computing Machinery.

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. *CoRR*, abs/2306.08896.

Polifonia Corpus. https://github.com/polifonia-project/Polifonia-Corpus. Accessed February 2025. [link].

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Hassan Shavarani and Anoop Sarkar. 2023. SpEL: Structured prediction for entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11123–11137, Singapore. Association for Computational Linguistics.

Author E. H. Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the royal statistical society series b-methodological*, 13:238–241.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? 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 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3025–3035, New York, NY, USA. Association for Computing Machinery.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605, Singapore. Association for Computational Linguistics.

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.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024. End-to-end beam retrieval for multi-hop question answering. 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 1718–1731, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Computational resources

EuroLLM-9b and Minerva-7b have been executed on two RTX A6000 GPUs with 48GB of RAM. LLAMA3.3 70B and GPT4 o-mini have been used through the APIs provided by OpenRouter[13].

B KE-MHISTO extended statistics

	MHERCL-EN	MHERCL-ITA
Documents	76	20
Sentences	875	533
Tokens	27,549	24,804
Mentions (all / unique)	2,370 / 1,805	2,431 / 1,625
# Types	58	33
NIL (%)	30	28

(a) MMHERCL

	DK-EN	DK-ITA
QA pairs	567	978
Question length (avg.)	12.15	11.06
Answer length (avg.)	5.77	12.50
Provenance length (avg.)	33.43	50.02
Wikipedia pages	208	706
Question per Wikipedia page (avg.)	2.73	1.39

(b) MDYNAKNOWLEDGE

Table 6: Statistics for the dataset in MMHERCL (Table 6a) and MDYNAKNOWLEDGE (Table 6b)

TYPE	#
PERSON	1,253
CITY	262
MUSIC	187
ORGANIZATION	93
WORK-OF-ART	85
COUNTRY	80
BUILDING	52
OPERA	52
THEATRE	42
WORSHIP-PLACE	41

(a) MMHERCL.

TYPE	#
PERSON	1,409
MUSIC	372
CITY	289
THEATER	120
NEWSPAPER	50
ORGANIZATION	36
SCHOOL	27
COUNTRY	25
BUILDING	13
COUNTRY REGION	12

(b) MHERCL-ITA.

Table 7: Top 10 NE types occurring in the benchmarks included in the MMHERCL dataset.

[13] https://openrouter.ai/

In this section, we report extended statistics of the KE-MHISTO benchmark. Table 6a compares the two datasets included in MMHERCL. Notably, MHERCL-ITA exhibits a higher average token count per sentence, reflecting structural differences between English and Italian historical texts. The top 10 NE types occurring in MMHERCL and MHERCL-ITA are reported in Table 7.

C MDYNAKNOWLEDGE examples

We report in Table 8 some samples from MDYNAKNOWLEDGE dataset.

D Further details on long-tail comparative analysis

D.1 Competing benchmark statistics and SotA results

We report dataset statistics and SotA results in Tables 9 and 10. For the EL task, we report F1 scores. For the QA task, we report Exact Match (EM) (Rajpurkar et al., 2016) for the compared datasets. EM measures the percentage of predictions that match any one of the ground truth answers exactly. For our proposed datasets, included in MDYNAKNOWLEDGE, answers are free-form, preventing direct EM computation. Instead, we report Accuracy (strict) (cf. 4), which we consider comparable to EM. As shown in Table 9, SotA results are systematically lower for datasets containing documents published in past centuries compared to datasets containing contemporary ones. This suggests that historical document datasets pose greater challenges for existing EL models, likely due to their long-tail nature, which is not adequately captured by contemporary training corpora.

D.2 Alternative popularity indices

We experiment with entity degree in the Wikidata KG as a proxy for popularity, calculated by counting all connections each entity has within the KG, treating it as undirected, following Arora et al. (2021). We report the analysis in Figure 6. Mainstream EL benchmarks show varying distribution characteristics (see Figure 6a). AIDA-ConLL-YAGO and WNED exhibit concentration around moderate-to-high degree values, while CWEB displays broader coverage. MHERCL-EN and MHERCL-ITA demonstrate extended tail distributions with substantial coverage of low-degree

[14] State-of-the-art methods achieve 90 EM. See https://rajpurkar.github.io/SQuAD-explorer/

Dataset	Lan	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
DK-EN	EN	Where was located the music school in which Gioachino Rossini studied?	In Bologna	Born in Pesaro to parents who were both musicians (his father a trumpeter, his mother a singer), Rossini began to compose by the age of twelve and was educated at music school in Bologna.
DK-ITA	IT	Per quali opere è principalmente ricordato Gioachino Rossini?	Il barbiere di Siviglia, L'italiana in Algeri, La gazza ladra, La Cenerentola, Il turco in Italia, Tancredi, Semiramide e Guglielmo Tell.	Fra i massimi e più celebri operisti della storia, la sua attività ha spaziato attraverso vari generi musicali, ma è ricordato principalmente per le sue opere celebri, quali Il barbiere di Siviglia, L'italiana in Algeri, La gazza ladra, La Cenerentola, Il turco in Italia, Tancredi, Semiramide e Guglielmo Tell.

Table 8: Example of question, answer, and provenance sample contained in DK-EN for the NE `Gioachino_Rossini` (`Q9726`).

Dataset	Lan	#entities (in-KB)	Source publication time	SotA F1
MHERCL-EN	EN	1,658	1823-1900	0.76 (Graciotti et al., 2025)
MHERCL-ITA	IT	1,734	1866-1921	0.51 (GPT4-o mini w/NIL, cf. Table 2)
AIDA CoNLL-YAGO	EN	27,642	Nowadays	0.88 (Shavarani and Sarkar, 2023)
WNED-WIKI	EN	3,396	Nowadays	0.90 (Ayoola et al., 2022a)
WNED-CWEB	EN	5,599	Nowadays	0.79 (Ayoola et al., 2022b)
NEEL-IT-TWITTER	IT	544	Nowadays	0.50 (Attardi et al., 2016)
HIPE2020	EN	311	19C-20C	0.76 (Graciotti et al., 2025)
AJMC	EN	176	19C	0.38 (Ehrmann et al., 2022b)
TopRes19th	EN	1,083	18C-19C	0.65 (Ehrmann et al., 2022b)

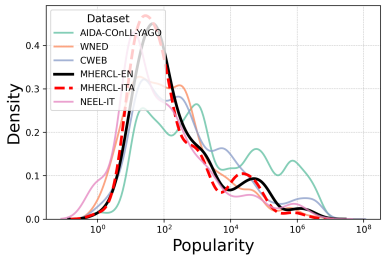
Table 9: Comparison of statistics and SotA performances on EL benchmarks compared to MMHERCL.

Dataset	Lan	#pages	SotA
DK-EN	EN	208	0.32 Acc. (Strict) (LLAMA3.3, cf. Table 4)
DK-ITA	IT	706	0.31 Acc. (Strict) (LLAMA3.3, cf. Table 4)
Natural Questions	EN	39,415	0.64 (EM) (Izacard et al., 2023)
HotpotQA	EN	88,287	0.72 (EM) (Zhang et al., 2024)
TriviaQA	EN	38,402	0.87 (EM) (Anthropic, 2023)
SQuAD2.0	EN	469	0.90 (EM)[14]
SQuAD-it	IT	445	0.64 (EM) (Croce et al., 2019)

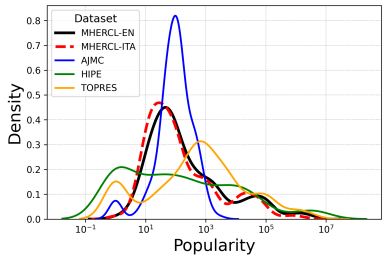
Table 10: Comparison of statistics and SotA performances on commonly used QA benchmarks compared to MDYNAKNOWLEDGE. EM stands for Exact Match. Pages are Wikipedia pages on the basis of which the questions contained in each dataset are created.

entities. In Figure 6b, historical benchmarks show mixed patterns: TopRes19th exhibits concentration at higher degrees, while AjMC, HIPE-2020, and MHERCL-EN and MHERCL-ITA show broader coverage across the degree spectrum with similar tail characteristics. In Figure 6c, QA datasets display distinct behaviors: Natural Questions, HotpotQA, and TriviaQA concentrate around moderate degree values, SQuAD-EN/IT show narrow distributions at higher degrees, while DK-EN/DK-ITA maintain substantial representation of low-degree entities.

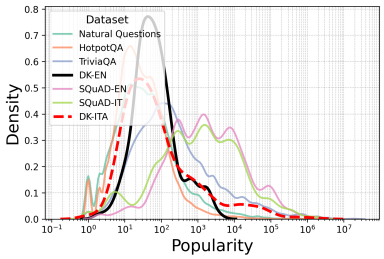
We experiment with Wikipedia article quality scores as a proxy for popularity, computed by extracting structural features from article wikitext and applying the language-agnostic weighting model by (Johnson, 2021) following Arora et al. (2024b). The quality scores derive from a weighted combination of several normalized features: article length, reference count, section count, wikilinks, media files, and categories. The analysis reported in Figure 7 reveals distribution patterns across dataset categories. For each plot, we report the minimum and the maximum Wikipedia page quality obtained for the group of datasets included in the graph. Mainstream EL benchmarks (AIDA-ConLL-YAGO, WNED, CWEB) concentrate near maximum quality scores (1.0), while MHERCL-EN/MHERCL-ITA demonstrate longer-tailed distributions spanning the quality spectrum (see Figure 7a). Historical benchmarks (TOPRES, HIPE2020) show concentration at high quality values compared to MMHERCL's coverage (see Figure 7b). Question answering datasets (SQuAD-EN, NaturalQuestions) display concentration around high quality scores, whereas DK-EN/DK-ITA maintains coverage of lower-quality entities (see Figure 7c).



(a) Comparison of MHERCL-EN/ITA, KILT's EL datasets (AIDA-COnLL-YAGO, WNED, CWEB) and NEEL-it.

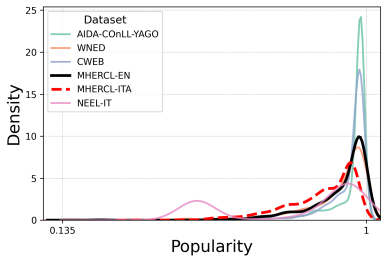


(b) Comparison of MHERCL-EN/ITA, and other historical EL benchmarks (AJMC, HIPE-2020, TopRes19th).

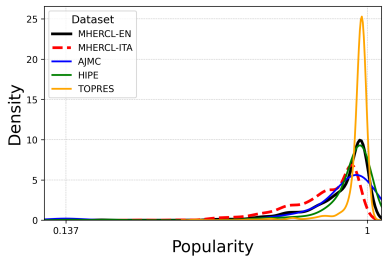


(c) Comparison of DK-EN, DK-ITA, Natural Questions, HotpotQA, TrivialQA, SQuAD2.0 and SQuAD-IT.

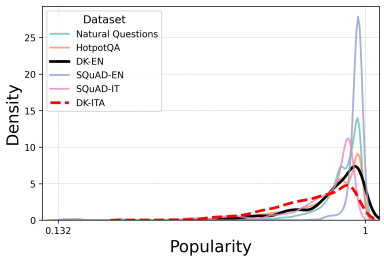
Figure 6: Popularity (Wikidata KG degree) distribution (computed using Kernel Density Estimation) of named entities across selected benchmarks in English and Italian. Only entities with a valid QID are considered. `NIL` entities are excluded.



(a) Comparison of MHERCL-EN/ITA, KILT's EL datasets (AIDA-COnLL-YAGO, WNED, CWEB) and NEEL-it.



(b) Comparison of MHERCL-EN/ITA, and other historical EL benchmarks (AJMC, HIPE-2020, TopRes19th).



(c) Comparison of DK-EN, DK-ITA, Natural Questions, HotpotQA, TrivialQA, SQuAD2.0 and SQuAD-IT.

Figure 7: Popularity (Wikidata page quality) distribution (computed using Kernel Density Estimation) of named entities across selected benchmarks in English and Italian. Only entities with a valid QID are considered. `NIL` entities are excluded.

D.3 Mann-Whitney U Test Analysis

We applied Mann-Whitney U tests, which measure the likelihood of two distributions being equal based on samples randomly sampled from them. In our case, we consider three popularity proxies (QRank, Wikipedia article quality, Wikidata degree) as distributions and measure the popularity distributions of each pair of datasets. Table 11 reports log p-values where lower values indicate higher similarity between distributions. The results confirm the observations discussed in Section 4.1, with particularly notable patterns showing WNED-MHERCL-EN similarity in QRank (-1.47) versus WNED-HIPE divergence (-19.09), and consistent MHERCL-ITA-NEEL-IT differences across all indices.

E Mention Detection and Classification

Table 12 reports a further analysis of the NER task. In particular, we separately evaluate Mention Detection (MD) – finding the spans containing a named entity in a sentence – and Entity Classification (EC) – predicting the type of each named entity found. A named mention is correctly identified when the span found exactly overlaps with the annotation. Similarly, a named mention is correctly classified if the predicted type matches the type annotation. Models are evaluated with the same span-based offset Micro-F1 as done for the full NER task. The analysis reveals that LLama 3.3 is the best-performing model in both tasks. GPT-4 performs equally well in the MD task, but is less accurate in EC, penalizing the ultimate NER score.

F OCR Error Analysis

Table 13 presents the distribution of OCR errors across correctly and incorrectly linked entities in MHERCL-ITA, demonstrating the impact of text quality on EL performance. The same analysis for MHERCL-EN is reported in (Graciotti et al., 2025). The results show that OCR errors are consistently more prevalent in wrongly linked entities across all models, with higher error rates for incor-

Dataset	WNED	NEEL-IT	MHERCL-EN	MHERCL-ITA
AjMC	-23.26/-2.43/-0.01	-0.65/-16.52/-9.42	-11.13/-1.25/-0.04	-46.05/-46.05/-33.22
HIPE	-19.09/-21.33/-20.75	-0.97/-45.25/-35.46	-12.20/-14.89/-12.28	-46.05/-46.05/-46.05
TopRes	-46.05/-46.05/-46.05	-8.15/-46.05/-46.05	-46.05/-46.05/-46.05	-46.05/-46.05/-46.05
AIDA	-46.05/-46.05/-46.05	-46.05/-46.05/-46.05	-46.05/-46.05/-46.05	-46.05/-46.05/-46.05
WNED	0.00/0.00/0.00	-46.05/-46.05/-18.06	-1.47/-1.85/-6.65	-46.05/-46.05/-46.05
CWEB	-46.05/-46.05/-46.05	-7.67/-46.05/-46.05	-46.05/-46.03/-29.09	-46.05/-46.05/-46.05
NEEL-IT	-46.05/-46.05/-18.06	0.00/0.00/0.00	-44.01/-46.01/-34.64	-46.05/-46.05/-27.90
MHERCL-EN	-1.47/-1.85/-6.65	-44.01/-46.01/-34.64	0.00/0.00/0.00	-46.05/-46.05/-46.05
MHERCL-ITA	-46.05/-46.05/-46.05	-46.05/-46.05/-27.90	-46.05/-46.05/-46.05	0.00/0.00/0.00

Table 11: Mann-Whitney U test results showing distribution similarity between datasets. Each cell reports log p-values as "QRank / Wikipedia article quality / Wikidata degree." Lower values indicate higher similarity.

	MHERCL-ITA		MHERCL-EN	
Model	**MD**	**EC**	**MD**	**EC**
GLiNER-multi	0.72	0.77	0.69	0.74
NuExtract 1.5	0.56	0.57	0.57	0.56
LLAMA 3.3 70B	**0.76**	**0.80**	**0.74**	**0.80**
GPT-4o mini	**0.76**	0.70	0.73	0.73

Table 12: Mention Detection and Entity Classification results on MHERCL-ITAand MHERCL-EN.

Dataset	Model	OCR errors over correctly linked entities	OCR errors over wrongly linked entities
MHERCL-ITA	LLAMA3-NIL	9.54%	19.11%
	mGENRE	4.08%	12.33%
	BELA	7.16%	11.98%
	GPT4-NIL	8.90%	17.97%

Table 13: OCR error rates for correctly vs. incorrectly linked entities in MHERCL-ITA, showing the impact of text quality on entity linking performance.

rect links compared to correct ones.

G Prompt Engineering

G.1 Named Entity Recognition

We prompted the LLM requesting to identify all named entities categorizable according to the mMHERCL annotated types. For each sentence, the LLM receives in input the sentence and the list of all possible valid entity types. We leveraged the same English prompt, detailed in Table 14, for both datasets.

You are a Named Entity Recognition Tool.

Given a sentence and a list of possible types, find the named entities mentioned in the sentence that belong to one of the input types.

Output a list structured as: mention1[type], mention2[type], ...

Print only the list and nothing else.

Possible types: `<LABELS>`

Sentence: `<SENTENCE>`

Table 14: Prompt used for zero-shot named entity recognition on LLMs.

G.2 Entity Linking

For models that provide a Wikipedia title, we obtain the corresponding Wikidata QID using Wikimapper[15]. For models that provide multiple ordered alternatives for a single sentence (e.g. BELA), we choose the one with the higher scores. If a model does not provide an answer, we consider it as a wrong prediction rather than a `NIL` one. In other words, we only consider `NIL` predictions when they are explicitly predicted by the model.

In the sentence: `SENTENCE` what is the Wikipedia page name of the entity `MENTION`? Answer with the page title only. Do not write anything else.

Table 15: Prompt used for zero-shot entity linking on LLMs.

[15] https://github.com/jcklie/wikimapper

Model	MHERCL-ITA	MHERCL-EN
GPT-4o mini	0.73	0.61
LLAMA 3.3 70B	0.36	0.35

Table 17: Recall of EL on NIL entities

Model	MHERCL-ITA	MHERCL-EN
GPT-4o mini	0.53	0.64
LLAMA 3.3 70B	0.67	0.82

Table 18: Precision of EL on NIL entities

In the sentence: SENTENCE what is the Wikipedia page name of the entity MENTION? Answer with the page title only. If no Wikipedia page is appropriate, answer 'NIL'. Do not write anything else.

Table 16: Prompt used for zero-shot entity linking on LLMs (with NIL option).

Similarly to the NER, we prompt the LLM with an English prompt for both MHERCL-EN and MHERCL-ITA. Prompts are reported in Table 15 and Table 16 including NIL as a possible prediction.

In Table 17 we report the recall of both LLMs on predicting NIL entities. These results evaluate how reliable a model is – i.e. how sensible it is in detecting whether an entity do not exist on Wikipedia. Surprisingly, each LLM has a very different behavior. While GPT-4 is generally correct when predicting NIL, LLAMA has a more conservative behavior.

Conversely, in Table 18 we show the precision of the LLMs on NIL entities. Differently than before, LLAMA 3.3 has an higher precision when compared to GPT-4. Hence, even though the performances of Table 2 generally suggest that GPT-4 is the model best suited for EL on long-tail knowledge, we argue that depending on the application different LLMs might be better suited. In this case, the more conservative approach of LLAMA might be beneficial to those domains where a reliable entity alignment process is preferred to a more extensive, but possibly approximate, coverage.

G.3 Question Answering

Zero-shot For zero-shot QA, we replicate the prompting strategy of Sun et al. (2024), and use the same prompt (modulo translation) for English (Table 19) and Italian (Table 20).

Answer the following question in as few words as possible. Say "unsure" if you don't know. <QUESTION>

Table 19: Zero-shot QA prompt in English.

Rispondi alla seguente domanda nel minor numero di parole possibile. Rispondi "non so" se non conosci la risposta. <QUESTION>

Table 20: Zero-shot QA prompt in Italian.

Retrieval Augmented Generation For RAG, we provide the context using the same ICL (Tables 21 and Table 22) and CoT (Tables 23 and Table 24) prompts (modulo translation) for both English and Italian datasets.

Given the following context:

* <SPAN> <SPAN TYPE> - <Abstract>

Answer the following question in as few words as possible. Say "unsure" if you don't know. <QUESTION>

Table 21: In-context learning QA prompt in English.

Dato il seguente contesto:

* <SPAN> <SPAN TYPE> - <Abstract>

Rispondi alla seguente domanda nel minor numero di parole possibile. Rispondi "non so" se non conosci la risposta. <QUESTION>

Table 22: In-context learning QA prompt in Italian.

Answer the following question in as few words as possible. Say "unsure" if you don't know.

<QUESTION>

* <SPAN> <SPAN TYPE> - <Abstract>

Consider this additional context, if you think it is useful to answer the question:

* <SPAN> <SPAN TYPE> - <Abstract>

Table 23: Chain-of-thought QA prompt in English.

Rispondi alla seguente domanda nel minor numero di parole possibile. Rispondi "non so" se non conosci la risposta.

<QUESTION>

* <SPAN> <SPAN TYPE> - <Abstract>

Considera queste informazioni aggiuntive se pensi possano essere utili per rispondere alla domanda:

* <SPAN> <SPAN TYPE> - <Abstract>

Table 24: Chain-of-thought QA prompt in Italian.

Given a linked entity, we retrieve the Wikipedia abstract of a Wikidata entity by taking the first section of the Wikipedia page of that entity. We rely on two resources to retrieve the abstracts. We first query DBPedia using the query

```
SELECT DISTINCT ?abstract WHERE {
    [] owl:sameAs <WIKIDATA_IRI> ;
       dbo:abstract ?abstract .
    FILTER (langMatches(
        lang(?abstract),
        <LANGUAGE>)
    )
}
LIMIT 1
```

where <WIKIDATA_IRI is the Wikidata IRI of an entity (e.g. https://www.wikidata.org/entity/Q207628) and <LANGUAGE> is en for DK-EN and it for DK-ITA. If an abstract is not available, we query Wikidata for the corresponding Wikipedia page and rely on the Wikipedia python library[16] to retrieve the abstract. We retrieve at least one abstract for 80% of the samples in DK-EN and for 51% of the samples in DK-ITA.

LLM as a judge To evaluate the LLM responses in the QA task, we adopted an "LLM-as judge" approach because we could not rely on exact matches between predicted and actual responses, as generated answers could have different phrasing or additional contextual information that did not invalidate the answer. As a result, also using traditional metrics like BLEU or METEOR would not have provided an accurate evaluation. Therefore, we used an LLM to assess responses based on three criteria: correct, incorrect, and partially correct, providing the model with clear instructions and examples in the prompt. The prompt is detailed in Table 25. To verify the reliability of the LLM as a judge, we also conducted a human evaluation to compare the LLM's judgments with those of human evaluators. This analysis revealed that the LLM made the correct judgment 95% of the time.

[16] https://github.com/goldsmith/Wikipedia

You are an impartial and highly accurate judge.

Your task is to evaluate the correctness of a predicted answer based on a given question and the ground-truth answer. Follow these steps carefully:

1. Analyze the question to understand what is being asked.

2. Compare the predicted answer to the ground-truth answer.

3. Determine if the predicted answer is correct or partially correct based on its relevance, accuracy, and completeness in addressing the question.

4. Provide a judgment as either "Correct", "Partially Correct" or "Incorrect"

5. Give a brief explanation about your judgement decision.

Provide your judgement and explanation.

Format your output as :

Judgement : <Correct/Partially Correct/Incorrect>

Explanation : <brief explanation>

In the following there are some examples for better instruct you on how to provide your judgements. Examples:<EXAMPLES>

Table 25: LLM-as-a-Judge prompt

H Error Examples

H.1 Question requiring temporal reasoning

In Table 26, we present a selection of questions from DK-EN that necessitate temporal reasoning. For each question, we specify the named entity referenced, the corresponding supporting sentence from Wikipedia, and the source documents in MMHERCL where the named entity appears. Additionally, we report the incorrect answers produced by all evaluated models. In Table 27, we do the same for DK-ITA.

H.2 Questions requiring lists of elements as answers

In Table 28, we present a selection of questions from DK-EN that necessitate lists of elements to be exhaustively answered. For each question, we specify the named entity referenced, the corresponding supporting sentence from Wikipedia, and

Named entity:	Niels Gade (Q154632, occurring in The Musical Times, 1901 - MHERCL-EN)
Analyst's question:	*The premiere of what musical work did Niels Gade conduct in 1845?*
Provenance sentence:	In 1845 Gade conducted the premiere of Mendelssohn's Violin Concerto in E minor with Ferdinand David (from: https://en.wikipedia.org/wiki/Niels_Gade)
Analyst's answer:	Mendelssohn's Violin Concerto in E minor
Dataset:	DK-EN

Model	Erroneous Answer
EuroLLM-9b	The premiere of "Elverskud" was conducted by Niels Gade in 1845.
Minerva-7b	The premiere of "Elverskud" was conducted by Niels Gade in 1845.
LLAMA 3.3 70B	Ossian
w/ ICL RAG	Overture to "Othello"
w/ CoT RAG	Ossian
GPT4 o-mini	The premiere of "Elverskud" was conducted by Niels Gade in 1845.
w/ ICL RAG	Elverskud.
w/ CoT RAG	The premiere of the opera "Elverskud."

Named entity:	Henry Purcell (Q9695, occurring in *The Musical Times*, 1885, *The Quarterly Musical Magazine and Review*, 1826, *The Harmonicon*, 1827 and 1831 - MHERCL-EN)
Analyst's question:	*What works did Purcell compose for Saint Cecilia's Day in 1694?*
Provenance sentence:	Purcell's Te Deum and Jubilate Deo were written for Saint Cecilia's Day, 1694, the first English Te Deum ever composed with orchestral accompaniment. (from: https://en.wikipedia.org/wiki/Henry_Purcell) **Analyst's answer:** Te Deum and Jubilate Deo
Dataset: DK-EN	

Model	Erroneous Answer
EuroLLM-9b	Dido and Aeneas.
Minerva-7b	Dido and Aeneas.
LLAMA 3.3 70B	Hail, Bright Cecilia.
w/ ICL RAG	Hail, Bright Cecilia.
w/ CoT RAG	Ode for St. Cecilia's Day.
GPT4 o-mini	Dido and Aeneas.
w/ ICL RAG	"Welcome to All the Pleasures" and "Hail! Bright Cecilia.".
w/ CoT RAG	"Come, ye Sons of Art, away" and "Hail! Bright Cecilia."

Named entity:	Giacomo Meyerbeer (Q105237, occurring in *The Musical World*, 1855, *Dwight's Journal of Music* - MHERCL-EN)
Analyst's question:	*Who appointed Meyerbeer 'Court Composer' on 12 February 1813?*
Provenance sentence:	On 12 February 1813 Beer received the first of the string of honours he was to accumulate throughout his life when he was appointed 'Court Composer' by Grand Duke Ludwig of Hesse-Darmstadt. **Analyst's answer:** Grand Duke Ludwig of Hesse-Darmstadt
Dataset: DK-EN	

Model	Erroneous Answer
EuroLLM-9b	King Frederick William III appointed Meyerbeer 'Court Composer' on 12 February 1813.
Minerva-7b	King Frederick William III appointed Meyerbeer 'Court Composer' on 12 February 1813.
LLAMA 3.3 70B	Frederick William III.
w/ ICL RAG	Frederick William III
w/ CoT RAG	Friedrich Wilhelm III
GPT4 o-mini	King Frederick William III appointed Meyerbeer 'Court Composer' on 12 February 1813.
w/ ICL RAG	King Friedrich Wilhelm III of Prussia.
w/ CoT RAG	The King of Prussia.

Table 26: Examples of questions from DK-EN requiring temporal reasoning wrongly extracted by all models.

Named entity:	Teatro Margherita (`Q83471820, occurring in L'arpa, 1888 - MHERCL-ITA`)
Analyst's question:	*Di quale città era il Teatro Margherita?*
Provenance sentence:	Il Teatro Margherita (per un periodo noto anche come Politeama Margherita) è stato un teatro di Genova, con sede nella centrale via XX Settembre. Per molte decadi è stato uno dei più importanti teatri nazionali. (from: `https://it.wikipedia.org/wiki/Teatro_Margherita_(Genova)`)
Analyst's answer:	Genova
Dataset:	DK-ITA

Model	Erroneous Answer
EuroLLM-9b	Il Teatro Margherita era situato a Bari.
Minerva-7b	Il Teatro Margherita era a Bari.
LLAMA 3.3 70B	Bari.
w/ ICL RAG	Bari.
w/ CoT RAG	Bari.
GPT4 o-mini	Bari.
w/ ICL RAG	Bari.
w/ CoT RAG	Bari.

Named entity:	Victor de Sabata (`Q464890`, occurring in *Rivista Nazionale di Musica*, 1921 - MHERCL-ITA)
Analyst's question:	*Dal 1918 al 1929, Victor de Sabata fu direttore di quale teatro d'opera?*
Provenance sentence:	Dal 1918 al 1929 fu direttore dell'Opera di Monte Carlo, e poi della Cincinnati Symphony Orchestra. (from: `https://en.wikipedia.org/wiki/Victor_de_Sabata`) **Analyst's answer:** Victor de Sabata fu direttore dell'Opera di Monte Carlo dal 1918 al 1929.
Dataset: DK-ITA	

Model	Erroneous Answer
EuroLLM-9b	Victor de Sabata fu direttore del Teatro alla Scala di Milano dal 1918 al 1929.
Minerva-7b	Teatro alla Scala di Milano
LLAMA 3.3 70B	Teatro Comunale di Trieste
w/ ICL RAG	La Scala.
w/ CoT RAG	La Scala.
GPT4 o-mini	Teatro alla Scala di Milano.
w/ ICL RAG	La Scala.
w/ CoT RAG	Teatro alla Scala.

Named entity:	Alessandro Scarlatti (`Q183087`, occurring in *Rivista Nazionale di Musica*, 1921, 1922 - MHERCL-ITA)
Analyst's question:	*A chi è stato intitolato il Conservatorio Alessandro Scarlatti nel 1915?*
Provenance sentence:	Nel 1915 è stato intitolato alla memoria di Vincenzo Bellini ed è un'istituzione di alta formazione artistica, musicale e coreutica (AFAM). **Analyst's answer:** Nel 1915 il Conservatorio Alessandro Scarlatti è stato intitolato a Vincenzo Bellini.
Dataset: DK-EN	

Model	Erroneous Answer
EuroLLM-9b	Il Conservatorio Alessandro Scarlatti è stato intitolato al compositore Alessandro Scarlatti nel 1915.
Minerva-7b	Il Conservatorio Alessandro Scarlatti è stato intitolato ad Alessandro Scarlatti nel 1915.
LLAMA 3.3 70B	Alessandro Scarlatti
w/ ICL RAG	Alessandro Scarlatti
w/ CoT RAG	Alessandro Scarlatti
GPT4 o-mini	Ad Alessandro Scarlatti.
w/ ICL RAG	Ad Alessandro Scarlatti.
w/ CoT RAG	Ad Alessandro Scarlatti.

Table 27: Examples of questions from DK-ITA requiring temporal reasoning wrongly extracted by all models.

Named entity:	Marcella Sembrich (Q265698, occurring in *The Musical Times*, 1884 - MHERCL-EN)
Analyst's question:	*What did Marcella Sembrich study at the Vienna Conservatory in 1875?*
Provenance sentence:	In the Autumn of 1875 she began her studies at the Vienna Conservatory, studying voice with Joseph Hellmesberger Sr., piano with Julius Epstein, and voice with Viktor Rokitansky. (from: https://en.wikipedia.org/wiki/Marcella_Sembrich)
Analyst's answer:	voice and piano
Dataset:	DK-EN

Model	Partially Correct Answer
EuroLLM-9b	Vocal technique.
Minerva-7b	Vocal technique.
LLAMA 3.3 70B	Singing
w/ ICL RAG	Singing.
w/ CoT RAG	Music.
GPT4 o-mini	Vocal technique.
w/ ICL RAG	Voice.
w/ CoT RAG	Voice and singing.

Table 28: Examples of question from DK-EN requiring lists of elements as answers partially correctly answered by all models.

Named entity:	Sanremo (Q33962, occurring in *L'arpa*, 1882 - MHERCL-ITA)
Analyst's question:	*A cosa deve la sua notorietà la città di Sanremo?*
Provenance sentence:	Deve la sua notorietà per ospitare l'arrivo della corsa ciclistica Milano-Sanremo (una delle "classiche monumento" del ciclismo) ed essere sede del Festival della canzone italiana, del Premio Tenco, dell'omonima manifestazione rallistica e del Challenge Sanremo (gara internazionale di triathlon); è anche sede di uno dei quattro casinò presenti in Italia. (from: https://it.wikipedia.org/wiki/Sanremo)
Analyst's answer:	La città di Sanremo deve la sua notorietà per ospitare l'arrivo della corsa ciclistica Milano-Sanremo ed essere sede del Festival della canzone italiana, del Premio Tenco, dell'omonima manifestazione rallisitica e del Challenge Sanremo, ed è anche sede di uno dei quattro casinò presenti in Italia.
Dataset:	DK-ITA

Model	Partially Correct Answer
EuroLLM-9b	Sanremo è famosa per il suo festival della canzone.
Minerva-7b	La città di Sanremo deve la sua notorietà principalmente al Festival della Canzone Italiana, che si tiene ogni anno nella città ligure dal 1951.
LLAMA 3.3 70B	Il Festival della Canzone Italiana.
w/ ICL RAG	La città di Sanremo deve la sua notorietà per ospitare l'arrivo della corsa ciclistica Milano-Sanremo ed essere sede del Festival della canzone italiana.
w/ CoT RAG	La sua notorietà è dovuta a diversi motivi tra cui il festival della canzone italiana, la corsa ciclistica Milano-Sanremo e la coltivazione dei fiori.
GPT4 o-mini	Al Festival della Canzone Italiana.
w/ ICL RAG	Festival della canzone italiana e corsa Milano-Sanremo.
w/ CoT RAG	Per il Festival della canzone italiana e la corsa ciclistica Milano-Sanremo.

Table 29: Example of question from DK-ITA requiring lists of elements as answers partially correctly answered by all models.

the source documents in MMHERCL where the named entity appears. Additionally, we report the partially correct answers produced by all evaluated models. In Table 29, we do the same for DK-ITA.