Beyond Single-Negative Preference: Multi-Negative DPO for LLM-Centric Historical Entity Linking

Tien Nam Nguyen[1] **Emanuela Boros**[1] **Ahmed Hamdi**[2]
Adam Jatowt[3] **Mickaël Coustaty**[1] **Antoine Doucet**[1,4]
[1]L3i, University of La Rochelle, La Rochelle, France
[2]University of Toulouse, IRIT, Toulouse, France
[3]University of Innsbruck, Austria
[4]FRI, University of Ljubljana, Ljubljana, Slovenia

Abstract

Large language models (LLMs) have recently shown promise for historical entity linking, but preference optimization for this task is often formulated with only one negative candidate per training instance. This discards information from the remaining candidates retrieved for the same mention. We introduce multi-negative direct preference optimisation (MDPO), a reference-based pairwise objective that compares the correct entity with all valid rejected candidates associated with each mention. MDPO preserves the Bradley-Terry formulation of DPO while exploiting the complete candidate set through masked, length-normalised sequence scores. We evaluate MDPO on `hipe-2020` and `newseye`, covering French, German, English, Swedish, and Finnish historical newspaper text. Experiments show that MDPO improves over supervised fine-tuning and single-negative DPO, with particularly strong gains for NIL mentions, semantic ambiguity, OCR noise, and historically difficult names. Further analyses disentangle candidate-generation and selection errors, showing that candidate retrieval remains a key bottleneck for end-to-end entity linking. These results demonstrate that incorporating all within-instance negative candidates is a simple and effective improvement for LLM-based historical entity linking.

1 Introduction

Entity linking (EL) maps textual mentions to their corresponding entities in a knowledge base and is a core component of information extraction, question answering, digital library access, and knowledge base population. Standard EL systems usually decompose the task into candidate generation and entity disambiguation or ranking (Sevgili et al., 2022): candidate generation retrieves plausible entities for a mention, while the ranking stage selects the most appropriate entity given the context. Recent neural approaches have improved both stages using dense retrieval, bi-encoders, cross-encoders, and generative models, including BLINK (Wu et al., 2020), ELQ (Li et al., 2020), GENRE (De Cao et al., 2021; De Cao et al., 2022), ReFinED (Ayoola et al., 2022), and LLMAEL (Xin et al., 2025).

Despite these advances, EL remains difficult in historical and archival documents, which often contain OCR noise, spelling variations, obsolete or multilingual forms, incomplete metadata, temporally ambiguous references, and mentions that are absent from the target knowledge base. These factors make candidate retrieval, disambiguation, and NIL prediction especially challenging. The `hipe-2020` dataset, introduced through the HIPE shared tasks, highlights the difficulty of robust named entity recognition (NER) and linking (EL) in multilingual historical documents (Ehrmann et al., 2020a,b). Complementarily, the `newseye` dataset provides multilingual historical newspaper material in French, German, Finnish, and Swedish. Together, these benchmarks show that historical EL requires methods that go beyond surface-form matching and account for noisy, incomplete, and historically situated contexts.

LLMs offer new opportunities for this setting because they combine broad parametric knowledge with contextual reasoning and generative capabilities. Recent work has explored LLMs for EL through prompting, instruction tuning, context augmentation, and selective reranking (Xiao et al., 2023; Ding et al., 2024; Vollmers et al., 2025; Xin et al., 2025; Li et al., 2025). In historical EL, LLMs have also been used to improve NIL prediction and candidate selection, often in combination with traditional retrievers (Santini et al., 2026). However,

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39970–39985
October 24-29, 2026 ©2026 Association for Computational Linguistics

most existing approaches still treat LLMs as auxiliary components: they augment contexts, rerank externally retrieved candidates, or predict from a fixed candidate set. This limits their ability to generate plausible candidates and does not directly model the multi-candidate nature of entity disambiguation.

In this paper, we propose an LLM-centric framework for historical entity linking that uses LLMs in both candidate retrieval and entity selection. First, we introduce an LLM-guided candidate retrieval strategy, where an LLM generates plausible entity candidates from a mention and its surrounding context. The generated candidates are validated against a knowledge base and combined with alias-based retrieval, allowing the system to exploit both contextual generation and high-recall lexical lookup. This is particularly useful for historical documents, where OCR noise, spelling variation, and archaic surface forms can make conventional retrieval brittle.

Second, we reformulate entity selection as a preference-learning problem. Standard DPO trains a model from one chosen–rejected comparison (Rafailov et al., 2023), whereas EL naturally supplies several hard competitors for the same mention. We introduce a multi-negative DPO objective that preserves the reference-based pairwise formulation while optimising all valid gold-versus-negative comparisons from each retrieved set.

Third, we evaluate the proposed framework on two multilingual historical-document benchmarks: `hipe-2020`, introduced for NER and EL in historical newspapers (Ehrmann et al., 2020a,b), and `newseye`, a multilingual historical newspaper dataset covering French, German, Finnish, and Swedish material (Hamdi et al., 2021). Our evaluation compares supervised fine-tuning, standard DPO, prompting-based baselines, and the proposed multi-negative DPO formulation, and provides ablation studies on retrieval, prompt design, context representation, and model choice.

2 Related work

LLM-based entity linking. EL is commonly formulated as a two-stage task: candidate generation followed by entity disambiguation. Traditional systems rely on lexical matching, prior probabilities, dense retrieval, and supervised ranking models, while recent work increasingly explores LLMs for candidate generation, context enrichment, and entity selection. INSGENEL (Xiao et al., 2023) shows that instruction-tuned decoder-only models, combined with a retriever, can perform effective entity linking while reducing the cost of full sequence generation. EntGPT (Ding et al., 2024) demonstrates the potential of LLMs for entity disambiguation through prompting and instruction tuning, using LLMs to generate auxiliary information and select the correct entity from a candidate set. Hybrid methods such as ARTER (Li et al., 2025) combine fast entity linkers with targeted LLM reasoning, routing only difficult mentions to the LLM to balance accuracy and efficiency.

Another closely related direction uses LLMs to enrich the input context. Vollmers et al. (2025) propose contextual augmentation, where ambiguous mentions are expanded into more explicit Wikipedia-like titles using surrounding context. Similarly, LLMAEL (Xin et al., 2025) uses LLMs as plug-and-play context augmenters and fuses the generated context with existing EL models such as BLINK (Wu et al., 2020; Li et al., 2020), GENRE (De Cao et al., 2021; De Cao et al., 2022), and ReFinED (Ayoola et al., 2022). These approaches show that LLMs can help resolve ambiguity by adding missing contextual or world knowledge. Our method shares this intuition, but instead of only augmenting the mention context or reranking candidates, we prompt the LLM to generate plausible entity candidates, validate them against the knowledge base, augment them with alias-based retrieval, and then train an LLM-based selector using preference learning.

Entity linking in multilingual, historical, and domain-specific settings. Entity linking is especially difficult in multilingual, historical, and domain-specific corpora, where mentions may be noisy, rare, temporally ambiguous, or absent from standard knowledge bases. BELA (Plekhanov et al., 2023) addresses multilingual end-to-end entity linking with a large Wikipedia- and Wikidata-based index, jointly modeling mention detection, disambiguation, and rejection. For historical EL, Santini et al. (2026) propose MHEL-LLaMo, an unsupervised multilingual framework that combines bi-encoder retrieval with an instruction-tuned LLM for NIL prediction and candidate selection, applying the LLM mainly to difficult cases. In cultural heritage domains, studies on musical heritage and museum collections highlight the importance of low-popularity entities, temporal constraints, NIL

prediction, OCR noise, and domain-specific adaptation (Graciotti et al., 2025; Cadavid-Sanchez et al., 2023).

Preference optimisation. DPO learns a reference-relative Bradley–Terry objective from one chosen–rejected pair (Rafailov et al., 2023), whereas SimPO replaces the reference model with a length-normalised, reference-free reward (Meng et al., 2024). Multi-response methods address richer supervision: MPPO supports arbitrary negative samples and studies pointwise, pairwise, and listwise variants using average-likelihood rewards (Xie et al., 2025), while LiPO couples candidates through a learning-to-rank loss over ranked response lists (Liu et al., 2025). Our EL data instead provide one gold entity and an unordered set of retrieved competitors. We therefore retain DPO's reference-relative score, preserve the shared mention-level candidate structure, and aggregate independent gold-versus-negative terms. Unlike LiPO, our loss has no shared listwise normalisation; unlike single-negative DPO, it does not discard the remaining competitors. Our contribution is this task-specific construction and implementation rather than a new listwise objective.

3 Methodology

Entity linking is naturally a multi-candidate decision problem: for each mention m appearing in context c, the model must select the corresponding entity e from a knowledge base $\mathcal{K}$ or abstain with a special NIL label when no valid entity exists. Formally, the task is defined as learning a function:

$$f(m, c; \mathcal{K}) \rightarrow e, \quad e \in \mathcal{K} \cup \{\text{NIL}\}. \tag{1}$$

This structure supplies multiple valid preference comparisons for each mention. We retain the gold entity and all hard negatives or NIL alternatives from the same retrieved set, rather than sampling only one negative. The training objective remains pairwise; "multi-negative" describes the within-instance data structure and aggregation, not a listwise normalisation.

We propose an LLM-centric entity linking pipeline that integrates candidate retrieval and entity selection through structured interaction with LLMs, as shown in Figure 1. The pipeline consists of two main stages: (1) LLM-guided candidate retrieval and (2) LLM-based entity selection.

3.1 LLM-guided Candidate Retrieval

Traditional candidate retrieval methods for entity linking usually construct a small set of possible entities using lexical or heuristic signals, such as alias dictionaries, mention entity priors, inverted indices, and approximate string matching (Bunescu and Paşca, 2006; Cucerzan, 2007; Ratinov et al., 2011; Robertson and Zaragoza, 2009; Zhang et al., 2022). While efficient, these methods are sensitive to surface-form mismatch. This is particularly problematic in historical corpora, where mentions may contain OCR errors, spelling variants, archaic forms, abbreviations, or semantic shifts (Provatorova et al., 2020; Hamdi et al., 2023). If the correct entity is not retrieved at this stage, the subsequent selection model might not find it.

To improve recall, we adopt an LLM-guided candidate retrieval strategy. Instead of relying only on lexical similarity, we prompt the LLM with the mention m and its surrounding context c to generate semantically plausible candidate strings, including canonical names, aliases, normalised spellings, and historically plausible variants. These generated strings are not treated as final predictions; rather, they are used to expand retrieval over the knowledge base.

Formally, we define the LLM-generated query set as:

$$\mathcal{C}_{\text{LLM}}(m, c) = \text{Retrieve}_{\mathcal{K}}(\text{LLM}_{\text{gen}}(m, c)), \tag{2}$$

where $\mathcal{C}_{\text{LLM}}(m, c)$ denotes the knowledge-base candidates retrieved from the generated strings. Because LLMs may generate hallucinated or non-existent entity names, we validate the retrieved candidates against the knowledge base:

$$\mathcal{C}_{\text{valid}}(m, c) = \{\, e \in \mathcal{C}_{\text{LLM}}(m, c) \mid \text{Valid}_{\mathcal{K}}(e) \,\}. \tag{3}$$

Here, $\text{Valid}_{\mathcal{K}}$ checks whether the candidate corresponds to a valid entity identifier in the knowledge base.

To further improve recall, we build an alias dictionary independently for each benchmark using only its training split. Each normalised mention surface form is mapped to all of its annotated QID(s); one-to-many mappings are retained, and no external or manually curated aliases are used. Duplicate alias and LLM candidates are removed before the two sources are merged. Thus, test-only mention–entity associations cannot enter the lookup table.

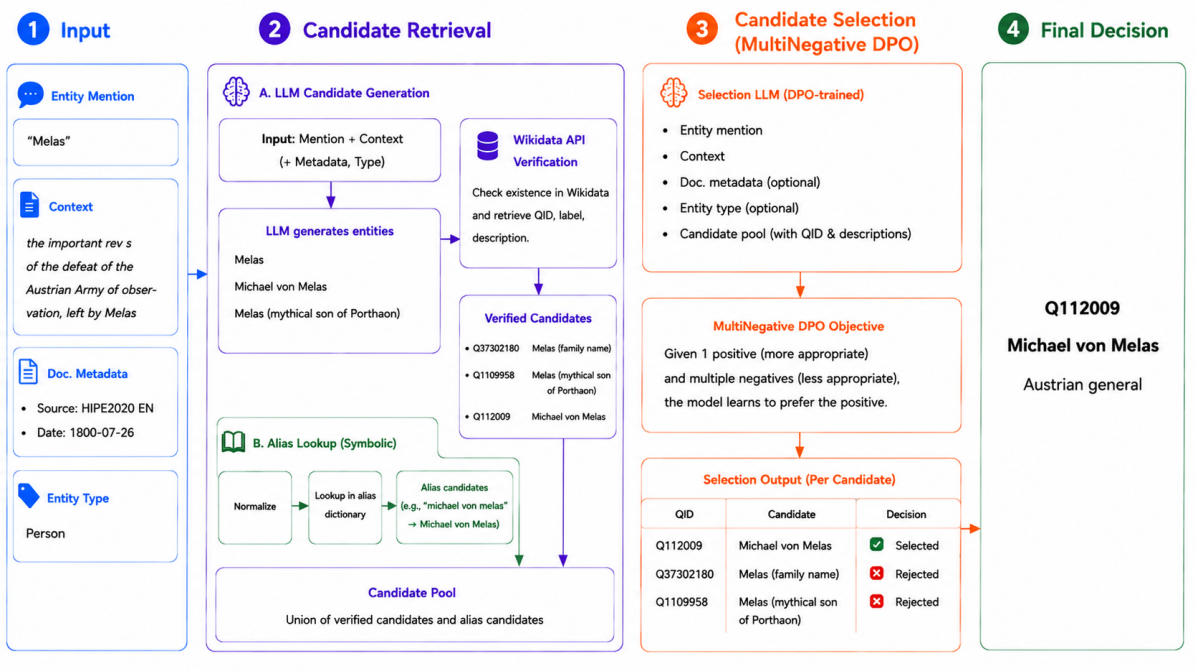


Figure 1: Overview of the proposed LLM-centric entity linking pipeline.

We define:

$$\mathcal{C}_{\text{alias}}(m) = \text{AliasLookup}_{\mathcal{K}}(m). \tag{4}$$

The final candidate set combines validated LLM-guided candidates with alias-based matches:

$$\mathcal{C}(m, c) = \mathcal{C}_{\text{valid}}(m, c) \cup \mathcal{C}_{\text{alias}}(m). \tag{5}$$

This approach uses the LLM to expand the search space while relying on the knowledge base to ensure candidate validity. Thus, the system improves recall without directly accepting unsupported LLM-generated entities.

3.2 LLM-based Entity Selection

In the second stage, the LLM performs joint reasoning over the candidate set to select the most appropriate entity. Given a mention m, its context c, and a candidate set $\mathcal{C}$ with associated descriptions, the model predicts the most likely entity:

$$\hat{e} = \arg\max_{e \in \mathcal{C}} P(e \mid m, c, \mathcal{C}) \tag{6}$$

EL as preference learning. We reformulate entity linking as a preference learning problem over the candidate set $\mathcal{C}$. Given a gold entity e^+ and a set of negative candidates $\{e_1^-, e_2^-, \ldots, e_n^-\}$, the objective is to learn preferences such that: $e^+ \succ e_i^-, \quad \forall i \in \{1, \ldots, n\}$.

Multi-negative DPO. Let $x_b = (m_b, c_b, \mathcal{C}_b)$ and $\bar{\ell}_\pi(e \mid x) = |e|^{-1} \log \pi(e \mid x)$ be the length-normalised sequence log-likelihood. For instance b with n_b valid negatives, we optimise:

$$\begin{aligned} r_\theta(e \mid x) &= \bar{\ell}_{\pi_\theta}(e \mid x) - \bar{\ell}_{\pi_{\text{ref}}}(e \mid x), \\ \Delta_{b,i} &= r_\theta(e_b^+ \mid x_b) - r_\theta(e_{b,i}^- \mid x_b), \\ \mathcal{L}_{\text{MDPO}} &= -\frac{1}{\sum_{b=1}^{B} n_b} \sum_{b=1}^{B} \sum_{i=1}^{N} M_{b,i} \log \sigma(\beta \Delta_{b,i}), \\ n_b &= \sum_{i=1}^{N} M_{b,i}. \end{aligned} \tag{7}$$

Here π_θ is the LoRA-adapted policy, π_{ref} is the same pretrained backbone with its adapters disabled, and $\beta > 0$ controls preference strength. The batch and tensor notation makes the implementation explicit. B is the number of mentions in a minibatch, N is the maximum number of valid rejected candidates in that minibatch, and L is the maximum token length used when scoring one entity response. The chosen responses therefore have shape $[B, L]$, while the rejected responses are padded to $[B, N, L]$. $M_{b,i} \in \{0, 1\}$ is a validity mask: it is one when the i-th negative belongs to mention b and zero for padding or an unavailable comparison. Hence n_b counts only real negatives, and the denominator averages over valid comparisons rather than padded tensor positions.

For each mention, the trainer computes the length-normalised chosen score once and reuses it for all valid negatives. It then computes one independent margin $\Delta_{b,i}$ for each gold–negative pair and applies the binary Bradley-Terry likelihood to that pair. A positive margin means that the policy improves the gold entity's reference-relative score more than the corresponding negative's. Normalisation by $\sum_b n_b$ gives every valid comparison equal weight, while an instance with more hard negatives contributes more pairwise evidence. Crucially, N and the $[B, N, L]$ tensor are implementation dimensions, not a listwise probability distribution: there is no softmax over candidates, no negative-to-negative interaction, and no shared candidate-set normalisation. Thus, "multi-negative" means that every retrieved competitor supplies supervision for the same mention, whereas the objective remains a sum of ordinary reference-based pairwise DPO terms. If $n_b = 1$, Eq. 7 becomes the single-negative version of this length-normalised, reference-based DPO loss.

4 Preference Data Construction

To train our model with the multi-negative DPO objective, we construct preference data automatically from standard entity linking annotations. Each instance consists of a mention m, context c, and a ground-truth entity e^+, which can be either a valid knowledge base entity (non-NIL) or NIL.

We first apply our LLM-based candidate generation pipeline to obtain a candidate set $\mathcal{C}$. Based on the relationship between e^+ and $\mathcal{C}$, we consider four cases:

Case 1: Non-NIL with Correct Candidate Present. If $e^+ \neq$ NIL and $e^+ \in \mathcal{C}$, we construct preference pairs where the gold entity is preferred over all other candidates:

$$(e^+, e^-), \qquad \forall e^- \in C \setminus \{e^+\}$$

Case 2: Non-NIL with Missing Candidate. If $e^+ \neq$ NIL but $e^+ \notin \mathcal{C}$, we explicitly add e^+ into the candidate set:

$$C \leftarrow C \cup \{e^+\}$$

and construct preference pairs as in Case 1. This ensures the model always observes the correct entity during training.

Case 3: NIL with Non-empty Candidate Set. If $e^+ =$ NIL and $\mathcal{C} \neq \emptyset$, we treat NIL as the correct choice and prefer it over all retrieved candidates:

$$(\text{NIL}, e^-), \qquad \forall e^- \in C$$

Case 4: NIL with NIL Candidate Present. If NIL is already included in $\mathcal{C}$, we retain it as the preferred candidate and construct preference pairs against all non-NIL entities.

DPO Formatting. Each record stores one prompt $(m, c, \mathcal{C})$, one *chosen* entity, and the complete list of valid *rejected* entities. These comparisons remain grouped through collation and loss computation (Appendix A). Records with no valid rejected entity, including empty-candidate gold-NIL cases, provide no preference comparison and are excluded from DPO training.

5 Experiments

5.1 Datasets

We conduct experiments on the following datasets:

`hipe-2020` (Ehrmann et al., 2020a) dataset provides a benchmark for NER and EL on multilingual historical newspapers. The dataset contains $17,553$ named entity mentions distributed across German, French, and English historical newspaper corpora, with annotations aligned to Wikidata for the entity linking task. The corpus includes $8,205$ location entities, $6,113$ person entities, $1,971$ organisation entities, 679 temporal expressions, and 585 product entities, reflecting a diverse range of semantic categories in historical documents.

`newseye` (Hamdi et al., 2021) dataset provides a suitable benchmark for the entity linking task, as it includes $30,580$ named entities, among which $6,704$ are aligned with Wikidata. This explicit grounding of entities in a knowledge base enables the evaluation of linking approaches in a realistic, multilingual setting. The datasets consider four languages (French, German, Finnish, and Swedish), covering a diverse range of linguistic characteristics. The corresponding subsets contain $12,473$ entities for French, $12,818$ for German, $2,572$ for Finnish, and $2,717$ for Swedish, mainly spanning persons, locations, and organisations. The `newseye` dataset enables robust cross-lingual comparisons with the `HIPE` dataset, which is also used in our study. Both of them ensure consistency in annotation schemes.

Approach	Setup	hipe-2020			newseye			
		FR	**DE**	**EN**	**FR**	**DE**	**SV**	**FI**
SBB (Labusch and Neudecker, 2020)	Wiki emb. + BERT/RF	59.6	50.6	39.3	44.4	43.1	–	–
L3i (Boros et al., 2020)	Neural EL + postproc.	60.2	48.1	54.6	–	–	–	–
MELHISSA (Linhares Pontes et al., 2022)	Multiling. EL + filters	63.0	57.3	59.7	54.2	54.7	59.9	**65.2**
BELA (Plekhanov et al., 2023)	Bi-enc. + score	58.0	55.0	41.8	42.2	30.0	41.0	30.0
MHEL-LLaMo (Santini et al., 2026)	BELA + LLM	69.2	62.0	72.3	66.2	55.6	52.1	50.9
Prompting	GPT-120B / GPT-20B	65.7	59.6	65.2	57.9	49.2	53.3	60.4
SFT	GPT-120B / Qwen3-14B	60.3	55.0	46.2	44.8	36.6	48.2	38.4
DPO	GPT-120B / GPT-20B	70.2	62.5	70.9	63.5	52.8	62.8	63.5
Multi-DPO	GPT-120B / GPT-20B	**72.4**	**65.6**	**73.8**	**68.7**	**59.2**	**65.0**	62.7

Table 1: Micro-F1 on hipe-2020 and newseye.

5.2 Models and Implementation

We employ multiple LLMs across different components of the pipeline to balance retrieval quality and computational efficiency.

For candidate generation, we use a diverse set of open-source and large-scale LLMs, including GPT-20B, 120B (Agarwal et al., 2025), Qwen3.5-9B and 27B (Team, 2026), Qwen3-A30B (Yang et al., 2025), Gemma4-31B (Team et al., 2026). These models are used in a zero-shot setting to generate semantically relevant and diverse candidate entity names given a mention and its context.

For entity selection, we finetune Qwen3-14B (Yang et al., 2025), GPT-20B (Agarwal et al., 2025), or Qwen3-32B (Yang et al., 2025) using the proposed multi-negative DPO. Details of the training and inference settings can be found in Appendix A.

This design choice decouples candidate generation from ranking, enabling the use of large models for high-recall retrieval while maintaining efficiency during downstream selection.

We implement structured prompting and output formatting using the OpenAI Agents framework[1], which supports schema-constrained generation and function calling for candidate generation, validation, and entity selection. Since the framework is not fully compatible with some fine-tuned Qwen-based models, we use direct prompting with manually enforced structured outputs while preserving the same interaction protocol and output schema. This ensures consistent structured outputs and reduces parsing errors during inference.

5.3 Compared Systems

We compare our approach against both prior historical entity linking systems and several training configurations built on the same retrieval-augmented selection framework.

Prior historical EL systems. We report published results for SBB, L3i, MELHISSA, BELA, and MHEL-LLaMo, which represent lexical, neural, multilingual, and LLM-assisted approaches for historical entity linking.

Our training configurations. All variants use the same retrieval-augmented setup and differ only in the selection strategy:

- LLM prompting. The LLM selects an entity directly from the retrieved candidate list without task-specific fine-tuning.
- Supervised fine-tuning (SFT). The base model is fine-tuned to generate the correct entity given the input $(m, c, \mathcal{C})$.
- DPO (standard). A pairwise DPO setting using one positive and one negative candidate for each training instance.
- DPO (multi-negative). Our proposed approach extends DPO by contrasting the gold entity against multiple negative candidates within the same instance, providing a richer supervision signal for entity selection.

5.4 Evaluation Metrics

Following the standard evaluation protocol of the HIPE 2020 shared task[2], we use micro F1 as the primary evaluation metric for a fair comparison with previous work. In addition, since our framework contains an explicit candidate retrieval stage, we also report entity-level retrieval recall to evaluate the effectiveness of the retrieval component independently of the final entity selection stage.

[1] https://github.com/openai/openai-agents-python

[2] https://hipe-eval.github.io/HIPE-2022/evaluation

6 Results

Main Results. Table 1 reports micro-F1 scores on hipe-2020 and newseye. We compare against prior historical EL systems reported in the literature, including SBB, L3i, MELHISSA, BELA, and MHEL-LLaMo, as well as our SFT, DPO, and Multi-DPO variants.

The results show that DPO substantially improves historical EL performance across both datasets. Standard DPO already outperforms SFT and original with prompting in most settings, suggesting that preference-based supervision is better suited than token-level generation loss or original knowledge of LLMs for retrieval-augmented entity selection. This is especially visible on hipe-2020-FR, hipe-2020-EN, and the newseye subsets.

Multi-DPO further improves upon standard DPO in most languages, achieving the best results on all hipe-2020 subsets and on three of the four newseye subsets. The only exception is newseye-FI, where MELHISSA remains the strongest. These results suggest that modeling multiple preference relations provides a stronger ranking signal and improves robustness when the candidate set contains several plausible but incorrect entities.

6.1 Oracle Entity Selection Analysis

Dataset	Lang.	Gold-in-set	Sel. acc.	Full	Oracle	Gain
hipe-2020	FR	87.2	67.9	61.2	82.8	+21.6
	DE	83.4	64.9	56.0	79.3	+23.3
	EN	82.6	84.1	74.0	81.7	+7.7
newseye	FR	88.5	67.1	55.7	71.9	+16.2
	DE	84.7	56.8	46.8	66.2	+19.4
	FI	88.2	53.0	63.1	85.1	+22.0
	SV	84.2	69.8	65.0	81.5	+16.5

Table 2: Conditional oracle analysis (%). **Gold-in-set**: non-NIL gold retrieval recall; **Sel. acc.**: accuracy conditioned on gold retrieval; **Full/Oracle**: entity-level accuracy.

We isolate the remaining selection headroom using the candidate sets from the full retrieval pipeline. For each non-NIL mention whose gold QID is retrieved, the oracle replaces the model prediction with that QID; all other predictions, including NIL mentions, remain unchanged. Table 2 reports gold-in-set recall among non-NIL mentions, selection accuracy conditional on successful retrieval, and full entity-level accuracy before and after the replacement.

Oracle gains range from 7.7 to 23.3 points. Thus, even with high non-NIL candidate recall, entity selection remains a substantial source of end-to-end error. The newseye-FI oracle uses all mentions, whereas conditional selection accuracy uses only retrieved non-NIL mentions; its high NIL fraction explains the apparent gap.

6.2 Retrieval Analysis

Backbone LLMs. Tables 3 and 4 compare candidate retrieval recall across backbone LLMs without alias augmentation, isolating each model's intrinsic candidate-generation ability. All models are evaluated using the same prompting strategy (Appendix C).

Candidate model	hipe-2020		
	FR	DE	EN
Qwen3.5-9B (Team, 2026)	68.8	63.5	58.8
GPT-20B (Agarwal et al., 2025)	66.2	63.3	62.4
Qwen3.5-27B (Team, 2026)	75.0	72.9	61.5
Qwen3-A30B (Yang et al., 2025)	72.2	66.7	60.9
Gemma4-31B (Team et al., 2026)	**76.1**	**73.6**	64.9
GPT-120B (Agarwal et al., 2025)	73.1	65.8	**66.2**

Table 3: Retrieval recall on hipe-2020 using different LLMs for candidate generation.

Candidate model	newseye			
	FR	DE	SV	FI
Qwen3.5-9B	55.0	46.2	55.7	58.6
GPT-20B	56.3	49.8	62.1	65.8
Qwen3.5-27B	55.8	50.1	62.8	59.3
Qwen3-A30B	**59.2**	**51.7**	11.4	24.4
Gemma4-31B	55.7	50.8	59.4	63.1
GPT-120B	58.9	51.0	**64.1**	**69.6**

Table 4: Retrieval recall on newseye using different LLMs for candidate generation.

The results show that model scale alone does not determine retrieval quality. On hipe-2020, Gemma4-31B achieves the strongest recall for French and German, while GPT-120B performs best for English. On newseye, GPT-120B is strongest for Swedish and Finnish, whereas Qwen3-A30B obtains the best recall for French and German but performs poorly on the Nordic languages.

Overall, we notice that the retrieval performance varies substantially across languages and datasets, suggesting that multilingual coverage, training data composition, and robustness to historical spelling variation are as important as model size.

Prompt Instruction and Alias Lookup. We further study the impact of prompt design, context representation, and alias augmentation. Table 5 shows that simple prompts consistently outperform more complex instructions, while chunk-based context is more effective than summary-based context. Summary representations occasionally omit useful information, leading to lower recall. Alias lookup provides the largest improvement, consistently recovering entities missed by the LLM retrieval stage. Thus, combining simple prompts, chunk context, and alias augmentation yields the best retrieval performance.

Prompt	Context	Alias	hipe-2020		newseye	
			FR	DE	FR	DE
Simple	Summary	✗	74.5	69.3	55.6	50.3
Complex	Chunk	✗	73.3	71.8	53.4	47.4
Simple	Chunk	✗	75	72.9	55.8	50.1
Simple	Chunk	✓	**76.5**	**74.4**	**57**	**53**

Table 5: Retrieval recall using different prompts, context information, and alias dictionary lookup.

Model	hipe-2020		
	FR	DE	EN
Qwen3-14B (Yang et al., 2025)	67.0	61.0	61.1
Finetuned MDPO Qwen3-14B	67.3	51.8	65.9
GPT-20B (Agarwal et al., 2025)	65.0	58.3	59.9
Finetuned MDPO GPT-20B	72.4	65.6	73.8
Qwen3-32B (Yang et al., 2025)	67.0	58.5	63.9
Finetuned MDPO Qwen3-32B	63.1	59.0	68.7

Table 6: Micro-F1 on hipe-2020 with different entity-selection backbones before and after fine-tuning.

Model	newseye			
	FR	DE	SV	FI
Qwen3-14B (Yang et al., 2025)	58.3	50.0	51.8	61.7
Finetuned Qwen3-14B	69.1	62.6	64.0	60.6
GPT-20B (Agarwal et al., 2025)	59.0	49.7	59.3	60.3
Finetuned GPT-20B	68.7	59.2	65.0	62.7
Qwen3-32B (Yang et al., 2025)	58.7	51.1	55.2	60.2
Finetuned Qwen3-32B	49.0	48.5	40.7	49.5

Table 7: Micro-F1 on newseye with different entity-selection backbones before and after fine-tuning.

6.3 Selection Analysis

Backbone LLMs. Tables 6 and 7 show that larger models do not necessarily yield better fine-tuned performance. While Qwen3-32B achieves competitive zero-shot results, its finetuned variant degrades substantially on several newseye languages, suggesting optimisation instability under noisy supervision. In contrast, GPT-20B consistently benefits from finetuning and achieves the best overall performance, indicating stronger compatibility with preference optimisation objectives.

Candidate model	hipe-2020		
	FR	DE	EN
Qwen3.5-9B (Team, 2026)	71.5	63.1	67.0
Qwen3.5-27B (Team, 2026)	68.7	64.1	72.4
Qwen3-A30B (Yang et al., 2025)	62.0	64.2	72.5
GPT-20B (Agarwal et al., 2025)	67.5	58.5	62.5
GPT-120B (Agarwal et al., 2025)	**72.4**	65.6	**73.8**
Gemma4-31B (Team et al., 2026)	72.1	**65.8**	73.0

Table 8: Micro-F1 on hipe-2020 with the same multi-negative finetuned DPO model

Candidate model	newseye			
	FR	DE	SV	FI
Qwen3.5-9B (Team, 2026)	67.5	56.5	62.1	**64.5**
Qwen3.5-27B (Team, 2026)	63.6	57.3	58.7	59.3
Qwen3-A30B (Yang et al., 2025)	65.0	58.9	62.5	60.9
GPT-20B (Agarwal et al., 2025)	65.9	55.5	61.5	63.1
GPT-120B (Agarwal et al., 2025)	**68.7**	**59.2**	**65.0**	62.7
Gemma4-31B (Team et al., 2026)	65.0	56.3	59.6	60.2

Table 9: Micro-F1 on newseye with the same multi-negative DPO model

Retrieval LLMs. Tables 8 and 9 evaluate the impact of retrieval quality while keeping the same GPT-20B multi-DPO ranker fixed. Stronger retrieval models generally improve end-to-end performance, with GPT-120B achieving the best results overall. However, the gains remain moderate, suggesting that the ranker can partially compensate for weaker candidate generation. Mid-sized models such as Gemma4-31B and Qwen3.5-27B remain competitive across languages.

6.4 Error Analysis

Retrieval errors. This section provides a detailed analysis of candidate generation and its interaction with downstream entity selection. Across the four subsets shown in Figure 2, non-NIL recall ranges from 83.4% on hipe-2020-DE to 88.5% on newseye-FR. In contrast, NIL recall ranges only from 17.0% on newseye-DE to 40.7% on hipe-2020-FR because the system often retrieves a plausible entity instead of abstaining. NIL handling therefore remains the main retrieval weakness. Figure 3 compares simple and complex prompts: simple prompts generally recover more

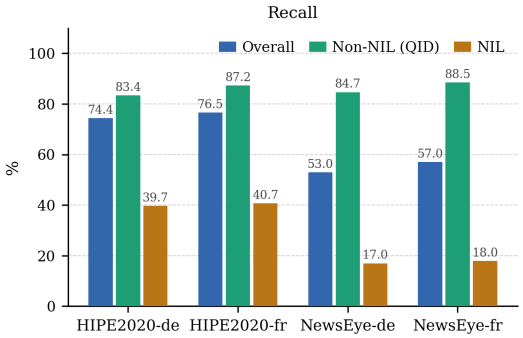
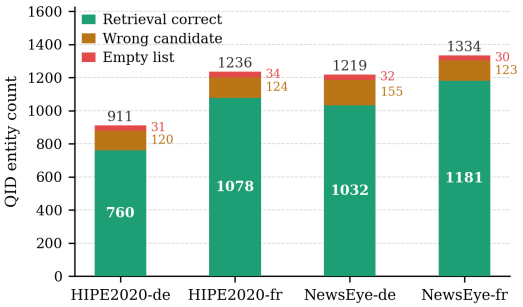
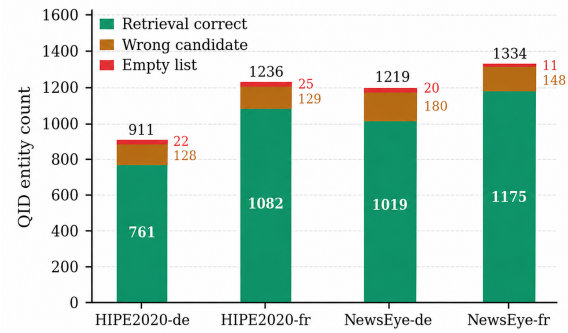


Figure 2: Retrieval recall (%) across datasets.



(a) Simple prompt



(b) Complex prompt

Figure 3: Retrieval error breakdown for simple and complex prompts.

correct candidates, whereas complex prompts may reduce empty outputs at the cost of generating additional incorrect candidates.

Fine-Grained Errors. We heuristically categorise errors across all seven subsets and report the absolute accuracy gain of the finetuned selector over the base model in Table 10. The largest gains occur for gold-NIL/knowledge-base coverage (+13.7%) and semantic ambiguity (+8.1%), followed by OCR noise. These are relative improvements, not solved cases: low absolute NIL retrieval recall and false NIL predictions on valid QID mentions remain important limitations. Appendix D provides additional examples involving historical spelling, stub entities, fine-grained geography, and evolving gold-NIL annotations.

Error type	N	Base	Finetuned	Gain (pp)
NIL / KB	3636	48.4	62.1	+13.7
Semantic	801	59.4	67.5	+8.1
OCR noise	79	53.2	60.8	+7.6
Other	2993	68.2	71.8	+3.6
Spelling / orthog.	1194	63.7	64.9	+1.2
Historical naming	717	56.5	57.5	+1.0

Table 10: Accuracy by heuristic error type across the seven evaluation subsets. "Other" contains cases not matched by the surface-form heuristics.

7 Conclusions

In this work, we introduced a retrieval-augmented framework for historical named entity linking that combines LLM-based candidate generation with preference-optimised entity selection. We showed that while modern LLMs provide strong retrieval capabilities, effective candidate ranking remains the primary challenge in multilingual and OCR-degraded historical documents. To address this, we proposed a multi-negative DPO objective that improves the model's ability to distinguish correct entities from noisy candidates. Experiments on `hipe-2020` and `newseye` demonstrate consistent improvements over both supervised and unsupervised baselines across multiple languages. Our ablation studies further show that retrieval quality, context design, and preference optimization all contribute to final performance, while moderate context windows and simple prompting strategies are generally the most effective.

Limitations

Although our LLM-guided retrieval framework improves candidate recall in noisy historical corpora, several limitations remain.

First, the approach still struggles with NIL entity prediction. The LLM tends to generate semantically plausible entities even when no correct entity exists in the knowledge base, leading to increased false positives. This behavior is particularly challenging in historical documents containing OCR errors, spelling variations, and incomplete contextual information.

Second, the overall performance remains highly dependent on the retrieval stage. If the correct entity is not retrieved or validated, the downstream

selection model cannot recover it. While LLM-guided query expansion improves recall, it may also introduce noisy or hallucinated candidates.

Third, our objective is a decomposable pairwise extension of single-negative DPO. We did not evaluate alternative listwise or contrastive objectives with shared candidate-set normalisation, so our results do not establish superiority over those formulations.

Finally, candidate generation is the dominant inference cost in the measured deployable configuration, and the reported latency is not directly comparable with published systems running on different hardware and software stacks.

Acknowledgments

This work has been co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 — AI4DH. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

This work has also been supported by the TERMITRAD (2020-2019-8510010) and ACTUDATA (2022-2021-17014610) projects funded by the Nouvelle-Aquitaine Region (France), and it has benefited from the computing resources of the L3i laboratory, operated and hosted by the University of La Rochelle, and funded by the French government and the Nouvelle-Aquitaine Region.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. 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.

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

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Sebastian Cadavid-Sanchez, Khalil Kacem, Rafael Aparecido Martins Frade, Johannes Boehm, Thomas Chaney, Danial Lashkari, and Daniel Simig. 2023. Evaluating end-to-end entity linking on domain-specific knowledge bases: Learning about ancient technologies from museum collections. *arXiv preprint arXiv:2305.14588*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 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.

Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. 2024. Entgpt: Linking generative large language models with knowledge bases. *arXiv e-prints*, pages arXiv–2402.

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 *European Conference on Information Retrieval*, pages 524–532. Springer.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020b. Extended overview of clef hipe 2020: named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696. CEUR-WS.

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

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

Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2023. In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Natural Language Engineering*, 29(2):425–448.

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

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *EMNLP*.

Yajie Li, Albert Galimov, Mitra Datta Ganapaneni, Pujitha Thejaswi, De Meng, Priyanshu Kumar, and Saloni Potdar. 2025. Leveraging the power of large language models in entity linking via adaptive routing and targeted reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 871–882.

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

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J Liu, and Xuanhui Wang. 2025. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2404–2420. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37.

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

Vera Provatorova, Svitlana Vakulenko, Evangelos Kanoulas, Koen Dercksen, and Johannes M. van Hulst. 2020. Named entity recognition and linking on historical newspapers: UvA.ILPS & REL at CLEF HIPE 2020. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Cristian Santini, Marieke van Erp, and Mehwish Alam. 2026. It's all about the confidence: An unsupervised approach for multilingual historical entity linking using large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3939–3954, Rabat, Morocco. Association for Computational Linguistics.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, and 1 others. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. Contextual augmentation for entity linking using large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8535–8545, Abu Dhabi, UAE. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2267–2282.

Shuo Xie, Fangzhi Zhu, Jiahui Wang, Lulu Wen, Wei Dai, Xiaowei Chen, Junxiong Zhu, Kai Zhou, and Bo Zheng. 2025. MPPO: Multi pair-wise preference

optimization for LLMs with arbitrary negative samples. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1545–1554. Association for Computational Linguistics.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2025. Llmael: Large language models are good context augmenters for entity linking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3550–3559.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Training and Inference Details

We subclass the TRL DPO trainer[3] and use a custom collator so that every EL instance retains one prompt, one chosen entity, and all rejected candidates from the same retrieval set. The trainer computes length-normalised policy and reference log-likelihoods, obtains the chosen score once, compares it independently with every valid rejected score, and masks padded entries before averaging. The frozen reference is the pretrained backbone with LoRA adapters disabled. Thus, the code implements Eq. 7 rather than flattening examples into unrelated one-to-one records.

LoRA is applied to the attention and feed-forward projection layers with rank $r = 8$ and $\alpha = 16$.

Training is performed for 3 epochs using AdamW 8-bit optimisation with a learning rate of 5×10^{-6}, linear learning-rate scheduling, and a warmup ratio of 0.1. We use a per-device batch size of 8 with gradient accumulation over 8 steps. Models are trained in BF16 precision with gradient checkpointing enabled. The model context window is 2048 tokens; prompts are capped at 1024 tokens and answers at 128 tokens. The LLM-generated branch is capped at five candidates; larger alias-augmented pools require candidate pruning or description truncation.

[3] https://huggingface.co/docs/trl/dpo_trainer

For inference, the merged LoRA model is quantised to 4-bit MXFP4 format and served using vLLM[4] for efficient batched decoding and memory-efficient inference. On one RTX A6000, a deployable GPT-20B generator/selector configuration requires 2.35–2.63 s for candidate generation and 0.35–0.50 s for selection (2.70–3.13 s combined). Candidate generation accounts for approximately 82–89% of latency. The strongest configuration in Table 1 instead uses a GPT-120B generator; because published systems use different hardware and implementations, we do not make a direct runtime claim against them.

We follow the official dataset splits for all experiments. For `hipe-2020`, we train on the French and German training sets and use the corresponding development sets for validation. The English subset is excluded from training and evaluated only in a zero-shot cross-domain setting to assess generalisation on unseen historical data. For `newseye`, we use the French, German, Swedish, and Finnish subsets following the same train/development split protocol.

For multi-negative preference construction, we use GPT-20B (Agarwal et al., 2025) as the candidate generation model, served through the vLLM inference engine for efficient large-scale decoding. For the standard DPO baseline, preference pairs are constructed using the gold entity as the positive sample and a randomly selected entity from the negative candidate pool as the negative sample.

Instances without a valid positive/negative comparison are excluded from DPO training. At inference, a single candidate is not automatically accepted because the generative selector may still emit NIL or an invalid response. In a controlled test, the selector chose the sole gold candidate in 44/50 cases (88.0%); for 50 empty-candidate gold-NIL cases, it returned NIL in every case.

Although GPT-20B (Agarwal et al., 2025) generates high-quality multi-negative comparison pairs, its SFT convergence was unstable. We therefore use Qwen3-14B (Yang et al., 2025) for the SFT baseline.

B Selection

B.1 Error Analysis Across Retrieval and Selection

Although the primary evaluation metric is micro-F1, we report entity-level accuracy to better anal-

[4] https://vllm.ai

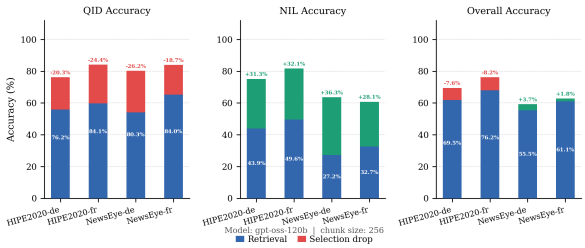


Figure 4: Entity-level accuracy before and after selection.

yse the transition from retrieval to selection. Figure 4 compares performance before and after the selection stage. We observe that the selection model consistently improves NIL prediction accuracy across all datasets, with substantial gains (up to +36%), indicating its effectiveness in identifying non-linkable mentions.

However, this improvement comes at the cost of reduced QID accuracy, where selection introduces a noticeable drop (up to -26%). As a result, overall accuracy shows only modest changes, with slight improvements in some cases and small degradations in others. These results highlight a key trade-off: while the selection model enhances robustness in handling NIL cases, it can also over-filter valid candidates, suggesting that balancing precision and recall in the selection stage remains critical.

B.2 Context length.

Context length	hipe-2020		newseye	
	FR	**DE**	**FR**	**DE**
128	64.4	59.1	64.1	53.9
256	**72.4**	**65.6**	**68.7**	**59.2**
512	69.8	64.7	67.9	56.6

Table 11: F1 score with different context lengths for backbone LLMs.

Table 11 shows that a moderate context window (256 tokens) yields the best overall performance. Shorter contexts lack sufficient disambiguating information, while longer contexts (512 tokens) often introduce additional noise and slightly degrade performance. This suggests a trade-off between contextual coverage and irrelevant information in retrieval-augmented selection.

C Instruction Prompts

C.1 Simple Retrieval Prompt

The following prompt in Tab 12 was used to generate candidate entity names for Wikidata entity linking.

Table 12: Complete retrieval prompt used for candidate generation.

```
RETRIEVAL_PROMPT = """
You are an expert in Wikidata entity linking and
historical linguistics.  Your goal is to generate
a concise list of potential candidate entity
names based on the given context.
Attention:

- The entity mention may include typos,
punctuation, extra spaces, or unusual
capitalisation.

- The surrounding context text is useful to help
identify the entity's historical or cultural
background.

Your task:

1.  Normalise the entity string, remove
unnecessary punctuation, spacing, or symbols.

2.  Use contextual clues, such as time period,
nationality, and occupation, to infer more
precise or alternate forms.

3.  Generate up to **5 likely candidate names**
that could correspond to valid Wikidata items.

4.  Each candidate must be a clean,
human-readable string and validated through the
wikidata_checker function.
Rules:

- Do not use markdown, extra commentary, or
natural language.

- If no plausible candidates are found, return
the original entity.

Mode:  No Reasoning
"""
```

C.2 Complex Retrieval Prompt

The following prompt in Tab 13 was used for handling complex historical multilingual text retrieval and OCR noise:

C.3 Selection Prompt

The following prompt in Tab 14 was used for Named Entity Linking, given the context and candidate strings:

D Detailed Error Analysis

To better understand the behavior of our approach, we conduct a qualitative error analysis across the seven historical entity linking datasets from hipe-2020 and newseye.

D.1 Success Cases and Comparison with Baseline LLM

Finetuning substantially mitigates several limitations observed in zero-shot LLM-based entity linking. In particular, baseline LLMs frequently struggle with historical spelling variation, ambiguity

Table 13: Complex retrieval prompt used for Wikidata candidate generation.

```
COMPLEX_RETRIEVAL_PROMPT = f"""

You are an expert in Wikidata entity linking for historical multilingual texts.

The entity mentions come from OCR-scanned historical newspapers (French, German)
published between 1797 and 1920, from Swiss, Luxembourgish, and Austrian
collections.

You will be given the entity mention, its type (PERS/LOC/ORG/PROD/TIME), and
context.

Generate up to {MAX_CANDIDATES} candidate name strings to search on Wikidata.

## By entity type:

- PERS → search the surname first, then the full name; drop honorifics and titles

- LOC → strip generic geographic words; search the proper name

- ORG → search the canonical short name of the organisation

- PROD → search the publication or doctrine title

- TIME → return empty (dates are not linkable to Wikidata)

## Common OCR noise in these texts:

- Broken hyphens, extra spaces, or stray characters from imperfect scanning

- Abbreviated honorifics and titles -- expand or drop them

- Old or non-standard spellings -- try the modern standard form

- Multi-token mentions where only part is the actual entity name

## Steps:

1.  Clean OCR noise from the mention

2.  Use the entity type to identify the core searchable part

3.  Generate variants:  clean form, sub-parts, expanded abbreviations, and
alternate spellings

4.  Order candidates:  most distinctive first, raw cleaned mention last

Rules:

- Validate each candidate through the wikidata_checker function

- Return clean human-readable strings only -- no QIDs and no markdown

- Use context clues such as dates, nationality, and profession

- If no candidates are found, return the cleaned raw mention

Mode:  No Reasoning

"""
```

Table 14: Selection prompt used for entity linking.

```
SELECTION_PROMPT = """

You are an expert Named Entity Linking (NEL)
specialist.  You are given an entity, its context,
and a list of candidate strings.
Return:

<entity>:  <QID>

Rules:

- If no candidate is suitable, return "NIL". -
Only use provided candidates; do not invent new
ones.
"""
```

introduced by duplicate or incomplete Wikidata entries, and fine-grained geographical disambiguation. For example, the baseline model often links *Bartenstein* to a deprecated or stub Wikidata entry, or resolves *Philadelphia* to the naval ship rather than the city. Similar errors occur for geographically ambiguous mentions such as *Berne* and *New-York*, where the model confuses cities with larger administrative regions.

As illustrated in Table 15, the finetuned model learns stronger context-sensitive linking preferences and better captures historical newspaper conventions, enabling more accurate selection of the gold-standard Wikidata entities.

Dataset	Context window	Gold entity	Baseline LLM without FT	Finetuned model
HIPE2020-DE	... Am 22 . Merz sah man auf der großen Parade zu **Petersburg** 4 . eroberte schwedische Fahnen...	Saint Petersburg (Q656)	NIL	Saint Petersburg (Q656)
HIPE2020-FR	... lieutenant d ' Avoyer de la Préfecture de Berne , à **Berne** le 30 Avril 1828...	Bern (Q70)	Canton of Berne (Q11911)	Bern (Q70)
HIPE2020-FR	... Les amateurs pourront se trouver rassemblés à l ' auberge du Cerf aux **Ponts** .	Les Ponts-de-Martel (Q68617)	Les Ponts-de-Cé (Q752975)	Les Ponts-de-Martel (Q68617)
HIPE2020-FR	... soignée du beau Traire des causes civiles de la Principauté de **Neuchâtel** , par Al . le châtelain Alonvert ...	Neuchâtel (Q69345)	Principality of Neuchâtel (Q3137802)	Neuchâtel (Q69345)
HIPE2020-EN	... ship Caledonia , arrived at **Philadelphia** on Monday , from Cadiz , states , that the French army ...	Philadelphia (Q1345)	USS Philadelphia (Q2288745)	Philadelphia (Q1345)
HIPE2020-EN	... respectable meeting in **New - York** , against the violation of the treaties with the...	New York City (Q60)	New York State (Q1384)	New York City (Q60)
HIPE2020-EN	... while that proposed by **Gen . Harrison** , with the present number ot our militia , would cost...	William H. Harrison (Q11869)	W. H. Harrison (politician) (Q8012023)	William H. Harrison (Q11869)
newseye-DE	... Plener wünschen , wogegen bei der Wahl im Jahre 1891 Herr **v . Plener** einen Gegenkandidaten hatte ...	Ernst von Plener (Q324918)	von Plener (family) (Q23866513)	Ernst von Plener (Q324918)
newseye-DE	... des Satanismus sadistische Befriedigung sucht . Der Satanismus wird Modesache . Unter **Ludwig XIV .** wird die schwarze Messe populär...	Louis XIV of France (Q7742)	NIL	Louis XIV of France (Q7742)
newseye-FR	... Brest , 14 janvier . — Le **conseil municipal** de Camaret vient de démissionner , parce qu ' il...	municipal council (Q701632)	city council (Q3154693)	municipal council (Q701632)

Table 15: Examples of named entity linking where the baseline LLM without finetuning fails (by predicting incorrect/stub QIDs or returning NIL), but our proposed finetuned model correctly links to the gold standard Wikidata QID.

D.2 Wikidata Annotation Discrepancies (Gold NIL vs. Predicted QID)

A recurring challenge in historical entity linking is the evolving nature of the underlying knowledge base. Benchmark datasets such as hipe-2020 and newseye rely on static annotations, where many mentions are labeled as NIL because no suitable Wikidata entity was available or verified at annotation time.

However, Wikidata has expanded considerably since the creation of these datasets. Combined with dense alias retrieval, our system is often able to retrieve plausible and contextually appropriate Wikidata entities for mentions annotated as NIL. Table 16 presents representative examples where the predicted entity appears historically consistent despite disagreement with the original annotation.

D.3 Key Qualitative Insights

- **Historical Orthographic Ambiguity:** In hipe-2020-DE, the mention *Bartenstein* refers to the historical East Prussian city now known as *Bartoszyce*. The baseline model is misled by duplicate or inactive Wikidata entries and predicts an incorrect QID (Q6467766), whereas the finetuned model correctly resolves the mention to *Bartoszyce* (Q809585).

- **Fine-Grained Contextual Disambiguation:** Baseline LLMs frequently prefer broader or more globally prominent entities. For instance, *Berne* is incorrectly linked to the *Canton of Berne* (Q11911) instead of the city (Q70), while *New-York* is resolved to *New York State* (Q1384) rather than *New York City*

Dataset	Context window	Gold label	Model prediction (Wikidata entity)
HIPE2020-FR	... curateur , ont été admis par arrêt du **Conseil d ' Etat** du 5 décembre 1797 , à solliciter une...	NIL	Conseil d'État (Q769657)
HIPE2020-FR	... tenoit feu Mlle . Eckard à la **rue des Moulins** , avisent le public qu ' elles le tiendront...	NIL	rue des Moulins (Q3451926)
HIPE2020-EN	... moft obedient , and very humble fervants , **RICHARD HENRY LEE** WILLIAM GRAYSON . PRESIDENT SULLIVAN...	NIL	Richard Henry Lee (Q725907)
NewsEye-DE	... Kramarsch hat heute erzählt , daß Fräulein **Marie Pospischil** , eine Schauspielerin des tschechischen Theaters in Prag...	NIL	Marie Pospíšilová (Q12035450)
NewsEye-DE	... Klausen . Historischer Roman von **Johann von Wildenradt** . Vor dem Rathhause stieg der siegreiche...	NIL	Johann von Wildenrath (Q55681870)
NewsEye-FR	... Steeg , gouverneur de l ' Algérie , et **M . Rault** , président de la commission du gouvernement de la Sarre...	NIL	Michel Rault (Q65597615)
NewsEye-FI	... kehotettu käymään Englannissa , Saksassa ja Venäjällä puhumassa **San Franciscon** näyttelyn puolesta .	NIL	San Francisco (Q62)

Table 16: Examples of Wikidata annotation discrepancies: mentions labeled as NIL in the official gold standard, which our proposed finetuned model successfully links to the active Wikidata QID.

(Q60). The finetuned model more reliably distinguishes municipalities from larger administrative regions using local contextual cues.

- **Resolution of Gold NIL Mentions:** Several mentions annotated as NIL can be linked to plausible Wikidata entities using updated knowledge base information. For example, in NewsEye-DE, the mention *Marie Pospischil* is linked by our model to *Marie Pospíšilová* (Q12035450), a Czech actress historically active in Prague theater. Similarly, in NewsEye-FR, the model links *M. Rault* to the French politician (Q65597615). These examples suggest that benchmark annotations may underestimate real-world linking performance in historically evolving knowledge bases.